MATH 536 Homework 4

Gustavo Esparza

9/25/2019

Textbook Problems

2.8

A research study estimated that under a certain condition, the probability a subject would be referred for heart catheterization was 0.906 for whites and 0.847 for blacks.

A

A press release about the study stated that the odds of referral for cardiac catheterization for blacks are 60 % of the odds for whites. Explain how they obtained 60 % (more accurately, 57 %).

Focusing on the odds ratios, we can see the following result:

$$\frac{\frac{\pi_b}{1-\pi_b}}{\frac{\pi_w}{1-\pi_w}} = \frac{.847}{.153} \times \frac{.094}{.906} \approx .60$$

Thus, the odds ratio was used for the percentage.

B

An Associated Press story that described the study stated "Doctors were only 60 % as likely to order cardiac catheterization for blacks as for whites." What is wrong with this interpretation? Give the correct percentage for this interpretation. (In stating results to the general public, it is better to use the relative risk than the odds ratio)

This interpretation is wrong because we want to compare two risk percentages but instead are comparing odds. A better proportion would be

$$\frac{\pi_b}{\pi_w} = \frac{.847}{.906} \approx 93\%$$

Thus, we would state that doctors were .93 times as likely to order catheterization for blacks as for whites.

2.9

An estimated odds ratio for adult females between the presence of squamous cell carcinoma (yes, no) and smoking behavior (smoker, nonsmoker) equals 11.7 when the smoker category consists of subjects whose smoking level s is 0 < s < 20 cigarettes per day; it is 26.1 for smokers with s ≥ 20 cigarettes per day.
Show that the estimated odds ratio between carcinoma and smoking levels (s ≥ 20, 0 < s < 20) equals 26.1/11.7 = 2.2.

Considering the below 20 a day group for smoking (denoted by 20), we have the following odds ratio (θ):

$$\frac{\theta_{20}}{\theta_0} = 11.7$$

Considering the above 20 a day group for smoking denoted by 20+, we have the following odds ratio

$$\frac{\theta_{20+}}{\theta_0} = 26.1$$

Then, the odds ratio when comparing the two smoking levels is denoted by:

$$\frac{\frac{\theta_{20+}}{\theta_0}}{\frac{\theta_{20}}{\theta_0}} = \frac{26.1}{11.7} = 2.2$$

2.10

Data posted at the FBI website stated that of all blacks slain in 2005, 91 % were slain by blacks, and of all whites slain in 2005, 83 % were slain by whites. Let Y denote race of victim and X denote race of murderer.

A

Which conditional distribution do these statistics refer to, Y given X, or X given Y ?

We are provided the probability of the murderer X given the race of the victim Y. Thus, we have the conditional distribution of X given Y

B

Calculate and interpret the odds ratio between X and Y.

Considering the following contingency table:

	B	W
B	.91	.09
W	.17	.83

Then the odds ratio for X and Y (M and V) is given as the ratio of the following ratios:

$$\frac{Odds_{M=B|V=B}}{Odds_{M=W|V=W}} = \frac{.91/.09}{.83/.17} \approx 2.07$$

Thus, the odds of a murderer being black given the victim was black are 2.07 times the odds of a murderer being white given the victim was white.

C

Given that a murderer was white, can you estimate the probability that the victim was white? What additional information would you need to do this? (Hint: How could you use Bayes's Theorem?)

By Baye's Theorem, we have

$$P(Y = W|X = W) = \frac{P(X = W|Y = W)P(Y = W)}{P(X = W)}$$

This requires us to know the probability of the murderer being white and the victim being white, both of which are unknown.

2.11

A 20-year study of British male physicians noted that the proportion who died from lung cancer was 0.00140 per year for cigarette smokers and 0.00010 per year for nonsmokers. The proportion who died from heart disease was 0.00669 for smokers and 0.00413 for nonsmokers.

A

Describe the association of smoking with lung cancer and with heart disease, using the difference of proportions, the relative risk, and the odds ratio. Interpret.

Lung Cancer

For the difference in proportions, we have

$p_1 - p_2$ = .00140 - .00010 = .0013.

The difference in proportions between smokers and non-smokers is 0.0013 for physicians who died of lung cancer.

For Relative Risk, we have

$$\frac{p_1}{p_2} = \frac{.00140}{.00010} = 14$$

Physician's death from lung cancer is 14 times more likely for smokers than non-smokers.

