Homework 1

Instructions: Before starting on Question 1, you should have completed both of the datacamp.com assignments for this week and the "Getting Started in R" handout. Your solution to Question 1 should be submitted as a .R script file on canvas.

Your answers for question 2 and the conceptual questions may be combined, and should be submitted as a .pdf file on canvas

Please feel free to discuss questions on the discussion board.

1. (2 points) Starting with the Homework1.R file from the "Getting Started in R" handout, add code and comments to complete the following tasks:

   (a) Calculate 3 squared, and include a comment indicating the operation performed.

   Solution:

   ```
   # Three to the second power
   3^2
   ```

   (b) Create one variable for each of the following R data types: numeric, logical, and character. Verify each is the correct class.

   Solution:

   ```
   one <- 31.28
   two <- TRUE
   three <- "Mike Trout"
     class(one)
     class(two)
     class(three)
   ```

   (c) Create a vector called "numbers" whose entries are 31282, 5, 1980, and 27. Name the elements of "numbers": "Date", "George", "Year", and "Trout".

   Solution:

   ```
   numbers <- c(31282, 5, 1980, 27)
   names(numbers) <- c("Date", "George", "Year", "Trout")
   ```

   (d) Now find the sum of "George" and "Trout" by subsetting from the "numbers" vector, and using the sum command.

   Solution:

   ```
   sum(numbers[c("Phone Number", "Birth Year")]
   # or sum(numbers[c(2, 4)]
   numbers[numbers > 100]
   ```

(e) Lastly, use a logical comparison operator to list the values from "numbers" that are greater than 100.

Solution:

```
numbers[numbers > 100]
```

(f) Restart your R session, and make sure your entire .R file "sources" without error.

(g) Submit your completed R script file to canvas.

2. (3 points) Find a news article reporting the results of a scientific study.

(a) Report the headline of the article and identify whether it implies population or causal inferences, neither or both.

(b) What inferences are justified by the study? Justify your answer by including parts of the article that report details of the study crucial to identifying the scope of inference. If the article doesn't provide enough information, specify what additional information is required.

Conceptual questions

Answer any three of the following six short answer questions.

3. (1 point) A study found that individuals who have large yards tend to have pets more often than individuals who do not have large yards.

(a) Can cause and effect be inferred? Why or why not?

Solution: No. It is an observational study, and there confounding factors.

(b) List two possible confounding factors that may be contributing to the difference.

Solution:

1. People with larger yards may have a higher income than those with smaller yards and thus be able to afford to own pets more often.
2. People with larger yards may live in areas without noise ordinances more often than those with smaller yards, and this may encourage them to own more pets.

4. (1 point) An experiment was performed in which mice were randomly assigned to two groups. One group was fed diet A and the other group was fed diet B. All environmental factors remained the same across both groups. After three months, the scientist measured the weight of the mice. It was found that the mice fed diet A weighed much less on averagre than the mice fed diet B. Can cause and effect be inferred? Why or why not?

Solution: Yes. This was a randomized experiment in which many environmental confounding factors were ruled out. There is still a possibility that by chance, the mice in one group are different than the mice in the other group with respect to some other variable that may be causing the difference. Later this term we will see how to quantify how likely it is that this difference happened by chance.

5. (1 point) Random samples of people from New York and Texas are invited to participate in a study comparing income of the two geographic groups. Volunteers participate in the study and their income for the last three years is recorded. In order to make inference to the population of all New Yorkers and all Texans, what must we assume? Why?

Solution: We must assume that the reasons for choosing to participate were not related to the response (income). If they were related to the response, the sample might not be representative of the populations.

6. (1 point) A random sample of monarch butterflies and a random sample of swallowtail butterflies were captured in Montana. Their weights were measured and recorded. We would like answer whether monarch butterflies are heavier on average than swallowtail butterflies in Montana. Explain which of the following best describes the goal(s) of this data analysis (description, estimation, hypothesis testing, or prediction)? Why is it important that the samples were randomly collected?

Solution: The goal of this data analysis is to perform a hypothesis test, and also perhaps to construct an estimate of the difference in average weights. It is important that the samples are random so that we they are representative of the populations to which we are making inference. If they were not representative we would not be making inference to the correct populations.

7. (1 point) Twenty ponderosa pine trees in Flagstaff, Arizona were randomly selected and their heights were measured. We would like to state what our best guess of the mean height is for the population of ponderosa pine trees in Flagstaff. We would also like to make our best guess of the height for the next randomly selected ponderosa pine tree in Flagstaff. Explain which of the following best describes the goal(s) of this data analysis (description, estimation, hypothesis testing, or prediction)? Would you expect your guess based on a new sample of twenty different ponderosa pine trees to be the same?

Solution: The goals of analysis are estimation and prediction. There will always be variability in our inference when we pick a random sample, so we would expect a guess based on a new sample to be different.

8. (1 point) Explain in two or three sentences where variability and uncertainty fit into statistics.

Solution: Answers may vary. Be sure to include both terms in the answer. We need statistics because we have variability in data. The goal of statistics is to describe or explain (or assign) this variability. When we would like to make statements or decisions about populations from which we do not have all the information, we do not ever know for sure if we are exactly correct. This is because when we do not have all the information about the population, we have uncertainty, and this must be incorporated into our statements or decisions.