EDS 222: Week 1: In-class Assignment

{STUDENT NAME}

2021-09-08

Sampling in R

We are going to learn two types of sampling methodologies here:

(i) Simple Random Sampling and;
(ii) Stratified Sampling.

For this we will make use of the county level data from the `usdata` package. This data frame contains data from counties in all 50 states plus the District of Columbia.

Step 1:

Use the `library(usdata)` to load the data.

Step 2:

You are tasked at the EPA to collect data on air quality from all 50 states in the US. However, the current president has initiated significant budget cuts in the EPA which only allows them to collect data from 150 out of the ~3000 counties in the United States. Take a random sample of 150 counties that you think can be representative of the US and store them in a tidy data frame called `county_srs` You are encouraged to write a function to do this.

Step 3:

Evaluate the distribution of sampled counties across different states in the data frame `county_srs`. Are the number of sampled counties in each state different? Would you call this a representative sample?

Step 4:

Using your answer from step 3, think about why or why not is the sampling strategy in step 2 representative of the population. Use the concepts of stratified random sampling learned in class to select an equal number of counties in each state. This can also be referred to as stratification on state, can you explain why or why not would this strategy make the sampling mor representative.

Sampling issues in real data:

How does sampling influence your data and what you can learn from it?

As we have learned in the lectures that statistic derived from the population data and the sample data differ. The statistic derived from the population is referred to as the estimand where the statistic derived from the sample data is referred to as the estimate.

In this exercise we will look at a case study from the developing world. We will take the case of air quality in South Asia. Out of the 6 South Asian countries, Pakistan was ranked to contain the second most air pollution in the world in 2020 (IQAIR, 2020). In 2019 Lahore, Pakistan was the 12th most polluted city in the world with a population of 11.1 million people.

We make use of two sources of data: (i) crowdsourced data from air quality monitors in people's homes; (ii) government data from official monitors installed at selected places in Lahore.

Step 1:

Load the data from dropbox and label it as `crowdsourced` and `govt` accordingly:

Step 2:

Explain in words what the ideal population dataframe for air quality in Lahore would look like. Use both the crowdsourced data and the government and try to identify how both of these datasets can be products of different sampling strategies.

Step 3:

For each of the government data and crowdsourced generate estimates of the mean, min, max of the data and discuss how the sampling strategy biases (or doesn't) those parameters as approximations of population statistics.

Step 5:

Identify why the government and corwdsourced data differ and how can they be related to biases generated as a result of the sampling process.

Designing an experiment

In this section, we will work through how a randomized controlled trial (RCT) is designed. We will go through how a data for a randomized trial can be simulated.

Step 1:

Generate a data frame with 100 observations. Add a random normal variable for potential outcomes for Y0 and Y1:

Step 2:

Assign a binary treatment variable `Z` which has equal probability of being in control and treatment:

Step 3:

Contruct observed outcome using Y0 and Y1:

Step 4:

Use a simple difference in means to calculate the treatment effect as the estimand and the estimate:

Step 5:

Now, we would like to do this in a simulation over multiple trials. First, start with doing steps 1-4 in a single pipeline and generating a function to execute them.

Step 6:

Using the function in 5 and replicate it times and report what you observe about the results:

Step 7:

add blocking within the sampling part:

Step 7.1:

First we need to identify what blocking is and why is it needed. For this we need to see what is the bias that our experiment suffers from due to lack of blocking. To do this, design an experiment, randomize the treatment status and see the bias in the simulated variable. Prove that in over 10,000 iterations this bias is reduced to 0.

Step 7.2:

Show the iterated bias generated in step 1 in a plot

Step 7.3:

Now we need to think about blocking. We will start to do this by changing the function to use block randomization. For this, first, we should construct blocks. First generate one observed covariate and construct blocks according to that.

Step 7.3.1:

we need to create a block indicator. Blocks are based on observed units being similar to each other in the greatest way possible. Make use of the function `arrange()` to sort the data in the order of the observed covariate and then mutate a block indicator where the the data resembles a series where the first two units are in block 1, the second two are in block 2, etc.

Step 7.3.2:

Now, once we have blocks, we need to randomize all the units within the blocks where one unit will go to treatment and one will go to control. randomize *within* blocks, exactly one to treatment and one to control.