# HW7: Repeated Sampling and Water Conservation

2023-11-07

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 unusual situation where we know the population parameter values.

In 2007, a water utility in Atlanta implemented an 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.

## Question 1 [9 pts]

### 1a [4 points]

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. What kind of variables are water use and whether or not a household is in the control group?

### 1b [3 points]

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?

**1c [2 points]**

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

1a : The proportion of people in the control group are 11675 which means its 10.94%. the mean of the population water use is 58.313 thousand gallons and the sd is 41.13.

1b. the mean water use in my sample is 59.82 thousand gallons and the sd is 38.15. the proportion of people in the control group is 11.33%.

1c. the mean of water use in 2006 for that sample is approximately 1 point away from the population mean. The sd is also at a 2.98 point difference. The proportion of people in the population for water use in 2006 is 0.4 percent points for the mean water use in the sample data. therefore the sample case variables are pretty close to the population mean, sd and proportions

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tinytex)

water <- read.csv("data0/water.csv")

 # <-- 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. If all students use the same seed number, you will all have the same random sample.)
pop_water_mean <- mean(water$WATER_2006)
pop_water_mean
```

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

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

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

```
prop <- proportions(table(water$group == "1"))
water <- water %>%
  mutate(control = if_else(group!=1, 0, 1))
pop_mean <- mean(water$control)
pop_mean
```

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

```
set.seed(1989) #update this number for your own lab.
my_samp <- sample_n(water, 900)
dim(my_samp)
```

```
## [1] 900   5
```

```
mean(my_samp$WATER_2006)
```

```
## [1] 59.82
```

```
sd(my_samp$WATER_2006)
```

```
## [1] 38.15391
```

```
proportions(table(my_samp$group == "1"))
```

```
##
##     FALSE      TRUE
## 0.8866667 0.1133333
```

## Question 2 [10 pts]

### 2a [2 points]

Using the code given below, draw 5000 random samples of size 900 and for each record the mean water use (this is what the code below does). In the code below you need to enter the name of the data set with data for the whole population (water), the size of each sample (900), the number of samples (5000), the name of the variable of interest (`WATER_2006`), and the sample summary statistic you want R to calculate and save for each sample (mean).

### 2b [5 points]

Make a histogram of the mean water use from the 5000 samples (each of size n = 900). Calculate summary statistics for the mean water use from the 5000 samples. Describe the distribution of the mean water use over repeated sampling. How does the mean of these 5000 sample means compare to the population mean? How does the standard deviation of these 5000 sample means compare to the standard deviation of water use in the population?

### 2c [3 points]

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 5000 samples had a mean of water use in 2006 that was within 1 standard error of the population proportion? What proportion of these 5000 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

2b. the mean water use over repeated sampling is 58.35 thousand gallons. The sd for the 5000 samples is 1.370. the mean and sd for the population data : 58.313 thousand gallons and the SD is 41.13. the mean differs by 0.02 thousand gallons. The mean of the 5000 sample means of water use is 58.35 thousand gallons. The median of the sample means of water is 58.30 thousand gallons. The standard deviation of the sample means of water is 1.371, so we expect the sample means of water to be about 1.371 away from the population mean of water for samples of size 900. distribution: the mean of the 5000 sample means of water use is ranging from 54.23- 64.60. THe iqr is 57.41 - 59.23 which means 50 percent of the water use lies between the Iqr.

the mean of the 5000 sample means of water are approximately the same for the population mean as well. while the sd is 1.37 for the 5000 sample means of 900 sample size and differs to the population mean which is 43.96.

2c. We see that 69.04 percent of sample mean of water use were within 1 standard error of the population mean and 95.28 were within 2 standard errors of the population mean. So even with sample sizes of 900, we get fairly precise estimates of the total water use sample mean almost all the time.

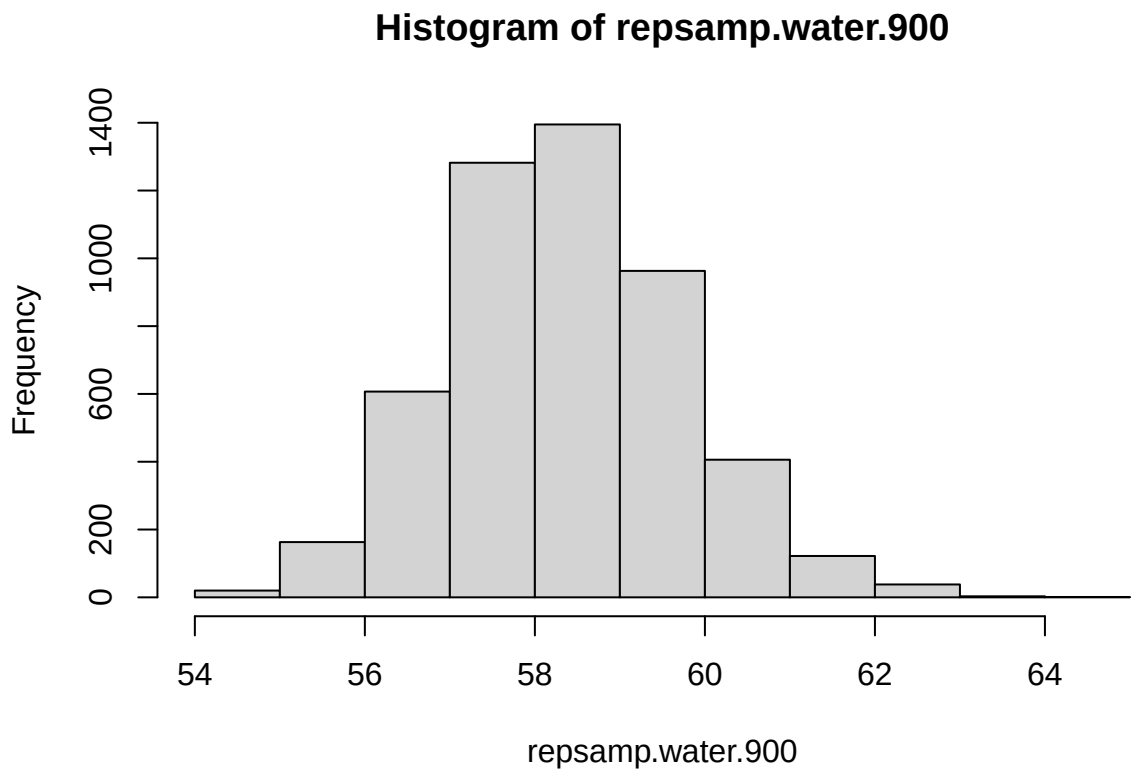
the mean of my sample was 1.09 standard errors away from the population proportion.