For the Odds Ratio, we have the following odds for cigarette smokers:

$$\frac{p_1}{1 - p_1} = \frac{.00140}{1 - .00140} = .001402$$

For the Odds Ratio, we have the following odds for non cigarette smokers:

$$\frac{p_2}{1 - p_2} = \frac{.00010}{1 - .00010} = .0001$$

Therefore our odds ratio becomes

$$\theta = \frac{\theta_{smokers}}{\theta_{non-smokers}} \frac{.001402}{.0001} = 14.02$$

The odds of a physician's death from lung cancer for those who smoke are 14.02 times the odds for those who don't.

Heart Disease

For the difference in proportions, we have

$p_1 - p_2$ = 0.00669 - 0.00413 = .00256

The difference in proportions between smokers and non-smokers is .00256 for physicians who died of heart disease.

For Relative Risk, we have

$$\frac{p_1}{p_2} = \frac{0.00669}{0.00413} = 1.62$$

Physician's death from heart disease is 1.62 times more likely for smokers than non-smokers.

For the Odds Ratio, we have the following odds for cigarette smokers:

$$\frac{p_1}{1-p_1} = \frac{0.00669}{1-0.00669} \approx .007$$

For the Odds Ratio, we have the following odds for non cigarette smokers:

$$\frac{p_2}{1-p_2} = \frac{0.00413}{1-0.00413} \approx .004$$

Therefore our odds ratio becomes

$$\theta = \frac{\theta_{smokers}}{\theta_{non-smokers}} \frac{.007}{.004} \approx 1.62$$

The odds of a physician's death from heart disease for those who smoke are 1.62 times the odds for those who don't.

B

Which response (lung cancer or heart disease) is more strongly related to cigarette smoking, in terms of the reduction in deaths that could occur with an absence of smoking?

We are considering reduction in death that could occur with an absense of smoking within each group, so we should focus on the difference in proprtions for this objective. We can see that the difference in proportions for the two groups when considering Heart Disease is greater than the difference for Lung cancer.
Thus, Heart Disease is more strongly related to cigarette smoking, in terms of the reduction in deaths that could occur with an absence of smoking.

2.13

Refer to Table 2.1 about belief in an afterlife.

	Yes	No	
F	509	116	625
M	398	104	502
	907	220	1127

A

Construct a 90 % confidence interval for the difference of proportions, and interpret.

The Confidence interval for the difference of proportions is as follows:

$$(p_1 - p_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Here, we have $p_1 = \frac{509}{625} = .8144$, $p_2 = \frac{398}{502} = .7928$, $n_1 = 625$, $n_2 = 502$. Then our confidence interval becomes

$$(.8144 - .7928) \pm 1.645\sqrt{\frac{.8144(.1856)}{625} + \frac{.7928(.2072)}{502}} = .0216 \pm .0392 = (-.017, .061)$$

It is important to note that this interval contains 0, which suggests there is no difference between the two proportions. From the interval, the difference in proportions $\pi_1 - \pi_2$ is estimated with 90% probability to be between these two values.

B

Construct a 90 % confidence interval for the odds ratio, and interpret.

Here, the confidence interval for the odds ratio is defined by

$$log(\hat{\theta}) \pm Z_{\alpha/2} \times \sigma_{log(\hat{\theta})}$$

Here, we have the necessary calculations:

$$log\hat{\theta} = log\left(\frac{509 \times 104}{398 \times 116}\right) = .059$$

$$\sigma_{log(\hat{\theta}} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{509} + \frac{1}{398} + \frac{1}{116} + \frac{1}{104}} = .151$$

Thus, we have the following confidence interval:

$$.059 \pm 1.645 \times .151 = (-.189, .307)$$

Raising our confidence interval to the power of e, we have a confidence interval of $(.828, 1.36)$.

We can see that the interval contains 1, and this implies there is no difference between the two odds ratios. Going forward we have 90% confidence that the odds of belief in afterlife in females is between .828 and 1.36 times the odds of belief in afterlife for males.

C

Conduct a test of statistical independence.Report the P-value and interpret.

H_0 : The two variables are independent VS H_A : The two variables are dependent.

Test statistic: $\chi^2 : \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(N_{ij}-E_{ij})^2}{E_{ij}}$

Here are the computations for E_{ij} using the marginal and total sum:

$$E_{11} = \frac{625 \times 907}{1127} = 502.9947$$

$$E_{12} = \frac{625 \times 220}{1127} = 122.0053$$

$$E_{21} = \frac{502 \times 907}{1127} = 404.0053$$

$$E_{22} = \frac{502 \times 220}{1127} = 97.9947$$

Thus, expanding our sum provides the following result:

$$\chi^2 : \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(N_{ij} - E_{ij})^2}{E_{ij}} = .8246$$

This χ^2 statistic has (2-1)(2-1) = 1 degree of freedom and therefore has a corresponding p-value of .3638. Therefore, we fail to reject the null hypothesis and conclude that belief in after life and gender are independent.

