Chapter 5 - Foundations for Inference

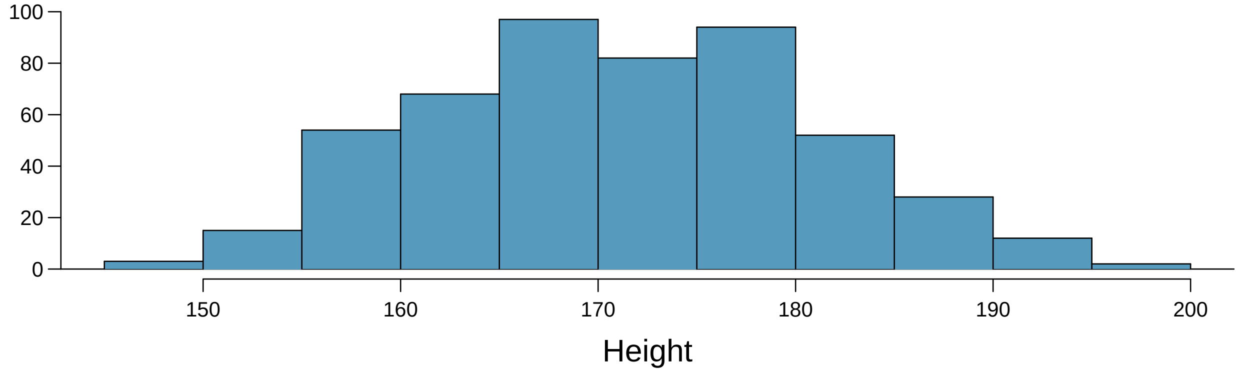
Zachary Palmore

```r
library(visualize)
library(DATA606)
```

```
##
## Welcome to CUNY DATA606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='DATA606') will list the demos that are available.
```

```r
library(openintro)
library(tidyverse)
# Data was downloaded from Github
tgSpending <- read.delim("https://raw.githubusercontent.com/jbryer/DATA606Fall2020/master/course_data/os
```

Heights of adults. (7.7, p. 260) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.



(a) What is the point estimate for the average height of active individuals? What about the median?

```r
# Calculating using summary stats
summary(bdims$hgt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   147.2   163.8   170.3   171.1   177.8   198.1
```

The average (Mean) is 171.1 and the median is 170.3. The point estimate for the average height of active individuals is the mean in this case because all individuals in the study were physically active.

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

```
hgt_sd <- sd(bdims$hgt)
hgt_IQR <- IQR(bdims$hgt)
hgt_mean <- mean(bdims$hgt)
hgt_sd
```

```
## [1] 9.407205
```