```
#' 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
```

```r
#' @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)
}

#Fill in each of the inputs below:
repsamp.water.900 <- SimulateSamplingDistribution(population_data = water ,
                                        number_samples = 5000 ,
                                        sample_size= 900,
                                        variable_name = "WATER_2006",
                                        statistic = mean,
                                        seed = 1989)
hist(repsamp.water.900)
```



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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.23   57.41   58.30   58.35   59.23   64.60
```

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

```
## [1] 1.371012
```

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

```
## [1] 1.371012
```

```r
prop.1e.q2 <- mean(repsamp.water.900 > pop_water_mean - se.900 &
                     repsamp.water.900 < pop_water_mean + se.900)
prop.2e.q2 <- mean(repsamp.water.900 > pop_water_mean - 2*se.900 &
                     repsamp.water.900 < pop_water_mean + 2*se.900)
prop.1e.q2
```

```
## [1] 0.6904
```

```r
prop.2e.q2
```

```
## [1] 0.9528
```

```r
(mean(my_samp$WATER_2006 - pop_water_mean))/se.900
```

```
## [1] 1.098562
```

## Question 3 [11 pts]

### 3a [2 points]

Using the SimulateSamplingDistributions function from Q2, draw 5000 random samples of size 900 and for each record the proportion of households the 900 selected households that are in the control group.

### 3b [5 points]

Make a histogram of the proportion of households in the control group from the 5000 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 5000 sample proportions compare to the population proportion? What is the standard deviation of the 5000 sample proportions?

### 3c [4 points]

What proportion of these 5000 samples had a share of households in the control group that is more than 3 percentage points away from the population proportion?