2.16

Table 2.12 comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In 20 hospitals in London, UK, patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a noncancer control patient at the same hospital of the same sex and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

	Cases	Controls
Y	688	650
N	21	59

A

Identify the response variable and the explanatory variable.

Here, the response variable is the patient's smoking habits and the explanatory variable is Lung Cancer, since we are already aware of which subjects have Lung Cancer.

B

Identify the type of study this was.

This study was determined by Cases and Controls that were conducted in retrospect, this we have a case-control study.

C

Can you use these data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?

Although we are provided with both the smoker status and the status of Lung Cancer, we can not use the data to compare smokers and non-smokers in terms of proportion who suffered from lung cancer.

In this study, smoking habits were directly measured from a set of two groups. If we were to determine the proportion of lung cancer given smoking habits, we would be assuming that no other factors could contribute to their health. This is clearly not the case, thus this exchange of explanatory and response variables would not be conclusive.

D

Summarize the association, and explain how to interpret it.

We will summarize the association using the odds ratio. Provided our $2X2$ table, we have the following result:

$$\theta = \frac{688 \times 59}{650 \times 21} = 2.9738$$

Thus, the odds of having lung cancer given being a smoker is 2.97 times higher than the odds of having lung cancer when not being a smoker.

2.21

Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (one or more) are responsible for increases in teenage crime: A, the increasing gap in income between the rich and poor; B, the increase in the percentage of single-parent families; C, insufficient time spent by parents with their children. A cross classification of the responses by gender is

Gender	A	B	C
Men	60	81	75
Women	75	87	86

A

Is it valid to apply the chi-squared test of independence to this 2×3 table? Explain.

Since there is more than one option that each subject can choose, we have dependent response variables and thus the Chi Square test does not apply.

B

Explain how this table actually provides information needed to cross- classify gender with each of three variables. Construct the contingency table relating gender to opinion about whether factor A is responsible for increases in teenage crime.

In this table, each factor could be constructed by gender to have a 2×2 contingency table where the factors are not compared to one another.

Here is the contingency table relating gender to opinion about whether factor A is responsible for increases in teenage crime:

A	Yes	No
Men	60	40
Women	75	25

2.33

In murder trials in 20 Florida counties during 1976 and 1977, the death penalty was given in 19 out of 151 cases in which a white killed a white, in 0 out of 9 cases in which a white killed a black, in 11 out of 63 cases in which a black killed a white, and in 6 out of 103 cases in which a black killed a black

A

Exhibit the data as a three-way contingency table.

Murderer's Race	Victim's Race	Penalty	No Penalty
White	White	19	132
	Black	0	9
Black	White	11	52
	Black	6	97

B

Construct the partial tables needed to study the conditional association between defendant's race and the death penalty verdict. Find and interpret the sample conditional odds ratios, adding 0.5 to each cell to reduce the impact of the 0 cell count.

Victim = Black

Murderer's Race	Penalty	No Penalty
White	0	9
Black	6	97

Conditional odds ratio for partial table of Black Victims:

$$\theta_B = \frac{0.5 \times 97.5}{6.5 \times 9.5} = .7895$$

Victim = White

Murderer's Race	Penalty	No Penalty
White	19	132
Black	11	52

Conditional odds ratio for partial table of White victims:

$$\theta_W = \frac{19.5 \times 52.5}{11.5 \times 132.5} = .6719$$

Both of the victim's race odds ratios are less than one. Based on our construction of partial tables, this implies that the odds of a white murderer getting the death penalty are .7895 times the odds of a black murderer when the victim is black and .6719 times the odds when the victim is white.

C

Compute and interpret the sample marginal odds ratio between defendant's race and the death penalty verdict. Do these data exhibit Simpson's paradox? Explain.

For this problem, we will need the marginal death penalty table provided as follows:

Murderer's Race	Penalty	No Penalty
White	19	141
Black	17	149

Thus, the marginal odds ratio between defendants race and the death penalty verdict are:

$$\theta_M = \frac{19 \times 149}{17 \times 141} = 1.181$$

Thus, the marginal odds of a white murderer getting the death penalty are 1.181 the odds of a black murderer getting the death penalty.