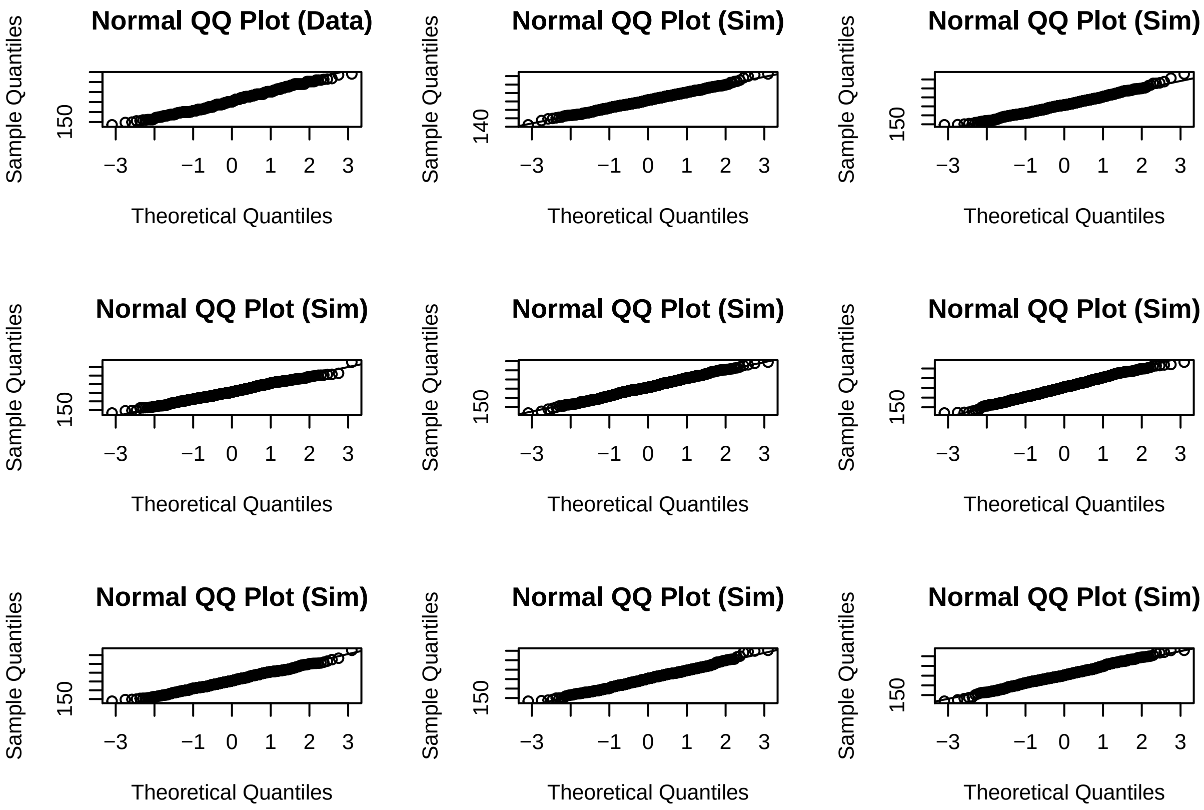
```
hgt_IQR
```

```
## [1] 14
```

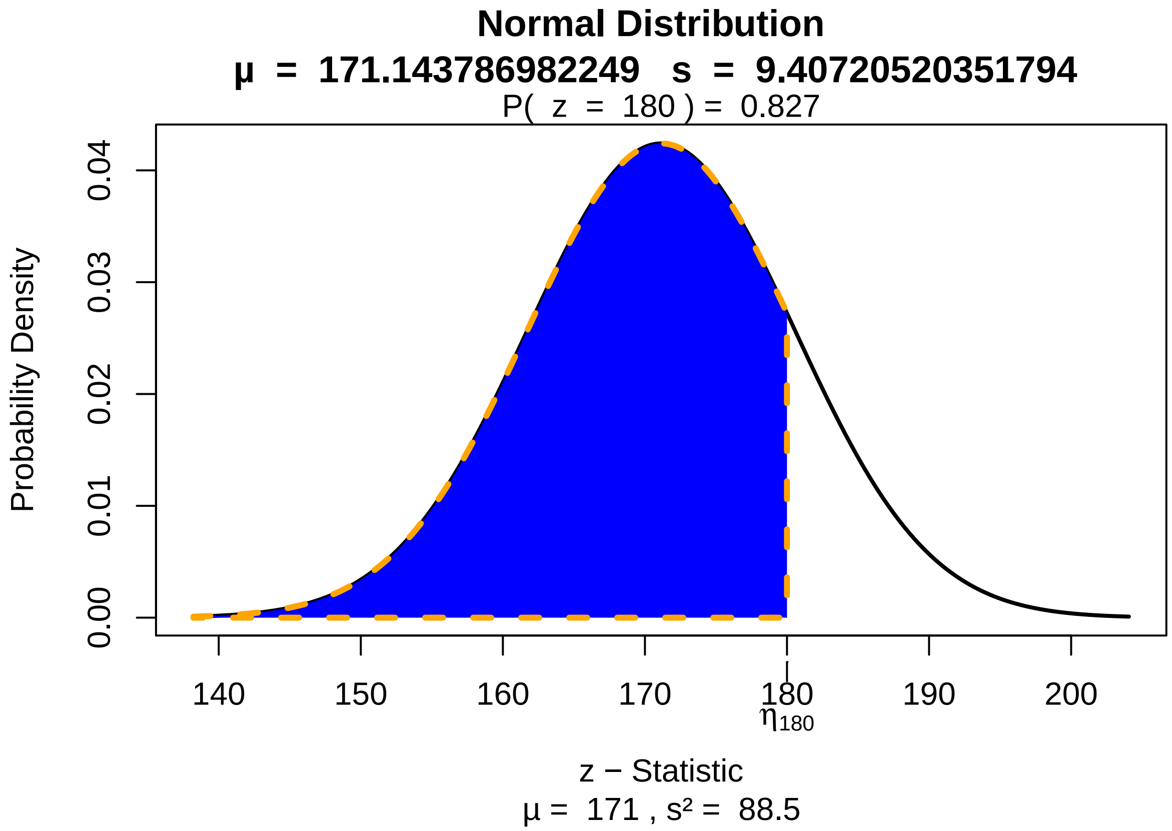
The point estimate for the standard deviation is approximately 9.41 for active individuals. The IQR is 14 for the same height data.

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

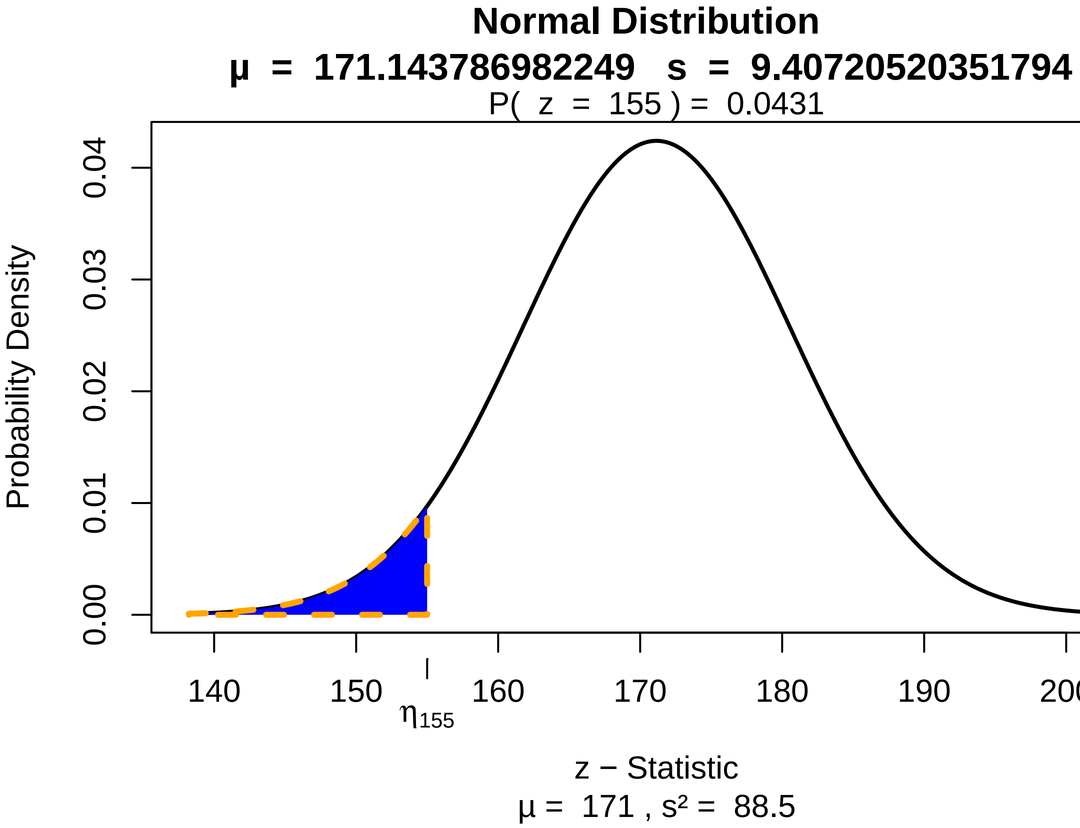
```
# Is this plot normal?
qqnormsim(bdims$hgt)
```



