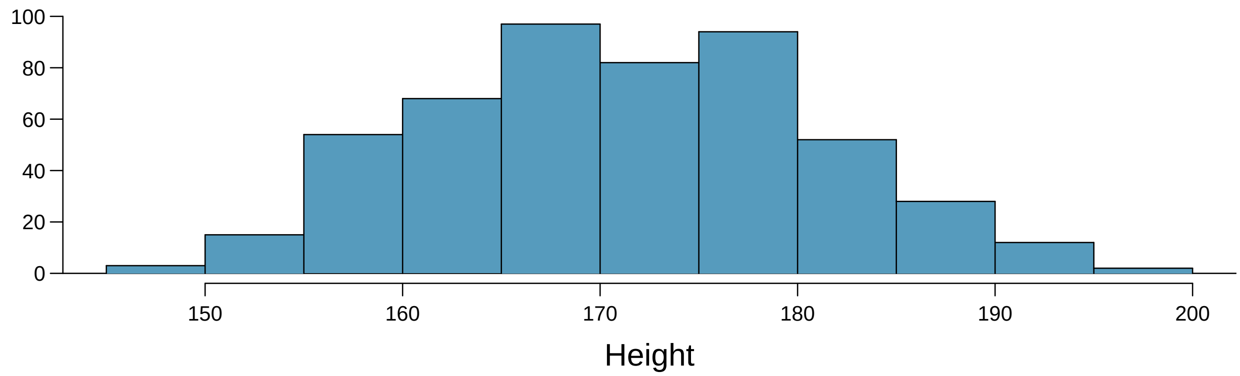
Chapter 5 - Foundations for Inference

Bryan Persaud

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?
(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?
(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.
(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.
(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 $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

(a)

```r
summary(bdims$hgt)
```

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

The point estimate for the average height is 171.1. The point estimate for the median is 170.3.

(b)

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

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

```r
# Calculate IQR by subtracting Q3 from Q1
177.8 - 163.8
```

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

The point estimate for the standard deviation of the heights is 9.4. The point estimate for the IQR is 14.

(c)

```r
# Calculate z-score for tall height
(180 - 171.1) / 9.4
```

```
## [1] 0.9468085
```

```r
# Calculate z-score for short height
(155 - 171.1) / 9.4
```

```
## [1] -1.712766
```

A person who is 180cm is not considered unusually tall. This is because they are only 0.95 above the standard deviation. A person who is 155cm tall is not considered unusually short. This is because they are only 1.71 below the standard deviation. Both are within two standard deviation from the mean.

(d)

I would not expect the mean and standard deviation to be the same as the sample above, if the researchers take another random sample, because since another random sample was taken the values for each individual in the data would be somewhat different.

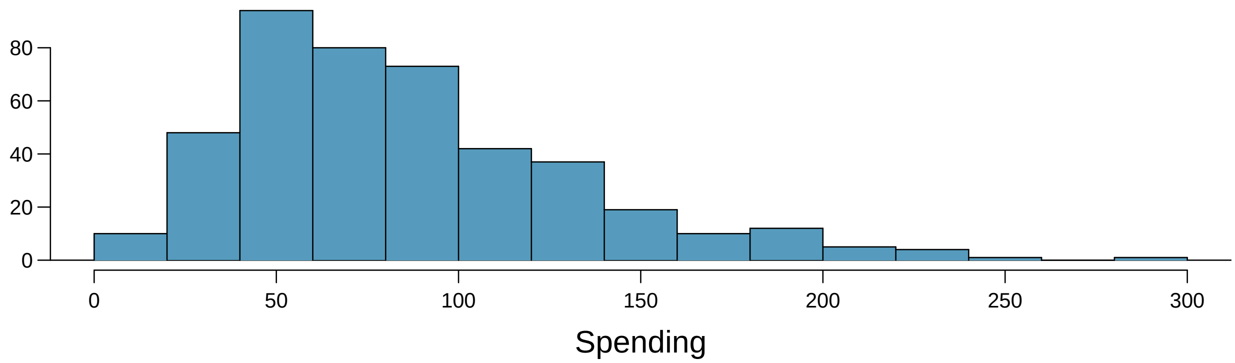
(e)

The measure we would use to quantify the variability is the standard error.

```r
9.4 / sqrt(507)
```

```
## [1] 0.4174687
```

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.
(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.
(c) 95% of random samples have a sample mean 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.
(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.
(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.
(g) The margin of error is 4.4.

(a)

This is false because the confidence interval is used for the population.

(b)

This is false because the sample size is 436 and it is large enough to not worry about the right skew.

(c)

This is false because the confidence interval is used for the population mean and not the sample mean.

(d)

This is true because the confidence interval is being used for the population mean.

