Lab Assignment: Inference for Categorical Data

Spencer Swartz

March 13, 2017

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

Chondromalatia patellae (CP) is a painful inflammation of the patella (kneecap). It can be diagnosed without error arthroscopically. However, a less invasive diagnostic test called Clarkeâs Sign test can also be used to diagnose CP but is not 100% accurate. The following contingency table contains data for 106 patients who were first diagnosed for CP using the Clarkeâs Sign test and then went on to have arthroscopic surgery to confirm or deny the diagnosis (Doberstein ST, Romeyn RL, Reineke DM , J Athl Train. 2008 Apr-Jun;43(2):190-6.).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Have CP | No CP | Total |
| Positive Clarke's Sign | 9 | 27 | 36 |
| Negative Clarke's Sign | 14 | 56 | 70 |
| Total | 23 | 83 | 106 |

### Part 1a

Conduct a chi-square test to determine if the outcome of the Clarke's Sign test and the presence of CP are associated. Let .

### Answer 1a

clarke <- matrix(c(9,27,14,56),nrow=2)  
clarke

## [,1] [,2]  
## [1,] 9 14  
## [2,] 27 56

stats::prop.test(clarke,correct = F)

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: clarke  
## X-squared = 0.34982, df = 1, p-value = 0.5542  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.1574690 0.2894753  
## sample estimates:  
## prop 1 prop 2   
## 0.3913043 0.3253012

: The distribution of the outcome is independent of the Clarke result. : There is a difference in the distribution of responses to the Clarke result.

At , we do not have enough evidence to reject (p = 0.5542). this means that the Clarke test and CP may not be associated.

### Part 1b

For the population of all patients who have CP, construct and interpret a 95% confidence interval for the proportion of positive Clarke's Sign tests.

### Answer 1b

prop.test(9,23,correct = FALSE)

##   
## 1-sample proportions test without continuity correction  
##   
## data: 9 out of 23, null probability 0.5  
## X-squared = 1.087, df = 1, p-value = 0.2971  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.2215762 0.5921448  
## sample estimates:  
## p   
## 0.3913043

We are 95% confident that population is between .22 and .60 standard deviations from the mean

### Part 1c

Compare the proportions of patients without CP who get a positive test to the proportion of patients with CP who get a positive test. Do this by computing a 95% CI for difference of proportions of positive tests of those who have CP and those who don't. Interpret the result.

### Answer 1c

posCl = c(27,9)  
CP = c(83,23)  
prop.test(posCl,CP,correct = F)

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: posCl out of CP  
## X-squared = 0.34982, df = 1, p-value = 0.5542  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.2894753 0.1574690  
## sample estimates:  
## prop 1 prop 2   
## 0.3253012 0.3913043

The difference in proportions between those without CP and tested positive and those who have CP and test positive is between -0.29 and 0.16.

### Part 1d

If one person is selected randomly, what is the risk of having CP given that the patient has a positive Clarkeâs Sign test? (âriskâ means probability)

### Answer 1d

CPpos= 9/36  
CPpos

## [1] 0.25

### Part 1e

If one person is selected randomly, what is the risk of having CP given that the patient has a negative Clarkeâs Sign test?

### Answer 1e

CPneg= 14/70  
CPneg

## [1] 0.2

### Part 1f

Compute the relative risk of having CP for those who have a positive Clarke's Sign test compared to those who have a negative test.

### Answer 1f

rrCP <- CPpos/CPneg  
rrCP

## [1] 1.25

### Part 1g

What are the odds of having CP given that the patient has a positive Clarkeâs Sign test?

### Answer 1g

oddsCPpos <- CPpos/(27/36)  
oddsCPpos

## [1] 0.3333333

### Part 1h

What are the odds of having CP given that the patient has a negative Clarkeâs Sign test?

### Answer 1h

oddsCPneg <- CPneg/(56/70)  
oddsCPneg

## [1] 0.25

### Part 1i

Compute the odds ratio (OR) of having CP for those who have a positive Clarke's Sign test compared to those who have a negative test.