Simpsons paradox states that the conditional odds ratios will be different than the marginal odds ratios. We can observe that the conditional odds ratios were less than one, while the marginal odds ratio is greater than one. This example does indeed exhibit Simpson's Paradox.

2.37

Based on murder rate in the United States, the Associated Press reported that the probability a newborn child has of eventually being a murder victim is 0.0263 for nonwhite males, 0.0049 for white males, 0.0072 for nonwhite females, and 0.0023 for white females.

A

Find the conditional odds ratios between race and whether a murder victim, given gender. Interpret.

The conditional odds ratio for Males can be found using the following table:

	Died	Survived
nonwhite	.0263	1 - .0263
white	.0049	1 - .0049

Thus, the odds ratio is defined as

$$\theta_{Males} = \frac{.0263 \times (1 - .0049)}{.0049 \times (1 - .0263)} = 5.49$$

The conditional odds ratio for Females can be found using the following table:

	Died	Survived
nonwhite	.0072	1 - .0072
white	.0023	1 - .0023

Thus, the odds ratio is defined as

$$\theta_{Females} = \frac{.0072 \times (1 - .0023)}{.0023 \times (1 - .0072)} = 3.15$$

Since both odds ratios are greater than one, we can conclude the following:
For both the Male and Female group, the odds of a nonwhite newborn becoming a murder victim is greater than the odds for a white newborn.

B

If half the newborns are of each gender, for each race, find the marginal odds ratio between race and whether a murder victim.

By taking 50% of each probability gender probability and summing over their race:

$P(Died|nonwhite) = .5 \times .0263 + .5 \times .0072 = .01675$
$P(Died|white) = .5 \times 0049 + .5 \times .0023 = .0036$

we can compute the following marginal table:

Both Genders	Died	Survived
nonwhite	.01675	1 - .01675
white	.0036	1 - .0036

Thus, the odds ratio is defined as

$$\theta_{Marginal} = \frac{.01675 \times (1 - .0036)}{.0036 \times (1 - .01675)} = 4.72$$

If half of the newborns are of each gender for each race, the odds that a nonwhite newborn will be a murder victim are 4.72times the odds of a white newborn.

Problem 1

When considering outcome categories, we typically like to describe the difference between the groups in a manner that is easy to interpret. There are different ways to explain this difference, and we will discuss three main methods.

The first and most basic method is known as *Difference of Proportions*. If we are specifically looking at a binomial distribution with 2 outcomes (a success and a failure) then we denote π_1 as the probability of success for the first group (or row) and π_2 as the probability of success for the second group. Likewise, we can denote the probability of a failure for group i as $1 - \pi_1$. Given this information, the difference in proportions is simply defined as $\pi_1 - \pi_2$ and falls between -1 and +1. It is useful to note that the difference is equal to zero when the proportions are equivalent, which implies that the response is independent of the defined groups.

Although Difference of Proportions is a very simple tool that describes a difference between two groups, it is not as useful method when the proportions of interest are in the middle range (IE closer to 0.50 than 0 or 1). This is due to the fact difference in proportions is only able to provide the distance between two proportions rather than a ratio between the two. In situations where more context is required for understanding the proportions of groups, *Relative Risk* is another method to be considered.

Again considering a 2 × 2 table, the *relative risk* is defined as the following ratio:

$$\text{relative risk} = \frac{\pi_1}{\pi_2}$$

From the ratio definition, it can be seen that the relative risk can equal any non negative value. If π_1 is greater than π_2, then our relative risk will be greater than one and vice-avers. This ratio can better explain differences in proportions that may have seemed negligible because their difference were minimal, but are greater multiples of one another in the context of their location between 0 and 1. Where difference in proportions implied independence when the difference was zero, here independence for relative risk occurs when the ratio is equal to 1 (IE when the proportions are once again equivalent.)

Once critical note for the relative risk pertains to constructing confidence intervals. The sampling distribution for relative risk is highly skewed for small sample sizes and thus requires complex formulas and software, whereas the difference in proportions has a relatively straight forward inference structure.

Now that we have covered two methods for comparing proportions for two proportions, we should make note of the fact that we have generally ignored the probability of failure in these methods. It is true that we are able to compare $1 - \pi$ in exchange for π for both methods, but we have yet to combine both success and failure into one comparison method. This is where the *Odds Ratio* is a much significant method of comparison.

Again considering our model for success and failure for two categories, we will define the odds as follows:

$$\text{odds} = \frac{\pi}{1 - \pi}$$

This definition gives a clear ratio for the success and failure rate for a given category. Then, for comparing two odds we have the following *Odds Ratio*, defined by θ:

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2(1 - \pi_2)}$$

When moving from the difference in proportions to the ratio of proportions, we witnessed the amount of increased context for the categories. Now, having included proportion of success and failure, we will more advantages once again.