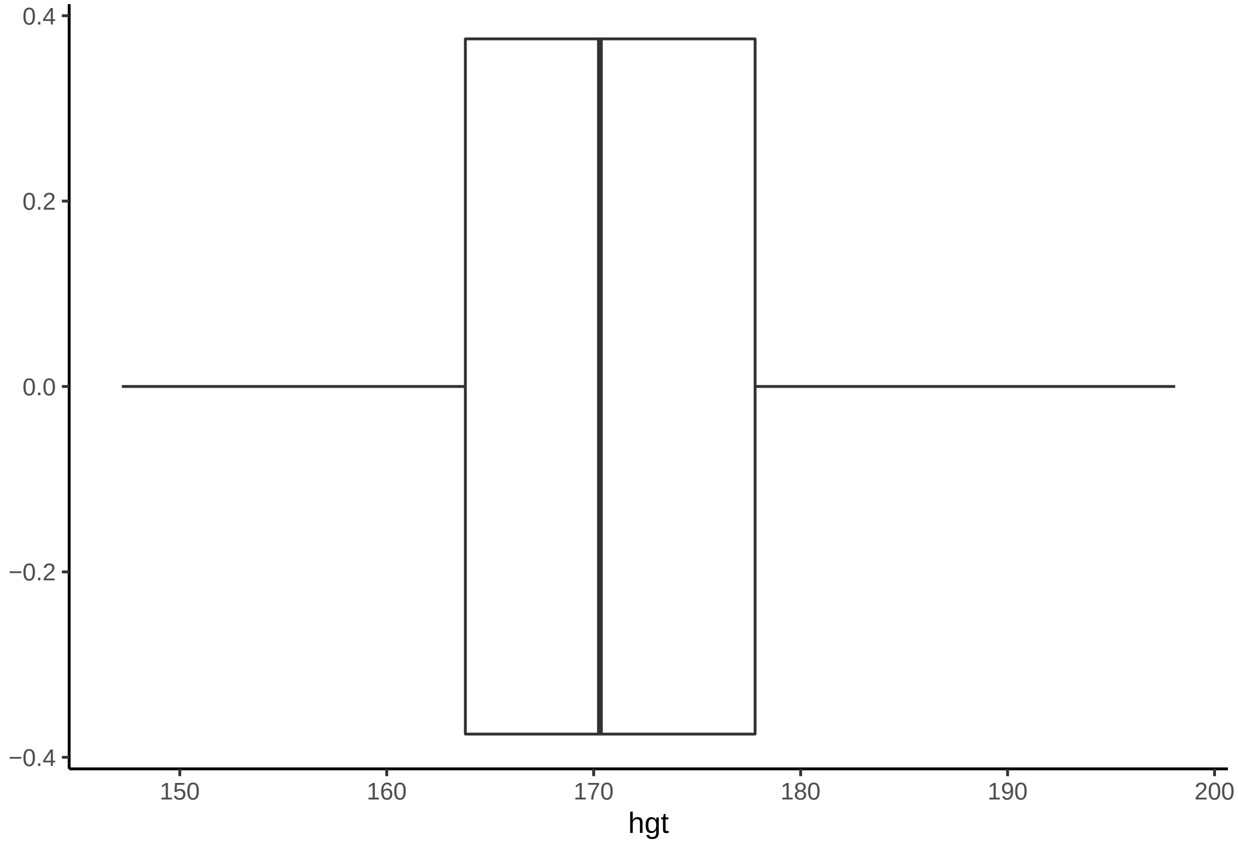
```
# Yes, it appears so.
# visualize.norm(stat = c(155,180), mu = hgt_mean, sd = hgt_sd, section = "bounded")
visualize.norm(stat = 180, mu = hgt_mean, sd = hgt_sd, section = "lower")
```



```
visualize.norm(stat = 155, mu = hgt_mean, sd = hgt_sd, section = "lower")
```



```
ggplot(data = bdims, aes(x = hgt)) +
  geom_boxplot() + theme_classic()
```



```
zscore <- function(x, m, s) {
  diff <- (x - m)
  z <- diff / (s)
  return(z)
}
zscore(155, hgt_mean, hgt_sd)
```

```
## [1] -1.716109
```

```
zscore(180, hgt_mean, hgt_sd)
```

```
## [1] 0.9414287
```

The height of 180 is not unusual. It falls within one standard deviation of the mean (specifically 0.941) and is only 2 cm from the upper IQR. Given that this is a normal distribution (as shown with the qqplots and normal simulations), the height of 180 falls within the range of 95% of the data.

The height of 155 is less common, but not unusual. It falls within 2 standard deviations of the mean (specifically 1.72 to the left of the mean). This point is still found among 99% of the data.

To me, for a height to be unusual, it has to fall outside of 2 standard deviations from the mean. An example would be the heights of 150 cm or 195 cm. Both have zscores less than or greater than -2 and 2 respectively which makes them unusual heights for this study.

```r
zscore(150, hgt_mean, hgt_sd)
```

```
## [1] -2.247616
```

```r
zscore(195, hgt_mean, hgt_sd)
```

```
## [1] 2.535951
```

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

No, they should not be identical to the ones given above. If done properly the random sampling will produce differences in the sample statistics because the data will have been selected at random a second time.