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 5000 samples had a share of households in the control group that was within 1 standard error of the population proportion? What proportion of these 5000 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

3b.it is a unimodal distribution with the majority at 0.11.it has a range from 0.07- 0.148. it has an Iqr of 0.102-0.11.The mean is 10.92 % of the households and the median is 10.88 of the households- the population proportion is 10.94 % households. therefore the difference is very low between both the population proportion and the samples control group mean/ median. 1.04 percent points is the standard deviation for the 5000 sample estimates.

The histogram tells us its unimodal and that the distribution is mostly symmetric . there is a slight Spike at the median.

3c.We see that 99.52 % percent of the proportion of the 5000 values were within 3 percent points away from the population proportion. 95.7 % were within 2 standard errors of the standard error and 66.22% were within 1 standard error of the population proportion.

the share of households in the control group in my sample was 0.37 standard errors away from the population proportion.

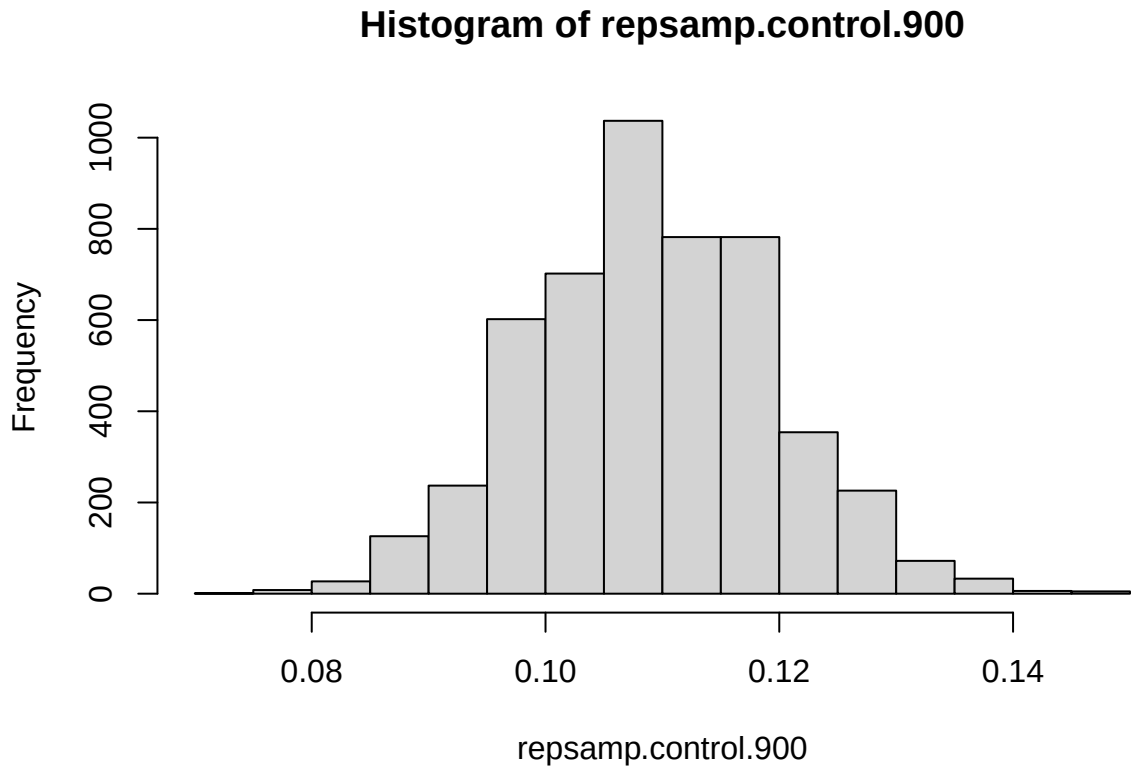
```
water <- water %>%
  mutate(control = if_else(group!=1, 0, 1))
repsamp.control.900 <- SimulateSamplingDistribution(population_data = water,
                                                    number_samples = 5000,
                                                    sample_size = 900,
                                                    variable_name = "control",
                                                    statistic = mean,
```

```
                                        seed =1989)


summary(repsamp.control.900)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07444 0.10222 0.10889 0.10942 0.11667 0.14889
```

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




```
sd(repsamp.control.900)
```

```
## [1] 0.01043237
```

