HW7: Repeated Sampling and Water Conservation

2022-11-09

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.

```r
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()
```

```r
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

```r
mean(water$WATER_2006)
```

```
## [1] 58.31386
```

```r
sd(water$WATER_2006)
```

```
## [1] 41.13629
```

```r
mean(water$control)
```

```
## [1] 0.1094507
```

The mean water usage in 2006 in the population is about 58.31 thousand gallons. The standard deviation of water use in the population is about 41.14 thousand gallons of water. The proportion of the households in the population that are in the control group is about 11 percent. ### Answer 1b

```r
set.seed(1988)
samp1 <- sample_n(water, 900)
mean(samp1$WATER_2006)
```

```
## [1] 60.86333
```

```
sd(samp1$WATER_2006)
```

```
## [1] 43.96847
```

```
mean(samp1$control)
```

```
## [1] 0.1277778
```

The mean water usage of *my* sample is 60.86 thousand gallon. This is a little higher than the population mean of 58.31. The mean proportion of the households in the population that are in the control group is 12.78 percent and is a little higher than the population proportion of the households in the control group which is 11%. Further, the sample standard deviation of water usage is 43.97 thousand gallons which is also a little higher than the population standard deviation of 41.14 thousand gallons of water. Overall, the sample statistics are little higher than the corresponding population parameter values.

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?

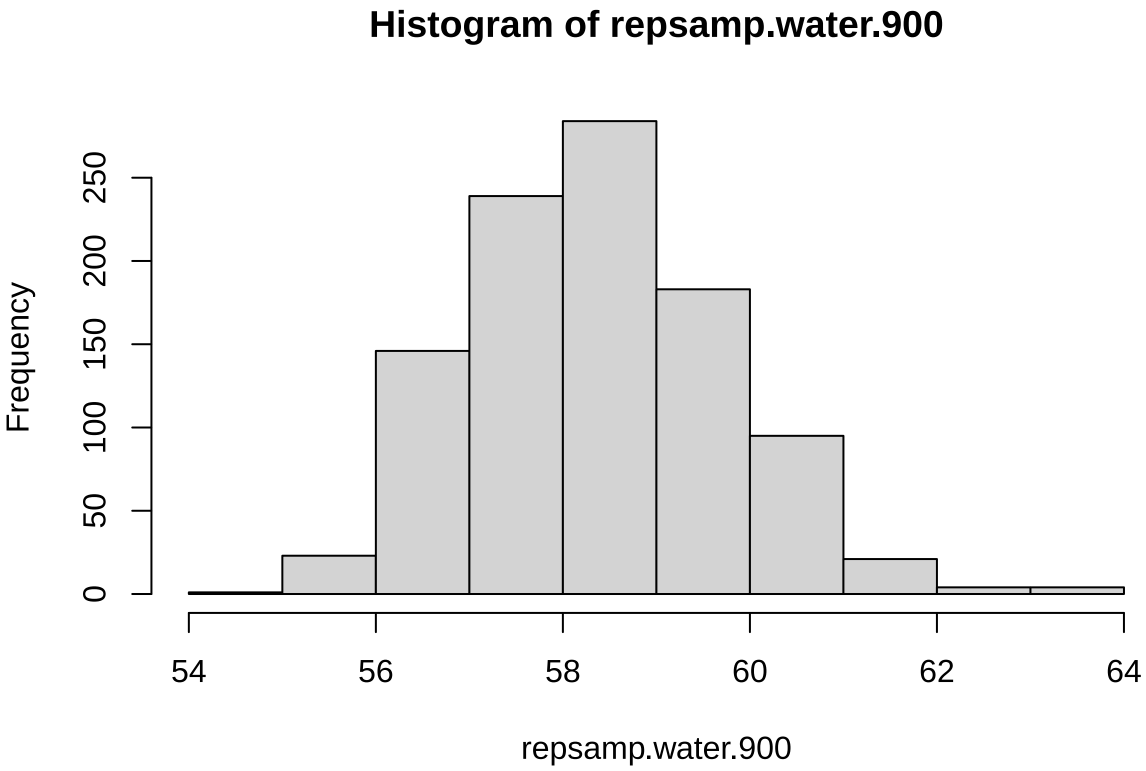
Answer 2

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 = water,
                                          number_samples = 1000,
                                          sample_size = 900,
                                          variable_name =  "WATER_2006",
                                          statistic = mean,
                                          seed = 10)
```

