STA 321 Spring 2016

Lecture 15 *Tuesday, March 31*

- **Summarizing Bivariate Data**
  - **Two categorical variables**
    - *Contingency Table*
    - *Row/Column Relative Frequencies*
    - *Relative Risk, Odds Ratio*

Confidence Interval for the Difference of Two Proportions: Example

- Famous five-year study on the effect of Aspirin to reduce heart disease
- Study subjects: 22,071 male physicians
- Every other day, participants took either an aspirin tablet or a placebo
- 11,034 who took placebo: 189 had a heart attack
- 11,037 who took aspirin: 104 had heart attacks
- *Estimate the heart attack rates for the two groups.*
- *Construct a 95% confidence interval to compare them.*
- *Interpret.*

$$(\hat{p_2} - \hat{p_1}) \pm z \cdot \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

Significance Test for the Difference of Two Proportions

- The large sample (see above) significance test for the null hypothesis that both population proportions are equal,

$$z_{obs} = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$\hat{p}$ is the "pooled" proportion of the total sample (both samples together) in the category of interest

Significance Test for the Difference of Two Proportions

- As above, most commonly, the alternative hypothesis is two-sided
- Then, the P-value is the two-tail probability of "anything at least as extreme as observed"
- The probability is taken from a z-score applet (normal distribution)

Significance Test for the Difference of Two Proportions: Example

- Effect of Aspirin to reduce heart disease
- Study subjects: 22,071 male physicians
- Every other day, participants took either an aspirin tablet or a placebo
- 11,034 who took placebo: 189 had a heart attack
- 11,037 who took aspirin: 104 had heart attacks
- *Test whether the rates are significantly different. Report the P-value and interpret.*

$$z_{obs} = \frac{\hat{p_2} - \hat{p_1}}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Confidence Interval and Hypothesis Test for the Difference of Two Proportions: Example

- A serious side effect of many allergy medicines is that they cause drowsiness, which makes them dangerous for industrial workers. In recent years, different nondrowsy allergy medicines have been developed. One of them was Hismanal. The manufacturer claimed that this was the first once-a-day nondrowsy allergy medicine.
- The nondrowsiness claim is based on a clinical experiment in which 1,604 patients were given Hismanal, and 1,109 patients were given a placebo.
- Of the first group, 7.1% reported drowsiness; of the second group, 6.4% reported drowsiness.
- Do these results allow us to infer at the 5% significance level that Hismanal's claim is false?
- Report a 95% confidence interval for the difference of the proportions.

Contingency Table

- The proportions are usually listed in a table called ***contingency table***
- How are the outcomes of the response variable *contingent* on the category of the explanatory variable
- Each row represents a category of one variable, and each column represents a category of the other variable
- The cells of the table contain frequency counts for the four possible combinations of outcomes

	Drowsy	Not Drowsy	
Hismanal			
Placebo			

Summary

Large Sample Significance Test for the Difference of Two Proportions

	One-Sided Tests		Two-Sided Test
Null Hypothesis	$H_0 : p_1 = p_2$		
Research Hypothesis	$H_1 : p_2 < p_1$	$H_1 : p_2 > p_1$	$H_1 : p_1 \neq p_2$
Test Statistic	$z_{obs} = \dfrac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}}$		
p-value	$P(Z < z_{obs})$	$P(Z > z_{obs})$	$2 \cdot P(Z > \lvert z_{obs} \rvert)$

QUIZ I

- What are the assumptions for t-test for a small sample? List all of them.

Quiz II

- In order to determine the *p*-value, which of the following is not needed?

a. The level of significance.

b. Whether the test is one-tail or two-tail.

c. The value of the test statistic.

d. All of these choices are true.

Describing the Relationship Between Two Nominal (or Ordinal) Variables

Contingency Table

- Number of subjects observed at all the combinations of possible outcomes for the two variables
- Contingency tables are identified by their number of rows and columns
- A table with 2 rows and 3 columns is called 2x3 table (“2 by 3”)

2x2 Table: Example

- 327 commercial motor vehicle drivers who had accidents in Kentucky from 1998 to 2002
- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)

		Accident Fatal		
		Yes	No	
Seat Belt	Yes	30	212	242
	No	33	52	85
		63	264	327

Contingency Table: Example, contd.

