HW7: Repeated Sampling and Water Conservation

2022-11-02

In this homework we will use R to simulate selecting a random sample from a population, repeatedly. We will do this under the unusual circumstance where we have data for the whole population. This means we can compare what we learn in each sample (estimates) to what we wanted to know about the population (population parameters) in the unusal situation where we know the population paramater values.

In 2007, a water utility in Atlanta implemented a natural field experiment using all of their water customers, which was just under 140,000 households The data we use for this HW gives the water use for that whole population. Although not our main focus for this assignment, the water utility randomized their customers (as households) into four treatment arms: a control group, a group that received technical advice, a group that received both technical advice and an appeal to pro-social preferences, and a group that received both technical advice and an appeal to pro-social preferences that included a social comparison (see Ferraro and Price 2009). In a later assignment we will look at the treatment effects we estimate in random sampling compared to the treatment effects in the whole population.

The data we analyze are available as the CSV file `water.csv`. The names and descriptions of variables in the data set are:

Name	Description
`group`	1 = control; 2 = treatment A, 3 = treatment B, 4 = treatment C
`WATER_2006`	Water use for a household in 2006.
`APR_MAY_07`	Water use for a household in April and May of 2007
`SUMMER_07`	Water use for a household in Summer (June - August) of 2007

Each observation in the data represents a household, and for each household the file contains information about its treatment status, its water use prior to the field experiment (2006), its water use during the field experiment (spring 2007), and its water use after the field experiment (Summer 2007).

For this HW we will use the `WATER_2006` and `CONTROL` variables only.

```
set.seed(1988) # <-- This will make your Rmd have the same output every time.
# Change the number in the command in line 42 when you do this HW (just change it once and keep it at
# that value)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
water <- read.csv("data/water.csv") %>%
  mutate(control = ifelse(group == 1,1,0))
```

Question 1 [8 pts]

1a

What is the mean water use (in 2006) in the population? What is the standard deviation of water use (in 2006) in the population? What proportion of households in the population are in the control group? These are the population parameters. Imagine we want to learn about these population parameters by gathering data for a random sample from that population.

1b

Draw one random sample of 900 observations. In this question and what follows consider this to be *YOUR* sample. What is the mean water use in your sample? What is the standard deviation of water use in your sample? What proportion of the households in your sample were in the control group?

How do the values in your sample - which we could use as estimates of these same features in the population - compare to the corresponding population parameter values? Are they larger/smaller? Do they seem close or far from the population proportion?

Answer 1

Answer 1a

Answer 1b

Question 2 [10 pts]

2a

Using the code given below, draw 1000 random samples of size 900 and for each record the mean water use.

2b

Make a histogram of the mean water use from the 1000 samples. Calculate summary statistics for the mean water use from the 1000 samples. Describe the distribution of the mean water use over repeated sampling. How does the mean of these 1000 sample means compare to the population mean? How does the standard deviation of these 1000 sample means compare to the standard deviation of water use in the population?

2c

The standard deviation of the sample means over a large number of samples of the same size is an estimate of the standard error of the sample mean. Use this estimate of the standard error of the sample mean in the next part of this question.

What proportion of these 1000 samples had a mean of water use in 2006 that was within 1 standard error of the population proportion? What proportion of these 1000 samples had a mean of water use in 2006 that was within 2 standard errors of the population proportion? How many standard errors away from the population proportion was the mean of *your* sample?

Answer 2

Answer 2a

```
#' SimulateSamplingDistribution: draws specified number of samples from a dataframe representing a popu
#'
#' @param population_data dataframe (or tibble) containing population data
#' @param number_samples number of samples to draw
#' @param sample_size number of observations to sample for each draw
#' @param variable_name variable of interest (e.g. a column within population_data)
#' @param statistic a statistic to calculate of the sampled variable of interest (e.g., mean, sd, media
#' @param seed fixes the samples so that the same samples are drawn each time. this is set to a default
#'
#' @return number_samples length vector of statistics for variable_name. this represents the sampling d

SimulateSamplingDistribution <- function(population_data, number_samples,
                                         sample_size, variable_name,
                                         statistic, seed = 10) {
  set.seed(seed)
  data_samples <- map(1:number_samples, ~sample_n(population_data, sample_size))
  res <- unlist(map(data_samples, ~statistic(.x[[variable_name]])))
  return(res)
}

#Uncomment and fill in each of the inputs below:

# repsamp.water.900 <- SimulateSamplingDistribution(population_data = ,
#                                                   number_samples = ,
#                                                   sample_size = ,
#                                                   variable_name = ,
#                                                   statistic = ,
#                                                   seed = )
```

Answer 2b

Answer 2c

Question 3 [11 pts]

3a

Using the SimulateSamplingDistributions function, draw 1000 random samples of size 900 and for each record the proportion of households that are in the control group.