Answer 2b

```
hist(repsamp.water.900)
```



```
summary(repsamp.water.900)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   54.18   57.36   58.33   58.36   59.22   63.16
```

```
sd(repsamp.water.900)
```

```
## [1] 1.355169
```

The histogram of distribution of sample means of water usage from 1000 samples of size 900 each is a little right skewed and appears concentrated around 59, with a range of about 54 to 64. There is no evidence from this figure of the outliers, spikes, or gaps. The summary statistics of the repeated sample of sample size 900 shows the mean of the means water usage in the sample is 58.36 thousand gallons with a range between 54.18 and 63.16, IQR between 57.36 and 59.22, and the median is 58.33. The mean of this sample means is much close to the population mean of 58.31 (if we took more than 1000 samples it would get increasingly closer). The standard deviation of the means is 1.36 while that of population standard deviation is 41.14 so they differ by almost a little less than 40 for a sample of 900. We expect the sample means of water usage to be about 1.36 thousand gallons away from the combined sample mean of water usage. Individual household water usage in the population may very much and it will impact the variation of individual sample means in these 1000 samples of sizes 900 from the combined sample mean i.e. higher the variation in population, greater the variation between means in the sample.

Answer 2c

```
se <- sd(repsamp.water.900)
se
```

```
## [1] 1.355169
```

```
pop_mean = mean(water$WATER_2006)
pop_mean
```

```
## [1] 58.31386
```

```
mean(repsamp.water.900 > pop_mean - se & repsamp.water.900 < pop_mean + se)
```

```
## [1] 0.674
```

```
mean(repsamp.water.900 > pop_mean - 2*se & repsamp.water.900 < pop_mean + 2*se)
```

```
## [1] 0.959
```

```
diff <- (mean(repsamp.water.900) - pop_mean)/se
diff
```

```
## [1] 0.03060719
```

We see that 67.4 percent of these 1000 samples had a mean of water use in 2006 that was within 1 standard error of the population mean of 58.31 and 96 percent were within 2 standard errors. Further, we find that the mean water use in this sample was 0.03 standard errors from the population mean.