(e)

This is true because the lower the confidence level, the narrower the interval gets.

(f)

This is false because to decrease the margin of error by a third would also mean decreasing the standard error by a third. To do this you would need a sample size that is nine times larger.

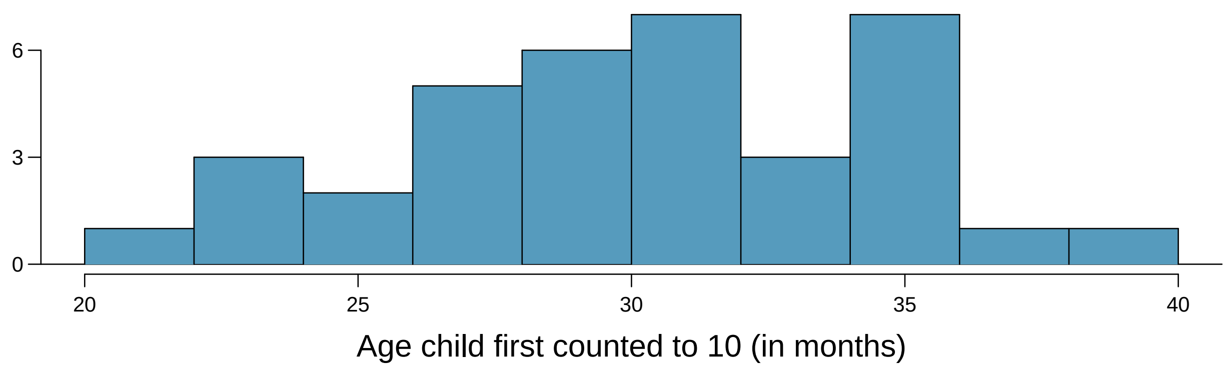
(g)

This is true because to calculate the margin of error you take the upper bound of the confidence interval and subtract it from the mean.

```r
89.11 - 84.71
```

```
## [1] 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?
(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.
(c) Interpret the p-value in context of the hypothesis test and the data.
(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.
(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

(a)

The conditions for inference are satisfied since a random sample was taken,independence is satisfied, the sample size is over 30, and the distribution is not extremely skewed.

(b)

```
mean <- 30.69
n <- 36
sd <- 4.31
x <- 32
se <- sd / sqrt(n)
Z_score <- (mean - x) / se
p <- pnorm(Z_score, mean = 0, sd = 1)
p
```

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

(c)

The p-value in context of the hypothesis and the data is since the p-value is 0.03 and is lower than the significance level of 0.10, we reject the null hypothesis that the average age at which gifted children first count to 10 is the same as the general population and accept the alternative hypothesis which is the average age at which gifted children count to 10 is less than the general population.

(d)

```
lower <- mean - 1.645 * se
upper <- mean + 1.645 * se
c(lower, upper)
```

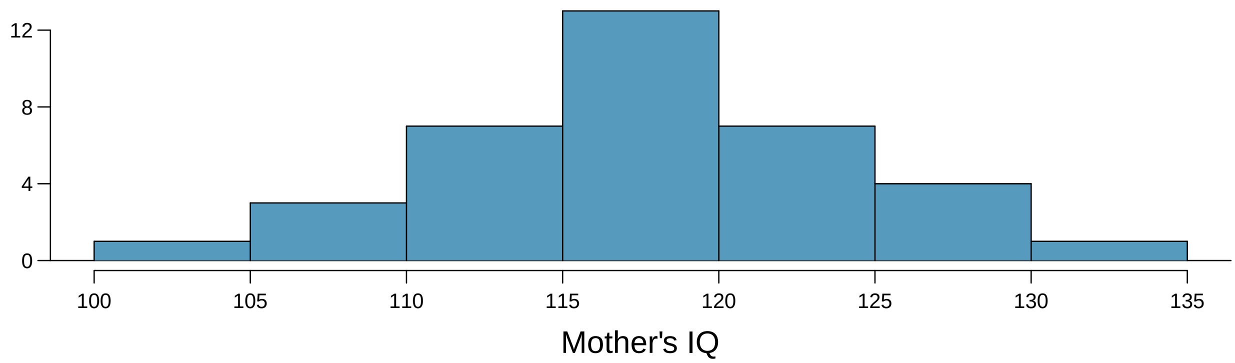
```
## [1] 29.50834 31.87166
```

The 90% confidence interval for the average age at which gifted children first count to 10 successfully is (29.51, 31.87).

(e)

My results from the hypothesis test and the confidence interval do agree because the hyothesis test tells us to reject the null hypothesis that the average age is 32 months and the 90% confidence interval does not include 32 months.

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) Performahypothesistesttoevaluateifthesedataprovideconvincingevidencethattheaverage 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.
(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.
(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

(a)

