HW8: The Moving to Opportunity Experiment & Confidence Intervals

Millions of low-income Americans live in high-poverty neighborhoods, which also tend to be racially segregated and sometimes have issues with community violence. While social scientists have long believed a lack of investment in these neighborhoods contributes to negative outcomes for the residents living in them, it is often difficult to establish a causal link between neighborhood conditions and individual outcomes. The Moving to Opportunity (MTO) demonstration was designed to test whether offering housing vouchers to families living in public housing in high-poverty neighborhoods could lead to better experiences and outcomes by providing financial assistance to move to lower-poverty neighborhoods.

Between 1994 and 1998 the U.S. Department of Housing and Urban Development enrolled 4,604 low-income households from public housing projects in Baltimore, Boston, Chicago, Los Angeles, and New York in MTO, *randomly assigning* enrolled families in each site to one of three groups: (1) The low-poverty voucher group received special MTO vouchers, which could only be used in census tracts with 1990 poverty rates below 10% and counseling to assist with relocation; (2) the traditional voucher group received regular section 8 vouchers, which they could use anywhere; and (3) the control group, who received no vouchers but continued to qualify for any project-based housing assistance they were entitled to receive. Today we will use the MTO data to investigate properties of confidence intervals. This exercise is based on the following article and the data is a subset of the data used for this article:

Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, J.R.K., and Sanbonmatsu, L., 2012. "Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults." *Science*, Vol. 337, Issue 6101, pp. 1505-1510.

The file `mto2.csv` includes the following variables for 3,263 adult participants in the voucher and control groups:

Name	Description
`group`	factor with 3 levels: `lpv` (low-poverty voucher), `sec8` (traditional section 8 voucher), and `control`

`econ_ss_zcore` | Standardized measure of economic self-sufficiency, centered around the control group mean and re-scaled such that the control group mean = 0 and its standard deviation = 1. Measure aggregates several measures of economic self-sufficiency or dependency (earnings, government transfers, employment, etc.)

`crime_vic` | Binary variable, `1` if a member of that household was the victim of a crime in the six months prior to being assigned to the MTO program, `0` otherwise based on self-report

The data we will use are not the original data as this dataset has been modified to protect participants' confidentiality, but the results of our analysis will be consistent with published data on the MTO demonstration.

```
mto2 <- read.csv("data1/mto2.csv")
library(tidyverse)

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

library(tinytex)
```

Question 1 [6 pts]

One of the baseline covariates in this dataset is crime victimization. We are going to use this variable (`crime_vic`) to learn about the coverage of confidence intervals created using information from a single sample. We will consider this dataset to be a complete population so we have the measurements of interest `crime_vic` for the entire population. Our parameter of interest is the proportion of households in this population where a household member experienced crime victimization in the last 6 months, μ_C.

How large is this population? Calculate μ_C, the proportion of households where a household member experienced crime victimization in the last 6 months in the population. Calculate the population level standard deviation of the `crime_vic` variable, σ_C. Is the `crime_vic` variable discrete or continuous? If discrete what kind of discrete variable is it? Make a histogram of the `crime_vic` variable in the population.

Answer 1

1a. 3263 particpants adn 10 variables. The proportion of individuals is 40.76 % - household member experienced crime victimization in the past 6 months. It is a discrete variable that is dichotomous and describes whether a household member experienced crime in the past 6 months or not.sd = 0.4914

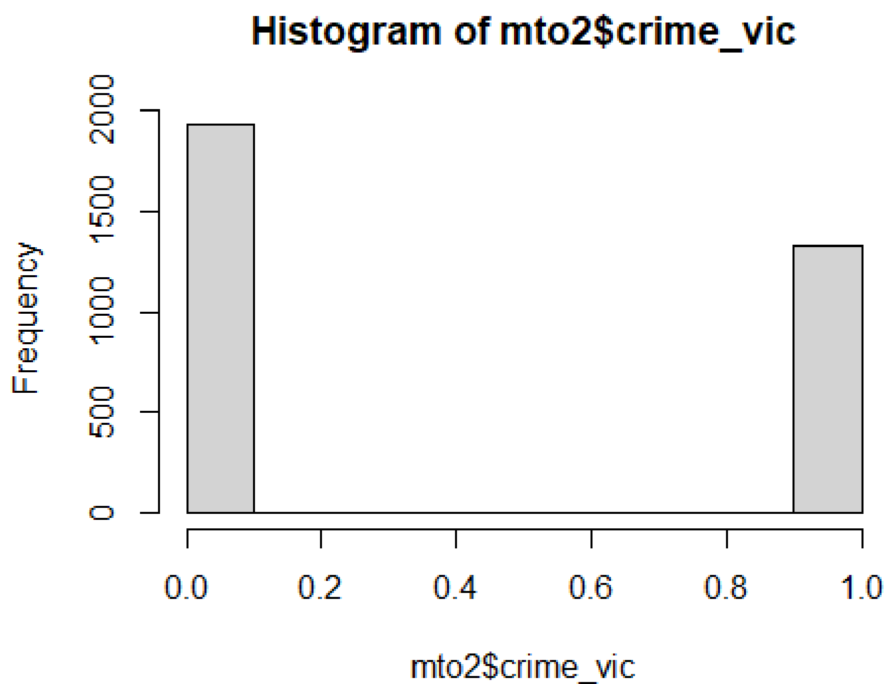
```
proportions(table(mto2$crime_vic == "1"))
```

```
## 
##     FALSE      TRUE 
## 0.5923996 0.4076004
```

```r
sd(mto2$crime_vic) -> sd_c
sd_c
```

```
## [1] 0.4914635
```

```r
hist(mto2$crime_vic)
```




```r
mean(mto2$crime_vic)
```

```
## [1] 0.4076004
```

```r
pop_mean_c <- mean(mto2$crime_vic)
```