## 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?

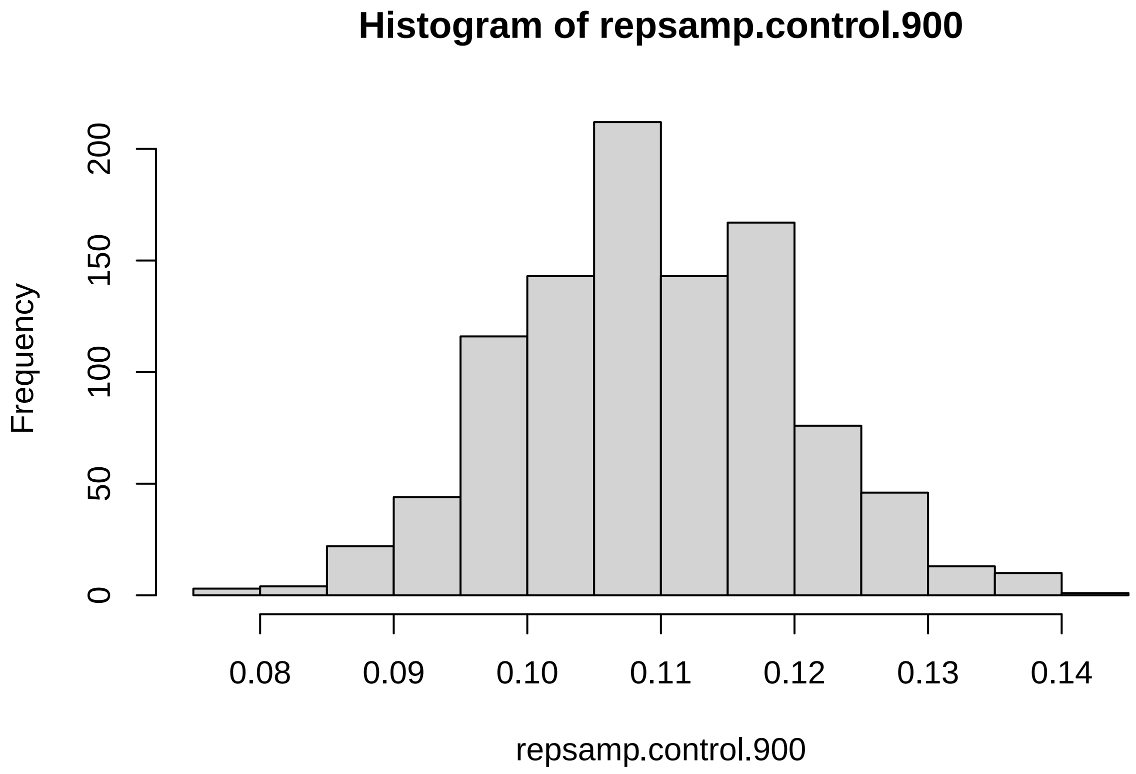
Answer 3

Answer 3a

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

 repsamp.control.900 <- SimulateSamplingDistribution(population_data = water,
                                          number_samples = 1000,
                                          sample_size = 900,
                                          variable_name = "control",
                                          statistic = mean,
                                          seed = 10)
```

Answer 3b

```
hist(repsamp.control.900)
```



```
mean(water$control)
```

```
## [1] 0.1094507
```

```
summary(repsamp.control.900)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07889 0.10222 0.10889 0.10962 0.11667 0.14111
```

```
mean(repsamp.control.900)
```

```
## [1] 0.1096178
```

```
sd.repsamp.c.900 <- sd(repsamp.control.900)
sd.repsamp.c.900
```

```
## [1] 0.01037381
```

```
sd(water$control)
```

```
## [1] 0.3122054
```

The histogram of distribution of sample means of households in the control group from 1000 samples of size 900 each is almost symmetric, unimodal and appears concentrated around 0.11, with a range of about .07 to .15. There is no evidence from this figure of the outliers or gaps; however, there is a little spike around .12. The summary statistics of the repeated sample of sample size 900 shows the mean of the sample means of the households in the control group is 0.11 with a range between 0.08 and 0.14 and IQR between 0.10 and 0.11. The mean of this sample means is pretty close to the population mean of 0.11 (if we took more than 1000 samples it would get increasingly closer). The standard deviation of the means is 0.01 while that of population standard deviation of people in the control group is .31. We expect the sample means of proportion of households in the control group to be about 0.01 away from the combined sample mean. ### Answer 3c

```
meanpop.control <- mean(water$control)
meanpop.control
```

```
## [1] 0.1094507
```

```
mean(repsamp.control.900)
```

```
## [1] 0.1096178
```

```
1-(mean(repsamp.control.900 > meanpop.control - 0.03 &
          repsamp.control.900 < meanpop.control + 0.03))
```

```
## [1] 0.004
```

Around 0.4% 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. ### Answer 3d

```
se.control <- sd(repsamp.control.900)
se.control
```

```
## [1] 0.01037381
```

```
meanpop.control
```

```
## [1] 0.1094507
```

```
mean(repsamp.control.900 > meanpop.control - se.control &
       repsamp.control.900 < meanpop.control + se.control)
```

```
## [1] 0.665
```

```
mean(repsamp.control.900 > meanpop.control - 2*se.control &
       repsamp.control.900 < meanpop.control + 2*se.control)
```

