Lecture 4

Paul Holaway, Abhi Thanvi

June 27th, 2022

Lecture 3 Review

Before we move on to the new material, we will do a quick review of Lecture 3 content. Last time we went over why we do data cleaning and introduced experimental design. Most data is not going to be clean. It will either be a disorganized mess or somewhat clean. Usually you will have to do some kind of cleaning even if it is not much. Recall the examples from last class.

Example 1; Data Cleaning Review

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Trains <- read.csv("~/Classes/DPISu22/Data Sets/AMTK_NJT Performance 5-20.csv", stringsAsFactors=TRUE)
Train = Trains %>% select(-c("train_id","to_id","from_id"))
head(Train,5)
```

```
##         date stop_sequence                 from                  to
## 1 2020-05-01             1 Newark Penn Station Newark Penn Station
## 2 2020-05-01             2 Newark Penn Station               Union
## 3 2020-05-01             3               Union        Roselle Park
## 4 2020-05-01             4        Roselle Park            Cranford
## 5 2020-05-01             5            Cranford           Westfield
##        scheduled_time         actual_time delay_minutes   status           line
## 1 2020-05-01 23:38:00 2020-05-01 23:40:09    2.15000000 departed Raritan Valley
## 2 2020-05-01 23:47:00 2020-05-01 23:47:01    0.01666667 departed Raritan Valley
## 3 2020-05-01 23:50:00 2020-05-01 23:51:04    1.06666667 departed Raritan Valley
## 4 2020-05-01 23:55:00 2020-05-01 23:55:31    0.51666667 departed Raritan Valley
## 5 2020-05-01 23:59:00 2020-05-01 23:59:01    0.01666667 departed Raritan Valley
##         type
## 1 NJ Transit
## 2 NJ Transit
## 3 NJ Transit
## 4 NJ Transit
## 5 NJ Transit
```

What we just did above was removing three irrelevant columns because the train and station IDs do not mean anything to us. They just serve as identifiers for the railroad.

