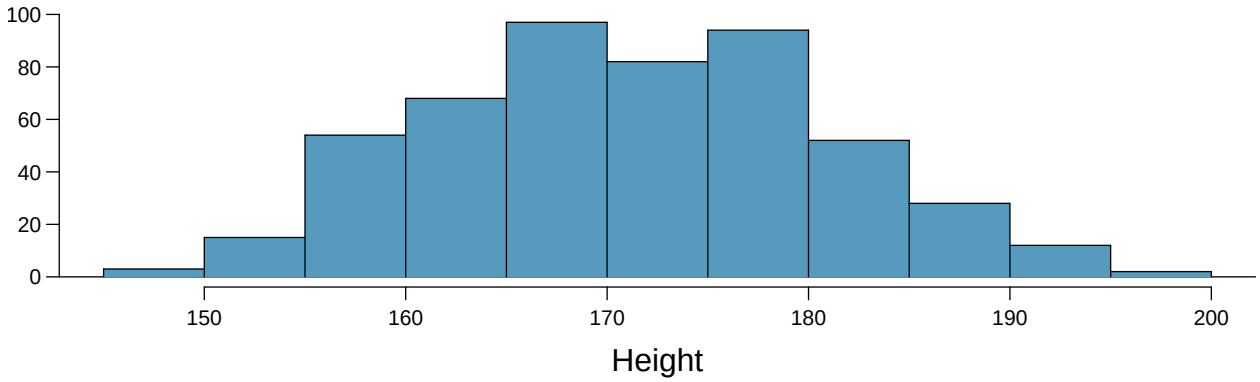
Chapter 5 - Foundations for Inference

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.

```
##
## 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.
```



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

```
mean(bdims$hgt)
```

```
## [1] 171.1438
```

```
median(bdims$hgt)
```

```
## [1] 170.3
```

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

```
sd(bdims$hgt)
```

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

```
IQR(bdims$hgt)
```

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

(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.

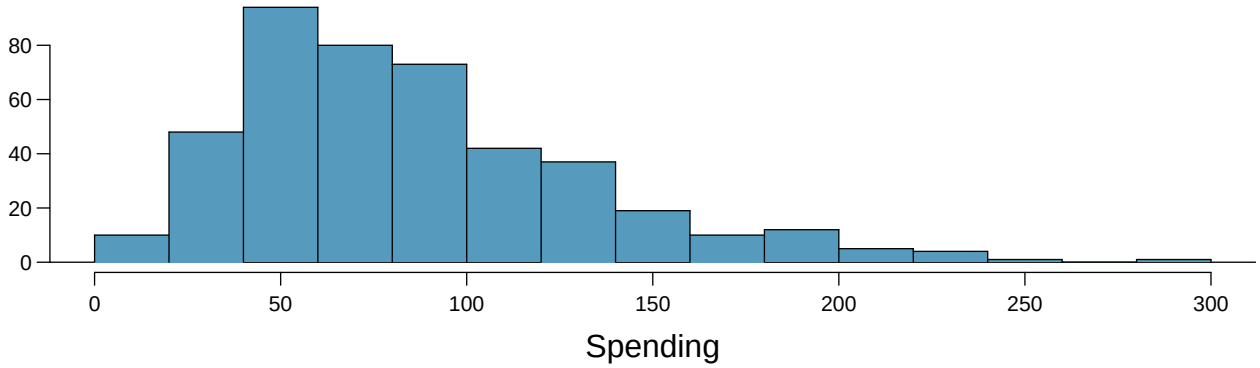
I would say that 180cm is not so unusual because it is within 1 standard deviation from the mean and median... However, 155cm is unusually short by the same logic.

(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.

I would expect them to be very similar, but not the same. There is some significant standard error depending on sample size, although the samples are taken from the same population.

(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.

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.



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

Well, yes and no. There was only one sample taken so the standard error among a sample necessarily smaller than 436 people will be significant.

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

Agreed.

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

Not necessarily! This confidence interval offers a range in which we might find the population mean. Other samples may be different.

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

Since we do not know the size of the sample, we cannot compute standard error!

(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.

Yes.

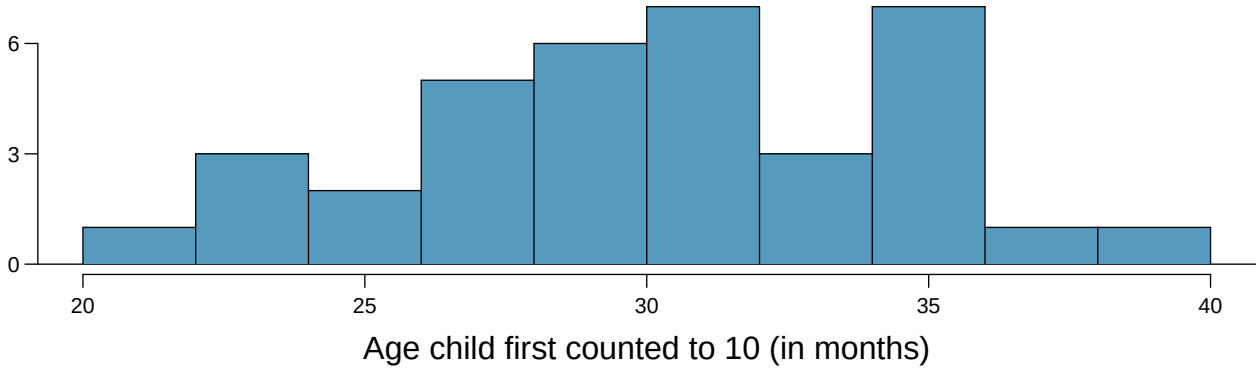
(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