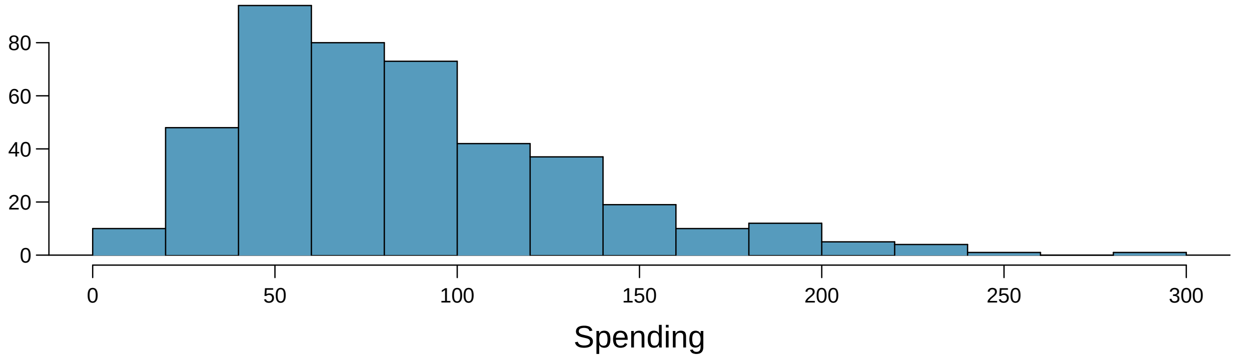
(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

The measure we use to quantify the variability of the sample means is the variance which is approximately 88.5.

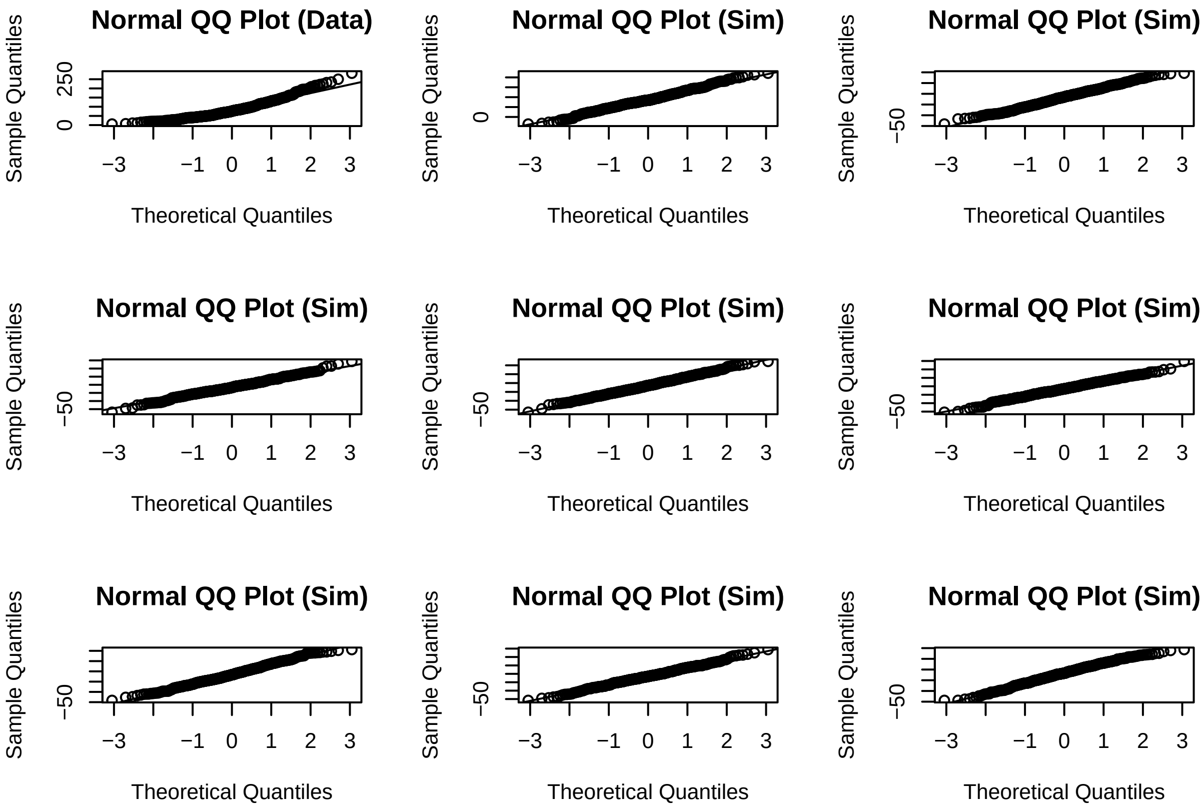
```r
var(bdims$hgt)
```

```
## [1] 88.49551
```

Thanksgiving spending, Part I. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged $84.71. A 95% confidence interval based on this sample is ($80.31, $89.11). Determine whether the following statements are true or false, and explain your reasoning.



```r
qqnormsim(tgSpending$spending)
```



```r
summary(tgSpending)
```

```
##     spending
##  Min.   :  5.719
##  1st Qu.: 49.177
##  Median : 75.792
##  Mean   : 84.707
##  3rd Qu.:112.255
##  Max.   :282.803
```

```
mean_tgSpending <- mean(tgSpending$spending)
sd_tgSpending <- sd(tgSpending$spending)
sd_tgSpending
```

```
## [1] 46.92851
```

```
z_tgSpending <- 1.96
n_tgSpending <- 436
upci_tgSpending <- mean_tgSpending + z_tgSpending *(sd_tgSpending/sqrt(n_tgSpending))
loci_tgSpending <- mean_tgSpending - z_tgSpending *(sd_tgSpending/sqrt(n_tgSpending))
ci_upper <- function(x, z, s, n) {
  ci <- x + z * (s/sqrt(n))
#   sdn <- s/sqrt(n)
  return(ci)
}
ci_upper(mean_tgSpending, z_tgSpending, sd_tgSpending, n_tgSpending)
```

```
## [1] 89.1118
```

```
ci_lower <- function(x, z, s, n) {
  ci <- x - z * (s/sqrt(n))
#   sdn <- s/sqrt(n)
  return(ci)
}
ci_lower(mean_tgSpending, z_tgSpending, sd_tgSpending, n_tgSpending)
```

```
## [1] 80.30173
```