```
TrainAMTK = Train %>% filter(type == "Amtrak")
head(TrainAMTK,5)
```

```
##         date stop_sequence                from                     to
## 1 2020-05-01            NA        Philadelphia          Philadelphia
## 2 2020-05-01            NA        Philadelphia               Trenton
## 3 2020-05-01            NA             Trenton        Newark Airport
## 4 2020-05-01            NA      Newark Airport   Newark Penn Station
## 5 2020-05-01            NA Newark Penn Station New York Penn Station
##   scheduled_time         actual_time delay_minutes   status     line   type
## 1                2020-05-01 15:57:05            NA departed REGIONAL Amtrak
## 2                2020-05-01 16:28:00            NA departed REGIONAL Amtrak
## 3                2020-05-01 17:00:00            NA departed REGIONAL Amtrak
## 4                2020-05-01 17:05:12            NA departed REGIONAL Amtrak
## 5                2020-05-01 17:38:05            NA departed REGIONAL Amtrak
```

What we just did above here also counts as us validating our data. We notice that `stop_sequence` and `delay_minutes` are noting but `NA` which means the data is missing. It is hard to work with a data set with key information missing. In this case there is not really anything we could fill these `NA` values with, we probably will not be able to do much, if at all, useful statistical analysis.

For experimental design, we looked at the four steps for doing any kind of statistical analysis.

1. Claim/Hypothesis: Any statement regarding the unknown population parameter you are studying (Maximum, Minimum, Average/Mean, etc.).
2. Experiment: Design the study

- Collecting the data set
- Designing the study

3. Likelihood: Analyzing the data set
4. Conclusion: Inference

You can do number 1. Using any kind of anecdotal data, you can formulate a claim/hypothesis. You probably have already done so. If you have ever observed something and asked yourself a question, you have done number 1. Numbers 3 and 4 will be coming in the next few weeks. Today we will focus on number 2. Remember you cannot do any analysis using anecdotal data, you must use available data that is collected. However, how do you go about collecting your data? You need to design an experiment or study to aid in the data collection process. This process is ***VERY*** important. A poorly designed study or experiment will usually result in inaccurate data, which will then usually result in inaccurate analysis results. Most survey designers when they go about designing their experiments or surveys do not design either for optimal analysis. This happens quite frequently. Today we will teach you some basic experiment and survey design.

Okay, now we can move onto the next portion of lecture content.

Experimental Design

Experimental vs. Observational

There are two different kinds of studies that can be done.

1. **Experimental:** We investigate the effects of certain conditions on individuals or objects in the sample.
2. **Observational:** We observe the response for a specific variable for each individual or object.

You can think of it as in an experiment, we control certain aspects and see what happens, while in an observational study, we do not control anything and see what happens. Now observational studies are pretty straight forward because all we have to do is record the data that we see. Experiments on the other hand can get more involved. For both though, we need to answer the following questions.

1. What are our experimental units? (Who/what are we trying to do inference on?)
2. What is the treatment that we are applying to these experimental units? (What are we doing during the experiment?)
3. What are the levels of the treatment? (Are we doing different things for different groups?)
4. What is the response variable? (What do we want to analyze/infer about who/what we are doing inference on?)
5. What is the statistical significance? (Is there an observed effect so large that it would rarely occur by chance?)

Again, these all need to be answered *before* the experiment/study is conducted.

Aspects of Designing a Study

There are three different aspects that can be in a study.

1. **Comparison:** Comparing all but one group with respect to the one group (control group).
2. **Randomization:** Allocation of members/participants into groups.
3. **Replications:** Repetition of participants within each group.

Ideally in a study, we want all three of these, however an observational study only has 1 & 3. An experimental study though does have all three. This makes experiments better than studies. So why do so many people choose to do studies? The answer is it is much cheaper and less time consuming to do an observational study. It is common for studies to be optimized based on the cost of doing it. (Money doesn't grow on trees.) Unless necessary, observational studies will be chosen due to the cost being more optimal.

Example 2a; Experimental vs. Observational?

The administration at the University of Nebraska is interested in student reaction to a planned parking garage on campus. A dormitory near the proposed site is selected and several Student Senate members volunteer to solicit responses. One Thursday evening, the volunteers each take a specific dorm wing, knock on doors, and record student answers to several prepared questions. Is this an observational or an experimental study?

Answer: Observational; Only one dorm is responding to the questions and the dorm was not randomly selected. It was selected due to its proximity to the proposed site.

Example 2b; Experimental vs. Observational?

Electric and plug-in electric cars are designed to save gasoline and help the environment. In addition, there are certain tax credits for these types of hybrid automobiles. Although there are certainly benefits to owning a hybrid car, many people complain about the slow acceleration, repair expense, and overall comfort. Thirsty-five passengers are randomly selected. Each is blind-folded and taken for a ride in a traditional combustion-engine automobile and in a comparably sized hybrid car (over the same route). The passenger is then asked to select the car with the most comfortable ride. Is this an observational or an experimental study?

Answer: Experimental; The people are randomly selected, are driven in both cars (along the same route), and the experiment can be repeated easily with different (randomly) selected people.

Other Kinds of Studies

- **Matched Pair Design:** This is another kind of study when each participant in the control group is matched with a participant from the rest of the other groups.
- **Block Design:** The random assignment of experimental units to treatments that are carried out within each block.
  - **Block:** A group of experimental units that are similar.

In statistics, you try to control what you can, block what you cannot control, and randomize to create comparable groups.

Sampling

Now we will move onto different types of sampling. There are a large variety out there, so we will just cover a select few.

- **Simple Random Sample (SRS):** A sample selected in such a way that every possible sample of size `n` has the same chance of being selected.
- **Stratified Random Sample:** Divide the population into smaller groups called strata and choose an SRS from each group.
  - Everyone/thing in the strata will have some common trait/attribute about them.

The selection of the SRS or SRS from strata is done using a probability sample. This is when each sample is chosen by chance. We can choose a probability sample in two ways. We can either do with or without replacement. With replacement means that observations could be selected multiple times. You randomly select an observation, and then put it back in the list of possible observations to be selected. Without replacement means that observations cannot be selected multiple times. You randomly select an observation, and then remove it from the list of possible observations to be selected. Let's try this and see how our results come out.

Example 3; Sampling Beach Data

Let's suppose you are have been asked to do a study about the beaches in Chicago along Lake Michigan. The city wants to find out what the average temperature (In Celsius) of the beaches are. They will then hire another person to look at whether climate change has affected the long-term average temperature or not. Your job is simply to design a study to collect the data. Pretend in this problem that you do not know what the data looks like even though we already have it and can calculate the estimated average temperature. Let's see how different kinds of sampling effect the results.

Here you choose to have the beaches hire people to take the temperature and the first 1,000 reported will be your sample to calculate from. We can replicate that below using the `tail()` function. It is like the `head()` function, only it chooses the last `n` observations in a data set. (The data is in order by data so we can do this.)
The syntax is just like the `head()` function, `tail(data, n)`. To calculate the mean (average) for a variable, you will use the `mean()` function.
The `mean()` function is **NOT** part of the `tidyverse` so the syntax will be a bit different. The syntax is `mean(data$variable)`.

