Homework 1

Student

Instructions

Use the R Markdown version of this file to complete and submit your homework. Items in **bold** require an answer. Make sure you change the author in the header to your own name.

Conceptual Questions

(1 point each)

1. 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?**
   No. This is an observational study because the "treatment" of large yards is not randomly assigned to individuals. Causal inference is not justified in observational studies because of the possibility of confounding factors.

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

      1. Income: 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. Noise ordinances: 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.

   Any reasonable confounders get points, but student must indicate the relationship between the confounder and response (more pets), ***and*** *the confounder and the explanatory variable (larger yards).*

2. 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 average than the mice fed diet B. **Can cause and effect be inferred? Why or why not?**

   Yes. This was a randomized experiment where the mice were randomly assigned to the treatment groups. This randomization protects against confounding variables.

   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 a confounding variable, but later this term we will see how to quantify how likely it is that this difference happened by chance.

3. 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?**

Despite starting with random samples from the New York and Texas populations, the actual recorded responses are only from those people who volunteered their income. 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 population.

Inference in the wild

Your task is to find a news article reporting the results of a scientific study and then complete the questions below based on this article.

Article headline: ____

1. **Does the headline of the article imply population inference, causal inference, neither or both?** Be specific about what language implies which inference. *(1.5 points)*
2. **What inference is 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. *(1.5 points)*

R foundations practice

(4 points total)

1. Consider the R function `sample()`.
   a) Open the help page for `sample()`. **What does the "Description" section say `sample()` does?**
   "`sample()` takes a sample of the specified size from the elements of x using either with or without replacement."
   b) **What are the names of the arguments to `sample()`?**
   `x`, `size`, `replace`, and `prob`.
   c) **Which arguments to `sample()` are optional and what are their default values?**
   `replace` and `prob` are optional. By default, `replace = FALSE` and `prob = NULL`.
2. **Create a numeric vector called `subjects` that contains the integers 1 to 100.**

```
subjects <- 1:100
```

3. **Use `sample()` to draw a sample of size 10 from `subjects` (without replacement) and save this sample in an object called `sampled_subjects`.**

```
sampled_subjects <- sample(x = subjects, size = 10)
```

4. **Use subsetting to extract the first five elements of `sampled_subjects`.**

```
sampled_subjects[1:5]
```

```
## [1] 53  1 43 25 72
```

5. **Create a character vector called `treatments` that contains ten values: "A" repeated 5 times and "B" repeated 5 times.**

```
treatments <- c("A", "A", "A", "A", "A", "B", "B", "B", "B", "B")
treatments
```

```
##  [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
```

6. **Use `sample()` to draw a sample of size 10 from `treatments` (without replacement) and save this sample in an object called `treatment_assignment`.**

```
treatment_assignment <- sample(x = treatments, size = 10)
treatment_assignment
```

```
##  [1] "B" "A" "B" "A" "A" "B" "B" "A" "A" "B"
```

7. Now consider the steps above in relation to the ideas of "selection of units at random" and "allocation of units to groups".

   a) **Which steps could be used to select units at random? Which object describes the population and which the sample?**
   Step 3 could be used to select units at random, where `subjects` represents the population and `sampled_subjects` represents the sample.

   b) **Which steps could be used to allocate units to groups at random? Which object describes which unit gets which treatment?**
   Step 5-6 could be used to allocate units to groups at random, by matching up `treatment_assignment` and `sampled_subjects`. For example, the first object (53) in `sampled_subjects` will get the first assignment in `treatment_assignment` ('B')