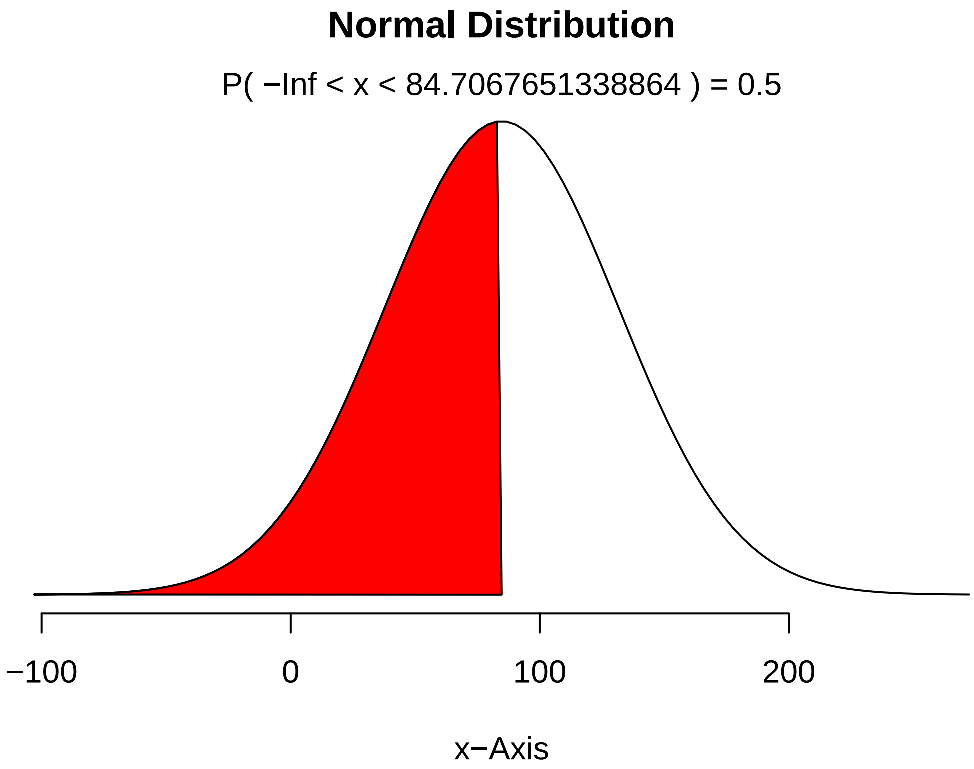
```
upci_tgSpending
```

```
## [1] 89.1118
```

```
loci_tgSpending
```

```
## [1] 80.30173
```

```
normalPlot(mean = mean_tgSpending, sd = sd_tgSpending, bounds = c(-Inf, mean_tgSpending))
```



(a) We are 95% confident that the average spending of these 436 American adults is between \$80.31 and \$89.11.

False, it is only referencing the sample not the population parameter. CIs are supposed to only be about the population parameter.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

False, although the CI is best calculated using a more normal distribution, this one has a large enough sample size that is valid.

(c) 95% of random samples have a sample mean between \$80.31 and \$89.11.

False, CI's are not a probability interpretation. There is also no reference to the parameter of interest. That is, this survey seeks to show that we have 95% confidence that the average spending of American Adults is between \$80.31 and \$89.11.

(d) We are 95% confident that the average spending of all American adults is between \$80.31 and \$89.11.

True. This statement is true because it uses proper language and is only about the population parameter.

(e) A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about our estimate.

True, a 90% CI would be narrower than the 95% CI. If you do not need to be as confident, or sure, about the estimate, then a 90% could be better.

(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

False, to decrease the margin of error, you could increase the sample size. However, to reduce it to a third of what it is now, you would need a sample size that is more than just 3 times larger (specifically about 9 times).

(g) The margin of error is 4.4.

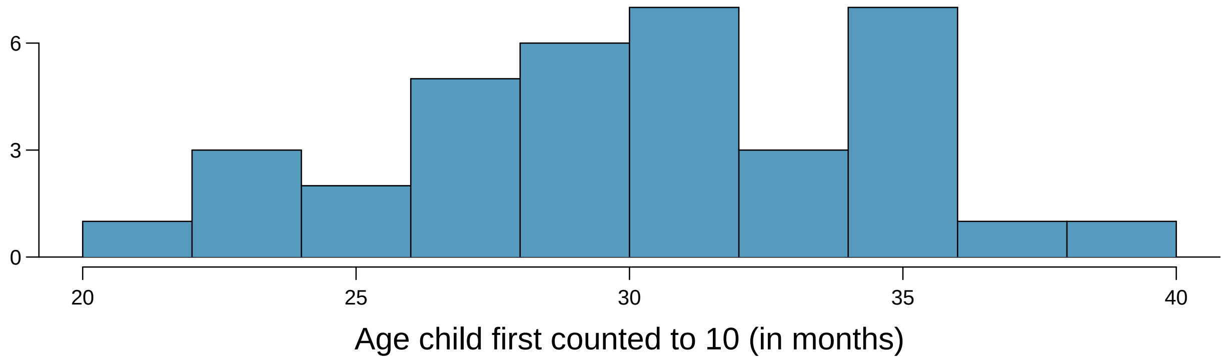
True, the margin of error is 4.4.

