DATA 557 - Homework 1

Will Wright

January 15, 2019

Homework Assignment 1

Instructions

Data set: 'iq.csv' (see Week 1 folder on canvas)

The data come from a study to examine the effect of lead exposure on IQ in children. The sample was randomly selected from the children in a large community (the "population") near a source of lead. IQ tests were administered to each child. The IQ scores were age-standardized using established normal values for the US population. Such age-standardized scores have a mean of 100 and a standard deviation of 15 in the US population.

Setup

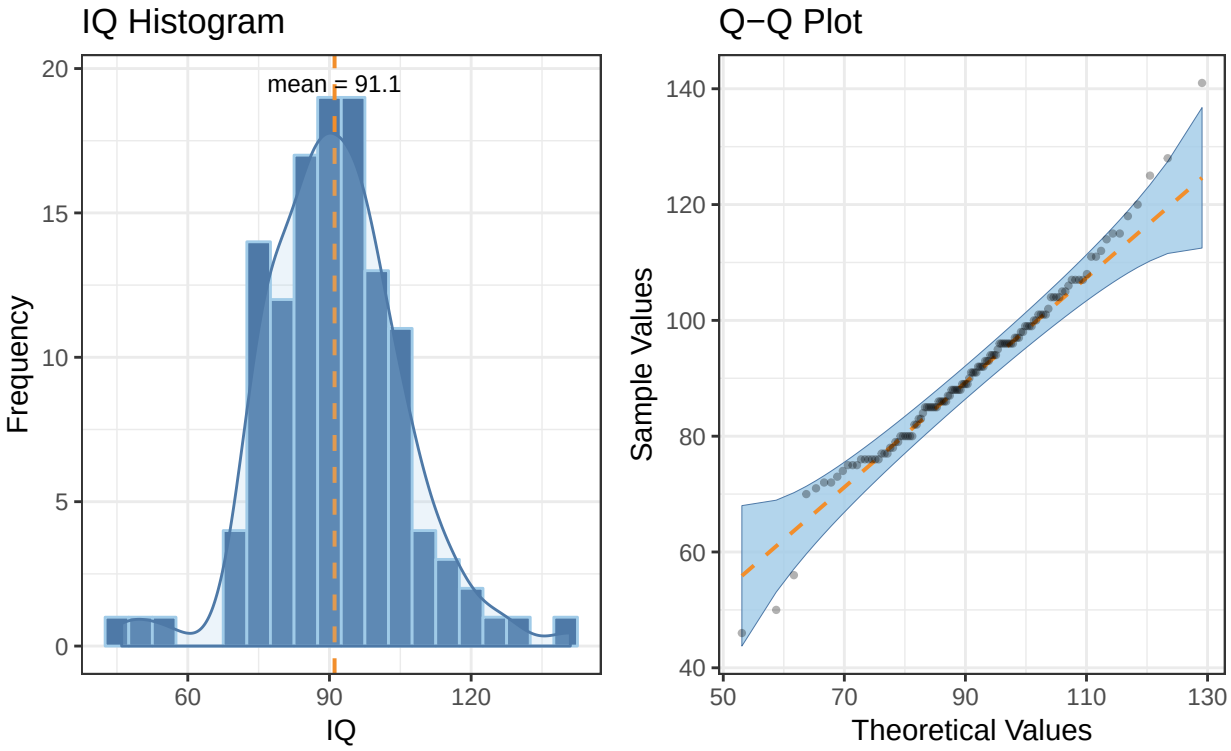
```
# Load packages
library(ggplot2)
library(dplyr)
library(scales)
library(ggthemes)
library(qqplotr)
library(gridExtra)
library(kableExtra)

# Get data
#setwd("~/UW/DATA557/WEEK01")
iqData <- read.csv("iq.csv")

# set up color palettes
colors <- ggthemes_data[["tableau"]][["color-palettes"]][["regular"]][[2]][[2]]
```

Problems

1. Create a histogram and normal q-q plot of the IQ variable. Using these plots comment on how well the distribution is approximated by a normal distribution.

```
distribution_visualizer <- function(data, title, x, y, binwidthInput){
  binwidthInput <- binwidthInput
  binCounts <- .bincode(data, seq(0,max(data), binwidthInput))
  xbar <- round(mean(data),1)
  sd <- round(sd(data),1)
  g <- ggplot(data.frame(data), aes(data)) +
    geom_histogram(fill = colors[1],
                   color = colors[2],
                   binwidth = binwidthInput) +
    geom_vline(aes(xintercept = mean(data)),
```

```
                color = colors[3],
                linetype = "dashed",
                size = 0.7) +
    geom_density(aes(y = binwidthInput * ..count..),
                 alpha = 0.2,
                 fill = colors[2],
                 color = colors[1]) +
    labs(title = title,
         x = x,
         y = y) +
    annotate("text", x = mean(data),
             y = max(binCounts, na.rm = TRUE)*0.75,
             label = paste0("mean = ",xbar),
             size = 3) +
    theme_bw()


  p <- ggplot(data.frame(data), aes(sample = data)) +
    stat_qq_band(color = colors[1], fill = colors[2]) +
    stat_qq_line(color = colors[3], linetype = "dashed", size = 0.7) +
    stat_qq_point(size = 0.8, alpha = 0.3) +
    labs(title = "Q-Q Plot",
         x = "Theoretical Values",
         y = "Sample Values") +
    theme_bw()

  grid.arrange(g, p, ncol = 2)
}

distribution_visualizer(iqData$IQ,"IQ Histogram","IQ","Frequency",5)
```



Given both the shape of the histogram's density and fact that only 5 of the 124 data points fall outside the 95% confidence band, it seems reasonable to say the data is well-approximated by the normal distribution.

2. Calculate the sample mean and sample SD of IQ. Give two different possible explanations for why these values differ from the corresponding values for the US population.

