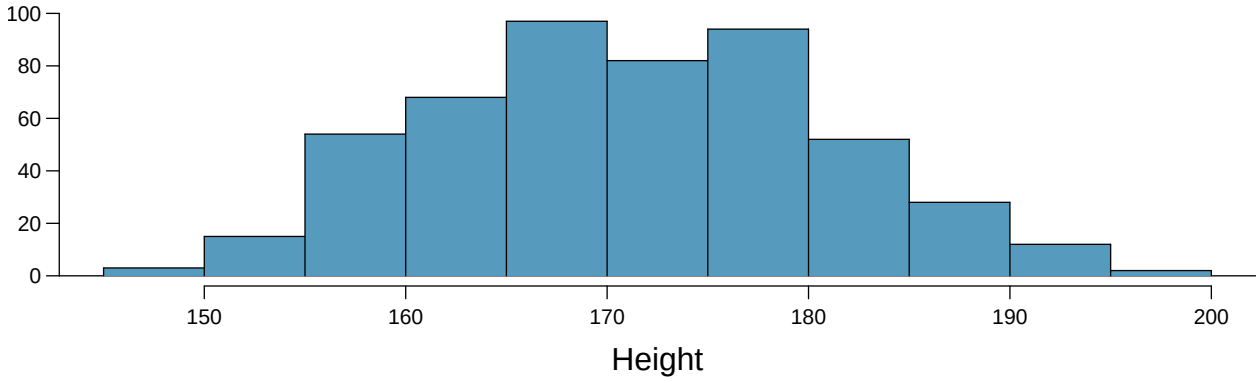
Chapter 5 - Foundations for Inference

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.



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

```
favstats(bdims$hgt)
```

min	Q1	median	Q3	max	mean	sd	n	missing
147.2	163.8	170.3	177.8	198.1	171.1438	9.407205	507	0

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

```
as.table(c(SD =sd(bdims$hgt), IQR= IQR(bdims$hgt)))
```

```
##        SD       IQR
##  9.407205 14.000000
```

(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.

Yes they would be. The 155 for sure as the value falls one standard deviation below from Q1. 180 is above Q3 but still closer than the other value. And it isn't close to the max. It's mor normal than the first value, but still on the high end.

(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.

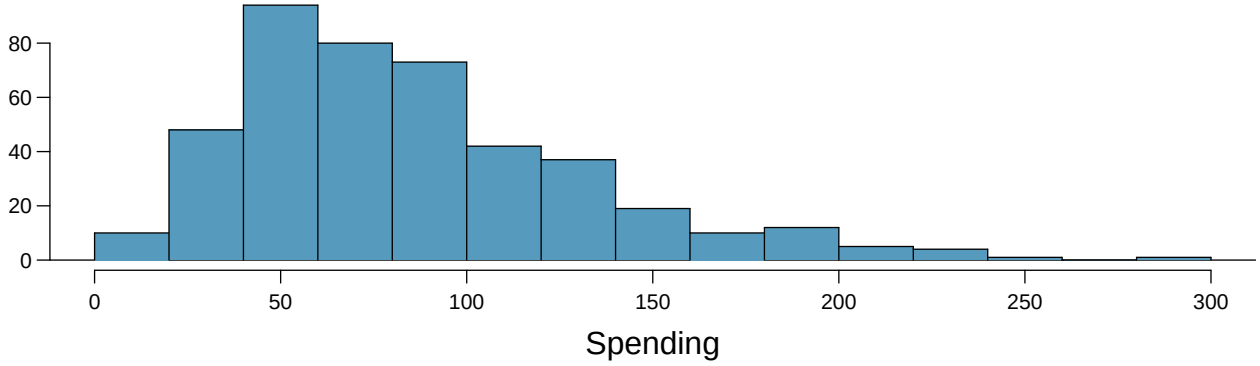
It is highly unlikely that the samples values will be exactly the same. However, we would expect to see a similar overall proportion of values, and a similar mean and standard deviation overall.

(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.

```
std.error(bdims$hgt)
```

```
## [1] 0.4177887
```

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.

False: We know that the sample mean is in the confidence interval. The confidence interval is for the overal population mean not the sample.

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

False: The skew is not significant enough to prevent this confidence interval from being valid. Majority of the data still lies within the confidence interval.

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

False: Samples often have different ranges.

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

True: As mentioned above, the confidence interval is reflection of the overall population.

(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: As shown in the lab!

