Inference for Categorical Data

Josh Jarvey

06/02/2020

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

## 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 a matrix to store the data within. first is bank by outstanding and not, then by credit union in the 2nd row.  
financial\_ratings = matrix(c(70,150,66,81),nrow = 2, byrow=TRUE)  
 #label the columns and rows accordingly.  
colnames(financial\_ratings) = c("Outstanding", "Not Outstanding")  
rownames(financial\_ratings) = c("Bank", "Credit Union")

### 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 -|-|-|-|-|-|-|-|-|-|-|-

#perform a prop.test to check for the true population of credit unions that recieve an outstanding rating.  
prop.test(66,81, correct = TRUE)

##   
## 1-sample proportions test with continuity correction  
##   
## data: 66 out of 81, null probability 0.5  
## X-squared = 30.864, df = 1, p-value = 2.767e-08  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.7098048 0.8893129  
## sample estimates:  
## p   
## 0.8148148

At a 95% significance level, the true population proportion of outstanding ratings for credit unions lies between 70.98% to 88.93%.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

#re-jigging the matrix: this puts credit union's outstanding first, then bank's outstanding  
outstanding = c(financial\_ratings[2,1],financial\_ratings[1,1])  
 #re-jigging the matrix: this puts credit union's not outstanding first, then bank's not outstanding  
not\_outstanding = c(financial\_ratings[2,2],financial\_ratings[1,2])  
  
 #prop.test for cu/bank outstanding, vs cu/bank not outstanding.  
prop.test(outstanding,not\_outstanding, correct = FALSE)

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: outstanding out of not\_outstanding  
## X-squared = 26.33, df = 1, p-value = 2.878e-07  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.2318293 0.4644670  
## sample estimates:  
## prop 1 prop 2   
## 0.8148148 0.4666667

At a 95% significance level, the population proportion of outstanding ratings for credit unions is 23.18% to 46.44% greater than the proportion of outstanding ratings for banks.

### Part 1d

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

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

#pull the number of not outstanding ratings for banks  
bank\_not\_outstanding = financial\_ratings[1,2]  
 #pull the number of outstanding ratings for banks  
bank\_outstanding = financial\_ratings[1,1]  
  
 #not outstanding / row total (i.e. outstanding + not outstanding). Multiply by 100 to get percent.  
bank\_not\_outstanding/(bank\_outstanding + bank\_not\_outstanding)\*100

## [1] 68.18182

A bank has a 68.18% chance of reciving a non-outstanding rating.

### Part 1e

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

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

#pull the number of not outstanding ratings for credit unions  
creditUnion\_not\_outstanding = financial\_ratings[2,2]  
 #pull the number of outstanding ratings for credit unions  
creditUnion\_outstanding = financial\_ratings[2,1]  
  
 #not outstanding / row total (i.e. outstanding + not outstanding). Multiply by 100 to get percent.  
creditUnion\_not\_outstanding/(creditUnion\_outstanding + creditUnion\_not\_outstanding)\*100

## [1] 55.10204

A credit union has a 55.10% chance of reciving a non-outstanding rating.

### Part 1f

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

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

#storing the proportion of "not outstanding" risk into variables.  
bank\_prop\_not\_outstanding = (bank\_not\_outstanding/(bank\_outstanding + bank\_not\_outstanding))  
creditUnion\_prop\_not\_outstanding = (creditUnion\_not\_outstanding/(creditUnion\_outstanding + creditUnion\_not\_outstanding))  
  
 #interpreting the relative risk ratio as a percentage.   
 #take the absolute value of 1 minus the resulting ratio calculation.  
 #then multiply this by 100, and you have a percentage.   
abs(1-(bank\_prop\_not\_outstanding/creditUnion\_prop\_not\_outstanding))\*100

## [1] 23.73737

There is a 23.73% greater chance for banks to be rated “not outstanding” relative to credit unions.

### Part 1g

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

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

#load mosaic package.  
library(mosaic)

## Warning: package 'mosaic' was built under R version 3.6.3

## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 3.6.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Loading required package: lattice

## Loading required package: ggformula

## Warning: package 'ggformula' was built under R version 3.6.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.3

## Loading required package: ggstance

## Warning: package 'ggstance' was built under R version 3.6.3

##   
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Warning: package 'mosaicData' was built under R version 3.6.3

## Loading required package: Matrix

## Registered S3 method overwritten by 'mosaic':  
## method from   
## fortify.SpatialPolygonsDataFrame ggplot2

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.  
##   
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.  
##   
## Have you tried the ggformula package for your plots?

