Lab 10: Chi Squared Testing

Felicia Liu

11/12/2020

- Due date: Friday, Nov 13th 23:59pm.
- Submission process: Follow the submission instructions on the final page. Make sure you do not remove any `\newpage` tags or rename this file, as this will break the submission.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration! **You must knit correctly before submitting.**
- If your code runs off the page of the knitted PDF then you will LOSE POINTS! To avoid this, have a look at your knitted PDF and ensure all the code fits in the file (you can easily view it on Gradescope via the provided link after submitting). If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

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.

```
library(dplyr)
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.

Part 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	XXXX	XXXX	XXXX	XXXX

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

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

2. What are the assumptions for completing a Chi Squared test? Are the conditions met for this example?

1) Fixed n 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 an 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.

3. 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, where k is head/neck, trunk, upper limbs, or lower limbs.

```
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
```

```
test_stat <- 107.83
p_value <-0


test_stat
```

```
## [1] 107.83
```

```
p_value
```

```
## [1] 0
```

```
check_problem3()
```

```
## [1] "Checkpoint 1 Passed: You found the correct test statistics!"
## [1] "Checkpoint 2 Passed: You found the correct p-value!"
##
## Problem 3
## Checkpoints Passed: 2
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

Part 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	XXXX	XXXXXX	XXXXXX	XXX	3326

4. State the null and alternative hypotheses. Fill in the expected counts for this statistic.

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

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

5. 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
reject_value <- qchisq(0.95, df = 3)


reject_value
```

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

```
check_problem5()
```

```
## [1] "Checkpoint 1 Passed: You found the correct reject_value!"
##
## Problem 5
## Checkpoints Passed: 1
## Checkpoints Errored: 0
```

```
## 100% passed
## --------
## Test: PASSED
```

6. Perform a chi-squared test for goodness of fit. Calculate and report your test statistic. Calculate and interpret your p-value. What are your conclusions?

```
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
```

The test statistic is 233.58. The p-value is approximately 0, so there is approximately no chance of observing a test statistic at least as extreme as 233.58. 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.

Part 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!

7. 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.

The null hypothesis is that anger levels and coronary heart disease outcomes are independent. The alternative 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:

8. 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]
   [new variable names: row_total, column_total, sample_size]

```
### Code your answer here. You will have to repeat this syntax a few times.
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()
#Alternative
chd_by_anger_level <-
  chd_by_anger_level %>%
  mutate(sample_size = 8284 + 190)

chd_by_anger_level
```

```
## # A tibble: 6 x 6
##   anger_level heart_disease actual_count row_total column_total sample_size
##   <chr>       <chr>                <dbl>     <dbl>        <dbl>       <dbl>
## 1 Low         CHD                     53       190         3110        8474
## 2 Low         No CHD                3057      8284         3110        8474
## 3 Moderate    CHD                    110       190         4731        8474
## 4 Moderate    No CHD                4621      8284         4731        8474
## 5 High        CHD                     27       190          633        8474
## 6 High        No CHD                 606      8284          633        8474
```

```
check_problem8()
```

```
## [1] "Checkpoint 1 Passed: Correct! It is a data frame"
## [1] "Checkpoint 2 Passed: Correct number of rows"
## [1] "Checkpoint 3 Passed: Correct number of columns!"
##
## Problem 8
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

	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

9. Use the following formula from lecture notes to create a column for expected counts:

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

```
# add expected counts (a new variable named "expected_count") to your dataframe here
chd_by_anger_level <- chd_by_anger_level %>%
  mutate(expected_count = row_total * column_total / sample_size)

check_problem9()
```

```
## [1] "Checkpoint 1 Passed: Correct! It is a data frame"
```

```
## [1] "Checkpoint 2 Passed: Correct number of rows and columns"
## [1] "Checkpoint 3 Passed: Correct value of expected_count!"
##
## Problem 9
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

10. Before moving forward with analysis, confirm two requirements for the Chi-Squared test of independence, namely:

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