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

```
## [1] 0.01043237
```

```
p.1e.q2 <- mean(repsamp.control.900 > pop_mean - secon.900 &
                    repsamp.control.900 < pop_mean + secon.900)
p.2e.q2 <- mean(repsamp.control.900 > pop_mean - 2*secon.900 &
```

```
                      repsamp.control.900 < pop_mean + 2*secon.900)
p.3e.q2 <- mean(repsamp.control.900 > pop_mean - 0.03 &
                      repsamp.control.900 < pop_mean + 0.03)
p.1e.q2
```

```
## [1] 0.6622
```

```
p.2e.q2
```

```
## [1] 0.9574
```

```
p.3e.q2
```

```
## [1] 0.9952
```

```
(mean(my_samp$control - pop_mean))/secon.900
```

```
## [1] 0.3721687
```

## Question 4 [9 pts]

### 4a [2 points]

Using the SimulateSamplingDistributions function, draw 5000 random samples of size 900 and for each record the the *standard deviation* of water use among the 900 households in the sample.

### 4b [5 points]

Make a histogram and a box-plot of the standard deviation of water use from the 5000 samples. Calculate summary statistics for the standard deviation of water use from the 5000 samples. Describe the distribution of the standard deviation of water use over repeated sampling. How does the mean of the 5000 sample standard deviations compare to the population standard deviation?

### 4c [2 points]

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

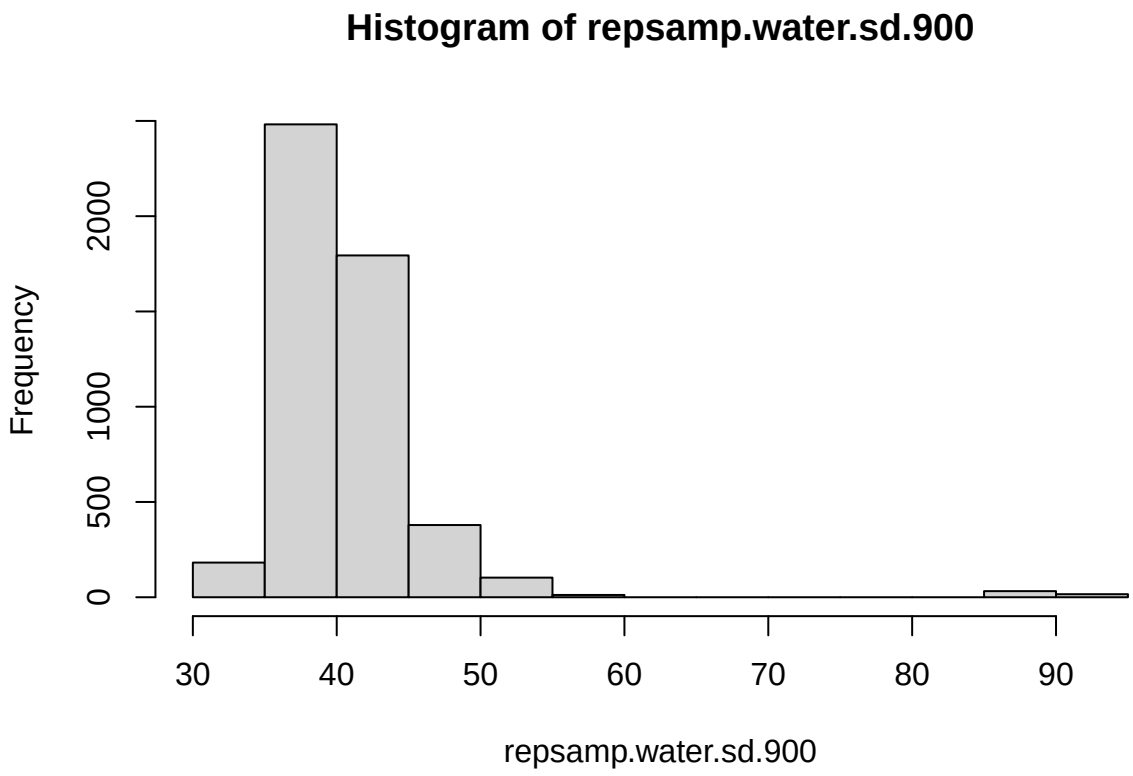
4b. the histogram is right skewed and is not symmetrical. The boxplot has a median at about 40 and there are outliers attached to the whiskers of the tail. and there is a gap observed between 60-85. there are outliers in the distribution which are more than 5 iqrs away. the range of the distribution is 31.43- 94.74. the mean is 40.76 and the median is 39.73. 50 percent of the data for standard deviation of water use lies between 37.82 and 42.24. 41.13 is the sd for the population mean of water use and it is only 1 percent point higher than the mean for the sd for the 5000 samples. therefore, it doesn't differ by a lot.