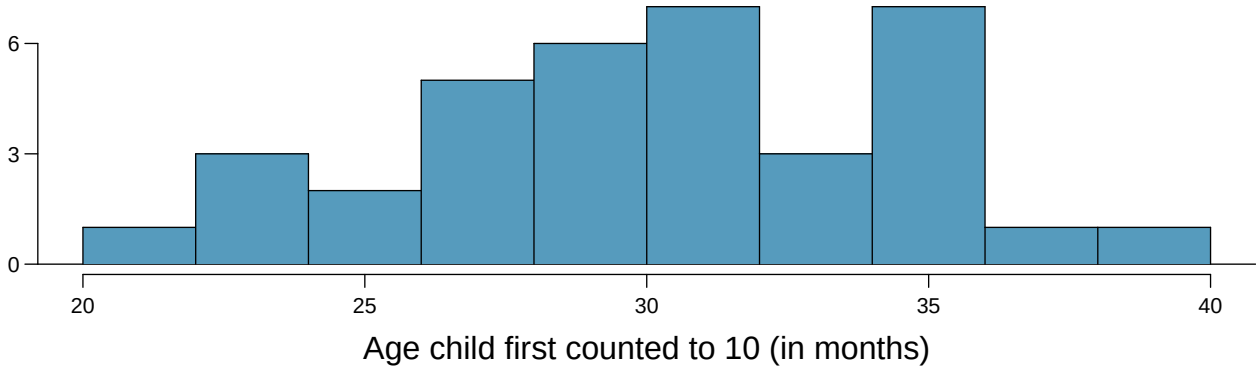
(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

False: Since we take the square root of the sample, we need it to be 9 times larger.

(g) The margin of error is 4.4.

True: The margin of error is z * se or, one half of the confidence interval, which would be $89.11 to $80.312 which equals $4.4

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, we have a large enough sample of randomly select variables from a larger population without a skewed distribution.

(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.

Since our Z is negative, and we want the left side, we can simply throw our Z into the pnorm function as it is, and we will get the same value as we would if we made it positive and then subtracted it from 1.

```
(mean(gifted$count)-32)/ std.error(gifted$count)
```

```
## [1] -1.81542
```

```
pnorm( (mean(gifted$count)-32)/ std.error(gifted$count) )
```

```
## [1] 0.03472969
```

(c) Interpret the p-value in context of the hypothesis test and the data. | We have a value way below .1, therefore we can reject the null hypothesis and conclude that

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

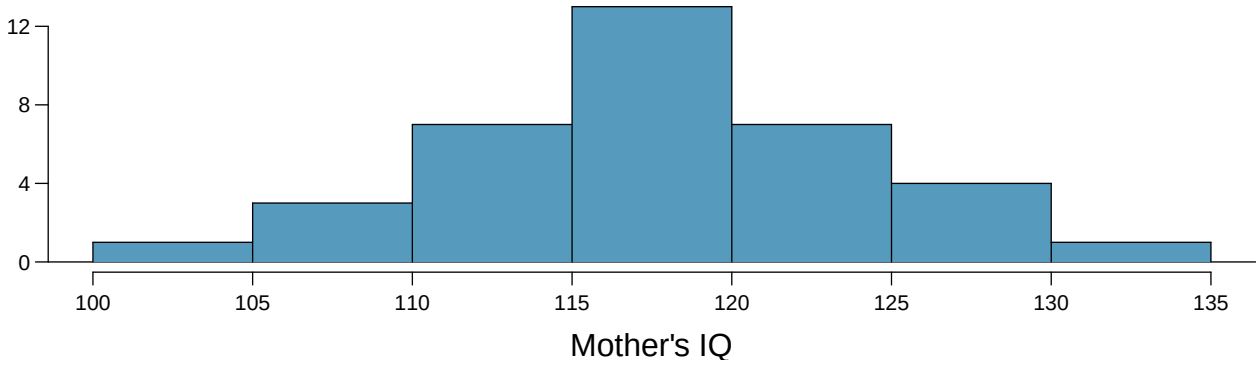
```r
# Using 1.645 for z of a 90% confidence interval

z_value <- 1.645



c(mean(gifted$count) -z_value *  std.error(gifted$count), mean(gifted$count) + z_value *  std.error(gif
```

```
## [1] 29.51145 31.87744
```

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

Yes! Both results point to the average of the gifted children being below the general average of children their age.

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.

```
((mean(gifted$motheriq)-100)/ std.error(gifted$motheriq))
```

```
## [1] 16.75649
```

```
1-pnorm(    (((mean(gifted$motheriq)-100)/ std.error(gifted$motheriq))) )
```

```
## [1] 0
```

Looking at the right for a positive z value, we subtract our z value from 1, giving us a p

of 0. We can reject the null hypothesis.

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

```
# Using 1.645 for z of a 90% confidence interval

z_value <- 1.645


c(mean(gifted$motheriq) -z_value *  std.error(gifted$motheriq),mean(gifted$motheriq) + z_value *  std.e
```

```
## [1] 116.3832 119.9501
```

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

Yes! As shown above, we see in the first test that the mothers IQ in this group is above the average, and in the second test we see that we are 90% certian that their IQ is between 116.38 and 119.95, which is well above the population average of 100.