Question 2 [11 pts]

2a [7 points]

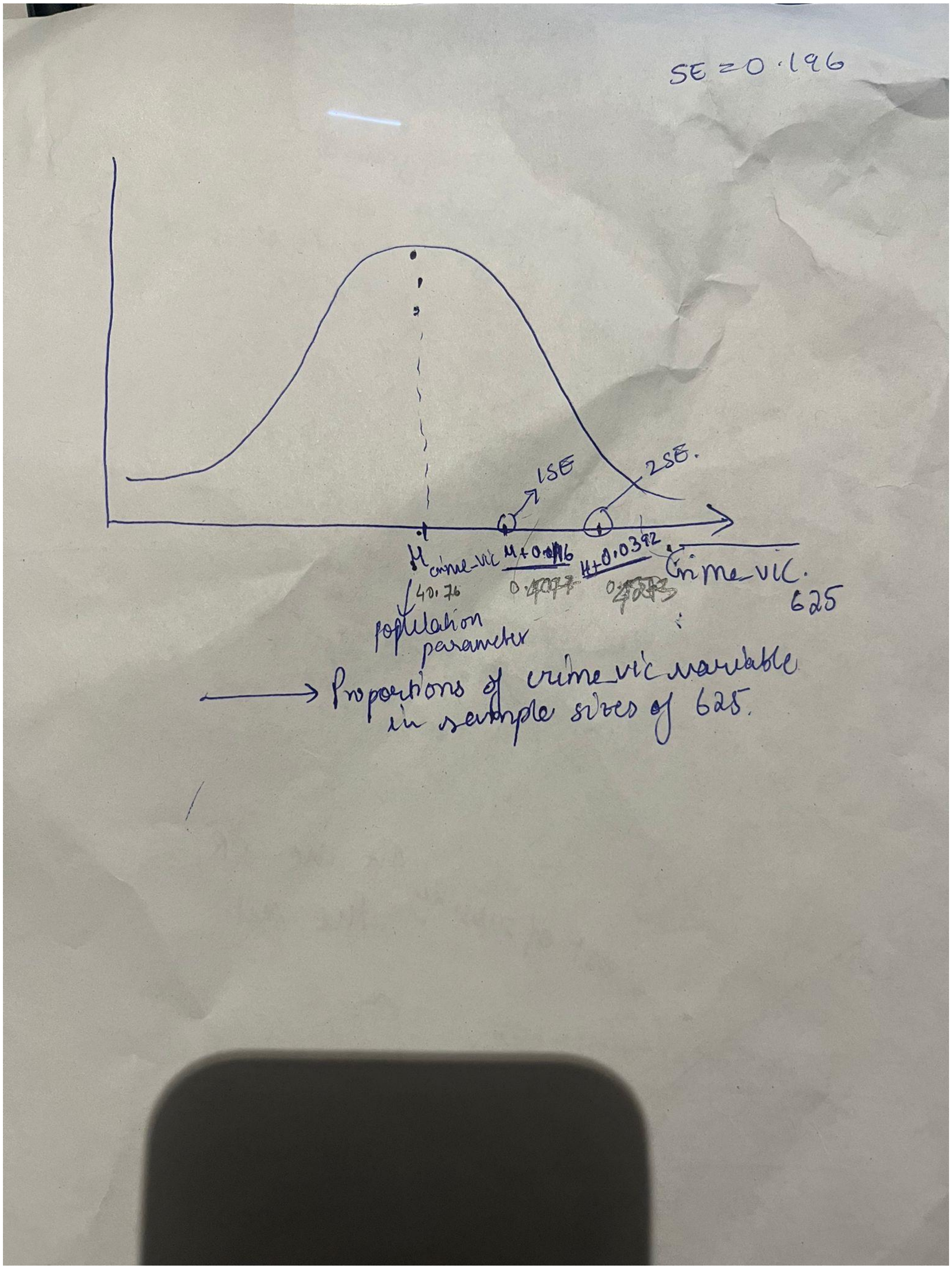
Consider samples of size n = 625 from this population. What summary statistic would we use from each of these samples to estimate the parameter of interest? Does the central limit theorem apply? Why does it or does it not apply? If it does apply, what are the implications of the CLT for what the sampling distribution of this summary statistic will be over repeated sampling (what are the mean, standard deviation, and shape of the sampling distribution)?

