Chapter 4: Sampling and Computing Statistics in R

Learning Outcomes: We are going to go over how to do the following:

- Produce n random numbers in R.
- How to select a random sample from a vector in R.
- How to set a seed in R.
- Access data that is built-in to R.
- Read a data file into R.
- Access a column of data.
- Compute sample statistics in R (including sample mean, sample proportion and sample variance).

Motivation: Why are we doing this?

Producing n random numbers in R:

Suppose we want to select n random integers between a and b inclusive. The following command in R will do this:

So, for example the code

Note: The default for the sample() command is to perform sampling **without replacement**. This means that the same individual cannot be in the sample more than once. If you want to change that to allow for an individual to be in the sample more than once (referred to as sampling **with replacement** then you would type in:

How to select a random sample from a vector in R: It would be nice to be able to create a sample of individuals directly in R (rather than labelling our individuals with numbers and then using the sample() function to pick a sample of numbers.

You can use the sample() function on a vector. So if you put your individuals inside a vector, then you can sample the individuals from the vector directly.

Exercise: We are going to work through an example which will have us take a random sample from a vector of individuals.

(a) Create a vector which contains the names of the individuals in a dinner party (this is a vector of characters) and call this vector $friends$.

(b) Define a variable called $friends.sample$ which is equal to a random sample of size 2 from $friends$.

(c) Print out the value of the variable $friends.sample$.

(d) Run the code again (choosing the sample and printing out the variable).

(e) Did the sample stay the same?

Setting a seed in R: In R, there is a function called set.seed() which sets the starting number used to generate a sequence of random numbers. It ensures that you get the same result if you start with that same seed each time you run the same process.

For example, if I use the sample() function immediately after setting a seed, I will always get the same sample.

So in the previous example, to get the same sample each time we run the code, we write the code as:

This is important for reproducibility of your results. If you are writing a paper, you want someone who runs your code to get the same results that you wrote about in your paper.

Accessing Data that is built into R: R has many built-in data sets which are there for you to use and practice with.

To see a list of the available data sets type in the command:

If you want to learn more about a certain data set. You can type a question mark followed by the name of the data set into R. For example:

If you want to load a certain data set into R, then you type in the command *data*(*data.set.name*). For example, to load the *sleep* data set into R, type:

How to access a single column of a data set:

If your data set has multiple columns (this is data frame), you might wish to compute some value using only one of the columns of the data set.

For example, consider the *trees* data set in R. This gives the diameter, height and volume for black cherry trees. If we want to compute the average height in our sample of cherry trees then we can access that specific column using the $ symbol:

Practice Question: Use the *trees* data set to answer the following questions:

1. What is the average diameter of the sample of cherry trees? (Round your answer to 2 decimal places)

   (A) 15.36 (B) 11.83 (C) 13.25 (D) 14.74

   Note: If you look into the data, it tells you that the diameter of the trees are labelled incorrectly as girth in the data set. This is why it is important to read up on the data you are using.

2. What does the value from the previous question (average diameter of the sample of cherry trees) represent?

   (A) A sample statistic.
   (B) An observed value of the sample statistic.
   (C) A population parameter.

3. What is the sample size?

   (A) 31 (B) 40 (C) 41 (D) 50

How to read an external data file into R: If you collect your own data, or download a data set from a website (for example Statistics Canada), then you need to be able to read the data into R so that you can work with the data set in the program.

For example, I downloaded a data set from Stats Canada which consists of the number of secondary education graduates in Canada between the years 2003 and 2019. I saved the data as a CSV file in my R Stat123 file. I called the file SecondaryGraduates. To read this file into R, I use the command:

If I want to call the data set something else in R, I can define a variable at the same time that I read in the data set:

It is often a good idea to take a quick look at the data set when you first read it into R. Some data sets are very large so you might not want to print the entire data set. To see what just the column headers and first 6 rows of the data set look like, you can use the command:

Practice Question: The following is related to the data set *SecondaryGraduates* which is available for download in Brightspace.

1. Download and save the SecondaryGraduates file to your computer (in your designated R folder). Read the file into R and name it *grad.data*.

   Note: Your working directory needs to be set to whatever folder you are saving this file in. Recall, to set your working directory in R click on Session − > Set Working directory − > Choose directory.

2. Using the head() function, determine the column names of the data set.

3. Determine the mean yearly number of secondary graduates between the years 2003 and 2019.

4. Assign the Year column from the data set to a vector called *yrs*. Be sure to force the years to be character variables, using the command as.character().

5. Assign the Number column from the data set to a vector called *num.grads*.

6. Use the names() function in R to assign the elements on the *num.grads* vector the names coming from the *years* vector.

7. Set the seed to be 21. Then, take a random sample of size 6 from the *num.grads* vector.

8. Find the sample average annual of graduates from secondary school.