```
mean_2 <- 118.2
n_2 <- 36
sd_2 <- 6.5
x_2 <- 100
se_2 <- sd_2 / sqrt(n_2)
Z_score_2 <- (mean_2 - x_2) / se_2
p_2 <- (1 - pnorm(Z_score_2, mean = 0, sd = 1)) * 2
p_2
```

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

(b)

```
lower_2 <- mean_2 - 1.645 * se_2
upper_2 <- mean_2 + 1.645 * se_2
c(lower_2, upper_2)
```

```
## [1] 116.4179 119.9821
```

The 90% confidence interval for the average IQ of mothers of gifted children is (116.42, 119.98).

(c)

My results from the hypothesis test and the confidence interval do agree because the hypothesis test tells us to reject the null hypothesis that the average IQ of mother is 100 and the 90% confidence interval doesn not include 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 collection of means taken from multiple samples of the population. AS the sample size increaes, the shape becomes that of a normal distribution, the center becomes closer to the true population mean, and the spread starts to decrease.

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?
(b) Describe the distribution of the mean lifespan of 15 light bulbs.
(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?
(d) Sketch the two distributions (population and sampling) on the same scale.
(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

(a)

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

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

The probability that a randomly chosen light bulb lasts more than 10,500 hours is 0.067 or 6.7%.

(b)

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

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

The distribution of the mean lifespan of 15 light bulbs is nearly normal with a mean of 9000 and a standard deviation of 258.20.

(c)

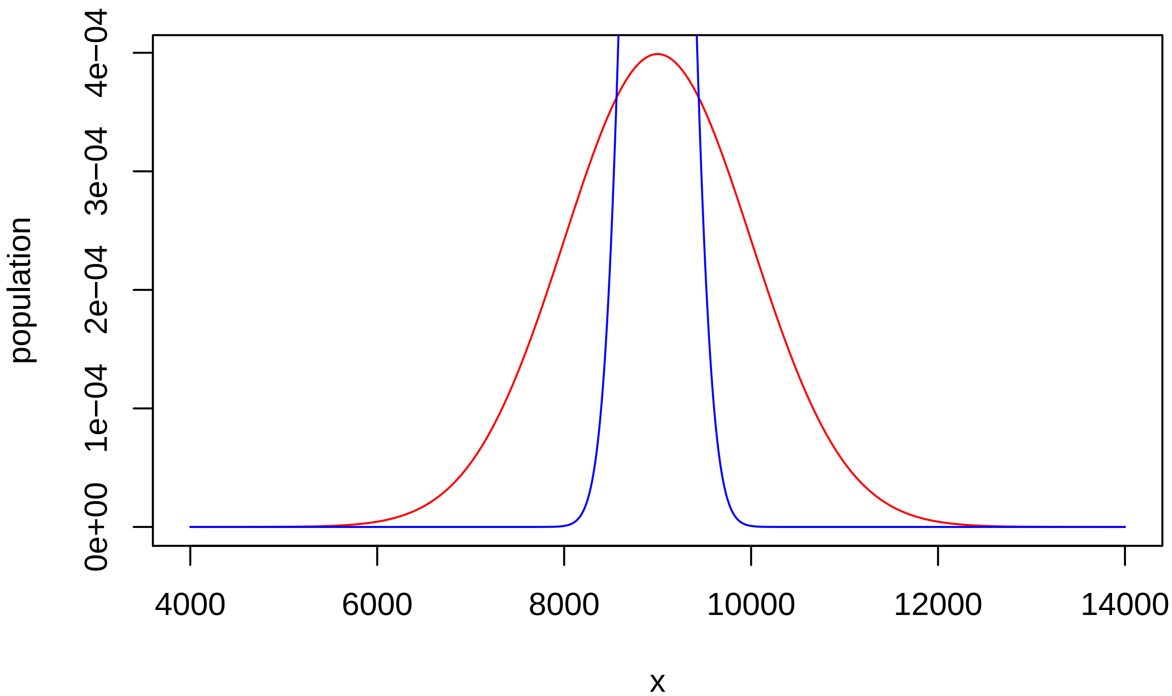
```
round(1 - pnorm(q = 10500, mean = 9000, sd = 258.20), 3)
```

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

The probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours is basically 0%.

(d)

```r
x <- 4000:14000
population <- dnorm(x, mean = 9000, sd = 1000)
sampling <- dnorm(x, mean = 9000, sd = 258.20)
plot(x, population, type = "l", col = "red")
lines(x, sampling, col = "blue")
```



(e)

You could not estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution because you would need a normal distribution and for part (c) the sanple size is too small.

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.

The p-value will decrease since you are increasing the sample size.