2b [4 points]

By hand, sketch the sampling distribution of this summary statistic and label the horizontal axis with regularly-spaced numbers and a label in words. From problem 1 we have the population standard deviation of the crime victimization variable, σ_C, and you know the sample size - use these to calculate the exact standard error. In your figure, also indicate the locations one and two standard errors away from the population parameter.

You can paste a sketch by taking a picture, converting it into a pdf, and paste it in with the rest of your Rmarkdown answers - see the R Guide for guidance on how to do this.

Answer 2



CAPTION

Answer 2a

2a. to estimate the mean of the crime_vic, we would use the mean of the sample means of all sample sizes of size 625. the population mean is: 40.76 %. the pop sd is :0.49. We would use the sample mean of crime_vic in each 625 person sample to estimate the population mean of crime victimization. we would use the mean to show the proportion of 625 samples that experienced crime victimization. The parameter of interest is mean of the crime_victimization in neighborhoods and the sample summary statistics is proportion of crime victimzation in neighborhoods in each sample. The distribution of the variable in the population is symmetric with no skew and no extreme outliers so a sample size of n = 625 is large enough and the CLT applies. The parameter of interest is proportion of population faced with victimization of crime and the sample summary statistics is proportion in each sample. The distribution of the variable in the population is not unimodal and is only spiking at 0 and 1.Size of n = 625 is large enough to offset the bimodal skews and outliers in the distribution and the CLT applies. we chose the mean to estimate the population mean of the crime_vic variable which is essential for CLT. If the crime_vic variable had a strongly skewed distribution, a strongly bi-modal distribution, and/or large outliers, the CLT may not hold for samples that are smaller. Based on the CLT, the sample mean ages from samples of size n = 625 will be similar to the bimodal distribution, with a mean (of the sample means) equal to the population mean, 40.76% , and with a standard deviation (of the sample means) of 0.0196.

```
true_se = 0.4914/sqrt(625)
true_se

## [1] 0.019656
```

Answer 2b

the standard error is: 0.0196. ## Question 3 [17 pts]

3a [3 points]

Consider 90% confidence intervals for the population parameter of interest. What is the formula for the 90% CI when the CLT applies?

3b [9 points]

For any single sample, we obtain a single summary statistic value that we use to estimate the parameter of interest. Using the single sample summary statistic, the exact standard error from Question 2, and the formula for the 90% CI we can create a single CI. We will repeat this process 20 times.

The code that follows below will draw the samples, calculate and record the summary statistic for each sample and calculate and record the 90% confidence interval for each sample (by calculating and recording the lower confidence limit and upper confidence limit). You need to fill in the number of samples, the sample size, the data set for the whole population, the variable name, the type of confidence interval (here use ‘normal’), and set the seed value to a value that only you use.

Create a histogram of the 20 sample means. Using the true standard error of the sampling distribution of sample proportions who experienced crime victimization for samples of size n = 625 (you calculated this in Question 2) determine how many of your 20 sample means are within 1 standard error of the population proportion who experienced crime victimization, how many are within 2 standard errors, and how many are within 3 standard errors.

3c [5 points]

Create a figure showing the value of the parameter of interest and the 20 confidence intervals created by the 20 samples (using R or make a sketch by hand). How many of these 20 90% confidence intervals do we expect to contain the population parameter? How many of your 20 confidence intervals contain the population parameter? If any of your confidence intervals did not contain the population parameter value, how far (in standard errors) were those samples' summary statistics from the population parameter value?

