COMPSCIX 415.2 Homework 6

Robert Clements

DUE DATE: Mar 13, 2018 @ 6:15PM

In this assignment. . .

. . . we will work with Simpson's paradox, factors, and sampling distributions.

Remember to make your document look good, which means you may need to change some settings on the figure sizes and locations, use markdown syntax to create headings or to format your text (use the cheatsheet), and you may want to play with the different themes.

Use complete sentences, and **divide your work in a logical way**. Remember that the whole point of doing reproducible analysis in R Markdown is so that a complete stranger can take your results, *understand them*, and reproduce them.

Remember to save and knit often. Commit when you've completed a big chunk of work or when you are done for the day and will be resuming later.

What to Turn In

For this assignment you have two choices:

You can upload a pdf document (you will have to install latex);
You can upload a standalone html file.

Prerequisite

Basic R Markdown knowledge
R and RStudio
tidyverse, and **mosaicData** package installed
Access to internet
git and Github

To start your assignment

1. Go to File -> Recent Projects

2. Click on the Project that you created during Homework 1. This project should be the one that is already under git version control.

3. RStudio will switch to that project and the Git pane should appear.

4. Go to File -> New File and choose either R Markdown or R Notebook.

5. Change the title ("COMPSCIX 415.2 Homework 6") and add your name and the date to the YAML header.

6. Save the file in the same folder (or create a subfolder) with your other HW assignments.

7. Knit your document into an html or pdf document.
8. Go to the Git pane and commit both your Rmd and html (or pdf) files by clicking on the checkboxes next to the file names and hitting the Commit button. Write a useful message, and hit the commit button.

Exercises

Here are the exercises for you to complete.

Answer these questions:

Exercise 1

Load the `Whickham` dataset (`data(Whickham)`). You will need to load the `mosaicData` package first, but I also included the data as an rds file on Canvas if you would rather download it there and load it with `readRDS()`.

Look at the help file on this dataset to learn a bit about it.

1. What variables are in this data set?
2. How many observations are there and what does each represent?
3. Create a table (use the R code below as a guide) and a visualization of the relationship between smoking status and outcome, ignoring age. What do you see? Does it make sense?

```
library(mosaicData)
library(tidyverse)
Whickham %>% count( _________ , _________ )
```

4. Recode the `age` variable into an ordered factor with three categories: age <= 44, age > 44 & age <= 64, and age > 64. Now, recreate visualization from above, but facet on your new age factor. What do you see? Does it make sense?

Exercise 2

The Central Limit Theorem states that the sampling distribution of sample means is approximately Normal, regardless of the distribution of your population. For this exercise our population distribution will be a Gamma(1,2) distribution, and we'll show that the sampling distribution of the mean is in fact normally distributed.

1. Generate a random sample of size n = 10000 from a gamma(1,2) distribution and plot a histogram or density curve. Use the code below to help you get your sample.

```
library(tidyverse)
n <- 10000

# look at ?rgamma to read about this function
gamma_samp <- tibble(x = rgamma(n, shape = 1, scale = 2))
```

2. What is the mean and standard deviation of your sample? They should both be close to 2 because for a gamma distribution:

mean = shape x scale
variance = shape x scale^2

```
mean_samp <- gamma_samp %>% .[['x']] %>% mean()
```

3. Pretend the distribution of our **population** of data looks like the plot above. Now take a sample of size n = 30 from a Gamma(1,2) distribution, plot the histogram or density curve, and calculate the mean and standard deviation.

4. Take a sample of size n = 30, again from the Gamma(1,2) distribution, calculate the mean, and assign it to a vector named `mean_samp`. Repeat this 10000 times!!!! The code below might help.

```
# create a vector with 10000 NAs
mean_samp <- rep(NA, 10000)

# start a loop
for(i in 1:10000) {
  g_samp <- rgamma(30, shape = 1, scale = 2)
  mean_samp[i] <- mean(g_samp)
}
# Convert vector to a tibble
mean_samp <- tibble(mean_samp)
```

5. Make a histogram of your collection of means from above (`mean_samp`).

6. Calculate the mean and standard deviation of all of your sample means.

7. Did anything surprise you about your answers to #6?

8. According to the Central Limit Theorem, the mean of your sampling distribution should be very close to 2, and the standard deviation of your sampling distribution should be close to $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{30}} = 0.365$. Repeat #4-#6, but now with a sample of size n = 300 instead. Do your results match up well with the theorem?

Turn in your completed assignment

Commit your changes with the comment "finished assignment 4" and push your R Markdown file and your html or pdf file to Github.

This week you should turn in your assignment by uploading it to Canvas.