```
## [1] 0.959
```

```
diff <- (mean(repsamp.control.900) - meanpop.control)/se.control
diff
```

```
## [1] 0.01610277
```

We see that 67 percent of these 1000 samples had a mean of households in the control group that was within 1 standard error of the population mean of 0.11 and 96 percent were within 2 standard errors. Further, we find that the mean of households in the control group in this sample was 0.02 standard errors from the population mean.

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?

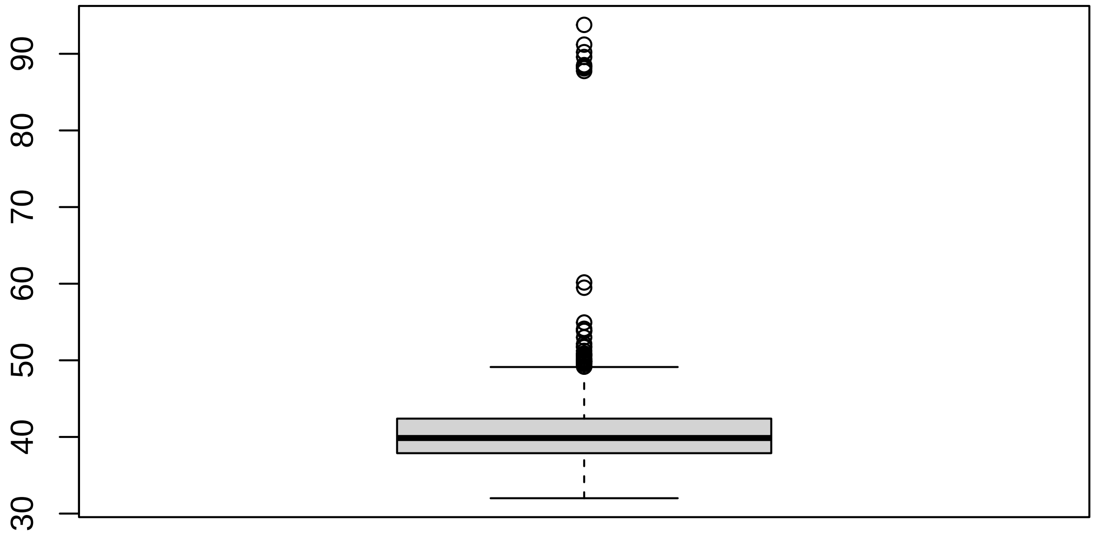
Answer 4

Answer 4a

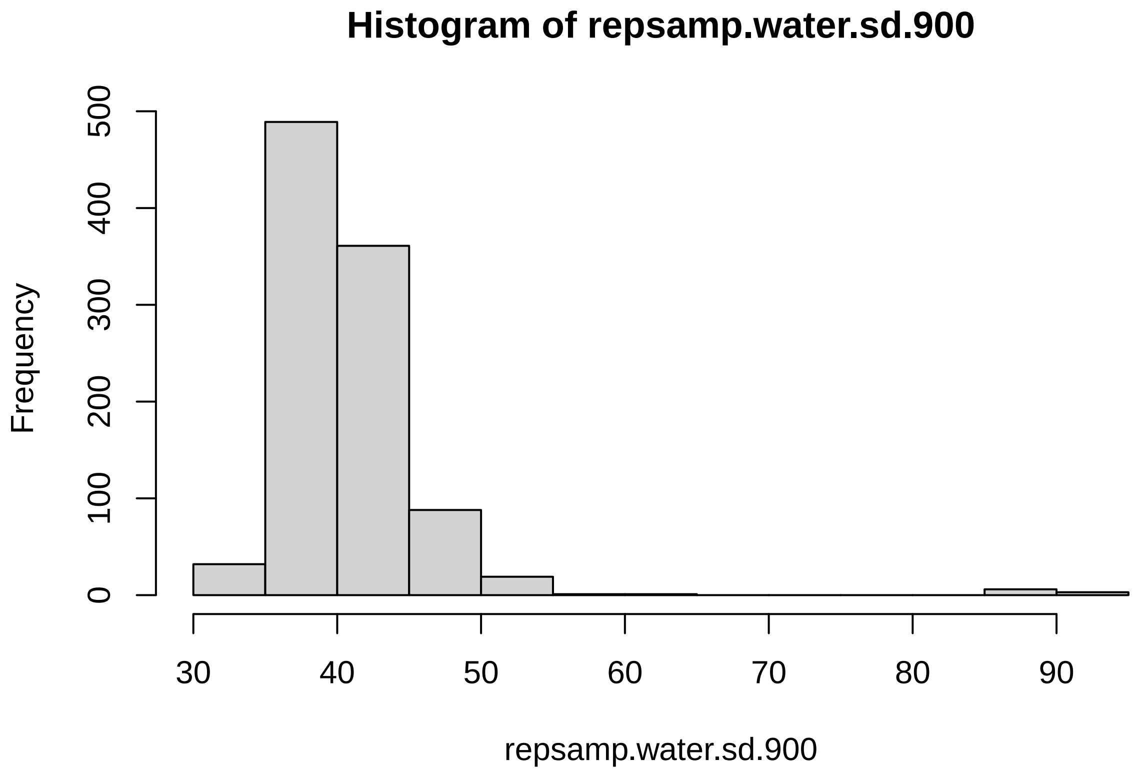
```
#Uncomment and fill in below