Answer 3

$$

$(\bar{x} - 1.64 * se, \bar{x} + 1.64 * se)$ $$

```
SimulateSamplingDistribution2 <- function(population_data,
                                          number_samples,
                                          sample_size,
                                          variable_name,
                                          distribution_type, # can be "t" or
"normal"
                                          seed = 10) {
        set.seed(seed)

        true_se =
sd(unlist(population_data[variable_name]))/sqrt(sample_size)

        if (distribution_type == 'normal') {q = qnorm(0.95)}
        else if (distribution_type == 't'){q = qt(0.95, sample_size-1)}
        else {stop("distribution_type must be 't' or 'normal'.")}

        repsamp.df <- data.frame(trial = 1:number_samples,
                             samp.mean = rep(0, number_samples),
                             samp.sd = rep(0, number_samples),
                             samp.lowci = rep(0, number_samples),
                             samp.highci = rep(0, number_samples))

        for (i in 1:number_samples){
                sample.rows <- sample(1:nrow(population_data), sample_size,
replace = FALSE)
                samp.variable <- unlist(population_data[sample.rows,
variable_name])
```

```
                repsamp.df$samp.mean[i] <- mean(samp.variable)
                repsamp.df$samp.sd[i] <- sd(samp.variable)
                if (distribution_type == 't') {se =
sd(samp.variable)/sqrt(sample_size)}
                else {se = true_se}
                repsamp.df$samp.lowci[i] <- repsamp.df$samp.mean[i] - q*se
                repsamp.df$samp.highci[i] <- repsamp.df$samp.mean[i] + q*se
                }

  return(repsamp.df)
}

# add the input values in the code below and remove the number signs at the
start of each line

 repsamp.q3 = SimulateSamplingDistribution2(population_data = mto2 ,
                              number_samples = 20 ,
                              sample_size = 625,
                              variable_name = 'crime_vic',
                              distribution_type = 'normal', # can be "t" or
"normal"
                              seed = 20)
 summary(repsamp.q3)

##      trial         samp.mean         samp.sd          samp.lowci
##  Min.   : 1.00   Min.   :0.3712   Min.   :0.4835   Min.   :0.3389
##  1st Qu.: 5.75   1st Qu.:0.3916   1st Qu.:0.4885   1st Qu.:0.3593
##  Median :10.50   Median :0.4016   Median :0.4906   Median :0.3693
##  Mean   :10.50   Mean   :0.4031   Mean   :0.4906   Mean   :0.3708
##  3rd Qu.:15.25   3rd Qu.:0.4128   3rd Qu.:0.4927   3rd Qu.:0.3805
##  Max.   :20.00   Max.   :0.4368   Max.   :0.4964   Max.   :0.4045
##   samp.highci
##  Min.   :0.4035
##  1st Qu.:0.4239
##  Median :0.4339
##  Mean   :0.4355
##  3rd Qu.:0.4451
##  Max.   :0.4691

 repsamp.q3

##    trial samp.mean   samp.sd samp.lowci samp.highci
## 1      1    0.4128 0.4927318  0.3804646   0.4451354
## 2      2    0.4112 0.4924455  0.3788646   0.4435354
## 3      3    0.3904 0.4882307  0.3580646   0.4227354
## 4      4    0.3920 0.4885877  0.3596646   0.4243354
## 5      5    0.4368 0.4963869  0.4044646   0.4691354
## 6      6    0.3968 0.4896257  0.3644646   0.4291354
## 7      7    0.4064 0.4915543  0.3740646   0.4387354
## 8      8    0.4336 0.4959684  0.4012646   0.4659354
```

```
## 9       9     0.3712 0.4835128  0.3388646   0.4035354
## 10     10     0.3952 0.4892852  0.3628646   0.4275354
## 11     11     0.3952 0.4892852  0.3628646   0.4275354
## 12     12     0.3728 0.4839368  0.3404646   0.4051354
## 13     13     0.4240 0.4945861  0.3916646   0.4563354
## 14     14     0.4048 0.4912465  0.3724646   0.4371354
## 15     15     0.3984 0.4899608  0.3660646   0.4307354
## 16     16     0.4048 0.4912465  0.3724646   0.4371354
## 17     17     0.4304 0.4955287  0.3980646   0.4627354
## 18     18     0.3904 0.4882307  0.3580646   0.4227354
## 19     19     0.3824 0.4863627  0.3500646   0.4147354
## 20     20     0.4128 0.4927318  0.3804646   0.4451354
```

