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.
   (b) Create one variable for each of the following R data types: numeric, logical, and character. Verify each is the correct class.
   (c) Create a vector called "numbers" whose entries are 31282, 5, 1980, and 27. Name the elements of "numbers": "Date", "George", "Year", and "Trout".
   (d) Now find the sum of "George" and "Trout" by subsetting from the "numbers" vector, and using the sum command.
   (e) Lastly, use a logical comparison operator to list the values from "numbers" that are greater than 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?
   (b) List two possible confounding factors that may be contributing to the difference.
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?

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?

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?

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?

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

Answers

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.

https://www.npr.org/2018/09/25/651618685/study-roundup-weed-killer-could-be-linked-to-widespread-bee-deaths

http://www.pnas.org/content/early/2018/09/18/1803880115

a. Study: Roundup Weed Killer Could Be Linked To Widespread Bee Deaths
The headline implies causal inference, as the herbicide in question is tentatively indicated as the cause of death of bees in the wild, via an impact of bees' gut bacteria.

b. The article itself is quite tentative in its conclusion, and is instead a summary of a recently published study, linked in the article, and the response from the herbicide's manufacturer.
The linked study, however, is much more detailed in its description of experiment, and posits a much narrower inference, namely that the herbicide impacts the size and composition of honey bee gut microbiome. This causal inference is supported by the description of experiment, included in the results and discussion section, and excerpted below. In addition, population inference is suggested by the repetition of the experiment with a different beehive, during a different season of the year.
"Hundreds of adult worker bees were collected from a single hive, treated with either 5 mg/L glyphosate (G-5), 10 mg/L glyphosate (G-10) or sterile sucrose syrup (control) for 5 d, and returned to their original hive. Bees were marked on the thorax with paint to make them distinguishable in the hive. ...To determine the effects of glyphosate on the size and composition of the gut microbiome, 15 bees were sampled from each group before reintroduction to the hive (day 0) and postreintroduction (day 3), and relative and absolute abundances of gut bacteria were assessed using deep amplicon sequencing of the V4 region of the bacterial 16S rRNA gene and quantitative PCR (qPCR)."
"This experiment was repeated using bees from a different hive and season, and similar trends were observed (SI Appendix, Fig. S2)."
Use of randomized sampling and inclusion of a control group provide a strong case for the causal inference of impact on the gut bacteria via a lower probability of confounding factors. The case for population inference, however, is somewhat weaker, in the sense that a stratified sampling approach was limited to two strata: the original hive and the repeated experiment.

The study continues with a second experiment to determine the susceptibility of treated bees to a common opportunistic pathogen, the method of which is excerpted below. Additional studies are referenced to make an argument based on subject matter expertise for the more general causal inference that the herbicide is responsible for honey bee mortality.
"Hundreds of late-stage pupae were removed from brood frames and allowed to emerge under sterile conditions in laboratory. (Experiment A) NEWs were exposed to bee gut homogenate for 5 d, then hand fed 1 mM glyphosate or sterile sugar syrup on 2 alternate days. Fifteen bees from each group were sampled 2 d after the last hand feeding. DNA was extracted from dissected guts, used as template for qPCR analyses, and submitted for Illumina sequencing at the GSAF, UT Austin. (Experiment B) NEWs were exposed to a bee gut homogenate or sterile sucrose syrup. Each group was divided into two subgroups and treated with 0.1 mM glyphosate or sterile sucrose syrup for 5 d. After that, half of the subgroups was exposed to the opportunistic pathogen S. marcescens kz19, whereas the other half was used as controls. Bees were exposed to similar amounts of glyphosate (~1.7 μg) in experiments A and B."

The causal inferences outlined above appear to be justified by the rigor of the experiments' designs, although the population inference is somewhat weaker, and more reliant on referenced studies.

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

a. This appears to be an observational study, which can't establish causal inference. Randomized methods are required for such a case.

b. While the proposed inference appears intuitive, possible confounding factors include wealth and housing restrictions. Owning a home with a large yard and owning a pet both have non-trivial demands on income, and hence could be confounded by that. Also, homes with little or no yard area could be dominated by apartments or condominiums, which would be more likely to have restrictions on the types of pets allowed.

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?

Yes, the use of controlling, randomization and replication support the causal inference, as randomization and controlling ensures low probability of confounders and replication provides increased statistical accuracy.

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?

It must be assumed that the voluntary participants in the study form a representative sample of the greater regional population, which is a risky proposition. This volunteer sample has a higher likelihood of confounding factors than a more thoroughly designed randomized sample.