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 term "sampling distribution" of the mean is the distribution of the values in a n sample from a given population. As that sample size increases, we will epxect to see the shape, center, and spead of the sampling distribution become more normal, having a far more pronounced center and less overall spread with a mean that will approach the original populations value.

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?

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

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

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

```
1000/sqrt(15)
```

```
## [1] 258.1989
```

Approximately normal with a mean of 9,000 and a standard error shown above.

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

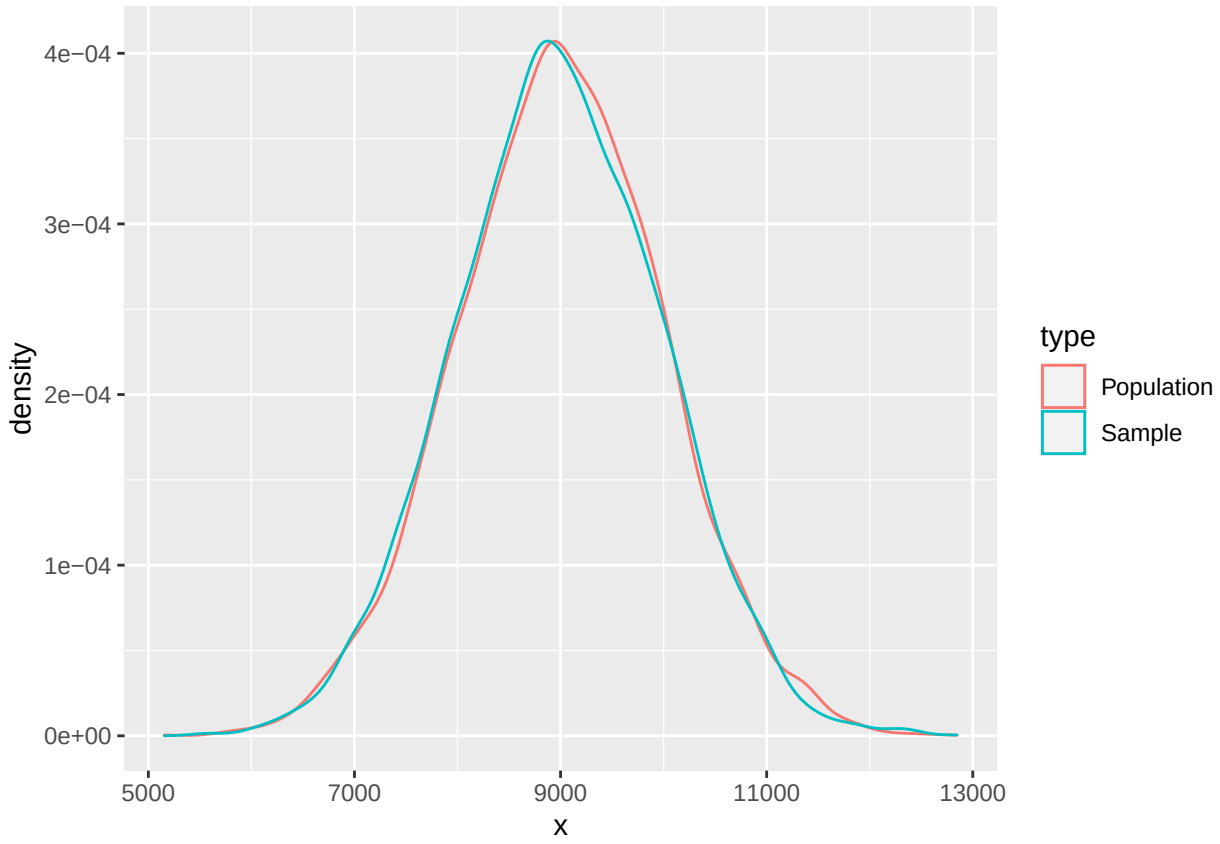
```
 1 - pnorm(q=10500, mean=9000, sd=258.20)
```

```
## [1] 3.13392e-09
```

The probability is so low that is approaching 0. Very unlikely to happen.

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

```
set.seed(123)

data <- as.data.frame(rnorm(10000,mean=9000,sd=1000))


samp <- data %>%
  slice_sample(n=950)

d = data.frame(x = c(data$`rnorm(10000, mean = 9000, sd = 1000)`, samp$`rnorm(10000, mean = 9000, sd = 
               type=rep(c("Population", "Sample"), c(length(data), length(samp))))
ggplot(d) +
  geom_density(aes(x=x, color=type))
```



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

No, with skewed data we cannot properly make inferences such as that. As there would be plenty of extreme values, it would skew any estimate we calculate and end up not being useful on every value in the population.

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.

Increasing the sample size should decrease the standard error. Which should increase our z value, and in the process, our p-value will also decrese.