- How can we compare fatality rates for the two groups?
- Relative frequencies or percentages within each row
- Two sets of relative frequencies (for *seatbelt=yes* and for *seatbelt=no*), called ***row relative frequencies***
- If seat belt use and fatality of accident are related, then there will be differences in the row relative frequencies

Row relative frequencies

- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)

		Accident Fatal		
		Yes	No	
Seat Belt	Yes			100
	No			100
				100

(No) Association

- If there is no association between two variables, then the row relative frequencies should be about the same

Fictitious Row Relative Frequency		VARIABLE Y		
		Yes	No	
VARIABLE X	Yes			100
	No			100
				100

Fictitious (Absolute) Frequency		VARIABLE Y		
		Yes	No	
VARIABLE X	Yes			242
	No			85
		63	264	327

- Note that the column relative frequencies should also be about the same when there is no association.

Association

- ***Association***: The distribution of the response variable changes in some way as the value of the explanatory variable changes
- Example: If all the Pattersonites love soccer, while the Communications students prefer basketball or baseball, then there is association between the two categorical variables “major” and “preferred sports”

Independence

- "*No association*" is called ***Independence***.
- For example, if 50% of ***every*** group (Patterson, Communications, …) prefer soccer, 30% prefer basketball, 20% prefer football, then the two categorical variables are independent
- In practice there is rarely data with perfect independence, and in samples, there is sampling variation

Measuring Association in 2x2 Tables: Example

- Case 1: Weak association

		Use Twitter	
		Yes	No
Gender	Male	15	25
	Female	25	15

- Case 2: Maximum association or strong association

		Use Twitter	
		Yes	No
Gender	Male	40	0
	Female	0	40

Measuring Association

- A measure of association is a statistic that summarizes the strength of the statistical dependence between two variables
- Common measures of association are
  - ***Difference between the group proportions*** for a given response level (Risk difference)
  - **Relative risk** (Risk ratio)
  - **Odds ratio**

Difference of Proportions (Risk Difference): Example

- **Case 1:**

Difference
37.5% – 62.5%
= – 25%

Proportions		Use Twitter		
		Yes	No	Total
Gender	Male	15/40=37.5%	25/40=62.5%	100%
	Female	25/40=62.5%	15/40=37.5%	100%

- **Case 2:**

Difference
100% – 0%
= 100%

Proportions		Use Twitter		
		Yes	No	Total
Gender	Male	40/40=100%	0/40=0%	100%
	Female	0/40=0%	40/40=100%	100%

Difference of Proportions (Risk Difference): Limitations

Proportions		Disease		
		Yes	No	Total
Exposure	A	37.5%	62.5%	100%
	B	34.5%	65.5%	100%

- Case 1:

Difference

37.5% – 34.5%

= 3%

Proportions		Disease		
		Yes	No	Total
Exposure	A	3.01%	96.99%	100%
	B	0.01%	99.99%	100%

- Case 2:

Difference

3.01% – 0.01%

= 3%

Solution: Ratio of Proportions (Relative Risk, Risk Ratio)

Proportions		Disease		
		Yes	No	Total
Exposure	A	37.5%	62.5%	100%
	B	34.5%	65.5%	100%

- Case 1:

Ratio

0.375 / 0.345

= 1.087

Proportions		Disease		
		Yes	No	Total
Exposure	A	3.01%	96.99%	100%
	B	0.01%	99.99%	100%

- Case 2:

Ratio

0.301 / 0.0001%

= 3010

Measures of Association

Proportions (not by row)		Disease		
		Yes	No	Total
Exposure	Yes	a	b	a+b
	No	c	d	c+d
	Total	a+c	b+d	a+b+c+d

- Difference of proportions (Risk difference): $\frac{a}{a+b} - \frac{c}{c+d}$
- Relative Risk (RR): $\frac{a}{a+b} / \frac{c}{c+d}$
- Odds ratio (OR): $\frac{a}{b} / \frac{c}{d}$

