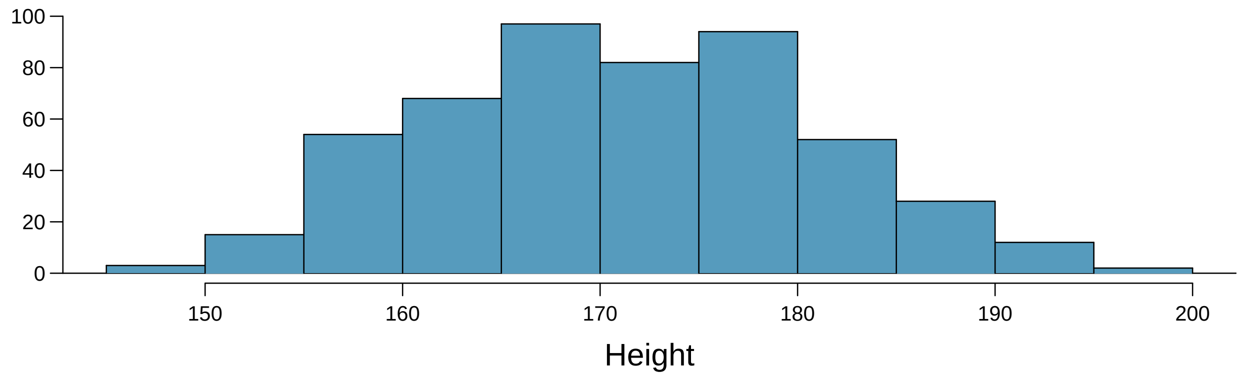
Chapter 5 - Foundations for Inference

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?

The point estimation for the average height of active individuals is 171.14 centimeters:

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

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

The point estimation for the median height of active individuals is 170.30 centimeters:

```
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?

The point estimate for the standard deviation of the heights of active individuals is 9.41 centimeters:

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

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

The point estimate for the IQR of the heights of active individuals is 14 centimeters:

```
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.

A person who is 180 cm tall would not be considered unusually tall since this falls within 1 standard deviation from the mean:

```
(180 - 171.14) / 9.41
```

```
## [1] 0.9415515
```

0.94 is the calculated z-score of this observation, which shows that it is even less than one standard deviation from them mean. A z-score greater than 2 or less than -2 would be considered unusual.

```
(155 - 174.14) / 9.41
```

```
## [1] -2.034006
```

However, a person that is 155 cm tall would be considered unusually short, since $Z = -2.03$. This falls outside of two standard deviations from the mean height of 174.14 centimeters.

(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 not expect the mean and the standard deviation of this new sample to be equal to the ones above. The mean and the standard deviation will most likely be close to the ones above, but due to a little bit of sampling error that naturally will occur between two different samples, they will not be the same.

(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.

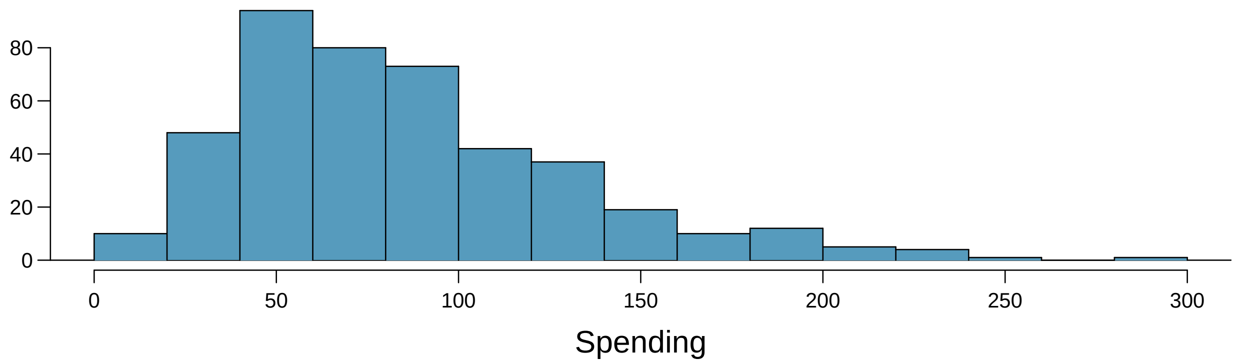
For this, we would use standard error to quantify the variability of the estimate. The calculation for this is:

```
9.41 / sqrt(507)
```

```
## [1] 0.4179128
```

The standard error is 0.418.

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.

This is false. The sample mean always falls within the 95% confidence interval. The 95% confidence interval would have to be placed on the *population*, not the sample of 436 American adults.