### Answer 1i

require("mosaic")

## Loading required package: mosaic

## Warning: package 'mosaic' was built under R version 3.3.3

## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 3.3.3

##   
## 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: ggplot2

## Loading required package: mosaicData

## Warning: package 'mosaicData' was built under R version 3.3.3

## Loading required package: Matrix

##   
## 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.

##   
## Attaching package: 'mosaic'

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

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

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cov, D, 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

clarke <- matrix(c(9,14,27,56),nrow=2)  
OR <- oddsRatio(clarke, verbose = TRUE)

##   
## Odds Ratio  
##   
## Proportions  
## Prop. 1: 0.25   
## Prop. 2: 0.2   
## Rel. Risk: 0.8   
##   
## Odds  
## Odds 1: 0.3333   
## Odds 2: 0.25   
## Odds Ratio: 0.75   
##   
## 95 percent confidence interval:  
## 0.3838 < RR < 1.668   
## 0.2886 < OR < 1.949   
## NULL

OR

## [1] 0.75

### Part 1j

Write an interpretation of the odds ratio as a percent.

### Answer 1j

The Odds of having CP given a Positive Clarke's Test is 0.75 times as large as the odds for a Negative Clarke's Test.

### Part 1k

Construct a 95% confidence interval for the OR of having CP for those who have a positive test compared to those who have a negative test. Interpret the interval, leaving the endpoints as a multiples.

### Answer 1k

OR

## [1] 0.75

With 95% Confidence, the odds of having CP given a Positive Clarke's test is between 63 and 90 percent as high as for having CP given a Negative Clarke's test.

## Exercise 2

There are said to be six general personality types for dogs (see <http://www.fl-k9.com/personalities.htm> for more details about dog personalities if you are interested). Suppose it has been hypothesized that the distribution of dog personalities is as follows.

|  |  |
| --- | --- |
| Personality Type | Hypothesized Proportion |
| Aggressive | 0.15 |
| Confident | 0.09 |
| Outgoing | 0.29 |
| Adaptable | 0.21 |
| Insecure | 0.12 |
| Independent | 0.14 |

### Part 2a

Consider a random sample of 125 dogs that have been categorized by personality in the table below.

|  |  |
| --- | --- |
| Personality Type | Observed frequency |
| Aggressive | 14 |
| Confident | 15 |
| Outgoing | 22 |
| Adaptable | 30 |
| Insecure | 19 |
| Independent | 25 |
| Total | 125 |

Create vectors for the counts and hypothesized probabilities, then conduct the chi-square test to determine if the distribution of dog personalities given above is correct. Include all parts of the test.

### Answer 2a

obs <- c(14,15,22,30,19,25)  
prop <- c(.15,.09,.29,.21,.12,.14)  
chisq.test(x=obs,p=prop)

##   
## Chi-squared test for given probabilities  
##   
## data: obs  
## X-squared = 12.872, df = 5, p-value = 0.02461

: At least one differs from another

At we do not reject (p-value = .025). The distributions of Personality types in dogs look to be accurate to those identified in the table above.

### 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

125\*prop

## [1] 18.75 11.25 36.25 26.25 15.00 17.50

All proportions of the personality types are above 5 when computing the expected cell counts, this means that our chi-square test is accurate, as identified in the lecture.

## 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 20 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

age <- c("teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","teen","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior","senior")  
belt <- c("always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","not always wear","not always wear","not always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","always wear","not always wear","not always wear","not always wear","not always wear","not always wear","not always wear","not always wear","not always wear")  
belts <- data.frame(age, belt)  
#belts <- data.frame(matrix(c(17,3,12,8),nrow=2))  
#rownames(belts) <- c("always wear", "not always wear")  
colnames(belts) <- c("agerange", "beltanswer")  
table(belt,age)

## age  
## belt senior teen  
## always wear 12 17  
## not always wear 8 3

### 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

chisq.test(age,belt)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: age and belt  
## X-squared = 2.0063, df = 1, p-value = 0.1567

results = chisq.test(age,belt)  
results$expected

## belt  
## age always wear not always wear  
## senior 14.5 5.5  
## teen 14.5 5.5

Age and Belt preference are independent

Age and Belt preference are associated

Conclusion:

Do not reject at There is insufficient evidence in this sample to claim that Age group and Belt preference are associated. (p-value = 0.16)

With all the expected values above 5 we can assume that the chi-squared test is valid as noted in the lecture.