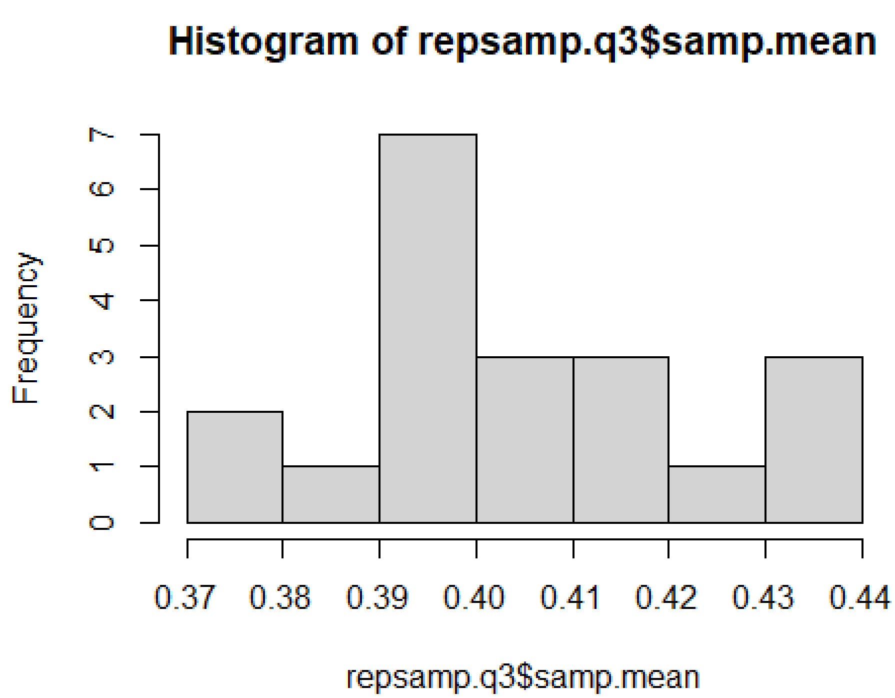
Answer 3a

90% CI: $\#X \pm 1.64 * SE$

Answer 3b

70 percent of the values are within 1 standard error of the population mean, 100 percent of the values are within 2 and 100 are within 3 standard errors.the histogram is not symmetrical, it is unimodal. there are no outliers.

```
hist(repsamp.q3$samp.mean)
```




```
mean(repsamp.q3$samp.mean > pop_mean_c - true_se &
       repsamp.q3$samp.mean < pop_mean_c + true_se)
```

```
## [1] 0.7

mean(repsamp.q3$samp.mean > pop_mean_c - 2*true_se &
       repsamp.q3$samp.mean < pop_mean_c + 2*true_se)

## [1] 1

mean(repsamp.q3$samp.mean > pop_mean_c - 3*true_se &
       repsamp.q3$samp.mean < pop_mean_c + 3*true_se)

## [1] 1
```

Answer 3c

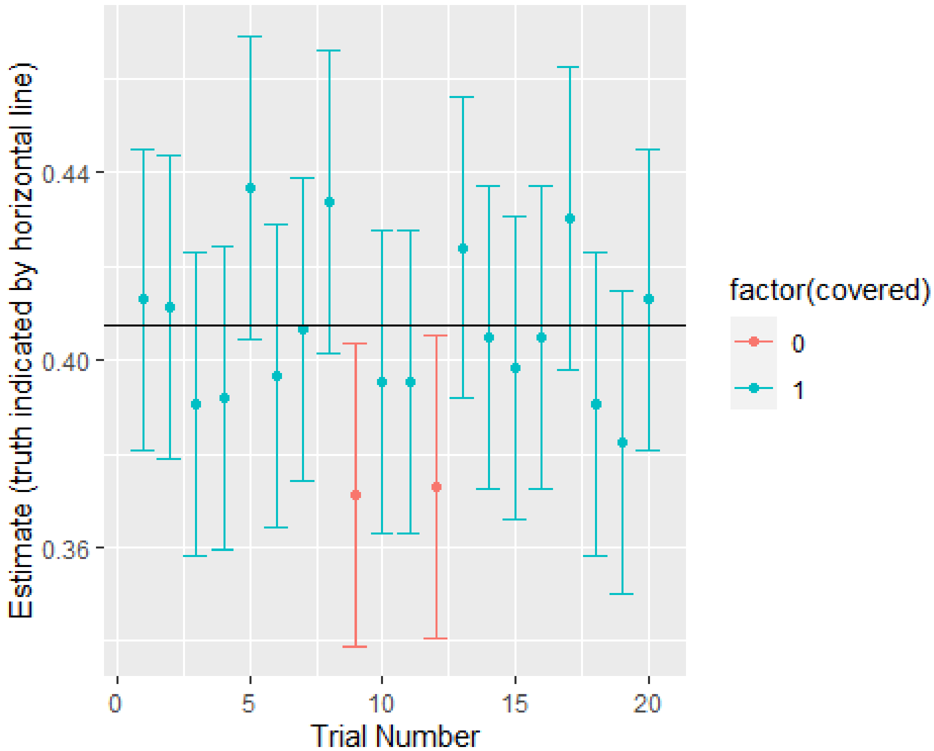
we can expect approximately 90 percent which is 18 out of 20 of the samples to cover the population mean. for these 20 samples 90% did contain the pop parameter which is 18/20. both of the samples that didn't contain the population mean and had the sample means as 0.372 and 0.371. 1.91 se below the pop mean

```
repsamp.q3 <- repsamp.q3 %>%
  mutate(covered = if_else(samp.lowci < pop_mean_c & samp.highci >
pop_mean_c, 1, 0))

mean(repsamp.q3$covered)

## [1] 0.9

repsamp.q3 %>%
  ggplot(aes(x = trial, y = samp.mean, ymin = samp.lowci, ymax = samp.highci,
color = factor(covered))) +
  geom_point() +
  geom_errorbar() +
  geom_hline(yintercept = pop_mean_c) +
  xlab("Trial Number") +
  ylab("Estimate (truth indicated by horizontal line)")
```




```
(0.37 - pop_mean_c)/true_se

## [1] -1.912921
```

Question 4 [8 pts]

4a [1 point]

Draw samples as in Question 3, but this time draw 10,000 samples each of size n = 625. Again estimate the 90% confidence intervals as you did in Problem 3 but now you will have 10,000 confidence intervals instead of 20. Use the same inputs as in Question 3 except for *number_samples*.