4c. the smallest sample is 1.308 times smaller than the population standard deviation. The highest value is 2.303 times greater than the population sd.

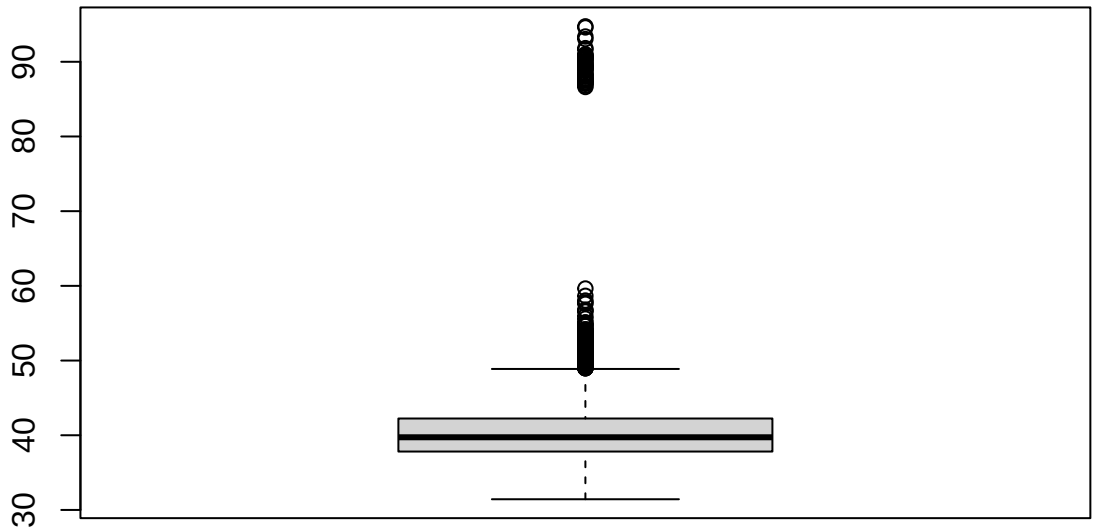
```
repsamp.water.sd.900 <- SimulateSamplingDistribution(population_data = water,
                                        number_samples = 5000,
                                        sample_size = 900,
                                        variable_name = "WATER_2006" ,
                                        statistic = sd,
                                        seed = 1989 )
summary(repsamp.water.sd.900)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.43   37.82   39.73   40.76   42.24   94.74
```

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




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




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

```
## [1] 1.371012
```

```
prop.1e.q2 <- mean(repsamp.water.900 > pop_water_mean - se.900 &
                     repsamp.water.900 < pop_water_mean + se.900)
prop.2e.q2 <- mean(repsamp.water.900 > pop_water_mean - 2*se.900 &
                     repsamp.water.900 < pop_water_mean + 2*se.900)
prop.1e.q2
```

```
## [1] 0.6904
```

```
prop.2e.q2
```

```
## [1] 0.9528
```

## Question 5 [9 pts]

### 5a [1 point]

Using the SimulateSamplingDistributions function, draw 5000 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 the output a different object name.

### 5b [5 points]

Calculate summary statistics of the 5000 sample means, now from samples of size 4000. What is the standard deviation of these 5000 sample means? Make a histogram of these 5000 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 [3 points]

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 5000 samples of size 4000 had a mean of water use in 2006 that was within 1 standard error of the population mean? What proportion of these 5000 samples had a mean of water use in 2006 that was within 2 standard errors of the population mean?

5b. the standard deviation is 0.63. The distribution of the mean of sample water use of sample size 4000 for 5000 samples ranges between 56.28 - 61.02 thousand gallons. the middle 50 percent of the values of sample water use lie between 57.90 - 58.75. the mean and median are 58.32 thousand gallons. the histogram is quite symmetrical and seems to have unimodal distribution. it takes the highest values between 58-58.5 thousand gallons of water use. the mean of the population data is very close to the mean of samples water use (4000 sample size). it is only differing by a 0.01 difference. the sd of the population was 41.13 and the sd of the (4000 sample size) water use is 0.63.therefore the standard deviation reduced from the population size in a smaller sample size.