Given the sample size, standard deviation, and mean, the margin of error can be calculated like this:

```
moe_tgspending <- z_tgSpending * (sd_tgSpending/sqrt(n_tgSpending))
moe_tgspending
```

```
## [1] 4.405038
```

Gifted children, Part I. Researchers investigating characteristics of gifted children col- lected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The following histogram shows the dis- tribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

Yes, it is a simple random sample and at a 5% significance level it passes the success-failure condition. It has a large enough sample size (>30) and is a nearly normal distribution too.

```
# Check if they are at least 10
n <- 36
n*(1 - 0.05)
```

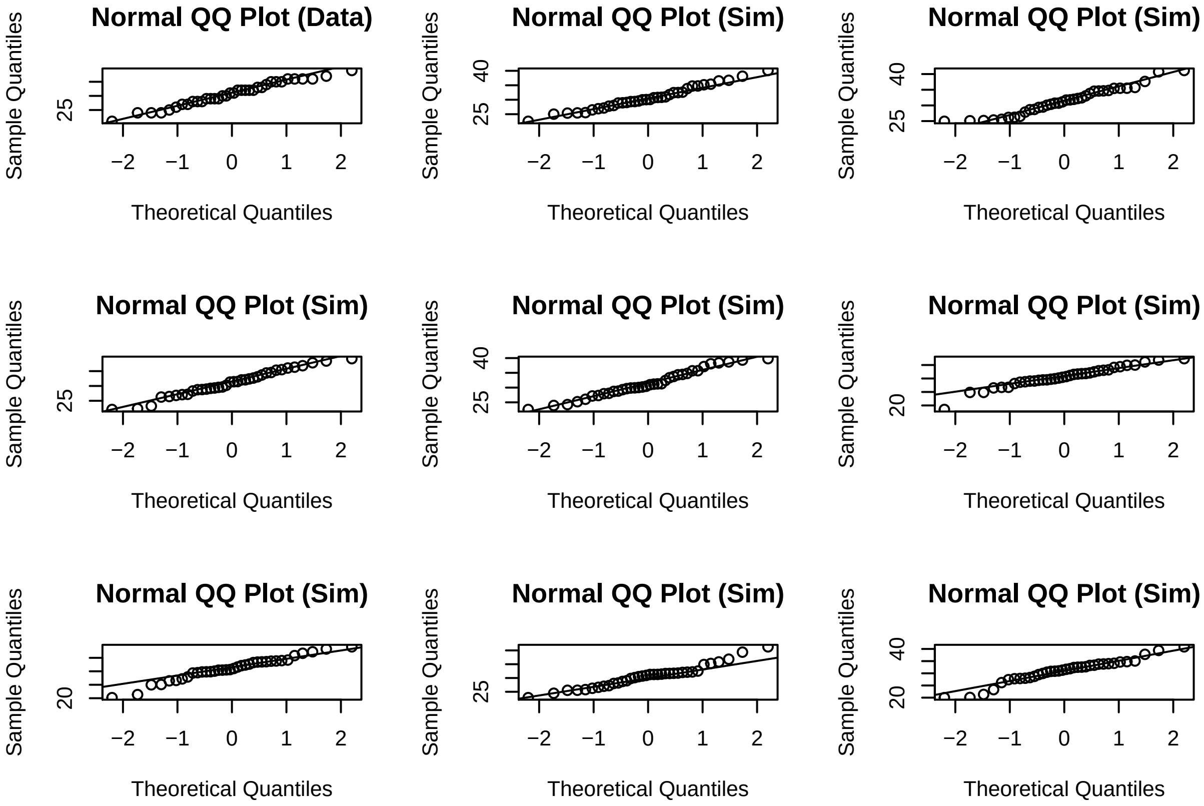
```
## [1] 34.2
```

```
n*0.05
```

```
## [1] 1.8
```

```
# Yes, 18 > 10
```

```
qqnormsim(gifted$count)
```



(b) Suppose you read online that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children fist count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

We are interested in understanding the average age that children learn to count to 10. The null hypothesis is that gifted children first learn to count to 10 successfully when they are 32 months old on average. The alternative hypothesis is that gifted children first learn to count to 10 earlier than the general average of 32 months.

Although we could add to the alternative hypothesis the statement that, *the gifted children first learn to count to 10 later than the general average of 32 months*, we will not. In this scenario, we will not consider if they learn after the general average to make this a one-tailed test rather than two.

In other words we have;

$H_o = 32 $

$H_a \neq 32$

```r
# Looking for confidence first
gif_mean <- mean(gifted$count)
gif_sd <- sd(gifted$count)
gif_n <- 36
# z score of 1.96 for 95% confidence
ci_upper(gif_mean, 1.96, gif_sd, gif_n)
```

```
## [1] 32.10397
```

```r
ci_lower(gif_mean, 1.96, gif_sd, gif_n)
```

```
## [1] 29.28491
```

We know the mean of the gifted children is 30.69 and median is 31.00. The standard deviation is 4.315. We were also given the number of children in the sample as 36. We are 95% confident that the average age when all gifted children learn to count to 10 is between ages 29.3 and 32.1. Because the null hypothesis of 32 is within this range, we cannot reject it. The null hypothesis is still plausible at this stage without more evidence.

```r
# Finding the p-value using the z-score
h <- 32
gif_z <- (h - gif_mean) / (gif_sd/(sqrt(gif_n)))
# Alternatively (32 - 30.69) / (4.3/(sqrt(36)))
```

Our z-score is 1.82, which corresponds with 0.9656 in the z-table.

```r
1 - 0.9656
```

```
## [1] 0.0344
```

The p-value is 0.0344. This is less than 0.10 (our significance level) so we should reject the null hypothesis.

(c) Interpret the p-value in context of the hypothesis test and the data.

The p-value is a measure of the probability that an observed value could have occurred by random chance in the data. We can interpret this p-value 0.034 to mean that, under the null hypothesis, the chance of observing this mean age is only about 3.4%.

(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.

29.51 to 31.88 is the interval with 90% confidence for the average age at which gifted children first counted to 10 successfully.

