**Lab Assignment 06**

1. Mentors
2. TA Profiles
3. TexStudio save session

**I. Objectives**: To gain experience with contingency tables and cross-tabulations. Students will explore questions of “independence” and will perform the two-way chi-square test.

**II. Datasets**: “cdc.R”

**III. Packages**: none

**IV. Preparation**

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Clear data from memory: > rm(list=ls())

3) To change R’s working directory, you will need to fill in the directory name *in quotes* inside the parentheses of this command: > setwd()

4) Download data: “” and place it in your working directory

5) Download R script “POLS 6480 Lab 06.R” and place it in your working directory.

6) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script.

**V. Instructions for Lab 06**

So far the labs have required you to examine data that are real-valued (i.e., allowed to take on any value in some interval of the real number line) or integer-valued. So far, the labs have required you to examine one variable at a time, rather than looking for relationships across variables. This lab instead focuses on making and analyzing contingency tables; these will require your response variable take on a small set of values and will require you to look for associations.

The first dataset you will use is a sample of 20,000 respondents from the Behavioral Risk Factor Surveillance System (BRFSS) in the year 2000. Explore this study at [*www.cdc.gov/bfrss*](http://www.cdc.gov/bfrss).

The second dataset you will use is one that you will create, examining whether students of different ethnic backgrounds are equally likely to be suspended from school.

**A. Cigarettes**

To load the dataset (this may take a few minutes) and make it active, type the following lines:

> source("http://www.openintro.org/stat/data/cdc.R")

> attach(cdc)

The first line of code is line 1 in the R script; the second line is optional, but if you choose not to type it, then in many lines of code below you will need to type cdc$ before a variable’s name.

The variable named smoke100 identifies whether the respondent has smoked more than 100 cigarettes during his or her lifetime. You can examine this variable’s unconditional distribution:

> table(smoke100)

You can also examine various *conditional* distributions, for example, dividing respondents into male and female and then looking at the distribution for each gender:

> table(gender, smoke100)

Based on what R displays, do male and female respondents appear to adopt the same habits with regard to smoking?

Let us shift our attention away from the causes of smoking to some potential effects. The variable *genhlth* measures whether the respondent perceives his or her own health to be generally excellent, very good, good, fair, or poor.

To see the effect of smoking on *genhlth* you can type:

> table(smoke100, genhlth) -> mytable

The code shown above creates a new object, a table with two rows and five columns, which you can display by typing mytable. The table that R Studio displays (shown in the **Console** window) only provides raw frequencies, but you can work with mytable to show the percentages instead.

If you type this code: > prop.table(mytable, 1)

… then R Studio will show you the row percentages. That is, it calculates what percentage of non-smokers falls into each of the five categories, and then it separately calculates what percentage of smokers falls into each of the five categories.

If you type this code: > prop.table(mytable, 2)

… then R Studio will show you the column percentages. That is, what percent of people in excellent health were non-smokers versus smokers; what percent of people with very good health were non-smokers versus smokers; and so forth.

Next, we will shift our attention away from smoking habits, and examine weight and height.

First, run lines 3–4 of the R script to create two data frames, one containing just male respondents and one containing just female respondents.

> m.cdc <- subset(cdc, cdc$gender == "m")

> f.cdc <- subset(cdc, cdc$gender == "f")

Now run lines 5–7 together; line 5 tells R to reshape the plot space so that the next two figures you ask for are presented as two rows aligned in a single column; line 6 and line 7 then ask for histograms of weight for the male and female respondents, respectively.

> par(mfrow=c(2,1))

> hist(m.cdc$weight, seq(50,500,25))

> hist(f.cdc$weight, seq(50,500,25))

The seq() command inside the parentheses is a convenient way to tell R what bins to create; in this case you are asking for bins that are 25 units wide, beginning at 50 and ending at 500.

Men are heavier than women (median weights 185 and 145, respectively), but men are also taller than women. To examine the weight-height relationship, run lines 8–10 together; line 8 tells R to reshape the plot space so the next two figures will be presented individually; line 9 and line 10 then create plots of weight against height, for the male and female respondents, respectively.

> par(mfrow=c(1,1))

> plot(jitter(m.cdc$height), m.cdc$weight)

> plot(jitter(f.cdc$height), f.cdc$weight)

The jitter() command adds a slight horizontal error to each observation, which can make the plot easier to see (except, it works better when you have far less than 10,000 observations…). Later in the semester you will discuss correlation and linear regression; for the moment it is only important to know that height and weight are positively associated.

In what follows, we will use just female respondents. Lines 12–13 generate a new variable for body mass index (BMI) and then graph the distribution of this variable as a histogram.

> f.cdc$bmi <- 703\*(f.cdc$weight/(f.cdc$height^2))

> hist(f.cdc$bmi , seq(12.5,74.5, 2))

Let us inspect the association of BMI with general health. As mentioned above, the variable *genhlth* is a five-level variable, from poor to excellent. You can look at the relationship between BMI and *genhlth* in a few ways; here is one:

> boxplot(f.cdc$bmi ~ f.cdc$genhlth)

Unfortunately, since BMI is so concentrated between 20 and 30 (as the histogram showed), the box-and-whisker plot is not very informative. Moreover, BMI may be too granular of a measure.

Physicians and physiologists categorize individuals into four categories based on BMI. Lines 15–18 create four ‘dummy’ variables indicating whether the respondent falls in the underweight category (BMI < 18.5), the regular weight category (BMI 18.5 – 25), the overweight category (BMI 25 – 30), or the obese category (BMI > 30). Line 19 then combines these into a single indicator variable, coded 1 = underweight, 2 = regular weight, 3 = overweight, and 4 = obese. The next two lines of code (21–22) create a factor variable and then names each of the levels.

