Lab 10 Solutions

PH 142 GSIs

11/8/2019

Instructions

1) We will be using data tidying functions and plotting functions to work through this lab. Load the required packages into our R session.

```
### SOLUTION
library(dplyr)
```

```
##
## 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
```

```
library(ggplot2)
```

Chi Squared Testing

As the textbook mentions, the chi-square statistic is a measure of how far the observed counts in a two-way table are from the expected counts. The formula for the statistic is:

$$X^2 = \sum \frac{(count_{observed} - count_{expected})^2}{count_{expected}}$$

The sum is over all cells in the table. That is, there are as many terms in the sum as there are cells in the table. Each term in the sum is called a X^2 component.

Question 1: Melanoma Adapted from Baldi and Moore Question 21.29

Melanoma is a rare form of skin cancer that accounts for the great majority of skin cancer fatalities. UV exposure is a major risk for melanoma. A question we would like to explore is if the body parts which have increased sun exposure are more susceptible to melanoma. A random sample of 310 women diagnosed with melanoma were classified according to the known location of the melanoma on their bodies. Here are the results:

Location	Head/Neck	Trunk	Upper Limbs	Lower Limbs
Count	45	80	34	151
Expected	77.5	77.5	77.5	77.5

(a) Assuming each of the four locations represent roughly equal skin areas, fill in the expected counts for the four areas of the body.

(b) What are the assumptions for completing a Chi Squared test? Are the conditions met for this example?

Your answer here:

Solution: 1. Fixed n of observations 2. All observations are independent of one another. 3. Each observation falls into just one of the k mutually exclusive categories. 4. The probability of a given outcome is the same for each observation. 5. At least 80% of the cells have an expected value of 5 or more and all cells have an expected value of at least 1.

Yes these conditions are met. (Discuss with the class)

(c) Perform a chi-squared test for goodness of fit. State the null and alternative hypotheses. Use R to calculate your test statistics and report this value. Calculate and report the p-value.

$H_0 : p_{H/N} = p_T = p_{UL} = p_{LL}$ H_A : At least one of the above specified P_k is different than the others. K being head/neck, trunk, upper limbs, or lower limbs.

```
#Replace with your code.
#Solution:
chisq.test(x = c(45, 80, 34, 151),
           p = c(0.25, 0.25, 0.25, 0.25))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(45, 80, 34, 151)
## X-squared = 107.83, df = 3, p-value < 2.2e-16
```

Question 2 Adapted from: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ChiSquare/BS704_HypothesisTesting-ChiSquare2.html

The National Center for Health Statistics (NCHS) provided data on the distribution of weight (in categories) among Americans in 2002. The distribution was based on specific values of body mass index (BMI).

Underweight was defined as BMI< 18.5, Normal weight as BMI between 18.5 and 24.9, overweight as BMI between 25 and 29.9 and obese as BMI of 30 or greater.

Americans in 2002 were distributed as follows: 2% Underweight, 39% Normal Weight, 36% Overweight, and 23% Obese. Suppose we want to assess whether the distribution of BMI is different in the Framingham Offspring sample.

Using data from the n = 3,326 participants who attended the seventh examination of the Offspring in the Framingham Heart Study we created the BMI categories as defined and observed the following:

BMI	Underweight	Normal Weight	Overweight	Obese	Total
Count	20	932	1374	1000	3326
Expected	66.5	1297.1	1197.4	765	3326

(a) State the null and alternative hypotheses. Fill in the expected counts for this statistic.

Hypotheses:H_0 p= 0.02, $p_2 = 0.39$, $p_3 = 0.36$, $p_4 = 0.23$ H_A: At least one of p_k is not equal to the proportion stated in the null hypothesis.

(b) For this test, we will use a 5% signficance level. For what value of the test statistic under Chi-Squared distribution will we reject the null hypothesis? (Hint: What are the degrees of freedom for this test?)

```
#Replace with your code
qchisq(0.95, df = 3)
```

```
## [1] 7.814728
```

Your answer here: 7.8147 We will reject the null for a test statistic at least as extreme as 7.8147 under the Chi Squared distribution with 3 degrees of freedom.

(c) Perfrom a chi-squared test for goodness of fit. Calculate and report your test statistic. Calculate and interpret your p-value. What are your conclusions?

```
#Replace with your code
chisq.test(x = c(20, 932, 1374, 1000),
           p = c(0.02, 0.39, 0.36, 0.23))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(20, 932, 1374, 1000)
## X-squared = 233.58, df = 3, p-value < 2.2e-16
```

#Solution:

Test statistic = 233.58

P-value is approximately equal to 0. There is approximately no chance of observing a test statistic at least as extreme as 233.58 so we reject the null hypothesis that the probabilities are the same as the study in 2002.

Chapter 22 - Chi Squared Test for Independence

The chi-square test for a two-way table with r rows and c columns uses critical values from the chi-square distribution with (r-1)(c-1) degrees of freedom.

- Side question: Think about how might we determine a p-value for a chi-square test statistic?