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

This is false, although skewness can affect the confidence interval, the sample size is much greater than 30. Therefore, a confidence interval can still be established.

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

This is false, we would need to adjust this to say that 95% of the confidence intervals built for the sample mean would contain the parameter.

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

This is true, this is the correct characterization of the confidence interval.

(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.

This is true, a 90% confidence interval would be narrower than the 95% confidence interval since we would be "casting a smaller net", according to the textbook. We would be less sure that we are catching the point estimate inside of our confidence interval.

(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.

This is false, the calculation below shows that the margin of error of a 95% confidence interval would only be reduced to about half of what it is now. It would have to be about 9 times larger to reach this.

```
# current margin of error
1.96 * (sd(tgSpending$spending) / sqrt(436))
```

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

```
# margin of error with a sample 3 times larger
1.96 * (sd(tgSpending$spending) / sqrt(436 * 3))
```

```
## [1] 2.54325
```

```
# margin of error with a sample 9 times larger - then checked by times 3
(1.96 * (sd(tgSpending$spending) / sqrt(436 * 9))) * 3
```

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

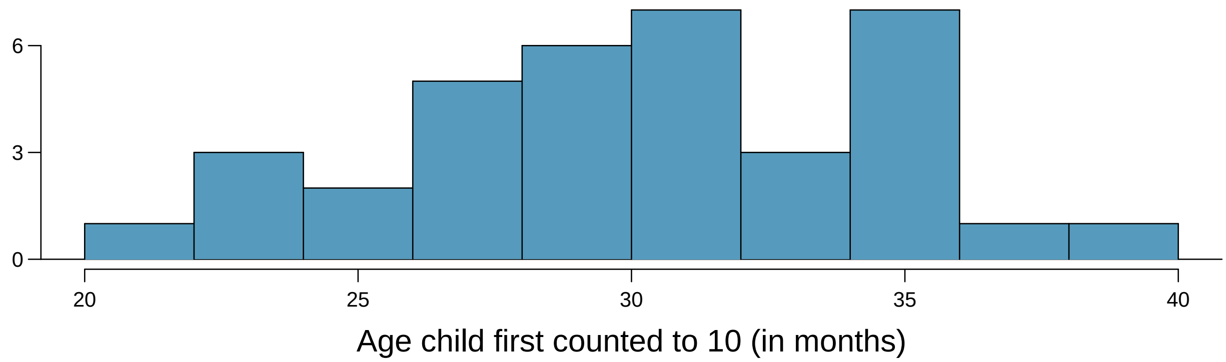
(g) The margin of error is 4.4.

This is true. By dividing the standard deviation of the population by the square root of the sample size, and then multiplying this by the z-score consistent with a 95% confidence interval, I received the same result.

```
1.96 * (sd(tgSpending$spending / sqrt(436)))
```

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

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?

Conditions for inference are satisfied, the sample size is greater than 30 individuals, it is a random sample, and the sample is less than 10% of the population.

(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 first count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

To perform a hypothesis test and evaluate these data, we'll do the following:

Null Hypothesis: The average age at which gifted children first count to 10 is the same as the average age of the general population of children at 32 months.

Alternative Hypothesis: The average age at which gift children first count to 10 is less than the average age of the general populaton of children.

Let's first calculate the z-score of the sample:

```
sample_mean <- 30.69
sample_sd <- 4.31
sample_size <- 36
population_mean <- 32

sample_z_score <- (sample_mean - population_mean) / (sample_sd/sqrt(sample_size))
sample_z_score
```

```
## [1] -1.823666
```

Then, we can find the p-value by the following:

```
p_value <- pnorm(sample_z_score)
p_value
```

```
## [1] 0.0341013
```

$p = 0.034$ **and** $p < 0.10$**, we can reject the null hypothesis in support of the alternative hypothesis.**

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

Since $p = 0.034$ **and** $p < 0.10$**, we can reject the null hypothesis in support of the alternative hypothesis. The average length of time until a gifted child counts to 10 is less than the general population of children, and this difference is not due to chance.**

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

To calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully, we'll have to do the following calculations:

```
point_estimate <- sample_mean
standard_error <- sample_sd / sqrt(36)
z_conf <- qnorm(1 - (1 - 0.90) / 2)

upper_conf <- point_estimate + z_conf * standard_error
lower_conf <- point_estimate - z_conf * standard_error

print(paste0('A 90% confidence interval is (', lower_conf, ', ', upper_conf, ')'))
```

