Lab 9: Sleep Study

Lab Goals

The goal of this lab is to practice analyzing a dataset in RStudio using reproducible methods. This lab will review the following concepts:

1) Creating new R Markdown documents
2) Reading in data and checking that it was read in correctly
3) Performing initial analysis with descriptive statistics
4) Creating a confidence interval for a population mean, μ
5) Performing a hypothesis test related to a population mean, μ

Similar to Lab 4, in this lab you will create a new `.Rmd` file from scratch. We will analyze data from a sleep study in which several variables related to sleeping habits, cognitive ability, alcohol use, and mental health were recorded for a random sample of students. The source of the `SleepStudy` data is "Onyper, S., Thatcher, P., Gilbert, J., and Gradess, S., Class Start Times, Sleep and Academic Performance in College: A Path Analysis, *Chronobiology International*, April 2002; 29(3): 318-335."

R Functions

R Function	Description
`barplot()`	Makes a bar graph
`boxplot()`	Creates one or more boxplots
`c()`	Combines values into a vector
`cbind()`	Combines columns of data frames, matrices, or vectors
`describe()`	Gives summary information for all variables in a data frame
`dim()`	Finds the dimensions of an object
`head()`	Finds the first part of a vector, matrix, table, data frame or function
`hist()`	Creates a histogram
`legend()`	Creates a legend for a plot
`length()`	Finds the length of a vector
`names()`	Finds the names of the variables in a data frame
`mean()`	Finds the mean of the input values
`plot()`	Makes a scatterplot
`prop.table()`	Tabulates proportions of categorical data
`read.csv()`	Reads a file in a table format and creates a data frame from it
`summary()`	Reports the minimum, 25th percentile, mean, median, 75th percentile, and maximimum of a numerical input variable
`table()`	Tabulates counts of categorical data
`tail()`	Finds the last part of a vector, matrix, table, data frame or function
`title()`	Adds a title to a plot
`t.test()`	Performs one and two sample t-tests on a vector of data.

Sleep Study Data

This dataset includes observations of the following variables from a random sample of 253 college students.

Variable	Description
`Gender`	1 = Male 0 = Female
`ClassYear`	Year in school, 1=first year,. . . , 4=Senior
`LarkOwl`	Early riser or night owl? Responses: `Lark`, `Owl`, or `Neither`
`NumEarlyClass`	Number of early classes each week (before 9am)
`EarlyClass`	Indicator for at least 1 early class
`GPA`	Grade Point Average
`ClassesMissed`	Number of classes missed in a semester
`CognitionZscore`	Z-score on a test of cognitive skills
`PoorSleepQuality`	Higher values indicate poorer sleep
`DepressionScore`	Measure of degree of depression
`AnxietyScore`	Measure of amount of anxiety
`StressScore`	Measure of amount of stress
`DepressionStatus`	`normal`, `moderate`, or `severe`
`AnxietyStatus`	`normal`, `moderate`, or `severe`
`Stress`	`normal` or `high`
`DASScore`	Combined score for depression, anxiety and stress
`Happiness`	Measure of degree of happiness
`AlcoholUse`	`Abstain`, `Light`, `Moderate`, or `Heavy`
`Drinks`	Number of alcoholic drinks per week
`WeekdaySleep`	Average hours of sleep on the weekdays
`WeekendSleep`	Average hours of sleep on the weekend days
`AverageSleep`	Average hours of sleep for all days
`AllNighter`	Had an all-nighter this semester? 1 = yes 0 = no

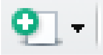
Investigation of the Speed Study Data

Part 1 - Reading in the Data

1. Use the `read.csv()` command in the console, creating a data frame named `SleepStudy` in the workspace for the console.
2. In the console, type `head(SleepStudy)`. Then open the raw `.csv` file in RStudio using the path *File > Open File.* Do the first few rows of the data frame match those of the actual data? You should also do the same for the last few rows of the data set. Does the last row number of `SleepStudy` correspond to the number of students selected for this study?

Part 2 - Creating a R Markdown File

Loading in data properly takes time and it is important to have a clear record of what we have done rather than just typing commands in the console and assuming we will later remember what we have done. This is especially important for published work since you want to make sure that someone else (or yourself in 5 years) will be able to exactly reproduce your analysis, resulting in the identical results.

1. Create a new R Markdown document by clicking on this arrow in the upper left corner of RStudio . Then choose the second option, *R Markdown. . . .* A window will pop up in which you can change the title of your document to *Lab 9* and denote yourself as the author. Also, you can choose a default knitting format. Note, all knitting options will still be available regardless of the chosen default.
2. This new `.Rmd` document needs to be renamed and saved in your folder for Lab 9. Choose the path *File > Save As..* to save this document as `LastF-Lab9.Rmd`.

3. Delete all of the default text in your new R Markdown document except for the header with the document title, author, date, and default knitting option.
4. Under the document header write a sentence or two that includes a description of the data set, where you found the data set (from blackboard), and some information on who collected the data.
5. Next, insert a code chunk to read the data into the workspace for the lab document.

Part 3 - Getting to Know Your Data

In the console, let's get some basic information on the `SleepStudy` data frame.