```
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
```

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.

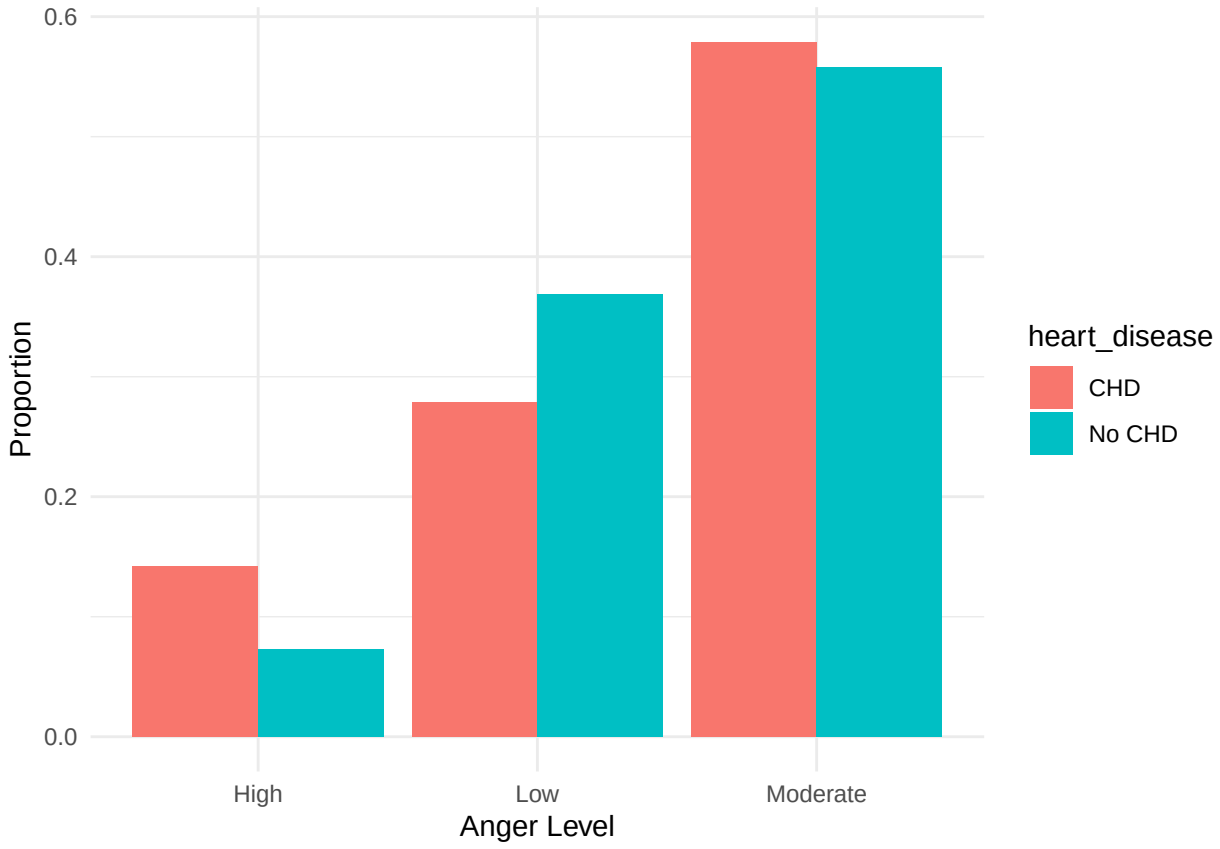
11. First, calculate the probability of anger level conditional on CHD Status

```
# add conditional probabilities (a new variable named "cond_prop_anger") to your dataframe here
chd_by_anger_level <-
  chd_by_anger_level %>%
  group_by(heart_disease) %>%
  mutate(cond_prop_anger = actual_count / sum(actual_count)) %>%
  ungroup()

check_problem11()
```

```
## [1] "Checkpoint 1 Passed: Correct! It is a data frame"
## [1] "Checkpoint 2 Passed: Correct number of rows and columns"
## [1] "Checkpoint 3 Passed: Correct value of cond_prop_anger!"
##
## Problem 11
## Checkpoints Passed: 3
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