4b [4 points]

Create a histogram of the 10,000 sample proportions of households where someone experienced crime victimization for samples of size n = 625. Calculate the percentage of the sample proportions that are within 1.65, 1.96, and 2.58 standard errors (use the exact standard error from Q2) of the population parameter.

4c [3 points]

Calculate the percentage of the 10,000 confidence intervals that contain the population parameter - this is a simulated but very close approximation of the *coverage level* of this type of confidence interval. It is a close approximation because we drew so many random samples that is closely approximates drawing all possible random samples. What is the length of each of the 10,000 confidence intervals?

Answer 4a

```
repsamp.q4 = SimulateSamplingDistribution2(population_data = mto2 ,
                              number_samples = 10000,
                              sample_size = 625,
                              variable_name = 'crime_vic',
                              distribution_type = 'normal' , # can be "t" or
"normal"
                              seed = 20 )
summary(repsamp.q4)
```

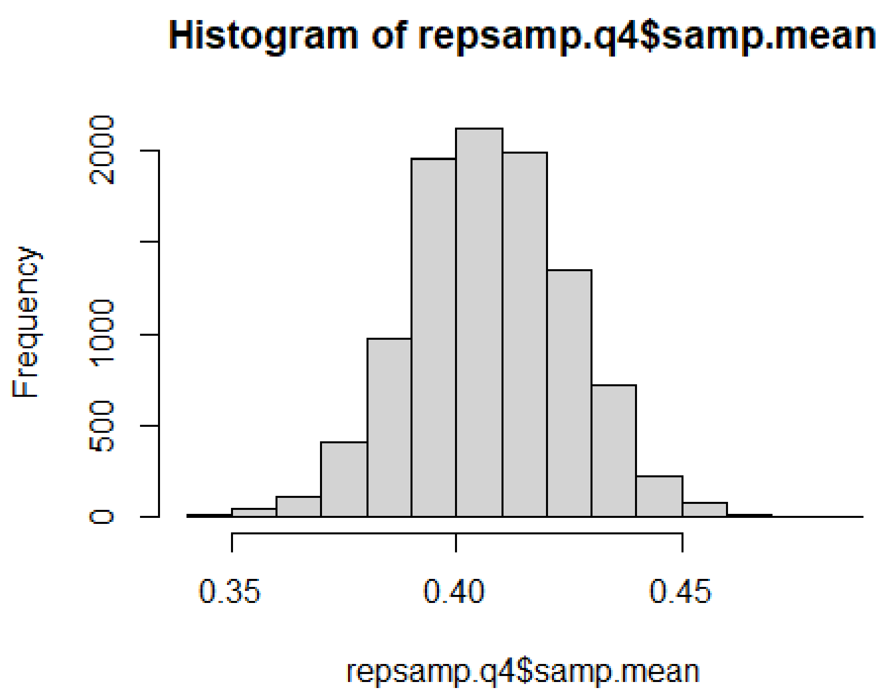
```
##      trial         samp.mean         samp.sd          samp.lowci
##  Min.   :    1   Min.   :0.3408   Min.   :0.4744   Min.   :0.3085
##  1st Qu.: 2501   1st Qu.:0.3968   1st Qu.:0.4896   1st Qu.:0.3645
##  Median : 5000   Median :0.4080   Median :0.4919   Median :0.3757
##  Mean   : 5000   Mean   :0.4077   Mean   :0.4915   Mean   :0.3753
##  3rd Qu.: 7500   3rd Qu.:0.4192   3rd Qu.:0.4938   3rd Qu.:0.3869
##  Max.   :10000   Max.   :0.4816   Max.   :0.5001   Max.   :0.4493
##   samp.highci
##  Min.   :0.3731
##  1st Qu.:0.4291
##  Median :0.4403
##  Mean   :0.4400
##  3rd Qu.:0.4515
##  Max.   :0.5139
```

$$(\bar{x} - 1.64 * se, \bar{x} + 1.64 * se)$$

Answer 4b

84.6 % of sample means of size 625 are 1.65 standard errors away from the population mean of crime_vic , 97% are 1.96 standard errors away from the population mean proportions of crime_vec and 99.5% are 2.58 Standard errors away from the population mean.

```
hist(repsamp.q4$samp.mean)
```




```
mean(repsamp.q4$samp.mean > pop_mean_c - 1.65*true_se &
       repsamp.q4$samp.mean < pop_mean_c + true_se)

## [1] 0.8461

mean(repsamp.q4$samp.mean > pop_mean_c - 1.96*true_se &
       repsamp.q4$samp.mean < pop_mean_c + 1.96*true_se)

## [1] 0.97

mean(repsamp.q4$samp.mean > pop_mean_c - 2.58*true_se &
       repsamp.q4$samp.mean < pop_mean_c + 2.58*true_se)

## [1] 0.9959
```