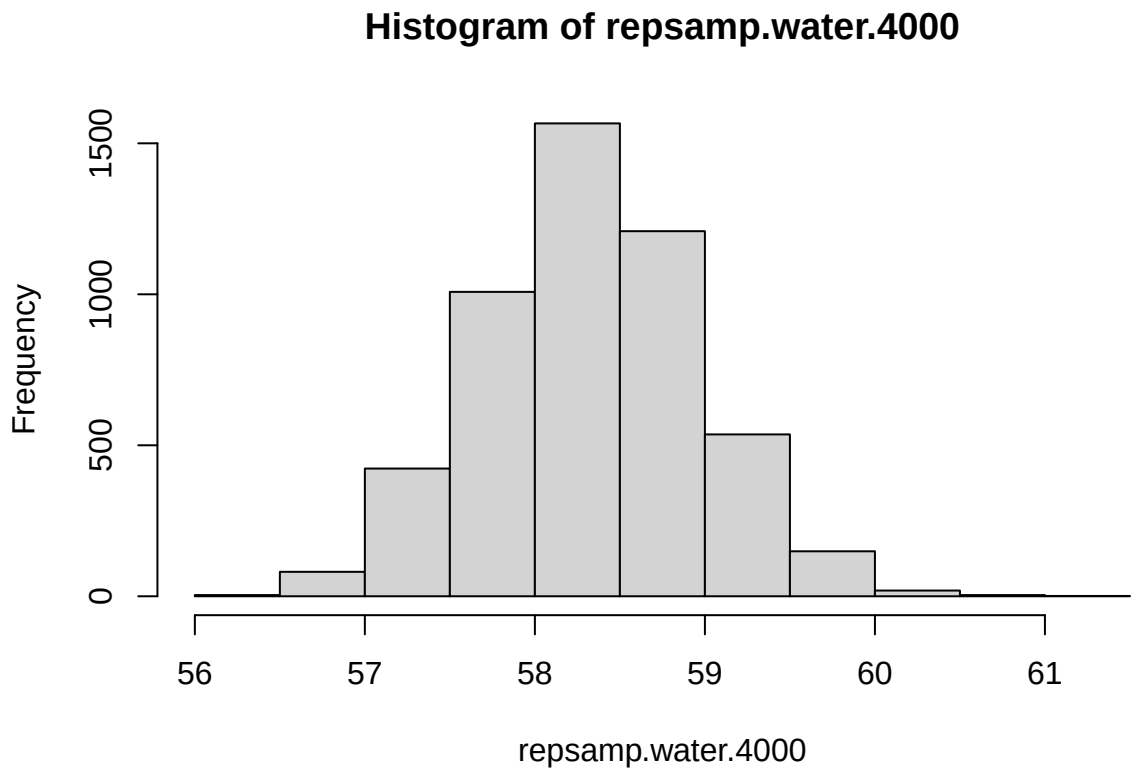
5c. We see that 68.94 percent of sample mean water use for 4000 sample size were within 1 standard error of the population mean and 95.4 were within 2 standard errors of the population mean . So even with sample sizes of 900, we get fairly precise estimates of the total water sample mean almost all the time.

## Answer 5

```
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)
}

#Fill in each of the inputs below:
repsamp.water.4000 <- SimulateSamplingDistribution(population_data = water ,
                                        number_samples = 5000,
```

```
                                        sample_size= 4000,
                                        variable_name = "WATER_2006",
                                        statistic = mean,
                                        seed = 1989)
hist(repsamp.water.4000)
```




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.28   57.90   58.32   58.32   58.75   61.02
```

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

```
## [1] 0.6368687
```

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

```
## [1] 0.6368687
```

```
prop.1fe.q2 <- mean(repsamp.water.4000 > pop_water_mean - se.4000 &
                    repsamp.water.4000 < pop_water_mean + se.4000)
prop.2fe.q2 <- mean(repsamp.water.4000 > pop_water_mean - 2*se.4000 &
                    repsamp.water.4000 < pop_water_mean + 2*se.4000)
prop.1fe.q2
```