3b

Make a histogram of the proportion of households in the control group from the 1000 samples. Calculate summary statistics for the proportion of households in the control group from the 1000 samples. Describe the distribution of the proportion of households in the control group over repeated sampling (the simulated estimate of the sampling distribution). How does the mean of the 1000 sample proportions compare to the population proportion? What is the standard deviation of the 1000 sample proportions?

3c

What proportion of these 1000 samples had a share of households in the control group that is more than 3 percentage points away from the population proportion?

3d

The standard deviation of the sample proportion over a large number of samples of the same size is an estimate of the standard error of the sample proportion. Use this estimate of the standard error of the sample proportion in the next part of this question.

What proportion of these 1000 samples had a share of households in the control group that was within 1 standard error of the population proportion? What proportion of these 1000 samples had a share of households that was within 2 standard errors of the population proportion? How many standard errors away from the population proportion was the share of households in the control group in *your* sample?

Answer 3

Answer 3a

```
## Uncomment the below chunk and fill in the numbers!

# repsamp.control.900 <- SimulateSamplingDistribution(population_data = ,
#                                                     number_samples = ,
#                                                     sample_size = ,
#                                                     variable_name = ,
#                                                     statistic = ,
#                                                     seed = )
```

Answer 3b

Answer 3c

Answer 3d

Question 4 [10 pts]

4a

Using the SimulateSamplingDistributions function, draw 1000 random samples of size 900 and for each record the the standard deviation of water use among the 900 households in the sample.

4b

Make a histogram and a box-plot of the standard deviation of water use from the 1000 samples. Calculate summary statistics for the standard deviation of water use from the 1000 samples. Describe the distribution of the standard deviation of water use over repeated sampling. How does the mean of the 1000 sample standard deviations compare to the population standard deviation? How much larger (as a ratio) is the largest sample standard deviation than the population standard deviation? How much smaller (as a ratio) is the smallest sample standard deviation than the population standard deviation?

Answer 4

Answer 4a

```
#Uncomment and fill in below

# repsamp.water.sd.900 <- SimulateSamplingDistribution(population_data = ,
#                                                      number_samples = ,
#                                                      sample_size = ,
#                                                      variable_name = ,
#                                                      statistic = ,
#                                                      seed = )
```

Answer 4b

Question 5 [10 pts]

5a

Using the SimulateSamplingDistributions function, draw 1000 random samples of size 4000 and for each record the mean water use in 2006. You will use very similar code to what you used in Q2, you just need to change the "sample_size" and give it a different object name.

5b

Calculate summary statistics of the 1000 sample means, now from samples of size 4000. What is the standard deviation of these 1000 sample means? Make a histogram of these 1000 sample means and describe it. How does the mean of these sample means compare with the population mean water use? How does the standard deviation of these sample means compare with the population standard deviation of household water use?

5c

The standard deviation of the sample means over a large number of samples of the same size is an estimate of the standard error of the sample mean. Use this estimate of the standard error of the sample mean for samples of size n = 4000 in the next part of this question.

What proportion of these 1000 samples of size 4000 had a mean of water use in 2006 that was within 1 standard error of the population proportion? What proportion of these 1000 samples had a mean of water use in 2006 that was within 2 standard errors of the population proportion?

Answer 5

Answer 5a

Answer 5b

Answer 5c

Question 6 [8 pts]

6a

Compare the distributions of sample mean water use between samples of size 900 and samples of size 4000. Make two SEPARATE histograms of the sample mean water use. ONE histogram from the 1000 samples of size 900 and ONE histogram from the 1000 samples of size 4000. How do their shapes compare? Their measures of central tendency? Their measures of spread? Any notable features?

6b

How did the proportion of sample means within 1 or 2 standard errors of the population mean compare for samples of size 900 and samples of 4000 (you calculated these proportions in Q2 and Q5, now compare them)?

Answer 6

Answer 6a

Answer 6b