Lab 1

Updated Saturday, October 1st at 5:15p

Submit via CCLE by Friday, October 7th, 10pm

Directions: Create an R Markdown document to complete the tasks below. Include all necessary lines of code and explain your work using complete sentences. Both the code you write and the outputs from `R` should be included in the compiled/knitted HTML document. Submit both the `.Rmd` and `.html` files to CCLE. Name the files `#########-lab01.Rmd` and `#########-lab01.html` where the `#########` are replaced by your Bruin ID.

1 Perform basic calculations in `R`

1.1 Use `R` as a calculator and show how to get answers for:

Note: You don't need to explain your work for this problem.

1. $3 \cdot (-2)^2 + 5 \cdot (-2) - 2$
2. $\sqrt{12}\left(1 - \frac{1}{3 \cdot 3} + \frac{1}{5 \cdot 3^2} - \frac{1}{7 \cdot 3^3}\right)$

1.2 Vectorized operations

Note: You don't need to explain your work for this problem.

1. Use a single line of code to create a vector of the squared values of all the even numbers between -50 and 50, i.e. $(-50^2, -48^2, \ldots, 50^2)$.

1.3 Sample vs. population standard deviation

The equation for calculating the standard deviation of a population is

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

where N is the size of the population, x_i are the individual values of the population, and $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the population mean.

The equation for calculating the standard deviation of a sample is

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where n is the size of the sample, x_i are the individual values from the sample, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

1. Create a numeric vector called `values` containing the values: $27, 36, 50, -24, 9, -38$.
2. Use `R`'s `sd()` function to calculate the standard deviation of `values`

3. Using R's `sum()`, `mean()`, `sqrt()` and `length()` functions (Use each of them once), calculate the population and sample standard deviation of the `values`.
4. Does R's `sd()` function calculate the sample or population standard deviation? Write down another source where we could have looked to answer this question without carrying out the calculations?

1.4 Vector classes

As we mentioned during the lecture, vectors in R are always a single class (Meaning if we mix numbers and strings/characters together in a vector, the numbers are coerced to strings).

1. Create a vector called `numbers` containing 4 unique numbers of your choosing. Create a vector called `strings` containing 3 unique strings of your choosing. Create a vector called `booleans` containing two `TRUE`s and two `FALSE` values (i.e. `TRUE`, `TRUE`, `FALSE`, `FALSE`). **Protip**: You can actually just use `T` and `F` for boolean (logical) values in R.
2. Write down the classes of:
    1. A vector containing `numbers` and `strings`.
    2. A vector containing `numbers` and `booleans`.
    3. A vector containing `strings` and `booleans`.
    4. A vector containing `numbers`, `strings` and `booleans`.
3. Why do you think R coerces the different combination of vectors (`numbers`, `strings` and `booleans`) into these classes?

2 Reading in different data file types

2.1 Childhood Respiratory Disease

The following link contains information about an observational study conducted to measure the effect of smoking on young people's lung capacities: http://www.statsci.org/data/general/fev.html

The link to the data for this study can be found here: http://www.statsci.org/data/general/fev.txt

1. Without downloading the data onto your laptop, use the `read.table()` function to read the data straight from the URL and name it `crd` (for Childhood Respiratory Disease). Hint: For this step, just use the `file` argument to tell R the URL where the data is located.
    - The first row of data in the file is actually the names of the variables. Be sure to find the specify the appropriate argument for `read.table()` so that the variable names are read in as variable names and NOT the values of the first observation. Hint: Check the help documentation for `read.table()` for the name of the appropriate argument.
2. Print the names of the variables using an R function.
3. Based on the following line of code, change the variable names so that (1) they're all lower-case letters and (2) rename the `FEV` variable as `lung_cap`.

```
names(data) <- c("new_name_1", "new_name_2", ..., "new_name_n")
```

4. Print the names of the variables again using an R function.
5. Run the line of code below to answer and then justify the question: *Which is larger, the proportion of female smokers or the proportion of male smokers?*

```
table(crd$sex, crd$smoker)
```

2.2 Reading `Stata`, `SAS` and `SPSS` files

`Stata`, `SAS` and `SPSS` are other statistical analysis softwares used in academia and industry (`Stata` is popular in economics, `SPSS` is popular in other humanities & psychology and `SAS` is a popular alternative to `R`).

These softwares export data in their own specialized formats, much like `R` exports data as `.Rda` files. Specifically, `Stata` exports data as `.DTA` files, `SAS` exports data as `.sas7bdat` files (among others) and `SPSS` exports data as `.sav` files.

Below are some links to various `Stata`, `SAS` and `SPSS` data files:

- http://www.sjsu.edu/people/carlos.e.garcia/courses/soci104/Course-Assignments/104data_2014.sav
- http://qcpages.qc.cuny.edu/~rvesselinov/statadata/WAGEPAN.DTA
- http://biostat3.net/download/sas/colon.sas7bdat

Perform the following tasks using the links provided above.

1. Install and load the `haven` package into `R`:
   - Use `install.packages("package_name")` to install packages. Remember, installing packages should happen in the console and not in `R` scripts nor `R` Markdown documents.
   - Use `library(package_name)` to load the functions in the package into `R`. Remember, to load packages in `R` Markdown files before using functions from the package to avoid errors.
2. In the *Packages* pane in `RStudio`, click the name of the package you installed to open a list of functions. Find and use the appropriate functions to load the following data sets from the links listed (much like you did for the Childhood Respiratory Disease, that is, read the data straight from the URLs.)
3. Use `R` functions to:
   1. Print the names of the variables in the `SPSS` data.
   2. Print the number of observations and variables in the `Stata` data.
4. Install and load the `knitr` package into `R`. Use the `kable` and `head` functions to print the first six rows in the `SAS` data as a table.

3 Create and interpret plots

1. Install and load the `ggplot2` package.
2. Use `ggplot2` functions to create a histogram for the ages of people in the Childhood Respiratory Disease data. Specify the argument `binwidth = 1`.
3. Use `ggplot2` functions to create a bargraph of whether people are current or non-smokers in the Childhood Respiratory Disease data. Let the `fill` of the bars be based on the sex of the people and use the argument `position = "dodge"` to make the bars side-by-side.
4. Use `ggplot2` functions to create a scatterplot where the height of people in the Childhood Respiratory Disease data is on the x-axis, their lung capacities is on the y-axis and the color of the points are colored based on whether they are a current or non-smoker. Include the code `+ theme_minimal()` to your plot to change the appearence of the plot.