##   
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
## quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

#credit union is the 'odds2' value here.   
oddsRatio(financial\_ratings, verbose = TRUE)

##   
## Odds Ratio  
##   
## Proportions  
## Prop. 1: 0.3182   
## Prop. 2: 0.449   
## Rel. Risk: 1.411   
##   
## Odds  
## Odds 1: 0.4667   
## Odds 2: 0.8148   
## Odds Ratio: 1.746   
##   
## 95 percent confidence interval:  
## 1.084 < RR < 1.837   
## 1.134 < OR < 2.688   
## NULL

## [1] 1.746032

The odds of a credit union being rated as Outstanding are 0.8148 to 1.

### Part 1h

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

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

#bank is the 'odds1' value here.   
oddsRatio(financial\_ratings, verbose = TRUE)

##   
## Odds Ratio  
##   
## Proportions  
## Prop. 1: 0.3182   
## Prop. 2: 0.449   
## Rel. Risk: 1.411   
##   
## Odds  
## Odds 1: 0.4667   
## Odds 2: 0.8148   
## Odds Ratio: 1.746   
##   
## 95 percent confidence interval:  
## 1.084 < RR < 1.837   
## 1.134 < OR < 2.688   
## NULL

## [1] 1.746032

The odds of a bank being rated as Outstanding are 0.4667 to 1.

### Part 1i

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

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

#!!!IS THIS NEEDED????!!!!   
 #swap outstanding and not outstanding columns. To get proper odds ratio for "outstanding  
#financial\_ratings = matrix(c(150,70,81,66),nrow = 2, byrow=TRUE)  
#colnames(financial\_ratings) = c("Not Outstanding", "Outstanding")  
#rownames(financial\_ratings) = c("Bank", "Credit Union")  
  
 #odds ratio is setting credit union as the numerator (odds2), and bank as the denominator (odds1).   
oddsRatio(financial\_ratings, verbose = TRUE)

##   
## Odds Ratio  
##   
## Proportions  
## Prop. 1: 0.3182   
## Prop. 2: 0.449   
## Rel. Risk: 1.411   
##   
## Odds  
## Odds 1: 0.4667   
## Odds 2: 0.8148   
## Odds Ratio: 1.746   
##   
## 95 percent confidence interval:  
## 1.084 < RR < 1.837   
## 1.134 < OR < 2.688   
## NULL

## [1] 1.746032

The odds of being rated as outstanding for credit unions are 1.746 times the odds of an outstanding rating for banks.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

The odds of being rated as outstanding for credit unions are 74.6% more than the odds of an outstanding rating for banks.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

#check the confidence interval of the OR.  
oddsRatio(financial\_ratings, verbose = TRUE)

##   
## Odds Ratio  
##   
## Proportions  
## Prop. 1: 0.3182   
## Prop. 2: 0.449   
## Rel. Risk: 1.411   
##   
## Odds  
## Odds 1: 0.4667   
## Odds 2: 0.8148   
## Odds Ratio: 1.746   
##   
## 95 percent confidence interval:  
## 1.084 < RR < 1.837   
## 1.134 < OR < 2.688   
## NULL

## [1] 1.746032

At a 95% significance level, the odds of an outstanding rating for a credit union are 1.134 to 2.688 times the odds of an outstanding rating for a bank.

### 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 confidence interval displayed above does not contain 1 which would be interpreted as being “equally likely”, therefore yes, there does appear to be significant evidence to support that if it is a rating is outstanding, it’s more likely to be a credit union than a bank.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

#create a matrix storing the data.  
market\_purchase = matrix(c(28,44,54,39), nrow = 1, byrow=TRUE)  
colnames(market\_purchase) = c("Under 18", "18 to 35", "36 to 50", "Over 50")  
rownames(market\_purchase) = c("Count")  
  
 #pull out the observations into a vector.  
observed = market\_purchase[1,]  
 #specify the stated probability proportions.  
prob = c(0.13,0.32,0.38,0.17)  
  
 #perform the chi-sq test with the observations and probabilities.   
chisq.test(x=observed,p=prob)

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

H\_null: Under18 = 0.13, 18 to 35 = 0.32, 36 to 50 = 0.38, Over 50 = 0.17 H\_Alternative: At least one of these proportions differs.