The mean is 91.1 and the standard deviation is 14.4 (calculated in the previous question). Given that lead is a known carcinogen that damages the brain and nervous system, reducing performance on IQ tests, one explanation is that the exposure to lead caused the lower IQ scores. Among other reasons, it is also possible that being near a source of lead had no impact on IQ, but that the locals were more prone to drink, causing fetal alcohol syndrome and reduced IQ.character(0)

3. Calculate a 95% confidence interval for the mean IQ score. For calculating the SE use the value 15 for the population SD of IQ. Is 100 in the confidence interval? Explain why it matters whether or not 100 is in the confidence interval.

```
pop_sd <- 15
n <- nrow(iqData)
se <- pop_sd/sqrt(n)
xbar <- mean(iqData$IQ)
ci <- c(xbar - se*qnorm(1- 0.05/2), xbar + se*qnorm(1- 0.05/2))
ci
```

```
## [1] 88.44050 93.72079
```

100 is not in the interval. This means that we can be at least 95% certain that the IQ of the sample is less than the IQ of the population.

4. Suppose that you were planning to repeat the study and that you wanted the 95% confidence interval for mean IQ to have width 30 or less. Assume that the distribution of IQ in the population is normal with mean 100 and standard deviation 15. What is the minimal sample size that would be needed for the new study?

The calculation for n for the given confidence interval, mean, and standard deviation, is give via the following:

```
width <- 30
oneTailWidth <- 0.5 * width
mean <- 100
sd <- 15
n <- ((sd*qnorm(1- 0.05/2))/oneTailWidth)^2
ceiling(n)
```

```
## [1] 4
```

5. Calculate the confidence interval for mean IQ using the sample SD instead of the population SD value of 15. In what way does the confidence interval differ from the one obtained previously? Explain the difference.

```
sample_sd <- sd(iqData$IQ)
n <- nrow(iqData)
se <- sample_sd/sqrt(n)
xbar <- mean(iqData$IQ)
ci <- c(xbar - se*qnorm(1- 0.05/2), xbar + se*qnorm(1- 0.05/2))
ci
```

```
## [1] 88.54541 93.61588
```

The bounds are smaller as a result of the standard deviation being smaller at 14.4 instead of 15 for the population.

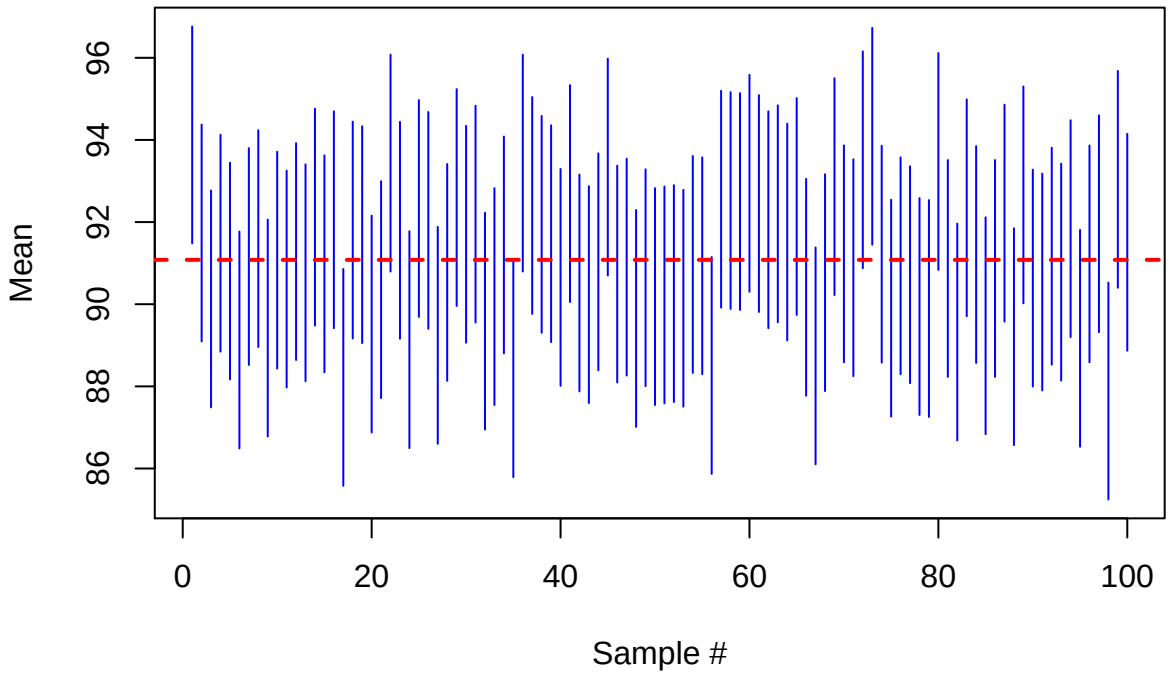
6. Perform a simulation study to estimate the coverage probability of the 95% confidence interval for mean IQ that uses the SD value of 15 to calculate the SE (as in Question 3). For your simulation, assume the distribution of IQ in the community has a normal distribution with mean 100 and standard deviation 15. Justify your choice of the number of replications (simulated samples). Do your results provide evidence that the coverage probability for this confidence interval is different than 0.95?

```
set.seed(7)

# set known parameters
sample_mean <- mean(iqData$IQ)
sample_sd <- 15
pop_mean <- 100
pop_sd <- 15
```

```r
# function with accepts sample and outputs upper and lower confidence bounds in a list
conf_intervalizer <- function(sample){
  n <- length(sample)
  sample_se <- sample_sd/sqrt(n)
  sample_mean <- mean(sample)
  upper <- sample_mean + sample_se*qnorm(1- 0.05/2)
  lower <- sample_mean - sample_se*qnorm(1- 0.05/2)
  list(lower=lower, upper=upper)
}

samples <- replicate(1000, sample(iqData$IQ, size = nrow(iqData), replace = T))
intervals <- apply(samples, FUN=conf_intervalizer, MARGIN=2)

get_lower <- function(inter) inter$lower
get_upper <- function(inter) inter$upper
n <- 100
lowers <- sapply(intervals[1:n], FUN = get_lower)
uppers <- sapply(intervals[1:n], FUN = get_upper)
print(matplot(rbind(1:n,1:n),t(cbind(lowers, uppers)),type="l",lty=1,lwd=1,col=4,xlab="Sample #",ylab="
```

```
## NULL
```

```r
abline(h=sample_mean, lty=2, col="red", lwd=2)
```



1000 replications were performed to ensure a large enough size to avoid bias when taking sample means. The results do not show that the coverage probability for this confidence interval is wrong since, out of the 100 samples, the confidence interval does not contain the true mean of the population 5 times, as expected.