In research, we are often interested in making the assumption that two explanatory variables are (mostly) independent. Independence is actually one condition which permits us to include multiple explanatory variables in a linear regression (i.e. the line of best fit model that you explored in the first part of the course). Thus, the Chi-Square test of independence can be quite useful as a tool to explore whether two categorical variables show substantial dependence.

In the second part of this lab, we proceed to walk through the data cleaning, visualization, and analysis required to carry out a Chi Square test for two-way tables.

3. Intro and Data are from the text (Ex 22.40 Do angry people have more heart disease??):

*NOTE: If at any point, you are unclear how to use dplyr to create a variable, feel free to manually calculate, and use the following code to add a manual variable:

```
# chd_by_anger_level <-
#   chd_by_anger_level %>%
#   ### input the 6 values below
#   mutate(new_variable = c( , , , , , ))
```

People who get angry easily tend to have more heart disease. That's the conclusion of a study that followed a random sample of 12,986 people from three locations for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data for the 8474 people in the sample who had normal blood pressure. 18 CHD stands for "coronary heart disease." This includes people who had heart attacks and those who needed medical treatment for heart disease.

	Low anger	Moderate anger	High anger
CHD	53	110	27
No CHD	3057	4621	606

Let's explore if these data support the conclusion of the study!

3a) Based on the two-way table above and the framework of the study, write out the null and alternative hypotheses that we will be exploring via a Chi-Squared test.

Write your answer here.

SOLUTION: The null hypothesis is that anger levels and coronary heart disease outcomes are independent. The alternate hypothesis is that anger level is related to heart disease outcomes.

We need to figure out what the expected counts of heart disease and anger levels would be if the two categorical variables are independent. Here is a data set of our two-way table values:

3b) Using our dplyr tools, add variables for row, column totals, and sample size, and fill out the two-way table below: [HINT: The code for computing row totals is given. Use this framework to compute column totals]

```
### Code your answer here.

### SOLUTION:
chd_by_anger_level <-
  chd_by_anger_level %>%
  group_by(heart_disease) %>%
  mutate(row_total = sum(actual_count)) %>%
  ungroup()

chd_by_anger_level <-
  chd_by_anger_level %>%
  group_by(anger_level) %>%
  mutate(column_total = sum(actual_count))  %>%
  ungroup()
```

```
chd_by_anger_level <-
  chd_by_anger_level %>%
  group_by(heart_disease) %>%
  mutate(sample_size = sum(column_total)) %>%
  ungroup()

### Alternatively, once filling out the table, can compute sample size manually.
chd_by_anger_level <-
  chd_by_anger_level %>%
  mutate(sample_size = 8284+190)
```

	Low anger	Moderate anger	High anger	Row Total
CHD	53	110	27	190
No CHD	3057	4621	606	8284
Column Total	3110	4731	633	8474

3c) Use the following formula from lecture notes to create a column for expected counts:

$$E_i = \frac{\text{row total} \times \text{col total}}{\text{overall total}}$$

```
### SOLUTION
chd_by_anger_level <-
  chd_by_anger_level %>%
  mutate(expected_count = row_total * column_total / sample_size)
```

3d) Before moving forward with analysis, confirm two requirements for the Chi-Squared test of independence, names:

- No more than 20% of the expected counts are smaller than 5.0, and
- All individual expected counts are 1.0 or greater.

```
### SOLUTION (you can check the data visually)
chd_by_anger_level %>%
  summarize(proportion_small_exp_counts = sum(expected_count < 5) / n(),
            very_small_exp_counts       = sum(expected_count < 1))
```

```
## # A tibble: 1 x 2
##   proportion_small_exp_counts very_small_exp_counts
##                         <dbl>                 <int>
## 1                           0                     0
```

SOLUTION: There are no expected counts smaller than 5, so we have satisfied the above conditions.