repsamp.water.sd.900 <- SimulateSamplingDistribution(population_data = water,
                                                     number_samples = 1000,
                                                     sample_size = 900,
                                                     variable_name = "WATER_2006",
                                                     statistic = sd,
                                                     seed = 10)
```

Answer 4b

```
boxplot(repsamp.water.sd.900)
```



```
hist(repsamp.water.sd.900)
```



```
summary(repsamp.water.sd.900)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   32.01   37.88   39.86   40.86   42.38   93.77
```

```
sd(water$WATER_2006)
```

```
## [1] 41.13629
```

```
(93.77/sd(water$WATER_2006))
```

```
## [1] 2.279496
```

```
1- (32.01/sd(water$WATER_2006))
```

```
## [1] 0.2218549
```

The boxplot of the standard deviation of water usage in the sample of 900 shows a median lies around 40 with outliers above the upper tail and there are gaps in the distribution. Similarly, the histogram of the standard deviation shows a huge right skew with gaps, spikes and many outliers on the right side while the mean center around 40 and it is unimodal. The summary statistics of the repeated sample of sample size 900 shows the mean of the standard deviations of water usage in the sample is 40.86 thousand gallons with a range between 32.01 and 93.77, IQR between 37.88 and 42.38, and the median is 39.86. The mean of this sample standard deviations is much close to the population standard deviation of 41.13 (if we took more than 1000 samples it would get increasingly closer). The largest sample standard deviation is around 228% larger than the population standard deviation. The smallest sample standard deviation is 22% smaller than the population standard deviation.

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?

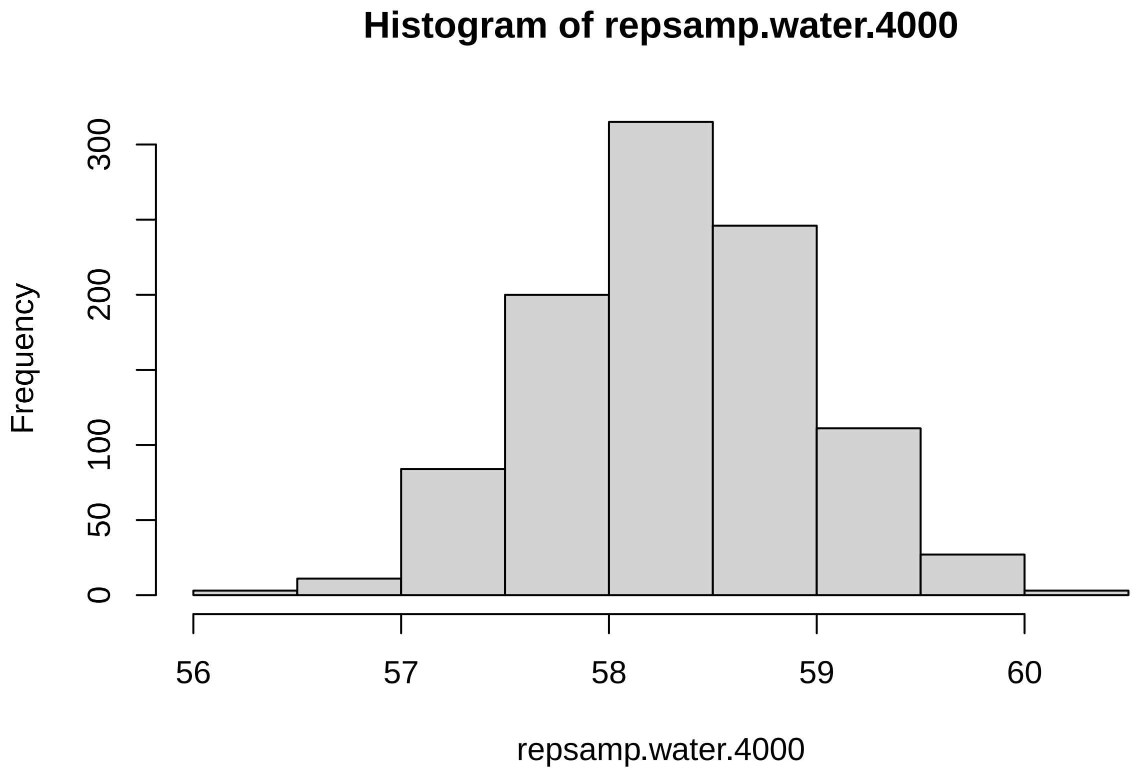
Answer 5

Answer 5a