No, you would need a sample 9 times larger!

(g) The margin of error is 4.4.

It would appear that way.

Gifted children, Part I. Researchers investigating characteristics of gifted children collected 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 distribution 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?

Barely, yes. I don't like it, though.

(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.

```
n <- 36
x_bar <- 30.69
sd <- 4.31
x <- 32
z_star <- 1.645

se <- sd / sqrt(n)

(lower <- x_bar - z_star/2 * se)
```

```
## [1] 30.09917
```

```
(upper <- x_bar + z_star/2 * se)
```

```
## [1] 31.28083
```

Since 32 does not fall inside the confidence interval, we reject H_0, and assume that the figures are distinct.

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

Around 62% of these gifted kids counted to 10 before 32 months.

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

```
(lower <- x_bar - z_star/2 * se)
```

```
## [1] 30.09917
```

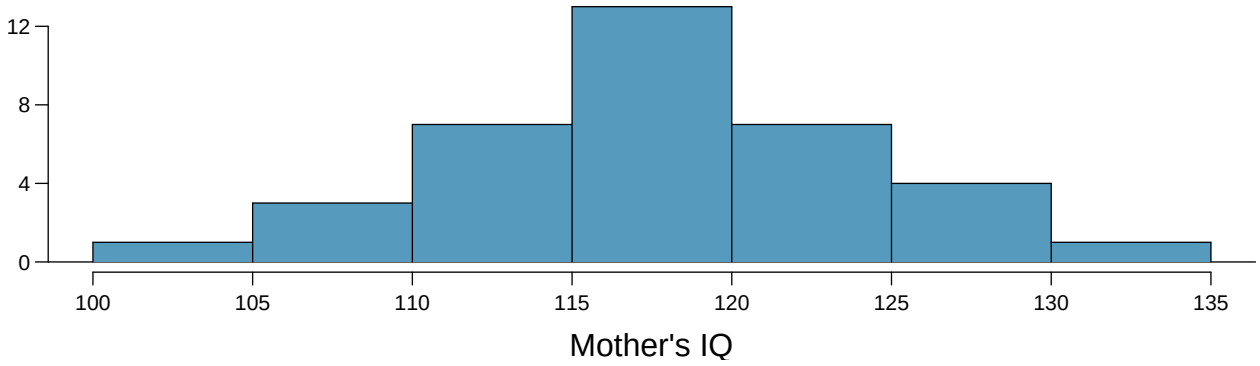
```r
(upper <- x_bar + z_star/2 * se)
```

```
## [1] 31.28083
```

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

No? They're the same thing.

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.

```
n <- 36
mu_hat <- 118.2
sd <- 6.5
se <- sd / sqrt(n)

x <- rnorm(n, mu_hat, sd)
mu <- 100

t.test(x, mu = mu, conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 15.858, df = 35, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 100
## 90 percent confidence interval:
##  118.5456 122.9687
## sample estimates:
## mean of x
##  120.7572
```

I am convinced that the population mean mu is far outside of the 90% confidence interval of the sample mean... We reject H_0, and conclude the sample of mothers with gifted children is distinct from random members of the population.

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

```
(lower <- x_bar - 1.645/2 * se)
```

```
## [1] 29.79896
```

```r
(upper <- x_bar + 1.645/2 * se)
```

```
## [1] 31.58104
```

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

Yes, they both show the same information.

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.

The distribution of sample proportions, taken from a group of samples of one population is called the sampling distribution. The center should reflect the mean of the population, and the spread is the variability of a point estimate. Standard error is taken instead of standard deviation for the sample distribution, and the spread should shrink, and the shape should resemble a normal distribution as sample size increases towards the size of the population.