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

```
dodged <- 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

dodged
```




```
check_problem12()
```

```
## [1] "Checkpoint 1 Passed: A ggplot has been defined"
## [1] "Checkpoint 2 Passed: You are using the correct dataset!"
## [1] "Checkpoint 3 Passed: Correct x variable plotted!"
## [1] "Checkpoint 4 Passed: Correct y variable plotted!"
## [1] "Checkpoint 5 Passed: You defined a Bar chart!"
## [1] "Checkpoint 6 Passed: You defined a dodged Bar chart!"
##
## Problem 12
## Checkpoints Passed: 6
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

There are some interesting differences but it's a close call that leaves us interested in whether the Chi-Square test wil be significant or not. We do see that higher-anger people are more likely to have CHD.

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

13. 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.]

```
# if you choose to use dynamic coding, you can do it here
chd_by_anger_level <- chd_by_anger_level %>%
  mutate(diff_sq = (actual_count - expected_count)^2,
         ratio = diff_sq / expected_count)

chi_square_test_statistic <- chd_by_anger_level %>%
  summarize(test_stat = sum((actual_count - expected_count)^2 / expected_count)) %>%
  pull(test_stat)

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

chi_square_test_statistic
```

```
## [1] 16.07676
```

```
check_problem13()
```

```
## [1] "Checkpoint 1 Passed: Correct chi square test statistics!"
##
## Problem 13
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

If you choose to calculate your statistic by hand, show your work here:

14. 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.

```
chi_sq_p_value <- pchisq(chi_square_test_statistic, deg_of_freedom, lower.tail = FALSE)


chi_sq_p_value
```

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

```
check_problem14()
```

```
## [1] "Checkpoint 1 Passed: Correct chi square p value!"
##
## Problem 14
```

```
## Checkpoints Passed: 1
## Checkpoints Errored: 0
## 100% passed
## --------
## Test: PASSED
```

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

15. 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?

In the probability section, we would have seen if $P(A|B)$ was exactly 4equal 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 probabilities are significantly different before we reject the null that A and B are independent.

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##                      Test Points_Possible          Type
## Problem 1  NOT YET GRADED               1 free-response
## Problem 2  NOT YET GRADED               1 free-response
## Problem 3          PASSED               1    autograded
## Problem 4  NOT YET GRADED               1 free-response
## Problem 5          PASSED               1    autograded
## Problem 6  NOT YET GRADED               1 free-response
## Problem 7  NOT YET GRADED               1 free-response
## Problem 8          PASSED               1    autograded
## Problem 9          PASSED               1    autograded
## Problem 10 NOT YET GRADED               1 free-response
## Problem 11         PASSED               1    autograded
## Problem 12         PASSED               1    autograded
## Problem 13         PASSED               1    autograded
## Problem 14         PASSED               1    autograded
## Problem 15 NOT YET GRADED               1 free-response
```

Submission

For assignments in this class, you'll be submitting using the **Terminal** tab in the pane below. In order for the submission to work properly, make sure that:

1. Any image files you add that are needed to knit the file are in the `src` folder and file paths are specified accordingly.
2. You **have not changed the file name** of the assignment.
3. The file is saved (the file name in the tab should be **black**, not red with an asterisk).
4. The file knits properly.

Once you have checked these items, you can proceed to submit your assignment.

1. Click on the **Terminal** tab in the pane below.
2. Copy-paste the following line of code into the terminal and press enter.

cd; cd ph142-fa20/lab/lab10; python3 turn_in.py

3. Follow the prompts to enter your Gradescope username and password. When entering your password, you won't see anything come up on the screen–don't worry! This is just for security purposes–just keep typing and hit enter.
4. If the submission is successful, you should see "Submission successful!" appear as output.
5. If the submission fails, try to diagnose the issue using the error messages–if you have problems, post on Piazza.

The late policy will be strictly enforced, **no matter the reason**, including submission issues, so be sure to submit early enough to have time to diagnose issues if problems arise.