```r
repsamp.water.4000 <- SimulateSamplingDistribution(population_data = water,
                                                   number_samples = 1000,
                                                   sample_size = 4000,
                                                   variable_name =  "WATER_2006",
                                                   statistic = mean,
                                                   seed = 10)
```

Answer 5b

```r
hist(repsamp.water.4000)
```



```
summary(repsamp.water.4000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.35   57.92   58.35   58.33   58.72   60.33
```

```
sd(repsamp.water.4000)
```

```
## [1] 0.6210769
```

```
mean(water$WATER_2006)
```

```
## [1] 58.31386
```

```
sd(water$WATER_2006)
```

```
## [1] 41.13629
```

```
# comparison of the standard deviation
```

The histogram of distribution of sample means of water usage from 1000 samples of size 4000 each is almost symmetric and appears concentrated around 58, with a range of about 56 to 60. There is no evidence from this figure of the outliers, spikes, or gaps. The summary statistics of the repeated sample of sample size

4000 shows the mean of the means water usage in the sample is 58.33 thousand gallons with a range between 56.35 and 60.33, IQR between 57.92 and 58.72, and the median is 58.35. The mean of this sample means is much close to the population mean of 58.31 (it is realtively more closer to the population mean than that of the 900 sample size mean). The standard deviation of the means in this sample is 0.62 while that of population standard deviation is 41.14 so they differ by almost a little more than 40 for 1000 samples of sample size 4000. We expect the sample means of water usage to be about 0.62 thousand gallons away from the combined sample mean of water usage. Individual household water usage in the population may very much and it will impact the variation of individual sample means in these 1000 samples of sizes 4000 from the combined sample mean i.e. higher the variation in population, greater the variation between means in the sample. ### Answer 5c

```
se.4000 <- sd(repsamp.water.4000)
se.4000
```

```
## [1] 0.6210769
```

```
pop_mean = mean(water$WATER_2006)
mean(repsamp.water.4000 > pop_mean - se.4000 &
       repsamp.water.4000 < pop_mean + se.4000)
```

```
## [1] 0.686
```

```
mean(repsamp.water.4000 > pop_mean - 2*se.4000 &
       repsamp.water.4000 < pop_mean + 2*se.4000)
```