At a 95% significance level, there is enough evidence to reject that the population proportion of all online shopping by customer’s age groups is equal to the following: Under18 = 0.13, 18 to 35 = 0.32, 36 to 50 = 0.38, Over 50 = 0.17 (p-value = 0.02998)

### 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 -|-|-|-|-|-|-|-|-|-|-|-

#to check expected counts, we must first take the sum of all the observations. We then multiple the sum against each claimed probabilty and ensure they are at least 5 or greater.   
sum(market\_purchase)\*prob

## [1] 21.45 52.80 62.70 28.05

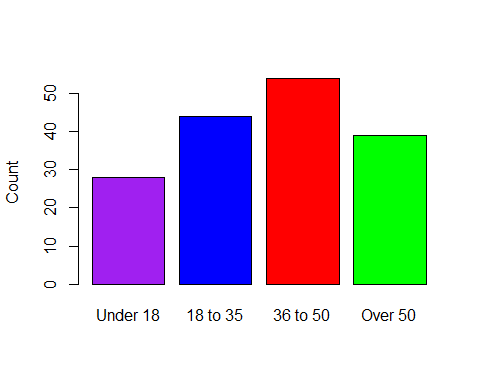
The expected cell counts are all greater than 5, therefore it is appropiate to perform the chi-squared goodness of fit test to this data.

### Part 2c

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

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

#create a barplot using the observed values sorted by age bracket. Set colors.   
barplot(observed, names.arg = c("Under 18", "18 to 35", "36 to 50", "Over 50"),  
 ylab = "Count", col = c("purple", "blue", "red","green"))



There is an increasing trend that as an individuals age increases, so does their likeliness of making online purchases - this is true until age 50 or over, which shows a decline.

## 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 a matrix storing the data.  
seatbelt = matrix(c(17,12,2,8), nrow = 2, byrow=TRUE)  
colnames(seatbelt) = c("Teen", "Senior")  
rownames(seatbelt) = c("Always Wear","Not Always Wear")

### 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 -|-|-|-|-|-|-|-|-|-|-|-

#perform a fisher exact test since the observations are too small for a chi-sq test.   
fisher.test(seatbelt)

##   
## Fisher's Exact Test for Count Data  
##   
## data: seatbelt  
## 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

H\_null: The proportion of Teens who wear seatbelts is equal to the proportion of Seniors who wear seatbelts. H\_alternative: The proportion of Teens who wear seatbelts is not equal to the proportion of Seniors who wear seatbelts.

At a 95% level of significance, there is not enough evidence to conclude that the proportion of Teens who wear seatbelts is equal to the proportion of Seniors that wear seatbelts (p-value = 0.06483).

## 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.

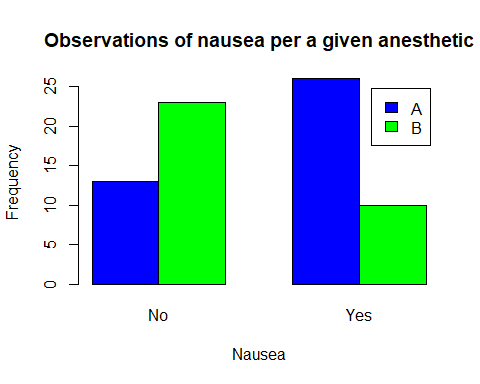
Comment on any potential relationships between nausea and type of anesthetic you see in the graph.

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

#load in the anesthesia data set.  
library(DS705data)  
data("anesthesia")  
  
 #pull a subset of the data - anesthetic and nausea, as a 2x2 table.  
drug\_nausea = table(anesthesia$anesthetic,anesthesia$nausea)  
  
 #display table with flanking total margins.  
addmargins(drug\_nausea)

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

#create a barplot using the 2x2 table. Each bar "set" is a column based on nausea. Within the set, we use side-by-side for the given anesthetic. We set distinct colors, and add a legend using the rownames from the table as the setting.   
barplot(drug\_nausea,xlab = "Nausea", ylab = "Frequency", col = c("blue", "green"), legend = rownames(drug\_nausea), beside = TRUE, main = "Observations of nausea per a given anesthetic")



It appears that the anesthetic A (blue) is more likely to cause nausea than anesthetic B (green).

### 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 -|-|-|-|-|-|-|-|-|-|-|-

#perform chi-sq() test for independence between the type of anesthetic used and if nausea is experienced. This uses the original dataset, and will automatically calculate sums and totals.   
chisq.test(anesthesia$anesthetic,anesthesia$nausea)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: anesthesia$anesthetic and anesthesia$nausea  
## X-squared = 8.0559, df = 1, p-value = 0.004535

H\_null: For patients undergoing reconstructive knee surgery, the type of anesthetic administered and having nausea are independent. H\_alternative: For patients undergoing reconstructive knee surgery, the type of anesthetic administered and having nausea are associated.

At a 90% significance level, there is enough evidence to reject claim that the type of anesthetic used is independent of experiencing nausea for patients undergoing reconstructive knee surgery (p-value = 0.004535).