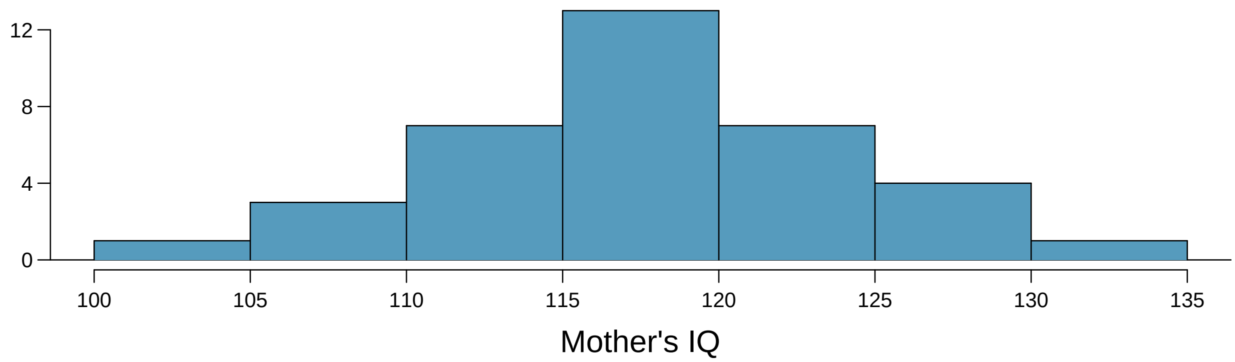
```
## [1] "A 90% confidence interval is (29.5084468113065, 31.8715531886935)"
```

By utilizing the formula $pointestimate \pm 1.65 \times SE$**, I found a 90% confidence interval of (29.51, 31.87).**

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

My results from the hypothesis test and confidence interval agree. Since the general average of 32 months is above the upper bound of the gifted children confidence interval, we can state that it would be an unlikely event that a gifted child would not count to ten prior to 31.87 months after birth.

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.

Similar to the last problem, to perform a hypothesis test and evaluate these data, we'll do the following:

Null Hypothesis: The average IQ of gifted childrens' mothers is the same as the average IQ of mothers in the general population.

Alternative Hypothesis: The average IQ of gifted childrens' mothers is greater than the average IQ of mothers in the general populaton of children.

Again, we'll first calculate the z-score:

```
sample_mean_mom <- 118.2
sample_sd_mom <- 6.5
sample_size_mom <- 36
population_mean_mom <- 100

sample_z_score_mom <- (sample_mean_mom - population_mean_mom) / (sample_sd_mom/sqrt(sample_size_mom))
sample_z_score_mom
```

```
## [1] 16.8
```

With a z-score of 16.8, we can then find the p-value:

```
p_value_mom <- pnorm(-abs(sample_z_score_mom))
p_value_mom
```

```
## [1] 1.22022e-63
```

We can see that $p < 0.001$, which is far less than $p < 0.10$. Therefore, we can reject the null hypothesis.

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

To calculate a 90% confidence interval for the average IQ of mothers of gifted children, we'll have to do the following calculations:

```
point_estimate_mom <- sample_mean_mom
standard_error_mom <- sample_sd_mom / sqrt(36)
z_conf_mom <- qnorm(1 - (1 - 0.90) / 2)

upper_conf_mom <- point_estimate_mom + z_conf_mom * standard_error_mom
lower_conf_mom <- point_estimate_mom - z_conf_mom * standard_error_mom

print(paste0('A 90% confidence interval is (', lower_conf_mom, ', ', upper_conf_mom, ')'))
```

```
## [1] "A 90% confidence interval is (116.418075237469, 119.981924762531)"
```

By utilizing the formula $pointestimate \pm 1.65 \times SE$, I found a 90% confidence interval of (116.42, 119.98).

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

My results from the hypothesis test and confidence interval agree. Since the average IQ of mothers of gifted children falls within the confidence interval, we can state that these two agree.

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 sample proportions from a population. As the sample size increases, the distribution spread narrows, the shape increasingly resembles a normal distribution, and the center becomes closer and closer to the population mean.

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?

To calculate this, we need to first calculate the z-score:

```
mean <- 9000
sd <- 1000
sample <- 10500

z_score_bulb <- (10500 - 9000) / 1000
z_score_bulb
```

```
## [1] 1.5
```

Then to calculate the probability, since it's nearly a normal distribution, we can do the following:

```
p_value_bulb <- pnorm(-abs(z_score_bulb))
p_value_bulb
```

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

The probability that a randomly chosen light bulb lasts more than 10,500 hours is about 0.067.