```
## [1] 0.952
```

We see that 68.6 percent of these 1000 samples of size 4000 had a mean of water use in 2006 that was within 1 standard error of the population mean of 58.31 and 95.2 percent were within 2 standard errors.

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?

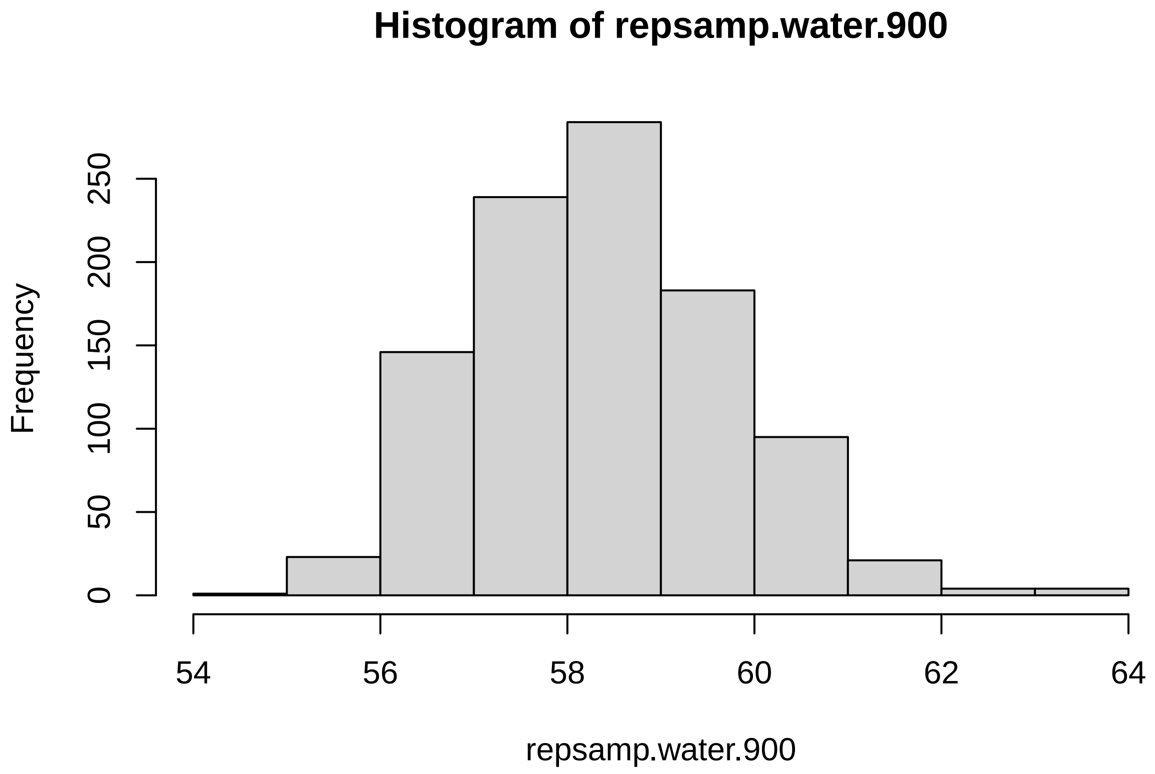
6b

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)?

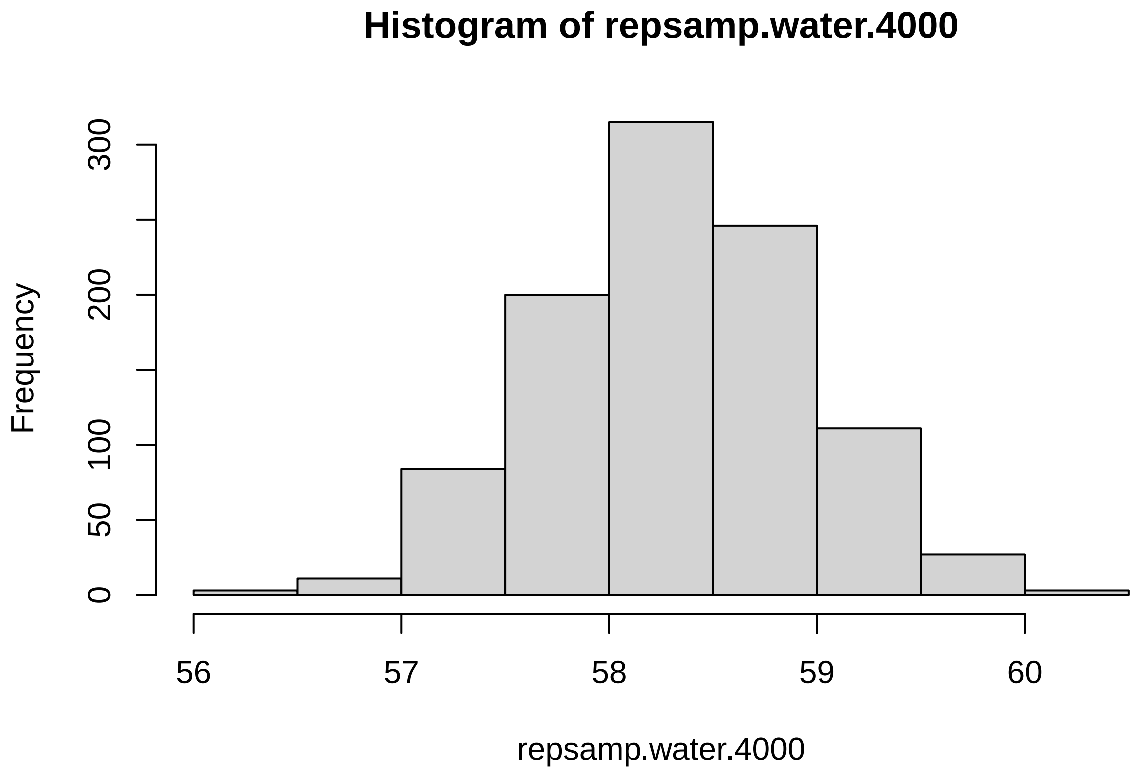
Answer 6

Answer 6a