```
## [1] 0.6894
```

```
prop.2fe.q2
```

```
## [1] 0.954
```

```
(mean(my_samp$WATER_2006 - pop_water_mean))/se.4000
```

```
## [1] 2.364916
```

## Question 6 [10 pts]

**6a [2 points]**

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 over repeated sampling. ONE histogram from the 5000 samples of size 900 and ONE histogram from the 5000 samples of size 4000. How do their shapes compare?

**6b [5 points]**

Refer back to the summary statistics you calculated of the 5000 sample means from samples of size 900 and samples of size 4000 in Q2 and Q5. How do their measures of central tendency compare? How do their measures of spread compare? Are there any notable features in the sampling distributions for sample means of size 900 and 4000?

**6c [3 points]**

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)? How do the standard errors from samples of size 900 and samples of size 4000 compare?

## Answer 6

6a. both the graphs look similar, however there is a much more narrow distinction in the modality of the 4000 samples graph as it gets more accurate. the distribution is much more symmetric in the case of the 4000 samples and the range has reduced. the standard deviation between the two has also reduced from 1.37 to 0.63. the mean and median have also slightly changed for the 4000 sample size distribution. the mean and median are the same for the 4000 sample size distribution. THe range of the iqr has also reduced and so have the min and max values of the 4000 sample size distribution.

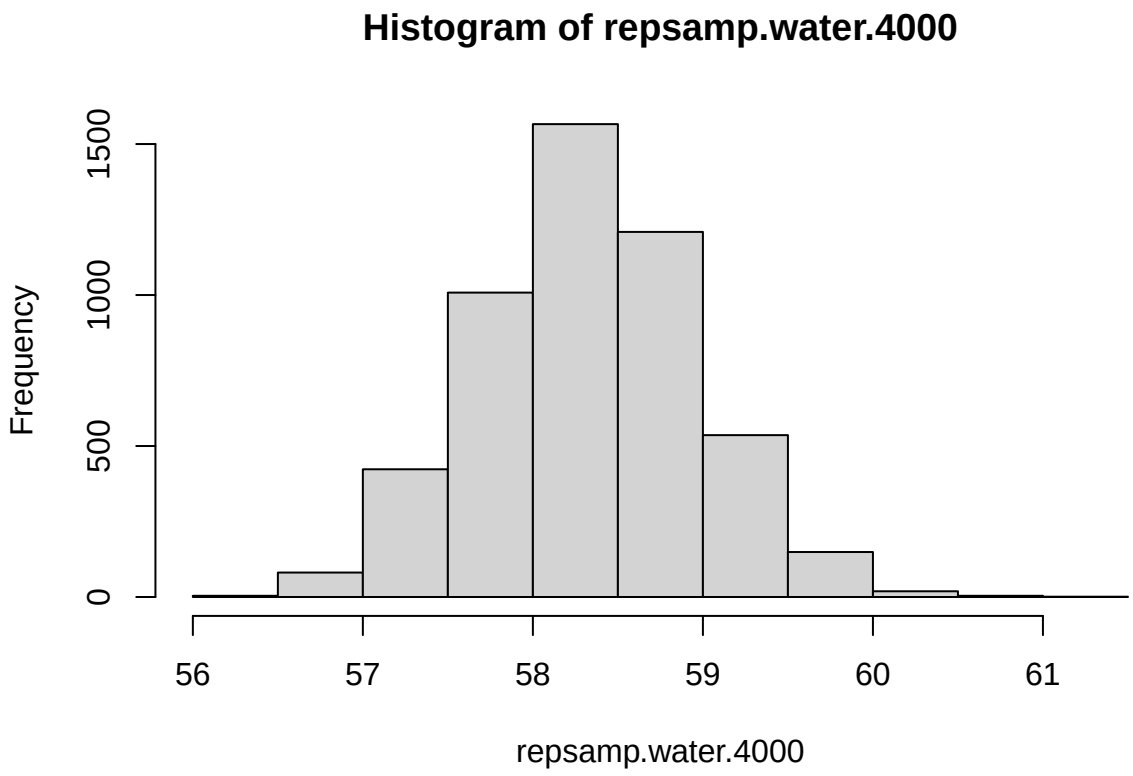
6b. their measures of central tendancy: for the measure on the 4000 sample size graph the median and mean are the same and the spread has reduced in its range as compared to the 900 sample size. 4000 sample size: 56.28 - 61.02 900 sample size: 54.23 - 64.60

