Inference for Categorical Data

Matt Allen

06/25/2018

Create a Word document from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word document via D2L Dropbox.

## Exercise 1

Suppose independent, random samples of credit unions and banks had the following frequencies for being rated as outstanding.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Outstanding | Not Outstanding |  |
| Bank | 70 | 150 |  |
| Credit Union | 66 | 81 |  |

### Part 1a

Create the table in R from the data and display it with the margins. Include the names for the rows and columns.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

# Create table and display it with margins  
  
institution <- c("Bank", "Credit Union")  
outstanding <- c(70,66)  
not\_outstanding <- c(150,81)  
  
ratings <- matrix(c(outstanding, not\_outstanding),ncol=2)  
colnames(ratings) <- c("Outstanding","Not Outstanding")  
rownames(ratings) <- institution  
ratings\_with\_margins <- addmargins(as.table(ratings))  
ratings\_with\_margins

## Outstanding Not Outstanding Sum  
## Bank 70 150 220  
## Credit Union 66 81 147  
## Sum 136 231 367

### Part 1b

For the population of all credit unions, construct and interpret a 95% confidence interval for the proportion of credit unions rated as Outstanding.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

# Construct 95% Confidence Interval for proportion of Credit Unions rated as Outstanding.  
prop.test(ratings\_with\_margins["Credit Union","Outstanding"],ratings\_with\_margins["Credit Union","Sum"],correct = FALSE)$conf.int

## [1] 0.3708931 0.5296648  
## attr(,"conf.level")  
## [1] 0.95

With 95% confidence, the true population proportion of outstanding credit unions is between [0.37, 0.53].

### Part 1c

Compare the proportions of credit unions that are rated as Outstanding to the proportion of banks that are rated as Outstanding. Do this by computing a 95% CI for difference in proportions of those rated as Outstanding for credit unions minus banks. Interpret the result.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

# Insert your R code here.  
prop.test(abs(ratings\_with\_margins["Credit Union","Outstanding"]-ratings\_with\_margins["Bank","Outstanding"]),abs(ratings\_with\_margins["Credit Union","Sum"]-ratings\_with\_margins["Bank","Sum"]),correct = FALSE)$conf.int

## [1] 0.0215125 0.1325900  
## attr(,"conf.level")  
## [1] 0.95

With 95% confidence, the true population proportion of credit unions rated as outstanding is between [.02,.13] greater than banks rated as outstanding.

### Part 1d

If one bank is selected randomly, what is the estimated risk of not being rated as Outstanding? (“risk” means probability)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

# what is the estimated risk of a bank not being rated as outstanding?  
bank\_risk\_not\_outstanding <- ratings\_with\_margins["Bank","Not Outstanding"]/ratings\_with\_margins["Bank","Sum"]  
bank\_risk\_not\_outstanding

## [1] 0.6818182

The estimated risk of a bank not being outstanding is equivalent to the proportion of banks rated not outstanding to all banks, which is equal to 0.68.

### Part 1e

If one credit union is selected randomly, what is the estimated risk of not being rated as Outstanding?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

# what is the estimated risk of a Credit not being rated as outstanding?  
credit\_union\_risk\_not\_outstanding <- ratings\_with\_margins["Credit Union","Not Outstanding"]/ratings\_with\_margins["Credit Union","Sum"]  
credit\_union\_risk\_not\_outstanding

## [1] 0.5510204

The estimated risk of a credit union not being outstanding is equivalent to the proportion of credit unions rated not outstanding to all credit unions, which is equal to 0.55.

### Part 1f

Compute the relative risk of not being rated as Outstanding for banks compared to credit unions.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

# Relative risk banks not outstanding compared to credit unions  
bank\_risk\_not\_outstanding/credit\_union\_risk\_not\_outstanding

## [1] 1.237374

A bank is 1.24 times more likely to be rated not outstanding.

### Part 1g

What are the estimated odds of a credit union being rated as Outstanding?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g -|-|-|-|-|-|-|-|-|-|-|-

The estimated odds of a credit union being rated outstanding is .81.

### Part 1h

What are the estimated odds of a bank being rated as Outstanding?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1h -|-|-|-|-|-|-|-|-|-|-|-

#manual calculation of bank being rated as outstanding  
odds\_bank\_outstanding <- (1-bank\_risk\_not\_outstanding)/bank\_risk\_not\_outstanding  
odds\_bank\_outstanding

## [1] 0.4666667

The estimated odds of a bank being rated outstanding is .47.