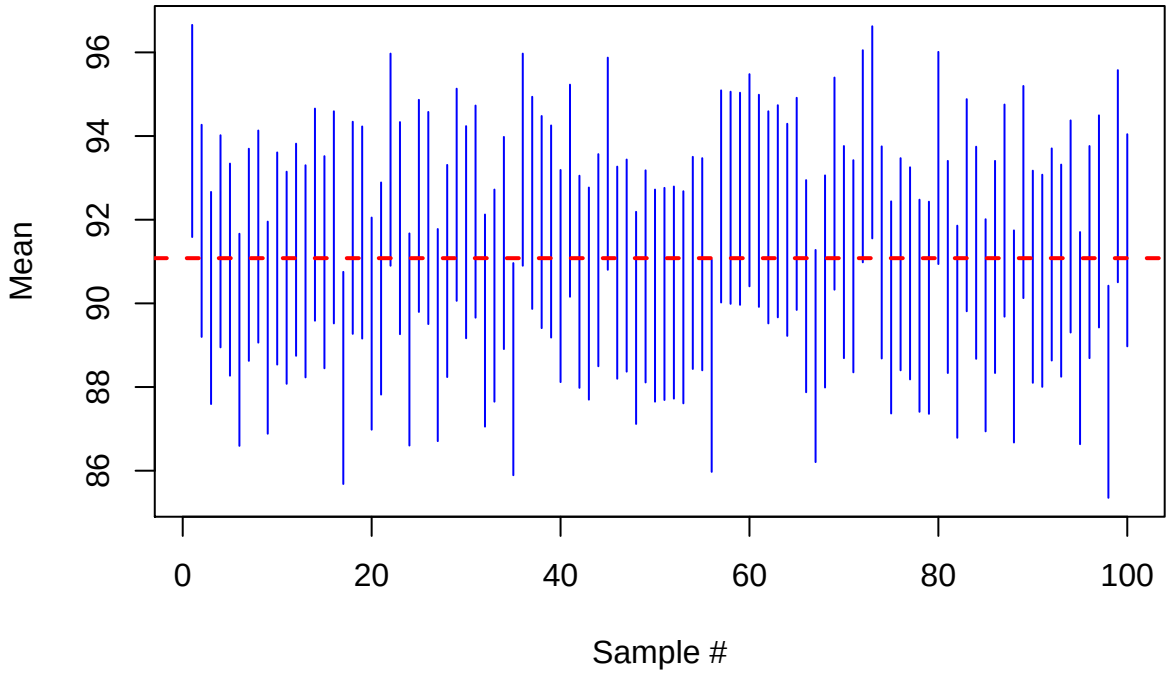
7. Perform a simulation study to estimate the coverage probability of the 95% confidence interval for mean IQ that uses the sample SD to calculate the SE (as in Question 5). (Note: save the sample means

and their SEs for answering the following questions.) As previously, assume the distribution of IQ in the community has a normal distribution with mean 100 and standard deviation 15 to generate your simulated samples. How do the results compare with those of the previous simulation?

```
set.seed(7)
# set known parameters
sample_mean <- mean(iqData$IQ)
sample_sd <- sd(iqData$IQ)
pop_mean <- 100
pop_sd <- 15

samples <- replicate(1000, sample(iqData$IQ, size = nrow(iqData), replace = T))
intervals <- apply(samples, FUN=conf_intervalizer, MARGIN=2)

get_lower <- function(inter) inter$lower
get_upper <- function(inter) inter$upper
n <- 100
lowers <- sapply(intervals[1:n], FUN = get_lower)
uppers <- sapply(intervals[1:n], FUN = get_upper)
print(matplot(rbind(1:n,1:n),t(cbind(lowers, uppers)),type="l",lty=1,lwd=1,col=4,xlab="Sample #",ylab="
```

```
## NULL
```

```
abline(h=sample_mean, lty=2, col="red", lwd=2)
```