Answer 4c

93.23% of these 10,000 90% confidence intervals contained the population mean Water-vic value. So our interpretation (over repeated sampling, 90% of confidence intervals constructed in the manner will contain the population mean of the proportion of households going through a criminal attack and 10% will not) seems to be correct.the mean length is 0.0647 for each of the samples of size 10000.

```
repsamp.q4 <- repsamp.q4 %>%
  mutate(covered = if_else(samp.lowci < pop_mean_c & samp.highci >
pop_mean_c, 1, 0))

mean(repsamp.q4$covered)

## [1] 0.9323

repsamp.q4 <- repsamp.q4 %>%
  mutate(CI_length = samp.highci - samp.lowci)

summary(repsamp.q4$CI_length)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06467 0.06467 0.06467 0.06467 0.06467 0.06467

mean(repsamp.q4$CI_length)

## [1] 0.06467084
```

Question 5 [12 pts]

5a [2 points]

Repeat the steps listed in Question 4 for a second type of CI - this time create 90% confidence intervals using the t-distribution rather than the standard normal distribution and also draw samples of size n = 81 instead of n = 625. Instead of using *1.65* standard errors (R computed this value using the command *qnorm(0.95)*) we will tell the code to use the quantiles of the t-distribution (which you can obtain from a table or from R using the command *qt(0.95, 80)*). In addition, we will tell the code to use the standard error *estimate* generated by each sample (where you use the *sample* standard deviation divided by the square root of the sample size) rather than the true standard error based on the population standard deviation. We will refer to these as T-distribution CIs. For the code, you'll use the same inputs as for Q4 except you will change the distribution_type from 'normal' to 't' and change the *sample_size*.

We do this because almost always, the population level standard deviation is not known (when we don't know the population mean) and instead has to be estimated using the sample standard deviation. We then need to use the adaptation to the Central Limit Theorem where the sampling distribution of the sample means expressed as z-scores is no longer Normal and instead follows a t-distribution with n-1 degrees of freedom. The t-distribution is very similar to the standard normal distribution but has thicker tails (less probability in the center and a bit more in the tails).

5b [4 points]

Create a histogram of the 10,000 sample proportions of households where someone experienced crime victimization with samples of size n = 81. Calculate the percentage of the sample proportions that are within 1.65, 1.96, and 2.58 (exact) standard errors of the population parameter. You will need to calculate the exact standard error using information from Question 1 and the sample size (n = 81) we are using for this Question. Are these proportions different than what you found in Question 4B? Why do you think they should or should not be different?

5c [2 points]

Calculate the percentage of the 10,000 t-distribution confidence intervals that contain the population parameter - this is a simulated but very close approximation of the coverage level of this type of confidence interval. It is a close approximation because we drew so many random samples that is closely approximates drawing all possible random samples.

5d [5 points]

How do the coverage levels you observe here compare to those you found in question 4C? What is the *average* length of the 10,000 confidence intervals? What are the *minimum* and *maximum* length among these 10,000 confidence intervals? What makes these confidence intervals have different lengths while those in Q4 all had the same length?

Answer 5a

```
repsamp.q5 = SimulateSamplingDistribution2(population_data = mto2 ,
                                number_samples = 10000 ,
                                sample_size = 81 ,
                                variable_name = 'crime_vic',
                                distribution_type = 't', # can be "t" or
"normal"
                                seed = 20)
summary(repsamp.q5)
```

```
##      trial          samp.mean          samp.sd          samp.lowci    
##  Min.   :    1   Min.   :0.1975   Min.   :0.4006   Min.   :0.1235  
##  1st Qu.: 2501   1st Qu.:0.3704   1st Qu.:0.4859   1st Qu.:0.2805  
##  Median : 5000   Median :0.4074   Median :0.4944   Median :0.3160  
##  Mean   : 5000   Mean   :0.4078   Mean   :0.4914   Mean   :0.3169  
##  3rd Qu.: 7500   3rd Qu.:0.4444   3rd Qu.:0.5000   3rd Qu.:0.3520  
##  Max.   :10000   Max.   :0.6173   Max.   :0.5031   Max.   :0.5269  
##   samp.highci    
##  Min.   :0.2716  
##  1st Qu.:0.4602  
##  Median :0.4988  
##  Mean   :0.4987  
##  3rd Qu.:0.5369  
##  Max.   :0.7077
```

Answer 5b

0.80 and 0.94 and 0.99 vs 84.6 and 97 99.9% in 4b. These proportions are different. they must be different as we are considering a smaller sample size of data - in 4b we considered a sample size of 625 and in 5b we consider sample sizes of 81 but the same number of samples. therefore, the smaller the sample size. The se found for 4b is 0.0196 and for 5b is 0.054 are the se's and it can be observed that the se has significantly increased between the sample size : 625 and sample size: 81.