Measures of Association: Example

Frequencies		Accident Fatal		
		Yes	No	
Seat Belt	Yes	30	212	242
	No	33	52	85
		63	264	327

Row Relative Frequencies		Accident Fatal		
		Yes	No	
Seat Belt	Yes	.124	.876	1
	No	.388	.612	1
		.193	.807	1

- Difference of Proportions: 0.124 – 0.388 = -0.264
- Relative Risk: 0.124 / 0.388 = 0.319 (=1/3.1)
- Odds Ratio: (30/212) / (33/52)

  = 0.1415 / 0.6346 = 0.223 (=1/4.5)

Measuring Association: Odds Ratio

- Within a row, the odds of success are defined to be

Odds = (Proportion of "event")/(Proportion of "no event")

- Example: Odds of fatal accident for seat-belt wearers were 30/212=0.1415, and for non-seat-belt wearers they were 33/52=0.6346
- Ratio of the odds from the two rows:***odds ratio***
- In this example, odds ratio = 0.1415 / 0.6346 = 0.223
- "For drivers wearing a seat belt, the odds of an accident being fatal were 0.223 times the odds for drivers not wearing a seat belt (or 'about 4.5 times smaller')."

Measuring Association: Odds Ratio

- There is a shortcut formula for the odds ratio:

  The odds ratio equals the ratio of the products of cell counts from diagonally opposite cells:
- (30 x 52) / (33 x 212) = 0.223
- When the odds ratio is greater than 1, the odds of "disease" are higher in row 1 than in row 2
- Values of the odds ratio farther from 1.0 in a given direction represent stronger association

Why the Odds Ratio?

- Isn't the Relative Risk more intuitive, easier to interpret?
- Yes, but there are situations where the Relative Risk does not make sense: Case-Control Studies
- Example
  - *Case control study of prostate cancer risk and male pattern balding.*
  - *Are men with certain hair patterns at greater risk of prostate cancer?*
  - *Roughly equal numbers of prostate cancer patients and (healthy) controls were selected.*
  - *Among cancer patients, 72 out of 129 had either vertex or frontal baldness, compared to 82 out of 139 among controls*

	Cases	Controls	Total
Balding	72	82	154
Hairy	55	57	112
Total	129	139	268

Why the Odds Ratio?

- You can calculate the proportion of bald men among the cases and among the controls, but it makes no sense to estimate the population proportion of cancer patients among bald men using this data because the prevalence of cancer was artificially inflated to about 50% because of the study design.
- However, you can *always* calculate and interpret the odds ratio in a case control study (as long as the prevalence of the outcome is relatively rare)
- The inflated prevalence *cancels out* in the odds ratio formula, but not in the relative risk formula.

	Cases	Controls	Total
Balding	72	82	154
Hairy	55	57	112
Total	129	139	268

Another Advantage of the Odds Ratio

- For every problem, there are two ways to calculate the relative risk.
- Example
  - *How much does a treatment intervention increase the probability of success? OR*
  - *How much does a treatment intervention decrease the probability of failure?*

	Success	Failure	Total
Treatment	19 (37.3%)	32 (62.7%)	51
Control	5 (8.8%)	52 (91.2%)	57
Total	24	84	108

- Risk Ratio = 0.373/0.088 = 4.2 [times more success] OR
- Risk Ratio = 0.627/0.912 = 0.7 [1.4 times less failure]
- Odds Ratio = (19x52)/(5x32)=6.2 [times higher odds for success] OR
- Odds Ratio = (32x5)/(19x52)=0.16 [6.2 times smaller odds for failure]

Summary: Measures of Association for Categorical Variables

- For events (diseases) with small prevalence, the difference of proportions is not informative.
- The relative risk is generally easier to interpret and more intuitive.
- However, sometimes it is not appropriate: In case-control studies, the odds ratio should be used. Also, there is potential ambiguity as to which relative risk should be compared.
- The odds ratio should only be used for events with small prevalence.
- When you read literature using any of these, be aware of the limitations.

Alcohol and Cigarettes

- Alcohol vs. Cigarette use of senior high school students in Dayton, OH

		Cigarette Use		
		Yes	No	
Alcohol Use	Yes	1449	500	
	No	46	281	