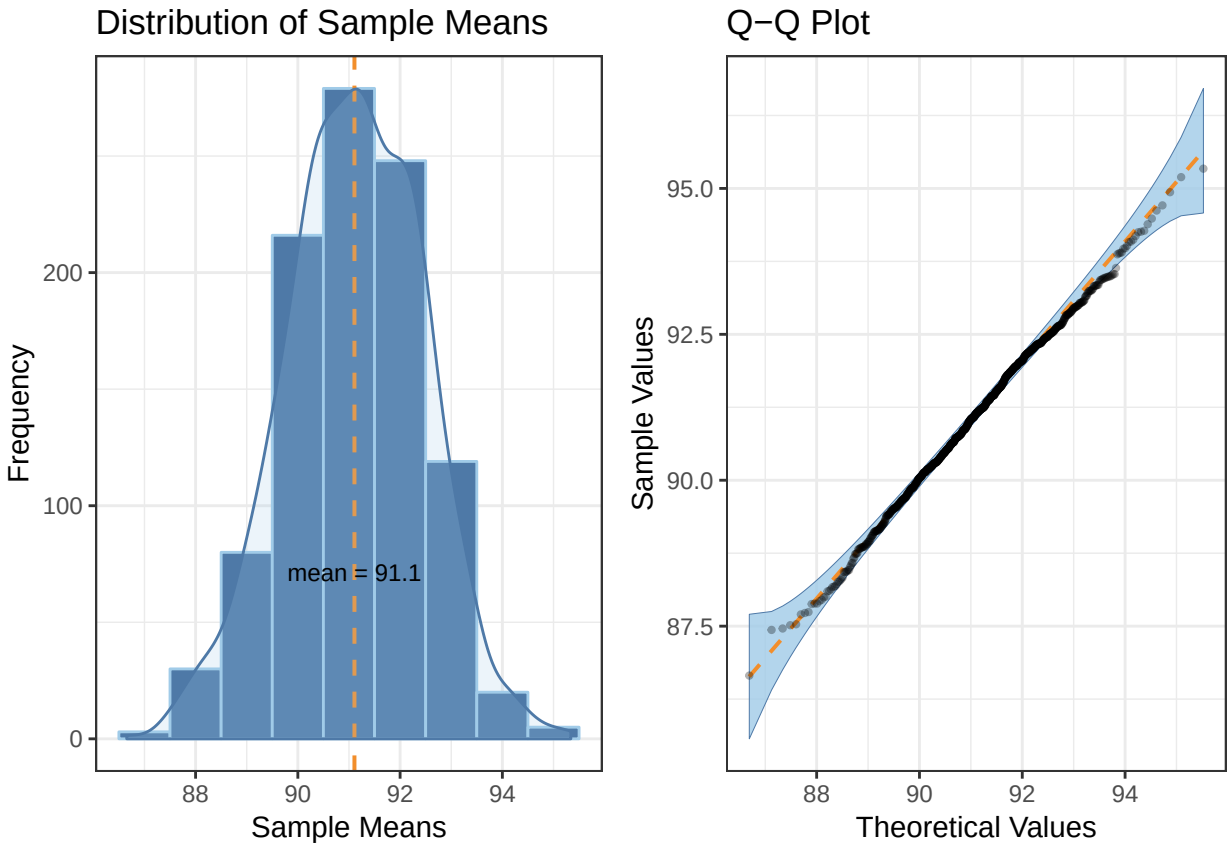


In this case, we again see 5 samples falling with confidence intervals that do not contain the true population mean. Because the difference in standard deviation is 14.4 vs 15.0, we shouldn't expect to see major differences between the coverage probability in the previous question and this one.

8. Display the distribution of the simulated sample means and describe how well it is approximated by a

normal distribution. Is this what you expected? Why or why not?

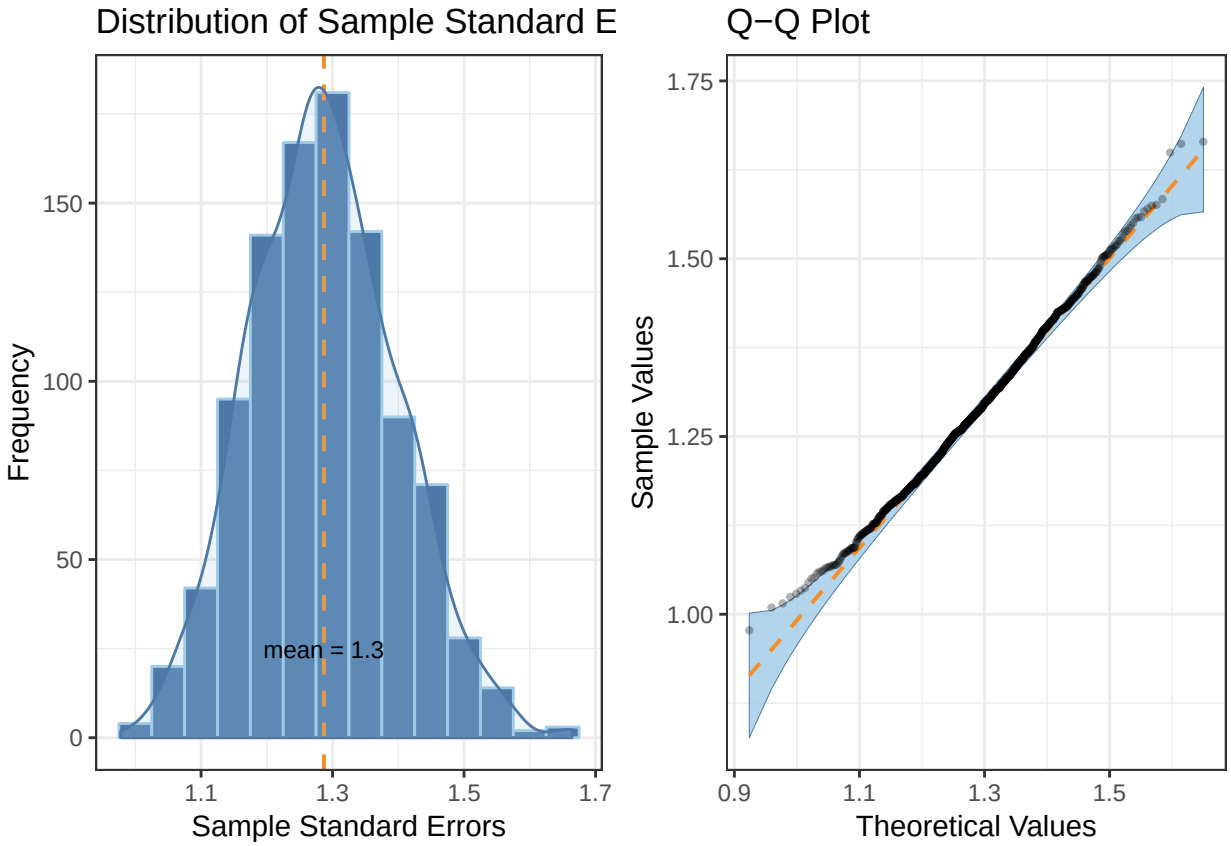
```
sample_means <- apply(samples, FUN=mean, MARGIN=2)
distribution_visualizer(sample_means, "Distribution of Sample Means","Sample Means","Frequency",1)
```



The sample means are very well represented by a normal distribution. This is expected since the Central Limit Theorem grants that the distribution of sample means from any distribution is approximately normal with sufficient sample.

9. Display the distribution of the estimated SEs. Compare the average value of the SEs to the SD of the sample means. Are they approximately equal? Explain why this matters.

```
standard_errorer <- function(sample) sd(sample)/sqrt(length(sample))
sample_errors <- apply(samples, FUN=standard_errorer, MARGIN=2)
distribution_visualizer(sample_errors, "Distribution of Sample Standard Errors","Sample Standard Errors
```