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.

```
mu <- 9000
sd <- 1000
```

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

```
(p <- 1 - pnorm(10500, mu, sd))
```

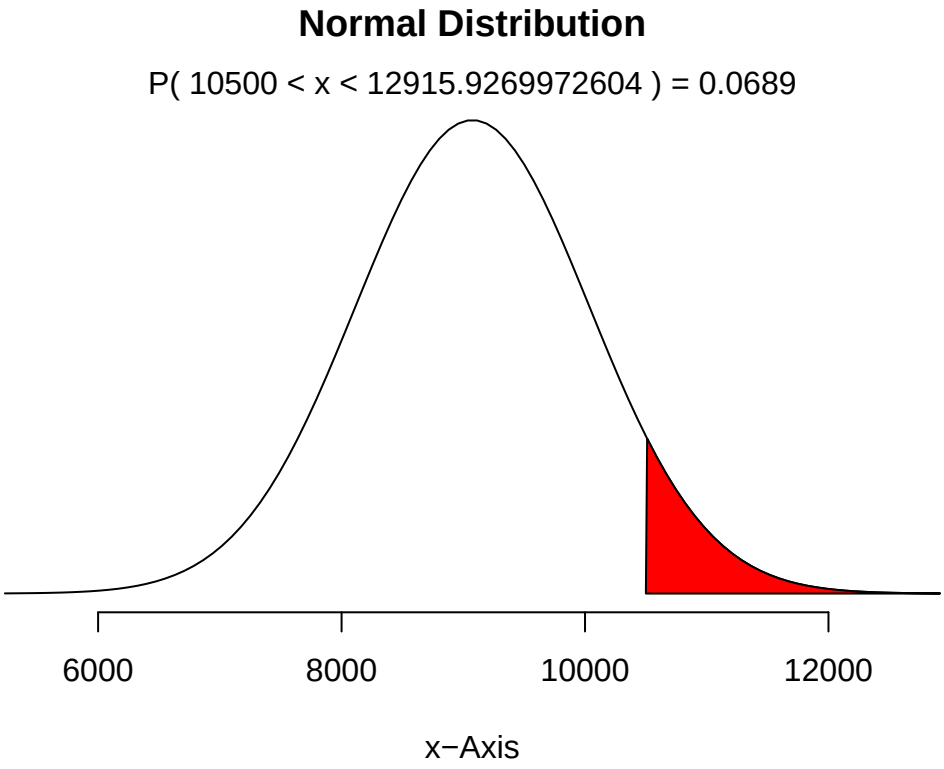
```
## [1] 0.0668072
```

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

```
samp <- rnorm(15, mu, sd)
m <- mean(samp)
s <- sd(samp)
```

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

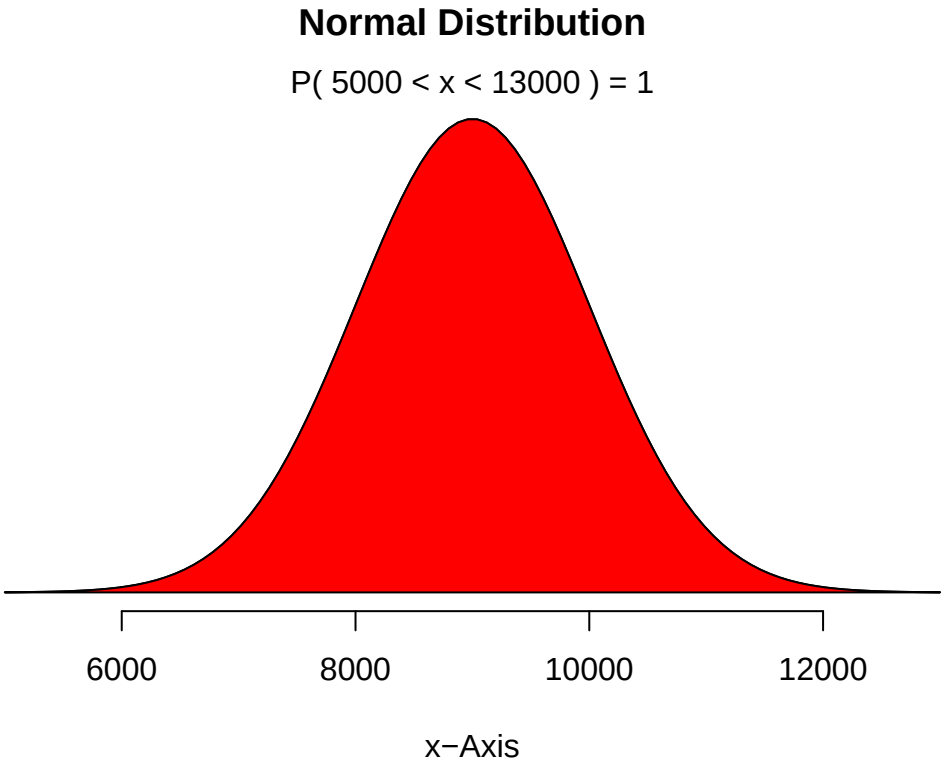
```
normalPlot(m,s, c(10500,m+(4*s)))
```



6.89% chance that the mean of the sample would be higher than 10500 hours.

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

```
normalPlot(9000, 1000, c(5000, 13000))
```



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

If you knew exactly how they were skewed, and you were able to normalize the distribution of the probabilities with some function, then yes. As long as the data stay skewed, these methods will not work.

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.

N is used to calculate p_0, the null proportion, in a hypothesis test... If n increases by a factor of 10, so might the p-value. This would change everything!