1. Type `dim(SleepStudy)` to determine the number of observations in these data as well as the number of variables.
2. In Lab 4, we used the `describe()` function to get a quick summary of all of the variables in a dataframe. To use this function, we will first need to load the `Hmisc` library. To do this, in the console type `library(Hmisc)`.
3. Now type `describe(SleepStudy)` into the console. Take a look of the summary data for each variable. Are there any missing values? Scroll through this summary data to become more familiar with the variables found in the `SleepStudy` data.
4. Now, let's return to our R Markdown document for this lab. Add a sentence in this document explaining that we will start by looking at the data before proceeding. Follow this by adding a code chunk that does the following:
   a) Finds the dimension of `SleepStudy`
   b) Uses `library(Hmisc)` to load `Hmisc` into the R Markdown document workspace (recall that when you knit your R Markdown file, a fresh R environment is spawned separate from the console)
   c) Describes the variables in `SleepStudy`

Part 4 - More Detailed Investigation of the Variables in `SleepStudy`

1. Left justified number signs (#) with a description behind them create different sections in your R Markdown document. Create a section dedicated to taking a closer look at some of the variables in `SleepStudy` by typing two # signs on the left margin of your R Markdown document followed by a space and the title, "Investigation of the Variables in `SleepStudy`."
2. Subsections within a section of your R Markdown document are added by including one more # sign before the title of the subsection. Under the code for creating the section, "Investigation of the Variables in `SleepStudy`", type 3 left justified # signs, a space, and the title, "`ClassYear` and `NumEarlyClass`."
3. In this subsection, first give a brief description of the two variables, `ClassYear` and `NumEarlyClass`.
4. How many students from each class were included in the study? Include a code chunk to create a table to answer this question.
5. It might be interesting to look at the relationship between `ClassYear` and `NumEarlyClass`. Since these can both be considered as qualitative variables, it makes sense to use a barplot to describe this relationship. Create a code chunk in your lab document that does the following:
   a) Create a table of `NumEarlyClass` counts by `ClassYear` where the rows of this table correspond to the student's number of early classes. Name this table `EarlyByClass`.
   b) Create a new table where the column corresponding to each class is divided by the total number of students in that class. This will give us the sample proportion of students in each class that belong to each level of `NumEarlyClass`. The `prop.table()` function in R will create this table for

you. To use this function for the EarlyByClass counts, put the following code in your code chunk: PropEbyC = prop.table(EarlyByClass,margin=2).

c) Include the following code in your code chunk to create the barchart.

```
par(mar = c(5.1, 4.1, 4.1, 7.1), xpd = TRUE)
barplot(PropEbyC, col = 2:7, width = 2)
legend("topright", inset = c(-0.25, 0), fill = 2:7,
       legend = rownames(EarlyByClass))
```

6. Knit your lab document to examine this barchart. Below the code chunk that creates this barchart, describe the relationships you see between the levels of ClassYear and NumEarlyClass.
7. Create a new subsection with three left justified number signs named, "Drinks and AlcoholUse"
8. Give a brief description under this title of the two variables Drinks and AlcoholUse.
9. Put a code chunk under this description that creates boxplots for Drinks ~ AlcoholUse. The margins of the plot need to be adjusted to properly label the different levels of AlcoholUse. Use the code provided below to create new margins using the par() function and create horizontal boxplots.

```
par(mar = c(5, 5, 4, 2) + 0.1)
boxplot(Drinks~AlcoholUse, data = SleepStudy, horizontal = TRUE, las = 1,
        main = 'Drinks per Week by Alcohol Use')
```

10. Knit your document to look at these boxplots and note any interesting or unusual findings in these data in your lab document.

Part 5 - Creating a Confidence Interval for the Mean Difference in Average Hours Slept on the Weekdays versus the Weekend Days

1. Create a new section in your lab document by left justifying two # signs, adding a space, and including the title, "Examining the Differences in Sleep Patterns on Weekdays Versus Weekend Days."
2. In this section, first briefly describe the two variables, WeekdaySleep and WeekendSleep.
3. It might be interesting to obtain a 95% confidence interval for the mean difference of the average amount of sleep students get per day during the weekdays and the average amount of sleep they get per day on the weekends. Using a code chunk in your lab document, add a column to the SleepStudy data that contains the differences WeekendSleep - WeekdaySleep. This can be done using the following code, SleepStudy$Differences = SleepStudy$WeekendSleep - SleepStudy$WeekdaySleep.
4. Create a confidence interval for the population mean of Differences using the following steps.
   i) Determine the sample mean of Differences
   ii) Determine the estimated standard error of Differences $(s/\sqrt{n})$.
   iii) Determine the 97.5th percentile of the standard normal distribution
   iv) Using (i) - (iii) determine the lower and upper endpoints of the 95% confidence interval for the mean difference between average hours slept per day on the weekdays versus the weekend days.
5. Interpret this confidence interval in your lab document. On average do students sleep a lot more per day on the weekends than on the weekdays?
6. This confidence interval can also be determined using the t.test() function in R. Add a code chunk, that includes the following code: t.test(SleepStudy$Differences).

Part 6 - Is the mean GPA for this college equal to the national average?

1. Create a new section in your lab document (with two # signs) titled, "Comparing the Mean GPA for This College to the National Average."

2. In this section, briefly describe the `GPA` variable.

3. The national average GPA for colleges in the U.S. is 3.0. We would like to see if the college from which these students are sampled from has a mean GPA equal to the national average. Use the following steps to perform this test in your lab document. Assume the significance level of this test is 0.01.

   i) First give a brief description of what you are testing. Then state the null and alternative hypotheses for this test.

   ii) In a code chunk, determine a p-value for this test.

   iii) What is your conclusion based on the p-value? Include this conclusion in the context of the problem in your lab document.

   iv) A two-sided hypothesis test is the default test for the `t.test()` function. The default significance level is 0.05. In a code chunk, perform this test using the `t.test()` function with the additional argument `conf.level = .99`. Setting the confidence level to 0.99 is equivalent to changinging the significance level to 0.01. Also, the default null hypothesis mean is 0. To change the null hypothesis mean to 3, include the argument `mu=3` for the `t.test()` function. Confirm that the p-value obtained in (ii) is equal to the p-value listed in the output from this test.