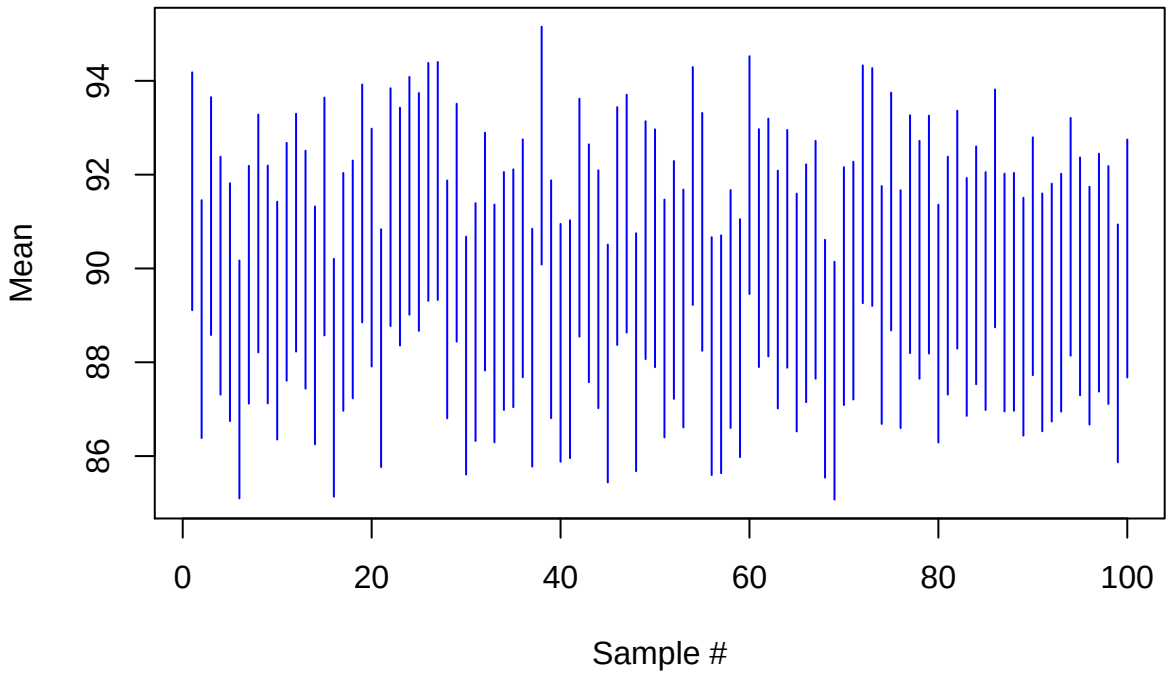
The average value of the SEs is 1.28 and the SD of the sample means is 1.34 so they are approximately equal. This matters because it means the SEs are an accurate reflection of the variation in our samples.

10. Now assume that the distribution of IQ in the community is a normal distribution with mean 90 and SD 15. Perform a simulation study of the 95% confidence interval for mean IQ (the version from Question 5). Calculate the proportion of the confidence intervals that are entirely below the value 100 (i.e., the upper limit of the confidence interval is lower than 100). Give an interpretation of the result.

```
set.seed(1)
# set known parameters
samples <- replicate(1000, rnorm(length(iqData$IQ), 90, 15)) # assuming our samples have size equal to
intervals <- apply(samples, FUN=conf_intervalizer, MARGIN=2)

get_lower <- function(inter) inter$lower
get_upper <- function(inter) inter$upper
n <- 100
lowers <- sapply(intervals[1:n], FUN = get_lower)
uppers <- sapply(intervals[1:n], FUN = get_upper)
print(matplot(rbind(1:n,1:n),t(cbind(lowers, uppers)),type="l",lty=1,lwd=1,col=4,xlab="Sample #",ylab="I
```

```
## NULL
```

```
abline(h=100, lty=2, col="red", lwd=2)
```



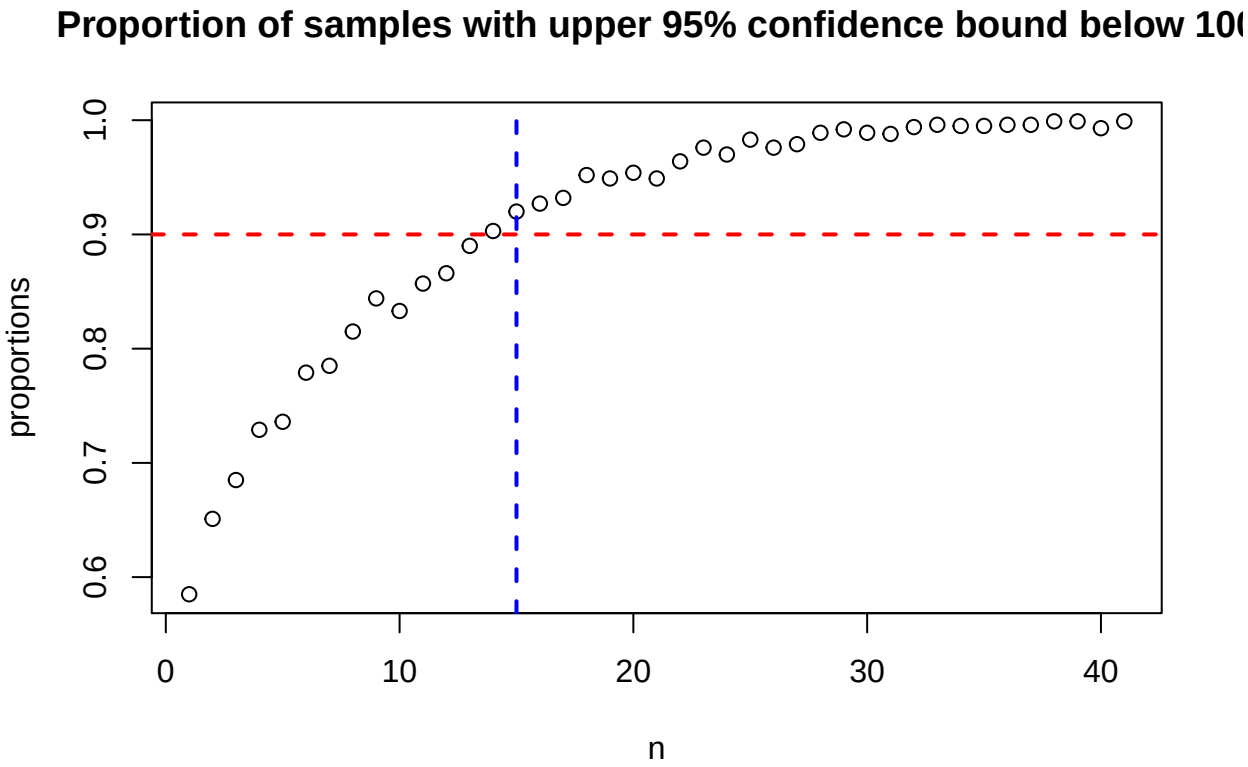

```
sum(uppers<100)
```

```
## [1] 100
```

All 100 samples do not have confidence intervals that reach 100. The reason this is the case for our similation is that a sample size of 124 yields fairly accurate means for each iteration such that it would be highly unlikely that 100 would fall within the bounds of any confidence interval.