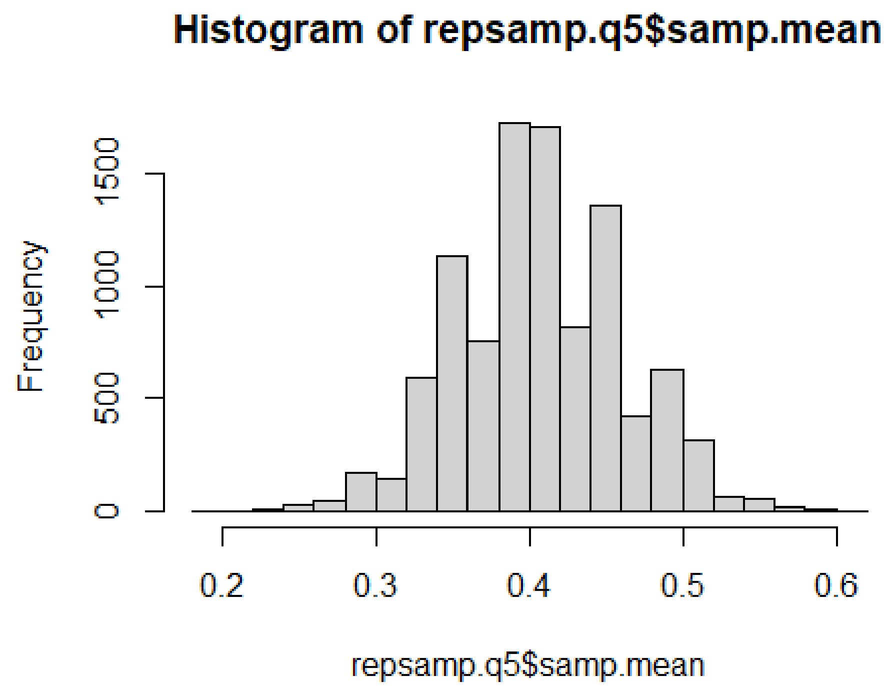
99.22% of sample means of size 625 are 1.65 standard errors away from the population mean of crime_vic , 80.91% are 1.96 standard errors away from the population mean proportions of crime_vec and 94.76% are 2.96 Standard errors away from the population mean.

the histogram seems fairly symmetrical with mode at 0.4. there are no outliers or gaps observed. The shape is at distribution with 80 degrees of freedom.

```
true_se_5 = 0.4914/sqrt(81)
true_se_5
```

```
## [1] 0.0546
```

```
hist(repsamp.q5$samp.mean)
```




```
mean(repsamp.q5$samp.mean > pop_mean_c - 1.65*true_se_5 &
       repsamp.q5$samp.mean < pop_mean_c + true_se_5)

## [1] 0.8091

mean(repsamp.q5$samp.mean > pop_mean_c - 1.96*true_se_5 &
       repsamp.q5$samp.mean < pop_mean_c + 1.96*true_se_5)

## [1] 0.9476

mean(repsamp.q5$samp.mean > pop_mean_c - 2.58*true_se_5 &
       repsamp.q5$samp.mean < pop_mean_c + 2.58*true_se_5)

## [1] 0.9922
```

Answer 5c

```
repsamp.q5 <- repsamp.q5 %>%
  mutate(covered = if_else(samp.lowci < pop_mean_c & samp.highci >
pop_mean_c, 1, 0))

mean(repsamp.q5$covered)

## [1] 0.914

repsamp.q5 <- repsamp.q5 %>%
  mutate(CI_length = samp.highci - samp.lowci)

summary(repsamp.q5$CI_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1482  0.1797  0.1828  0.1817  0.1849  0.1860
```

```r
mean(repsamp.q5$CI_length)
```

```
## [1] 0.1817046
```

answer 5d

the coverage is better in the 4 th question with a bigger sample size(625) at 93.2 %. the current coverage for this size of 81 and a T distribution is 91.4 %. the coverage seemed to decrease with a decrease in sample size. the average length is 0.181 for this case. the min length is 0.148 and the max length is 0.186. it is different because in Q4 we are using the true se to calculate confidence intervals while in this part we calculate standard error estimate that is generated by each sample of size 81.