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

Since the population is nearly normal, the distribution of the mean lifespan of 15 light bulbs will also be normal with a center around 9,000 hours. The standard error can be calculated $SE = \frac{1000}{\sqrt{15}}$**.**

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

To do this, we need to again calculate the z-score first:

```
sample_mean_bulb <- 10500
sample_size_bulbs <- 15

z_score_samp_bulb <- (sample_mean_bulb - mean) / (sd/sqrt(sample_size_bulbs))
z_score_samp_bulb
```

```
## [1] 5.809475
```

Then, with a z-score of 5.81, we can calculate the probability:

```
p_value_samp_bulb <- pnorm(-abs(z_score_samp_bulb))
p_value_samp_bulb
```

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

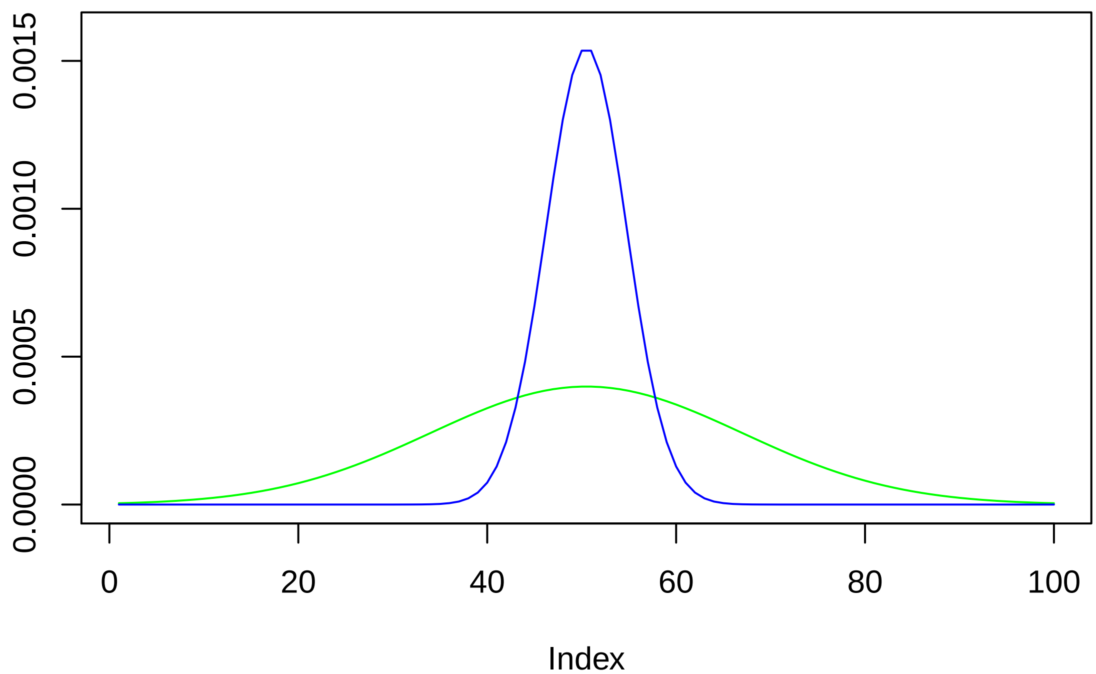
The probability that the mean lifespan of 15 randomly chose light bulbs is more than 10,500 is very close to zero - an extremely unlikely event.

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

```
xfit<-seq(min(mean - 3 * sd),max(mean + 3 * sd),length=100)

# we can plot the population curve in green
plot(dnorm(xfit, mean = mean, sd), type = "l", col = "green", ylim=c(0,0.0016), ylab = "")

# the sample curve will be in blue
lines(dnorm(xfit, mean = mean, sd = sd/sqrt(sample_size_bulbs)), col = "blue")
```



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

We should not estimate the probabilities from parts (a) and (c) if the lifespans of the light bulbs had a skewed distribution. In part (a), we would need to use the skewed distribution to calculate the probabilities, and in part (c), since the sample size is less that 50, we wouldn't be able to use the Central Limit Theorem to assume a normal distribution.

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.

If your sample size increases, such as it does in this example, then your p-value will decrease. The reason this will occur is that the p-value is dependent on the z-value, as we could see from previous calculations in this problem set. An increase in the sample size would cause a decrease in the standard error calculation ($SE = \sigma/\sqrt{n}$), since the denominator would be larger. Therefore, this would affect the z-score value, causing an increase, which would then cause a decrease in p-value.