11. For the previous question, estimate the minimum sample size required in order that the proportion of 95% confidence intervals lying entirely below the value 100 is at least 0.9. (There is no exact answer.)

```
set.seed(10)
ns <- seq(10, 50, 1)
sample_collection <- matrix(NA, nrow = 1000,ncol = length(ns))

for(i in 1:ncol(sample_collection)){
  samples <- replicate(1000, rnorm(ns[i], 90, 15))
  intervals <- apply(samples, FUN=conf_intervalizer, MARGIN=2)
  uppers <- sapply(intervals, FUN = get_upper)
  sample_collection[,i] <- uppers
}

proportionizer <- function(uppers) sum(uppers<100)/1000
proportions <- apply(sample_collection, FUN = proportionizer, MARGIN = 2)

plot(proportions, main = "Proportion of samples with upper 95% confidence bound below 100", xlab = "n")
abline(h=0.9,lty=2,col="red",lwd=2)
abline(v=15,lty=2,col="blue",lwd=2)
```



Seems like a sample size of about 15 will ensure that 90% of samples will have an upper 95% confidence interval below 100.

12. Provide an estimate and a 95% confidence interval for the proportion of children in the community with IQ score less than 100. What does the result tell you about differences between the distribution of IQ in the community and in the US population?

```
phat <- sum(iqData$IQ < 100) / length(iqData$IQ)
se_phat <- sqrt((phat*(1-phat))/length(iqData$IQ))
ci <- c(phat - se_phat*qnorm(1- 0.05/2), phat + se_phat*qnorm(1- 0.05/2))
phat
```

```
## [1] 0.7580645
```

```
ci
```

```
## [1] 0.6826873 0.8334417
```

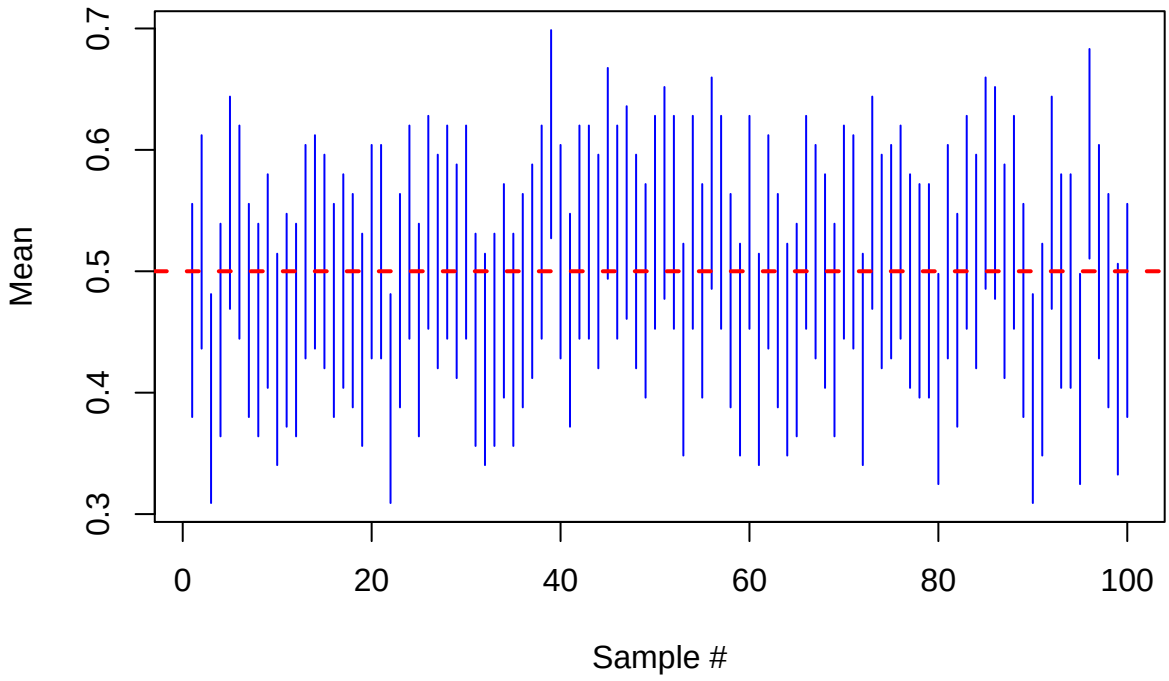
Because the age-standardized US population should have a mean IQ of 100 with a normal distribution that is symetrical around its' mean, we should expect 50% of the population to be below 100. In the case of the near-lead community, we're seeing that 76% have an IQ below 100; 24pts lower than expected.

13. Perform a simulation study to assess the performance of the 95% confidence interval for the proportion of children in the community with IQ less than 100. For the simulation, assume that the true population proportion is equal to 0.5. Does the confidence interval have approximate 95% coverage? (Note: save the sample proportions and their estimated SEs for answering the following questions.)

```r
set.seed(1)

proportion_studier <- function(sample_size, nb_samples, p) {
  phat <- rep(NA, nb_samples)
  for(i in 1:nb_samples) phat[i] <- mean(rbinom(sample_size, size=1, prob=p))
  sample_se <- sqrt(phat*(1-phat)/sample_size)
  se <- sqrt(0.5*(1 - 0.5)/sample_size)
  lower <- phat - sample_se*qnorm(1- 0.05/2)
  upper <- phat + sample_se*qnorm(1- 0.05/2)
  data.frame(phat, sample_se, se, lower, upper)
}

results <- proportion_studier(sample_size=length(iqData$IQ), nb_samples=100, p=0.5)


print(matplot(rbind(1:100,1:100),t(results[,c("lower","upper")]),type="l",lty=1,lwd=1,col=4,xlab="Sample
```

```
## NULL
```

```r
abline(h=0.5,lty=2,col=2,lwd=2)
```