Formally:

We can say with 90% confidence that the average age at which gifted children learn to count is between 29.51 and 31.88 months.

```r
# Z is 1.645 at this 90% confidence interval
ci_lower(gif_mean, 1.645, gif_sd, gif_n)
```

```
## [1] 29.51145
```

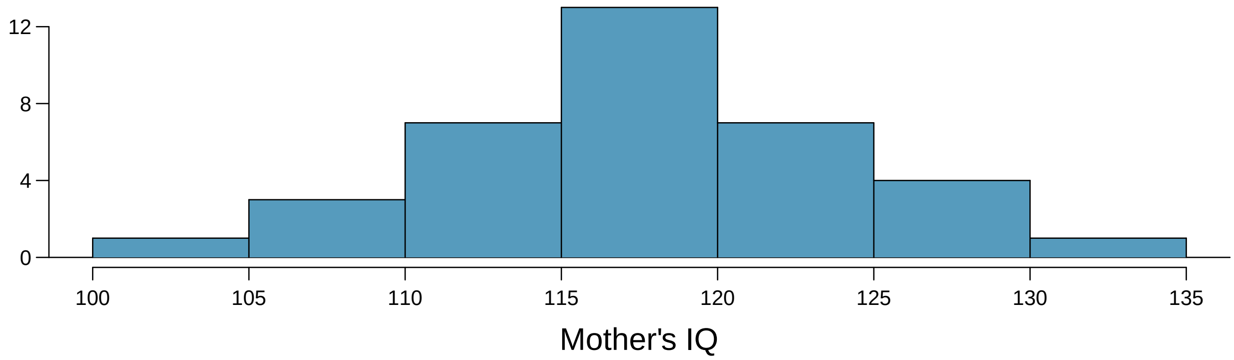
```r
ci_upper(gif_mean, 1.645, gif_sd, gif_n)
```

```
## [1] 31.87744
```

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

Yes, the results agree because the chance of observing a gifted child count to 10 at the age of 32 months is not very likely since in this case it is just over the upper 90% confidence interval.

Gifted children, Part II. Exercise above describes a study on gifted children. In this study, along with variables on the children, the researchers also collected data on the mother's and father's IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.



n	36
min	101
mean	118.2
sd	6.5
max	131

(a) Perform a hypothesis test to evaluate if these data provide convincing evidence that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100. Use a significance level of 0.10.

Null hypothesis: The mothers of gifted children have IQ scores that are the same as the average score for the population of all mothers.

Alternative hypothesis: The mothers of gifted children have IQ scores that are greater than or less than the average score for the population of all mothers.

```
# Using the given values
# Finding the p-value using the z-score
(100 - 118.2)/6.5
```

```
## [1] -2.8
```

Our z-score is -2.8.

Using the z-table of score the p-value is 0.002555. This is less than 0.10 (our significance level) so we should reject the null hypothesis.

(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.

```
# Z is 1.645 at this 90% confidence interval
ci_lower(118.2, 1.645, 6.5, 36)
```

```
## [1] 116.4179
```

```
ci_upper(118.2, 1.645, 6.5, 36)
```

```
## [1] 119.9821
```

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

Yes, the results from the hypothesis test and the confidence interval agree. There is very little chance that a mother with a gifted child would have an IQ of 100, which is equivalent to the population of mothers at large.

CLT. Define the term "sampling distribution" of the mean, and describe how the shape, center, and spread of the sampling distribution of the mean change as sample size increases.

A sampling distribution of the mean is a relative frequency distribution in which samples are gathered, means of those samples calculated, and plotted repeatedly until a large enough number of samples can create a distribution.

When sample size increases the spread of this distribution narrows. It slims itself as it centers closer to the mean of the distribution which is generally symmetric about the mean. As sample size continues to increase, the distribution becomes more symmetric. As sample size approaches positive infinity, values that were more discrete become indistinguishable as the distribution appears to smooth itself into a tight bell-shape.

CFLBs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?

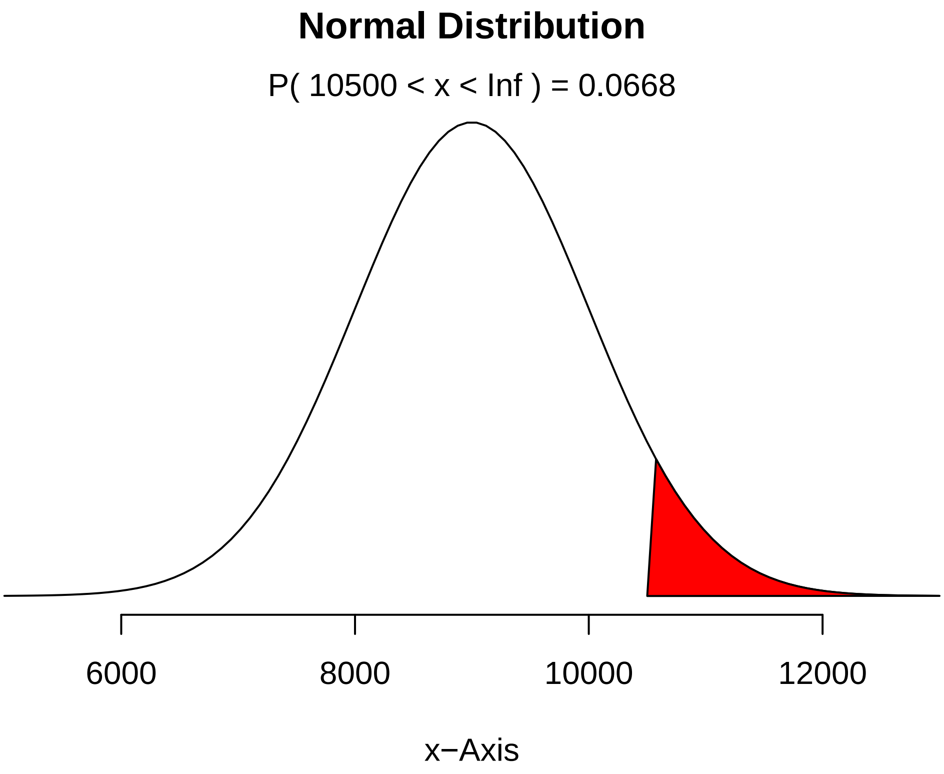
```
round(1 - pnorm(10500, mean = 9000, sd = 1000),4)
```

```
## [1] 0.0668
```

The probability that a randomly selected bulb lasts more than 10,500 hours is about 6.68%.