Once again, independence for X and Y would imply that the odds ratio is equivalent to 1. Moving forward, values of $\theta > 1$ imply that the odds of success in the first category are greater than the odds of success in the second category. In addition, more extreme values of θ (farther from 1) represent stronger association and more dependence.

There are a few qualities for the odds ratio that are extremely beneficial when constructing tables and making inference. The first is that the odds ratio holds to be the same, regardless of the construction of rows and columns. We do not see this quality in the relative risk. Continuing with the consideration of the contingency table, we are able to see the following simple definition for the odds ratio:

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

Thus, we are able to construct the Odds Ratio by cross multiplying our contingency table, a simple technique that still provides a great amount of context.

Problem 2

Typically, we test for an association between categorical variables using the chi-square test, which measures how much the counts in each cell differ from what we would expect the counts to be if the null hypothesis were true. However, if the expected counts is less than 5 for some cells in the table, then this chi square distribution may not be accurate.

Consider the following example on baldness and myocardial infarction, self assessed baldness. A study was completed to learn about the relationship between male-pattern baldness and the risk of cardiovascular disease. Cases were men less than 55 years of age hospitalized for a heart attack. Cases were excluded if they had a prior history of serious heart problems. Controls were men with no history of heart disease admitted to the same hospital for nonfatal, non-cardiac problems.

Question: Is there an association between between baldness and risk of cardiovascular disease? To access this create your own permutation distribution for chi-square and test the hypothesis of independence.

Here is the provided table:

```
##           Baldness
## Treatment Extreme Little Much None Some
##   Control       1    221   34  331  185
##   Disease       2    165   50  251  195
```

Here are the results from our usual χ^2 test for independence:

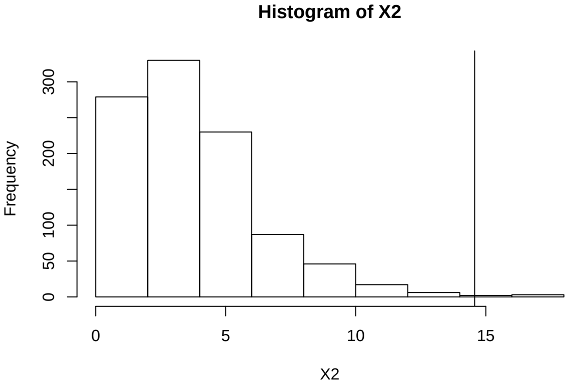
```
## Chi-Square Test Statistic =  14.56965
```

```
## P-value corresponding to test statistic =  0.006996502
```

We can observe that the standard χ^2 test statistic corresponds to a p-value that suggests that there is an association between baldness and risk of cardiovascular disease.

Now, we must perform a permutation test. For this purpose, we will sample from our observed values (without replacement) and compute a new χ^2 test statistic. We will perform this step 1000 times and create a histogram for the test statistics. Finally, we will compare the distribution of χ^2 values and compute a permutation test p-value by observing the proportion of permuted χ^2 values that exceed our original χ^2 statistic.

Here is the Histogram for our permutation χ^2 values, along with a line that marks the initially observed test statistic:



```
## Permutation test p-value =  0.004
```

We can see that the permutation test p-value agrees with our original p-value, thus we can agree that there is evidence to suggest a relationship between baldness and cardiovascular disease.

Appendix

2

```
Treatment = c(rep("Disease",663),rep("Control",772))
Baldness = c ( rep("None",251),rep("Little",165),rep("Some",195),rep("Much",50),rep("Extreme",2),
               rep("None",331),rep("Little",221),rep("Some",185),rep("Much",34),rep("Extreme",1))


df = as.data.frame(cbind(Treatment,Baldness))

data_freq = table(df)
data_freq
```

```
test = as.numeric(chisq.test(data_freq,simulate.p.value = T)$statistic)
pval = as.numeric(chisq.test(data_freq,simulate.p.value = T)$p.value)
```

```
X2 = rep(0,1000)

for(i in 1:1000) {
  permuted_data = as.data.frame(cbind(Treatment, sample(Baldness,replace = F)))
  perm_freq = table(permuted_data)
  X2[i] = as.numeric(chisq.test(perm_freq,simulate.p.value = T)$statistic)
}
```

```
hist(X2)
abline(v=test)
sim_pval = sum(X2>test)/1000
```