### Part 1i

Compute the estimated odds ratio of being rated as Outstanding for credit unions compared to banks.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1i -|-|-|-|-|-|-|-|-|-|-|-

# Compute the estimated odds ratio of being rated as outstanding for credit unions compared to banks.  
odds\_credit\_union\_outstanding/odds\_bank\_outstanding

## [1] 1.746032

The estimated odds ratio of being rated as Outstanding for credit unions compared to banks is 1.75.

### Part 1j

Write an interpretation of the odds ratio of being rated as Outstanding for credit unions compared to banks as a percent.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1j -|-|-|-|-|-|-|-|-|-|-|-

For every 4 banks rated as outstanding, there are 7 credit unions rated as outstanding.

### Part 1k

Construct a 95% confidence interval for the population odds ratio of being rated as Outstanding for credit unions compared to banks. Interpret the interval, leaving the endpoints as a multiples.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1k -|-|-|-|-|-|-|-|-|-|-|-

# Computed above in 1g.

A 95% confidence interval for the population odds ratio of being rated outstanding for credit unions compared to banks is between 1.134 and 2.688.

### Part 1l

Based on the 95% CI for the odds ratio, is there significant evidence of an association between being rated as Outstanding and whether or not an institution is a bank or credit union? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1l -|-|-|-|-|-|-|-|-|-|-|-

The interval is [1.134,2.688] which does not include 1. If it did include 1, we would fail to reject the null hypothesis that the odds are the same. Since 1 is not included in the interval, we can say there is significant evidence that credit unions are 1.134 to 2.688 times more likely to be rated as outstanding compared to banks.

## Exercise 2

Marketing Research reported results of a study of online purchases where demographic information was collected on customers. The age group of the customer (under 18, 18 to 35, 36 to 50, or over 50) purchased by each of 165 consumers was recorded.

### Part 2a

A leading internet market research company claims that 13% of all online purchases are made by customers under 18, 32% by customers between 18 and 35, 38% by customers between 36 and 50, and the remaining 17% by customers over 50 years of age.

Test this claim if sample data shows that 28 customers in the sample were under 18, 44 were 18 to 35, 54 were 36 to 50, and 39 were over 50.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age Group | Under 18 | 18 to 35 | 36 to 50 | Over 50 |
| Count | 28 | 44 | 54 | 39 |

Use . State the hypotheses for the test and use R to compute the test statistic, df, and P-value. State the conclusion, including a practical interpretation in the context of the problem. Include the P-value in the conclusion.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

# Run Chi Squared test  
observed <- c(28,44,54,39)  
proportions <- c(.13,.32,.38,.17)  
chisq.test(x=observed,p=proportions)

##   
## Chi-squared test for given probabilities  
##   
## data: observed  
## X-squared = 8.9486, df = 3, p-value = 0.02998

We reject the null hypothesis at a 5% level of significance that online purchases are distributed among age groups as the internet market research company claims. The p-value reported is 0.02998. From p. 519 of the Ott & Longnecker textbook, a p-value in this range indicates a poor fit.

### Part 2b

Compute the expected cell counts and verify that they are all large enough for the chi-square test to be valid. Include a reference to the criterion you are using to determine if expected cell counts are large enough.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

# Compute expected cell counts  
expected\_counts <- sum(observed) \* proportions  
expected\_counts

## [1] 21.45 52.80 62.70 28.05

# Verify that they are large enough

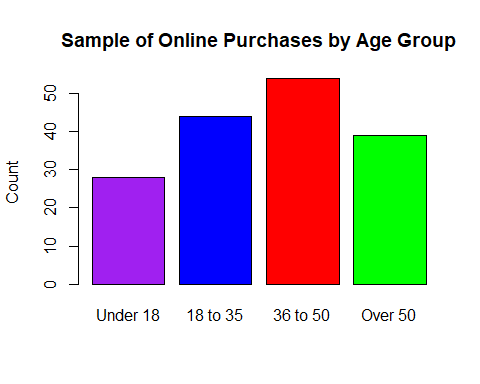
Based on the exprected cell counts they are large enough for the chi-square test to be valid. From Chapter 10 of the Ott & Longnecker textbook p. 516 Cochran’s guidelines, no expected count should be less than 1 and no more than 20% are less than 5.

### Part 2c

Display the data in a bar graph and comment on its features.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

#   
barplot(observed,names.arg=c("Under 18", "18 to 35", "36 to 50","Over 50"),  
 ylab="Count",col=c("purple","blue","red","green"), main = "Sample of Online Purchases by Age Group")



People between ages 36 to 50 are making most of the online purchases.