```
Beach <- read.csv("~/Classes/DPISu22/Data Sets/bwq.csv", stringsAsFactors=TRUE)
Beach = Beach %>% drop_na() #Removing rows with missing data.
sample = Beach %>% tail(1000)
mean(sample$Water.Temperature)
```

```
## [1] 23.0278
```

While you have your results here, there is an issue. This is a convenience sample, which is **NOT** a probability sample. You have done an observational study with no randomization, which invites a large amount of bias. We will cover this in a bit. Needless to say, this is bad because your result may not be accurate. Let's now try doing probability samples. While you would think this is difficult or tedious, good news, `R` has a built in `sample()` function. The syntax for this is a bit more complex. It is `data[sample(nrow(data), size = n, replace = ...),]`. Here, `n` is the desired sample size and `replace = ...` is where you decide if you are going to sample with or without replacement. If you want to do with replacement, do `replace = TRUE`, otherwise put `replace = FALSE`. In this case, it makes sense to use without replacement. `nrow(data)` tells `RStudio` to select all the columns in the data set for the sample. We do not want to pick observations twice that could potentially cause bias in our estimate.

```
set.seed(143572) # To be able to replicate results.
sampleSRS = Beach[sample(nrow(Beach), size = 1000, replace = FALSE),]
mean(sampleSRS$Water.Temperature)
```

```
## [1] 18.9372
```

Interesting, notice how now our estimate for the average temperature is much lower than before. By almost 3 degrees Celsius. Random sampling versus convenience sampling as you see can cause massive differences in your results. Just for fun, let's see what happens if you do random sampling with replacement. It will not have as much weight here because it makes more sense to use without replacement, but let's just see what happens.

```
set.seed(143572) #To be able to replicate results.
sampleSRSwR = Beach[sample(nrow(Beach), size = 1000, replace = TRUE),]
mean(sampleSRSwR$Water.Temperature)
```

```
## [1] 19.0362
```

Interesting, even with replacement the average is much closer to our SRS than convenience sample. Now let's try doing a stratified sample. Remember for stratified samples, you need to find a trait to divide the population into groups with. For this example I have chosen to divide up by the different beaches since that information is in the data set. First, let's see how many different beaches there are. We can do this using the `table()` function. This will tell us all possible values of a categorical variable. It will also tell us how often they occur in the data set. This function is also not a part of the `tidyverse`, so you will again have that slightly different syntax. The syntax is `table(data$variable)`.

```
table(Beach$Beach.Name)
```

```
##
## 63rd Street Beach     Calumet Beach    Montrose Beach Ohio Street Beach
##               934              2128              2226              2003
##    Osterman Beach     Rainbow Beach
##              1809               934
```

Notice how we have six different Chicago beaches here. So we will have to create six different strata. We can do that using our conditionals that we have learned. A way to check if you did this correctly is if the number of observations in the new filtered data sets is the same as our table numbers. To keep our sample sizes similar, we then adjust how many we take from each sample. Unfortunately, 1,000 is not divisible by 6, so we will have to round some samples up and others down.

```
set.seed(143572) #To be able to replicate results.
Beach63rd = Beach %>% filter(Beach.Name == "63rd Street Beach")
BeachCalumet = Beach %>% filter(Beach.Name == "Calumet Beach")
BeachMontrose = Beach %>% filter(Beach.Name == "Montrose Beach")
BeachOHst = Beach %>% filter(Beach.Name == "Ohio Street Beach")
BeachOsterman = Beach %>% filter(Beach.Name == "Osterman Beach")
BeachRainbow = Beach %>% filter(Beach.Name == "Rainbow Beach")
sample1 = Beach63rd[sample(nrow(Beach63rd), size = 166, replace = FALSE),]
sample2 = BeachCalumet[sample(nrow(BeachCalumet), size = 167, replace = FALSE),]
sample3 = BeachMontrose[sample(nrow(BeachMontrose), size = 167, replace = FALSE),]
sample4 = BeachOHst[sample(nrow(BeachOHst), size = 166, replace = FALSE),]
sample5 = BeachOsterman[sample(nrow(BeachOsterman), size = 167, replace = FALSE),]
sample6 = BeachRainbow[sample(nrow(BeachRainbow), size = 167, replace = FALSE),]
```

Now we have the SRS for each strata. To combine them together into a single sample, we can use the `bind_rows()` function. The syntax for this is easy, `bind_rows(x1, x2, ...)`.

```
sampleStr = bind_rows(sample1, sample2, sample3, sample4, sample5, sample6)
mean(sampleStr$Water.Temperature)
```

```
## [1] 18.8765
```

Now notice how the average is lower than the SRS, but not by that much. Also, it is not *nearly* as much of a change as the non-probability sample to the SRS. Let's see what the actual average is.