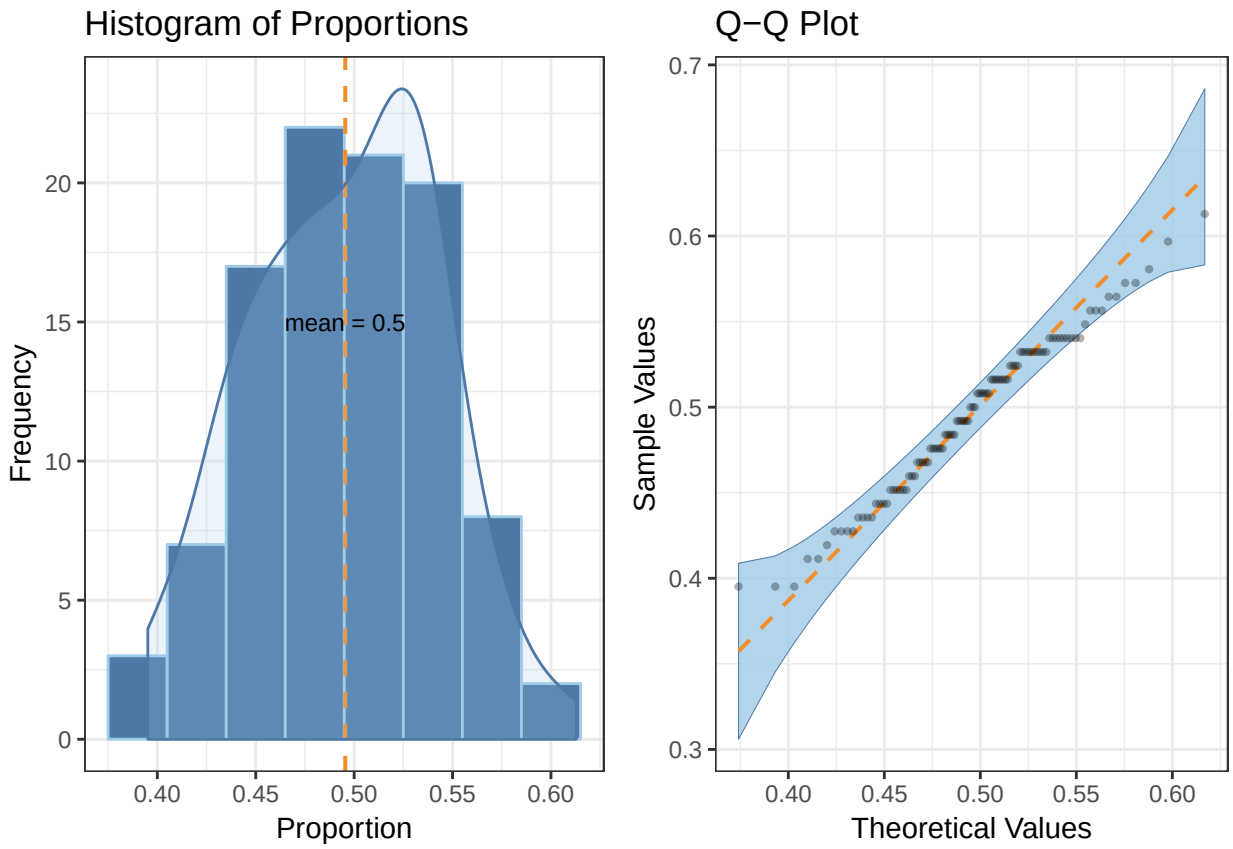



```r
coverage <- sum(results$upper >= 0.5 & results$lower <= 0.5) / length(results$upper)
```

Given that 5 of the 100 samples shown have 95% confidence intervals that don't include the mean, we have no reason to reject the idea that the coverage is as expected.

14. Display the distribution of the simulated sample proportions and describe how well it is approximated by a normal distribution. Is this what you expected? Why or why not?

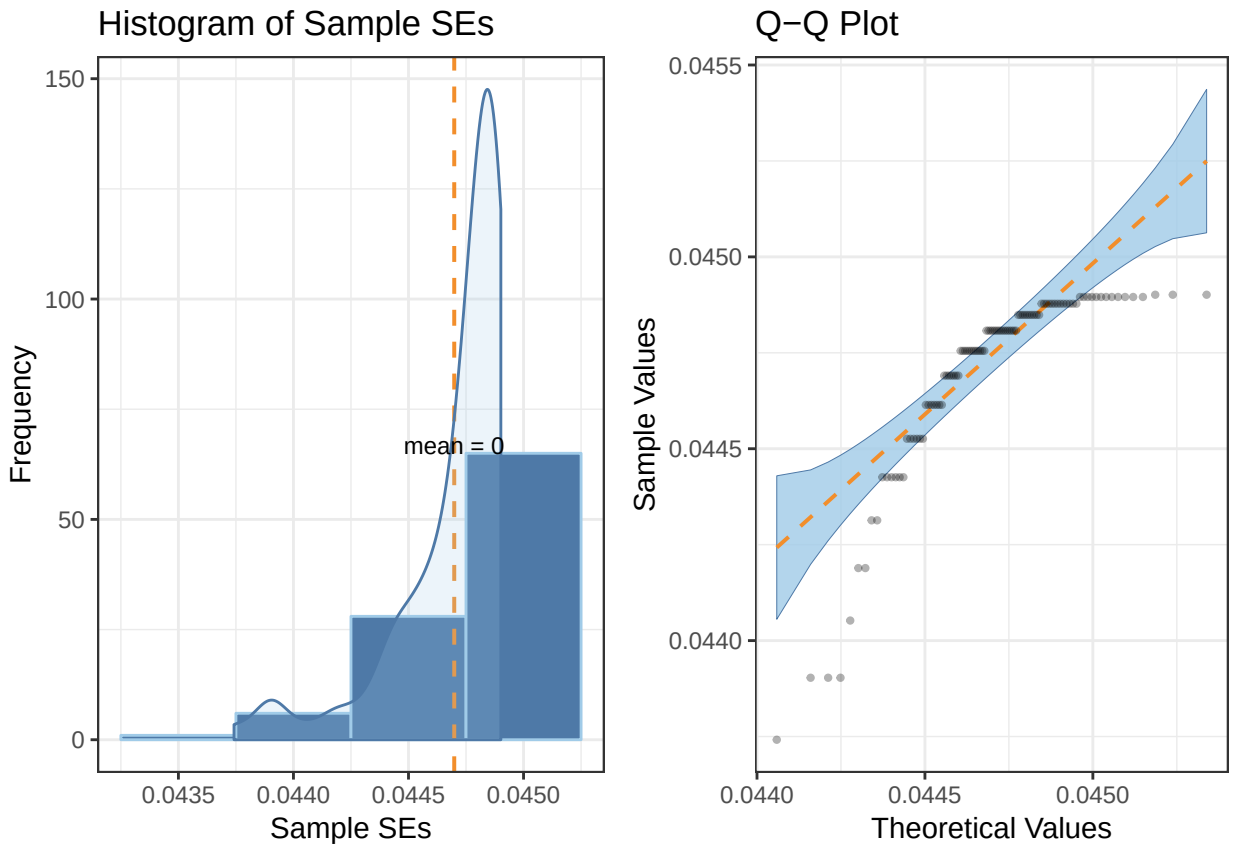
```
distribution_visualizer(results$phat, "Histogram of Proportions", "Proportion", "Frequency", 0.03)
```



