Lab Assignment: Inference for Categorical Data

Paul Isaacson

Mar. 9, 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

cp <- matrix(c(9,14,27,56), nrow=2)  
chisq.test(cp)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: cp  
## X-squared = 0.11742, df = 1, p-value = 0.7318

### 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  
## 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 for the population with CP, The Clarke's Sign test will yield a positive result between 22.16% and 59.21% of the time.

### 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

prop.test(c(9, 27), c(23, 83), correct=FALSE)

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: c(9, 27) out of c(23, 83)  
## 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

We are 95% that the difference in proportions for the population of patients without CP who get a positive test and for the population of patients with CP who get a positive test is -0.18 to 0.29.

### 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€œriskâ€ means probability)

### Answer 1d

prop1 <- 9/36  
prop1

## [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

prop2 <- 14/70  
prop2

## [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

prop1 / prop2

## [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

odds1 <- 9/27   
odds1

## [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

odds2 <- 14/56   
odds2

## [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

odds1 / odds2

## [1] 1.333333

### Part 1j

Write an interpretation of the odds ratio as a percent.

### Answer 1j

The odds of having CP is 33% higher when the Clarke's Sign test is positive versus when it is negative.

### 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

oddsRatio(cp, 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

## [1] 0.75

We are 95% confident that the odds of having CP when the test is negative is 28.90% to 195% that of having CP when the test is positive. (Note: the odds ratio function in R is computing the odds ratio as Odds 2 / Odds 1 and the confidence interval is based on that)

## 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

observed <- c(14, 15, 22, 30, 19, 25)  
proportions <- c(0.15, 0.09, 0.29, 0.21, 0.12, 0.14)  
chisq.test(x=observed, p=proportions)

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

At , we fail to reject the null hypothesis test (p = 0.02461). That is to say, we cannot say that the observed values differ significantly from described proportions.

### 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\*proportions

## [1] 18.75 11.25 36.25 26.25 15.00 17.50

To get the expected values we are multiplying the total sample size with the proportions. We have a requirement that all expected values (not observed values) have a value greater than 5. This sample passes that requirement.

## 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

seatBelt <- matrix(c(17,3,12,8), nrow=2)

### 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

result <- chisq.test(seatBelt)  
result$expected

## [,1] [,2]  
## [1,] 14.5 14.5  
## [2,] 5.5 5.5

result

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

At , we fail to reject the null hypothesis (p-value = 0.1567). We do not have enough evidence to say that the age group and seat belt wearing are associated. Also, we check that all expected values are > 5. In this case they are, so we will stick with the chi-square test vs the Fisher's Exact test.