## Exercise 3

A researcher is studying seat belt wearing behavior in teenagers (ages 13 to 19) and senior citizens (over 65). A random sample of 19 teens is taken, of which 17 report always wearing a seat belt. In random sample of 20 senior citizens, 12 report always wearing a seat belt. Using a 5% significance level, determine if seat belt use is associated with these two age groups.

### Part 3a

Create a 2x2 matrix with the correct cell counts. Arrange it so that the columns represent the age group (teen vs senior) and rows contain the seat belt usage (always wear vs not always wear).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

# Create 2X2 matrix for seat belt data  
age\_group <- c("Teen", "Senior")  
always\_wear <- c(17,12)  
not\_always\_wear <- c(2,8)  
  
seat\_belt\_usage <- matrix(c(always\_wear, not\_always\_wear),ncol=2,byrow = TRUE)  
colnames(seat\_belt\_usage) <- age\_group  
rownames(seat\_belt\_usage) <- c("always\_wear", "not\_always\_wear")  
seat\_belt\_usage

## Teen Senior  
## always\_wear 17 12  
## not\_always\_wear 2 8

### Part 3b

With the small cell counts in mind, use the appropriate test to determine if proportions of those who claim to “always wear” a seat belt is the same for these two age groups. Use a 5% significance level. Include all parts for the hypothesis test.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

#p.511-512 Fisher Exact Test  
#fisher.test  
fisher.test(seat\_belt\_usage,alternative="two.sided")

##   
## Fisher's Exact Test for Count Data  
##   
## data: seat\_belt\_usage  
## p-value = 0.06483  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.8666289 61.3203702  
## sample estimates:  
## odds ratio   
## 5.420308

At a 5% level of significance, we fail to reject the null hypothesis that those who claim “alway wear” a seat belt is the same for teens and seniors. Looking at the confidence interval [0.8666289, 61.3203702] for the odd ratio, we see that 1 is included in the interval.

## Exercise 4

A study was conducted whereby the type of anesthetic (A or B), nausea after the surgery (Yes or No), the amount of pain medication taken during the recovery period, and age for a random sample of 72 patients undergoing reconstructive knee surgery.

The data frame is the anesthesia in the DS705data package.

### Part 4a

Create and display a contingency table with the type of anesthetic defining the rows and nausea after the surgery as the columns. Display the margins for this table as well.

Also make a side-by-side bar graph showing the nausea (Yes vs No) on the horizontal axis and color-coded bars to indicate the type of anesthetic.

Comment on any potential relationships between nausea and type of anesthetic you see in the graph.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4a -|-|-|-|-|-|-|-|-|-|-|-

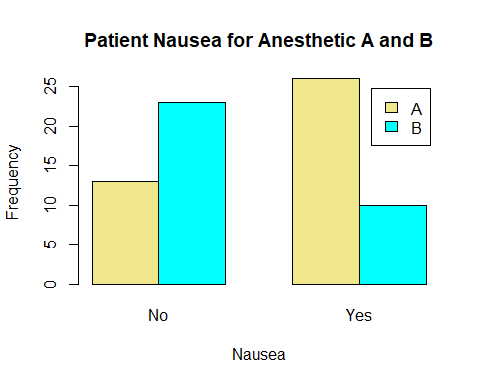
# Create contingency table from anesthesia data  
require(DS705data)

## Loading required package: DS705data

data("anesthesia")  
anesthetic\_nausea <- with(anesthesia,table(anesthetic,nausea))  
addmargins(anesthetic\_nausea)

## nausea  
## anesthetic No Yes Sum  
## A 13 26 39  
## B 23 10 33  
## Sum 36 36 72

#anesthetic\_nausea  
#create barplot  
barplot(anesthetic\_nausea,xlab="Nausea", ylab="Frequency", main = "Patient Nausea for Anesthetic A and B",  
 col=c("khaki","cyan"),legend=rownames(anesthetic\_nausea),beside=T)



From inspection of the bar graph, it appears anesthetic A is much more likely to cause nausea than anethetic B.

### Part 4b

Conduct a chi-square test for independence at the 10% level.

State the hypotheses (in words) and conclusion of the test. Use R to compute the test statistic, degrees of freedom, and P-value. Include the P-value in your written conclusion.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4b -|-|-|-|-|-|-|-|-|-|-|-

Ho: The type of anesthetic and whether or not patients become nauseated are independent of each other. Ha: There is a dependency between an anesthetic and nausea.

#p 521 chisq test for independence  
chisq.test(anesthetic\_nausea, correct=FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: anesthetic\_nausea  
## X-squared = 9.4545, df = 1, p-value = 0.002106

The p-value for the test is 0.002106. Thus at a 10% level of significance, we can conclude that there is evidence that an anesthetic is associated with nausea.