This data is well-represented by a normal distribution. This aligns with expectations given that the size of the samples and the CLT.

15. Display the distribution of the estimated SEs of the sample proportions. Compare the average value of the estimated SEs to the SD of the sample proportions. Are they approximately equal? Explain why this matters.

```
distribution_visualizer(results$sample_se, "Histogram of Sample SEs", "Sample SEs", "Frequency", 0.0005)
```




```
mean(results$sample_se)
```

```
## [1] 0.0446988
```

```
results$se[1]
```

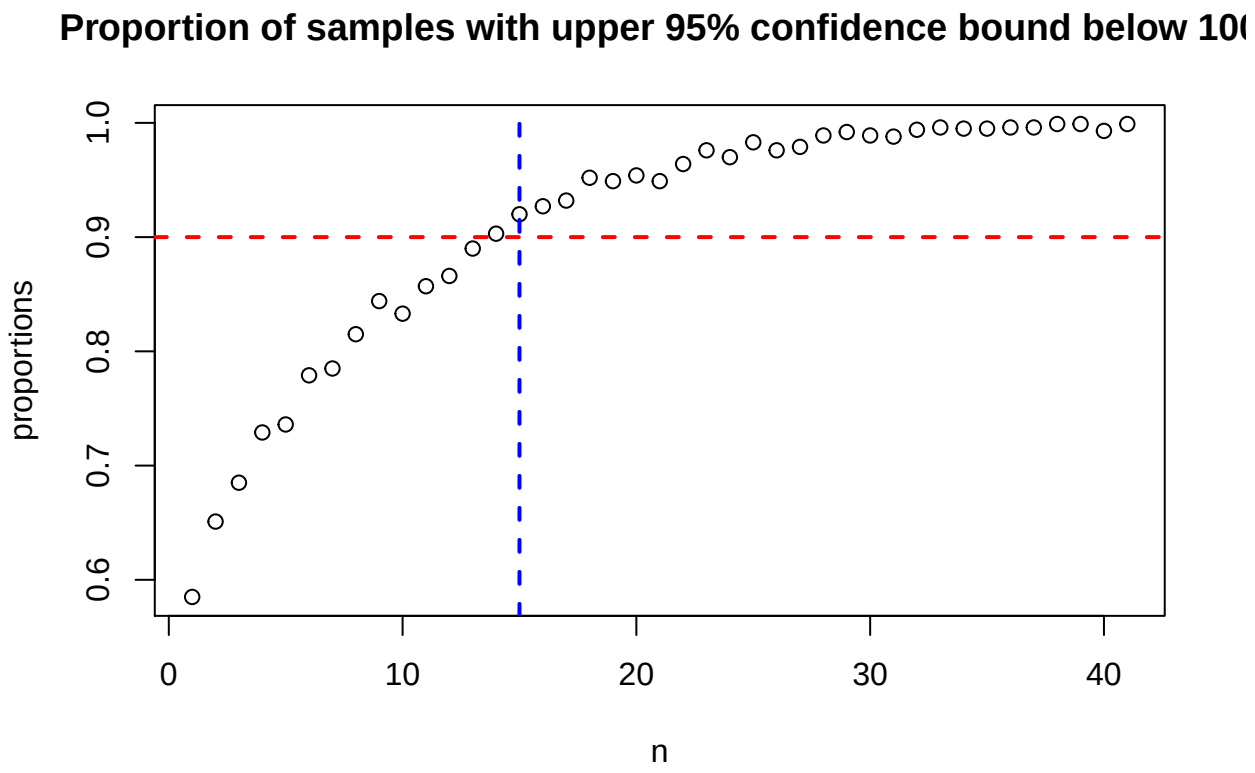
```
## [1] 0.04490133
```

Both are approximately equal. This means that the SEs are a good estimate of the variability in the samples.

16. Using simulation, determine an approximate minimal sample size that yields a valid 95% confidence interval for the proportion of children with IQ less than 100 if the true population proportion is equal to 0.5. (Note: the key word here is "approximate"; there is no precise answer to this question).

```
# needs updating for this problem

set.seed(10)
ns <- seq(10, 50, 1)
sample_collection <- matrix(NA, nrow = 1000,ncol = length(ns))

for(i in 1:ncol(sample_collection)){
  samples <- replicate(1000, rnorm(ns[i], 90, 15))
  intervals <- apply(samples, FUN=conf_intervalizer, MARGIN=2)
  uppers <- sapply(intervals, FUN = get_upper)
  sample_collection[,i] <- uppers
```

```
}

proportionizer <- function(uppers) sum(uppers<100)/1000
proportions <- apply(sample_collection, FUN = proportionizer, MARGIN = 2)

plot(proportions, main = "Proportion of samples with upper 95% confidence bound below 100", xlab = "n")
abline(h=0.9,lty=2,col="red",lwd=2)
abline(v=15,lty=2,col="blue",lwd=2)
```



17. Using the formula for the SE of a sample proportion, calculate the minimal sample size that would yield a 95% confidence interval for the sample proportion of width 0.1 or less. The confidence interval should have length 0.1 or less for any value of the population proportion.

```
# ran out of time... spent too much time working on that pretty plotter, haha.  Lesson learned
```