in the 4000 sample size we can observe that the spread has reduced, leading to a more accurate understanding of the central tendency and making more accurate conclusions. with the 900 sample size , the spread is a little more than the 4000 sample size. 900 histo has a slight skew.

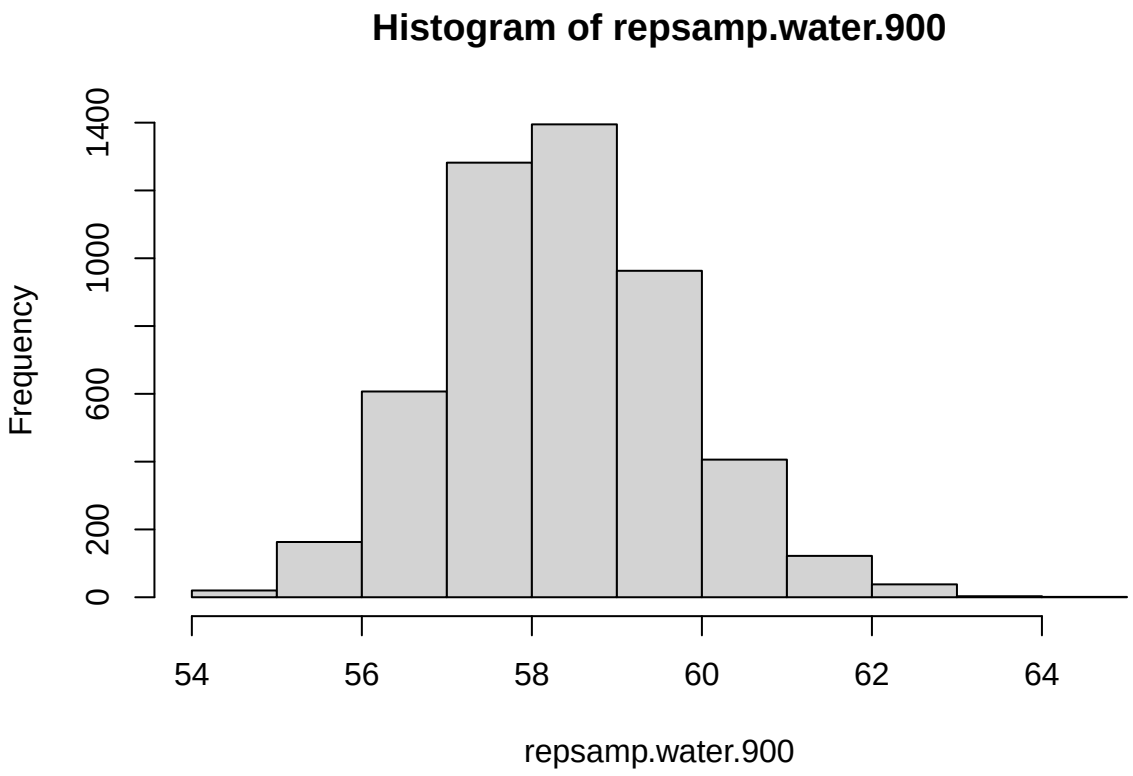
6c. for the 4000 sample size: 68.94 % of the sample means were within 1 standard error of the population mean and 95.4% of the sample means were within 2 standard errors of the population mean. for the 900 sample size : 69.04% of the sample means were within 1 standard error of the population mean and 95.28% were within 2 standard errors of the population mean.

the percentages for both sample sizes 4000 and 900, are very close and almost the same but what the inference for the percentage in that particular range would be will be different for both. the percentage for the 4000 sample size would describe a different outcome and lead to a more accurate understanding of the sample means deviation while the 900 sample size, the sample means being the same percentage as the 4000 would still mean a different inference as the range of the values and the spread is greater than the 4000 sample size samples . the standard error calculate for the 900 sample size was 1.09 SE and the for the 4000 sample size it was 2.36. it has increased meaning that the mean water use of my sample was 1.09 SE away from the population mean water use and 2.36 SE away from the population mean water use.

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




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




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.28   57.90   58.32   58.32   58.75   61.02
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.23   57.41   58.30   58.35   59.23   64.60
```

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

```
## [1] 1.371012
```

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

```
## [1] 0.6368687
```