We can see this in the plot as well.

```
normalPlot(mean = 9000, sd = 1000., bounds = c(10500, Inf))
```



(b) Describe the distribution of the mean lifespan of 15 light bulbs.

The distribution of 15 light bulbs would not appear as normally distributed as a larger sample size (>30). It would have more discrete values than the larger sample size, may be more spread out, and might not have a symmetric shape. Although we are aware this data is normally distributed about the mean of 9,000, having a smaller sample size increases the probability of deviating from that mean.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours is extremely small, very close to 0% (specifically 3.13*10^-9).

```
x <- 10500
mean_bulbs <- 9000
sd_bulbs <- 1000
n_bulbs <- 15
samp_bulbs_15 <- sd_bulbs/sqrt(n_bulbs)
Pbulbs_15 <- 1-pnorm(x, mean_bulbs, samp_bulbs_15)
Pbulbs_15
```

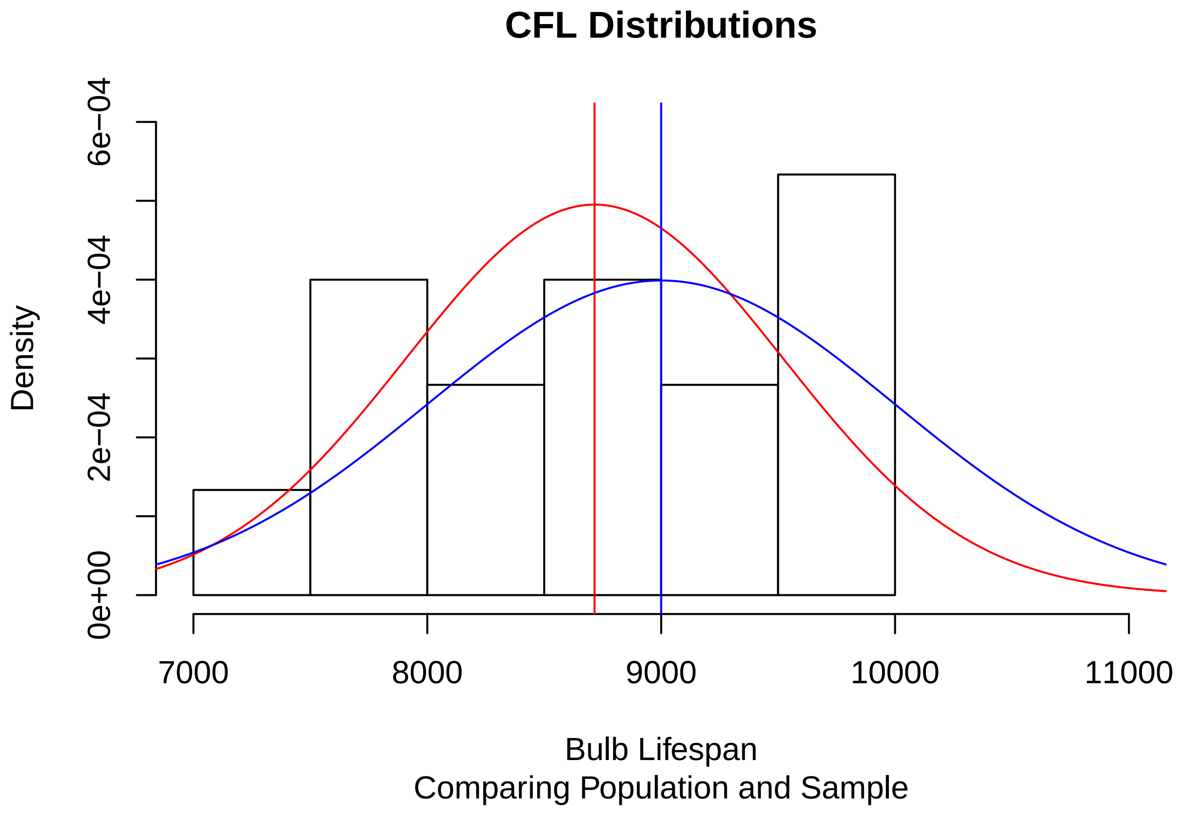
```
## [1] 3.133452e-09
```

(d) Sketch the two distributions (population and sampling) on the same scale.

Using lines on a histogram:

```
set.seed(0924020)
normbulbs <- rnorm(n_bulbs, mean = mean_bulbs, sd = sd_bulbs)
mean_normbulbs <- mean(normbulbs)
sd_normbulbs <- sd(normbulbs)

hist(normbulbs, probability = TRUE,
     xlim = c(7000,11000),
     ylim = c(0,0.0006),
     xlab= "Bulb Lifespan",
     ylab= "Density",
     main= "CFL Distributions",
     sub = "Comparing Population and Sample",
     col = "white",
     )
s <- 0:15000
y15 <- dnorm(x = s, mean = mean_normbulbs, sd = sd_normbulbs)
y <- dnorm(x = s, mean = mean_bulbs, sd = sd_bulbs)
lines(x = s, y = y15, col = "red")
abline(v= mean_normbulbs, col="red")
lines(x = s, y = y, col = "blue")
abline(v= mean_bulbs, col="blue")
```



(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

No, we could not. It must be a normal distribution for us to perform these calculations.

Same observation, different sample size. Suppose you conduct a hypothesis test based on a sample where the sample size is n = 50, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been n = 500. Will your p-value increase, decrease, or stay the same? Explain.

When sample size increases, the p-value should decrease. As the data available to calculate estimates increases, the accuracy of the mean, z-score, and accuracy of other statistics increases. Therefore the probability of error (either due to random chance or mistakes in the data) decreases, thereby reassuring the probability of statistical significance in the p-value.