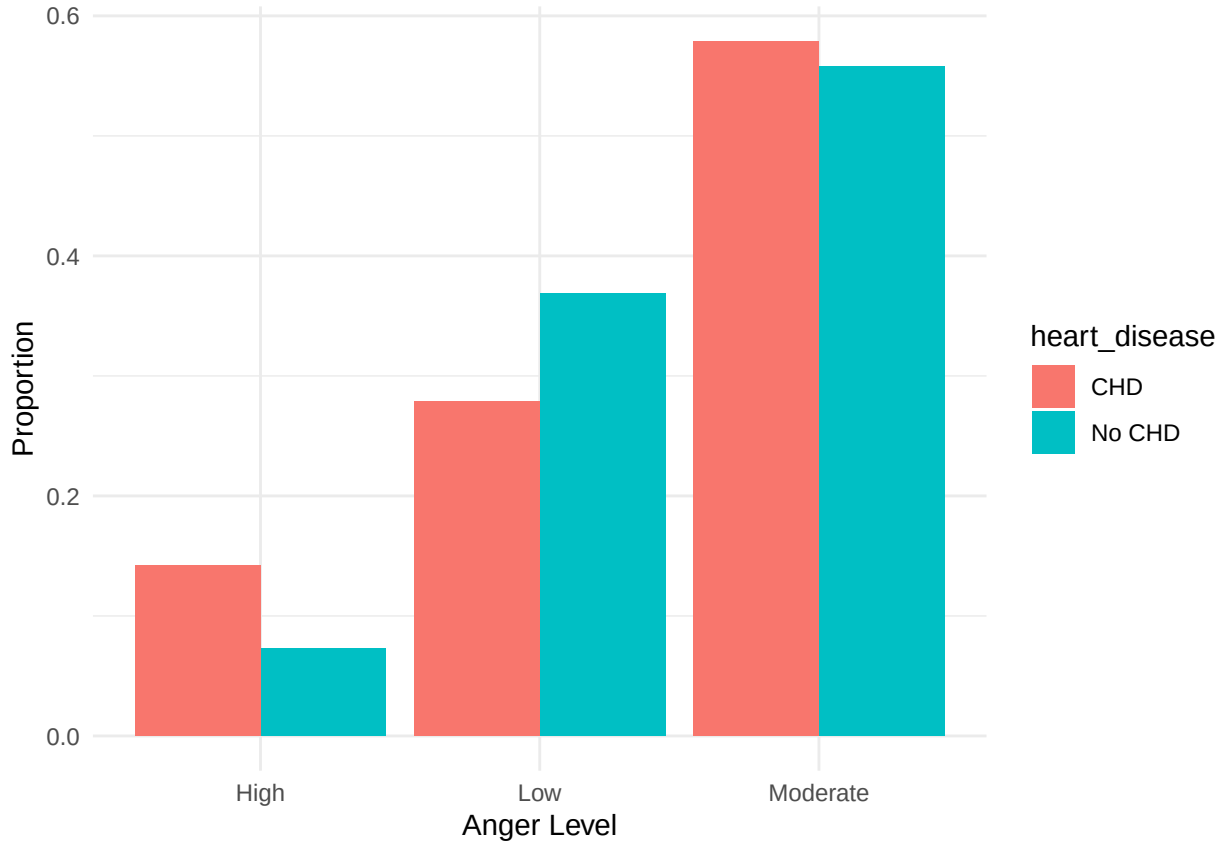
While, we are set to move forward with a Chi-Square test, let's practice visualizing our data to see if there may be a significant relationship between heart disease and anger.

3e) First, calculate the probability of anger level conditional on CHD Status

```
### SOLUTION:
chd_by_anger_level <-
  chd_by_anger_level %>%
  group_by(heart_disease) %>%
  mutate(cond_prop_anger = actual_count / sum(actual_count)) %>%
  ungroup()
```

3f) Reference your Ch 22 notes and create a dodged bar graph of anger probabilities, dodged by CHD status. Interpret the results.

```
### Solution:
chd_by_anger_level %>%
ggplot(aes(x = anger_level, y = cond_prop_anger)) +
  geom_bar(aes(fill = heart_disease), stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(y = "Proportion", x = "Anger Level", main = "Conditional distribution of anger level by CHD statu
```



Write your interpretation here. Solution: There are some peculiar differences but it's a close call that leaves us interested in whether the Chi-Square test will be significant or not. We do see that high-anger people are more likely to have CHD.

Now, we are ready to move forward with our Chi-Square test of independence.

3g) Compute the Chi-Square test statistic. [Optional: Practice dynamic coding. Assign important variables to your environment once, and only call the variable names when computing the final test statistic.]

```r
chd_by_anger_level <-
  chd_by_anger_level %>%
  mutate(diff_sq = (actual_count - expected_count)^2,
         ratio = diff_sq / expected_count)

### SOLUTION:
chi_square_test_statistic <-
  chd_by_anger_level %>%
  summarise(test_stat = sum((actual_count - expected_count)^2 / expected_count)) %>%
  pull(test_stat)

deg_of_freedom <-
  (3-1)*(2-1)
```

3h) Determine the p-value of your Chi-Square test statistic and interpret the p-value for a 5% level Chi-Square test in the context of this problem.

```r
### SOLUTION:
pchisq(chi_square_test_statistic, deg_of_freedom, lower.tail = FALSE)
```

```
## [1] 0.0003228312
```

SOLUTION: The p-value is extremely low, so we reject the null hypothesis that anger level is independent of coronary heart disease.

3i) How might we have tested for independence between anger and heart disease prevalence during the probability section of the course? Would we have found that anger and heart disease are independent using our old problem-solving process? How does this method differ from comparing conditional probabilities?

SOLUTION: In the probability section, we would have seen if $P(A|B)$ was exactly equal to $P(A)$ for any two variables A and B. If the probabilities were not exactly equal, we would claim that A and B were not independent. With the Chi-Squared test, we require that the probabiltiies are significantly different before we reject the hypothesis that A and B are independent.

Submission

Please submit your lab *directly* to Gradescope. You can do this by knitting your file and downloading the PDF to your computer. Then navigate to Gradescope.com to submit your assignment. Here is a tutorial if you need help: https://www.gradescope.com/get_started. Scroll down on that page to "For students:submitting homework".