```
mean(Beach$Water.Temperature)
```

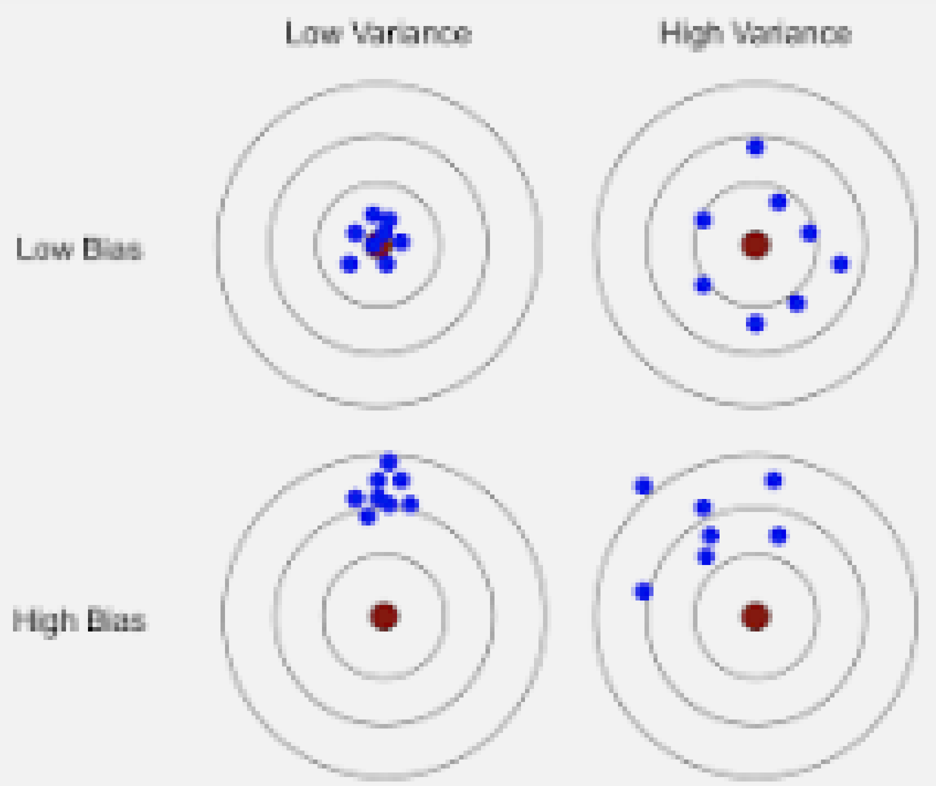
```
## [1] 19.07552
```

It looks like the random sampling procedures did a much better job than the non-random sampling one.

There are many more ways to do sampling and experimental design. Such that there are entire college courses for both. However, these few options above are just to get your feet wet.

Bias vs. Variablility

- **Bias:** The design of a study is biased if it systematically favors certain outcomes.
- **Variability:** The lack of consistency or a fixed pattern.



Low bias and variance is the best possible case for your study. This will give you the best results. To ensure this, you can use random sampling procedures to reduce bias, and take as large a sample as possible to reduce variability.

Example 4; Potential Bias

This about the last example. It appears that your first sample (most recent observations) is biased as it is not even close to the other three estimates. There is probably some high bias there. What are potential sources of bias here?

- Recency; If you take all your results too close together, you risk getting the short-term average instead of the long-term one.
- Time Convenience; You are relying on people to take the average temperature at different times of day to get a long term average. However, you could end up with people taking it at the beginning or end of their shifts, which would cause bias in the results as most shifts are going to be starting and ending around the same time. You ideally want the temperature taken at a variety of times.

More Potential Issues

- **Confounding Variables:** This is when two variables are associated in such a way that their effects on a response variable can not be distinguished from each other.
- **Lurking Variables:** This is a variable that is not among the explanatory or response variables in a study, but that may influence the variables in the study.

Example 5; Lurking Variables

A study was done once (don't ask me when, I can't remember) where the researchers found a strange correlation. There was a massive increase in drownings when there was a massive increase in ice cream sales. Obviously an increase in ice cream sales does not lead to an increase in drowning, so there is a lurking variable here. What is it?

Answer: The season (Summer)

While this is a simple example, it shows how you have to look for lurking variables. Remember, correlation does not equal causation. Scientists use the following steps to establish causation between two variables.

1. Perform an experiment
2. Find a strong association
3. Repeat experimental trials a large amount of times
4. Find that the association is consistent
5. The alleged cause always precedes the effect

Hopefully this gives you a nice introduction to experimental design. While we did not design a specific experiment here, you have practiced doing everything you will need to design one (in lab). Remember to sample your data using a random procedure if you can and design your experiment optimally. You will have to use your own judgement, background knowledge, and common sense to do so. We will get into some more advanced sampling procedures later on in the course.

End of Lecure 4 Notes