```
hist(repsamp.water.900)
```



```
hist(repsamp.water.4000)
```



```
summary(repsamp.water.900)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.18   57.36   58.33   58.36   59.22   63.16
```

```
summary(repsamp.water.4000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.35   57.92   58.35   58.33   58.72   60.33
```

Histogram of sample of sample size 900 is a little right skewed, unimodal, concentrated around 59, no gaps, spikes and outliers, and ranging between 54 and 64; while the histogram of sample size of 4000 seems symmetric, bell shapped, unimodal, concentrated around 58 and ranging between 56 and 60.5 with no gaps, spikes and outliers. The overall distribution of of the sample size of 900 shows that the mean is 58.36, median is 58.33, range between 54.18 and 63.16, and IQR between 57.36 and 59.22 (all units in thusand gallons). The overall distribution of of the sample size of 4000 shows that the mean is 58.33, median is 58.35, range between 56.35 and 60.33, and IQR between 57.92 and 58.72 (all units in thusand gallons). Conclusively, the means of sample size 4000 are more tightly concentrated around the population mean (i.e. have a smaller standard error). This shows that with a larger sample (4000 here), the estimated sample mean gets closer to the population mean as compared to the smaller sample size(900 here). In terms of notable features, both sampling distributions do not have outliers, gaps and spikes which demonstrates that at both sample sizes it is rare but possible to get a random sample with a sample summary statistic much farther from the population parameter value than expected for samples of that size.

Answer 6b

in 1000 samples of sample size 900, 67.4 percent of these 1000 samples had a mean of water use in 2006 that was within 1 standard error of the population mean of 58.31 and in 1000 samples of sample size 4000, we see that 68.6 percent of these 1000 samples of size 4000 had a mean of water use in 2006 that was within 1 standard error of the population mean of 58.31. Larger the sample size, closer the sample mean gets to the population mean and therefore, more concentration of values falling within 1 standard deviation as is evident here. Also, the difference is how large those standard errors are (for samples of size 900 the standard error is 0.34 while for samples of size 4000 the standard error is 0.62). in 1000 samples of sample size 900, 96 percent were within 2 standard errors and in 1000 samples of sample size 4000 each, 95.2 percent were within 2 standard errors.