Line 22 of the R script creates a table, named *contingency*, which you can display by typing contingency. Examine the table: does there appear to be an association between BMI category and overall health?

Perhaps looking at the raw frequencies isn’t the most useful way to examine the data. If you type:

> prop.table(t(contingency),1)

R will show you the contingency table with row percentages. For example, for people with excellent health, you can find what share are underweight ( = \_\_\_\_\_), what share have regular weight ( = \_\_\_\_\_), and so on. You should see that as health outcomes decline, fewer and fewer come from the regular weight category, while more and more come from the obese category.

If you run lines 24–26, then R will show you the contingency table with column percentages – allowing you to see what the health outcomes are for each category of weight – and will plot it using two versions of bar plots (stacked and side-by-side).

Finally, if you run lines 28–33 *together* (i.e., select all six rows, and click on the Run button at the top of the script window) then R will show you four pie charts. Line 28 reformats the **Plots** window to allow four figures to be plotted in quadrants (two rows, two columns).

Before moving on, you should clear the Environment by typing rm(list=ls()) and you should reset the plot space to a single figure by typing par(mfrow=c(1,1)). Both commands are in line 35 of the R script. The second command returns the **Plots** window to its original format.

**B. Explusion**

A recent story in the [*Houston Chronicle*](http://www.chron.com/news/education/article/Which-race-is-suspended-the-most-in-your-school-6521219.php) showed data – without any analysis! – from the Texas Education Agency; the data will allow you to examine whether students of different ethnic backgrounds are suspended from school at different rates. The article presented data from 24 Houston-area school districts, but we will focus on the largest urban district, Houston ISD, and the largest suburban district, Cypress-Fairbanks ISD.

(The article notes that when looking at aggregate data, one cannot distinguish between different students being suspended or the same student being suspended multiple times. For the analysis below, we will assume that each suspension is for a unique student.)

The data are shown at the top of the next page.

|  |  |
| --- | --- |
| **Houston ISD**(suspensions | enrolled)  **Alaskan/American Indian** | 39 | 483  **Asian** | 114 | 7394  **Black Or African American** | 19980 | 53238  **Hispanic/Latino** | 15681 | 131004 **Native Hawaiian/Other Pacific** | 16 | 204  **Two Or More Races** | 168 | 1857  **White** | 904 | 17372 | **Cy-Fair ISD** (suspensions | enrolled)  **Alaskan/American Indian** | 66 | 647 **Asian** | 230 | 9568 **Black Or African American** | 5719 | 18443 **Hispanic/Latino** | 7578 | 48561  **Native Hawaiian/Other Pacific** | 13 | 88  **Two Or More Races** | 321 | 2851 **White** | 2421 | 31552 |

Looking at raw data rarely provides much insight, so you will code the data and examine tables and figures, and then perform a statistical analysis called the *chi-square test of independence*.

The first step is to enter this data into R. Although there are means of ‘web scraping’ that would allow you to efficiently create large datasets, we are going to type in four variables, indicating suspensions and enrollments in each district. The rows will indicate the different ethnic groups. Run lines 37–38 to enter the data for Houston ISD (the .h suffix is for HISD); run line 39 to plot the proportion of students who are suspended in each ethnic group. Do the proportions look even?

Run line 41 to create a variable subtracting the suspended students from the enrolled students. *Assuming* each suspended student was only suspended a single time, this will tell you the number of students who were not suspended. Lines 42–44 turn the data into a table, named *table.h*, and then line 45 displays it.

If you type margin.table(table.h, margin = 2), as shown in line 47, then R will sum each row, which you can use to calculate the proportion of all students who are suspended.

The chi-square test is a test of the degree to which the columns in a contingency table deviate from the pattern established by the whole table. That is, if 17.5% of all students are suspended, then 17.5% of white students, 17.5% of black students, 17.5% of Latino students, and so on, should be suspended; a large enough deviation from this pattern violates the assumption that ethnicity is irrelevant.

The way that the chi-square test is calculated, you use the proportion generated by the marginal (in this case 17.5%) and then generate expected cell frequencies based on that. The chi-square test is the cell-by-cell sum of the *squared* difference between the observed and expected cell frequency, divided by the expected cell frequency. You should practice working this out by hand for 2x2, 2x3 and 3x2 tables. For larger tables, like the 2x7 table we are dealing with, you type:

> **chisq.test**(table.h)

A statistically significant p-value (i.e., *p* < .05) indicates that the difference between expected and observed cell frequencies is larger than we would have expected by chance alone, under the null hypothesis of independence. In this case, what is the p-value? \_\_\_\_\_

Degrees of freedom for a chi-squared test of independence equals (# rows – 1)\*(# columns – 1); with 2 rows and 7 columns, our degrees of freedom is 6. The critical value for a chi-square test with 6 df and 95% confidence is 12.59; the chi-squared statistic computed here equals \_\_\_\_\_. If the test statistic exceeds the critical value, then we reject the null hypothesis of independence; in this case, do you reject or retain the null hypothesis? \_\_\_\_\_\_

On your own, repeat the process for Cypress-Fairbanks ISD. The data are provided for you as lines 49–50; you will need to create a barplot of the frequency of students who are suspended, create a table that matches the frequency table, compute the chi-square statistic, and then find the p-value to determine whether to reject or retain the null hypothesis of independence.

To clear the **Environment**, type: > rm(list=ls())

To clear the **Console** window, type Ctrl-*l*