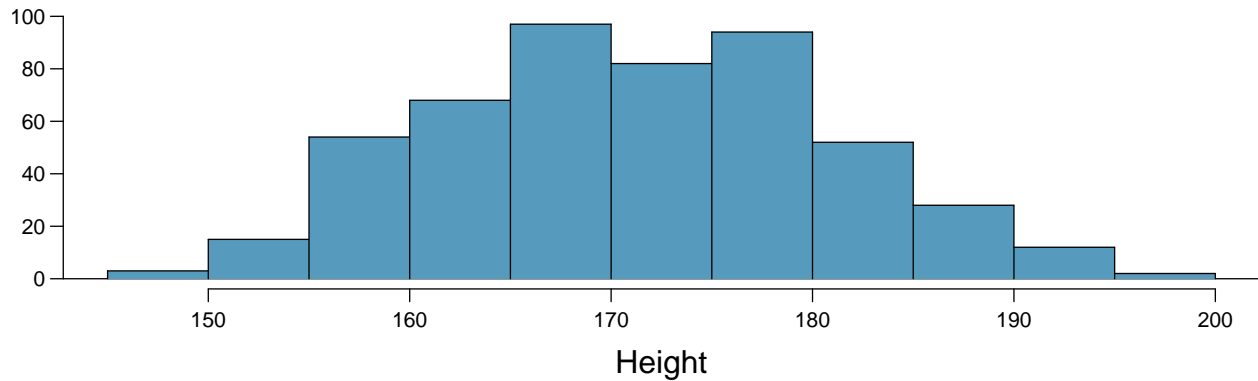# Chapter 5 - Foundations for Inference

Leticia Salazar

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 estimated average height of active individuals is 171.14 and the median is 170.3

```
#Calculate Mean for heights
mean_hgt <- mean(bdims$hgt)
mean_hgt
```

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

```
#Calculate Median for heights
median_hgt <- median(bdims$hgt)
median_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 and the IQR is 14.

```
#Calculate standard deviation for heights
sd_hgt <- sd(bdims$hgt)
sd_hgt
```

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

```
#Calculate IQR for heights
iqr_hgt <- IQR(bdims$hgt, na.rm = TRUE)
iqr_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.

Based on the z-score values below, a person who is 180 cm tall falls between the first (-1, 1) standard deviation, meaning it is not considered to be unusually tall. The person who is 155cm falls further than -1, 1 standard deviations, therefore they are unusually short based on these results.

```
#Calculate z-score for 180cm
zscore = (180 - mean_hgt) / sd_hgt
zscore
```

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

```
#Calculate z-score for 155cm
zscore2 = (155 - mean_hgt) / sd_hgt
zscore2
```

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

(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.

With a new random sample, I would expect the mean and standard deviation will be different than the ones given above since it is a new sample. I do however, expect the mean and standard deviation to be similar because the random sample will be taken from the physically active individuals.

(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.

We would use standard error which is 0.42.

```
#Calculate standard error
sd_2 <- sd_hgt / sqrt(nrow(bdims))
sd_2
```

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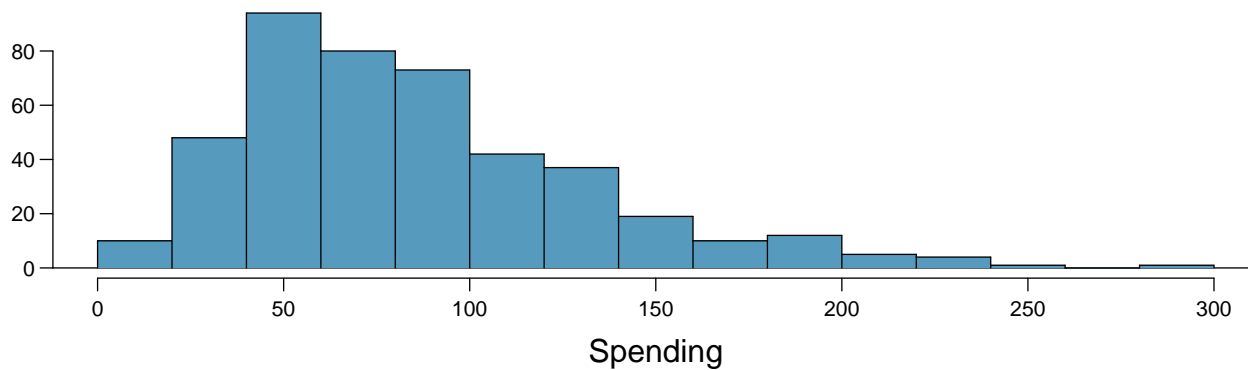
(a) We are 95% confident that the average spending of these 436 American adults is between $80.31 and $89.11.

True, the 95% confident interval is based on the sample size of 436 American adults matching the dollar amounts.

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

False, the skewness of the sample does not play a role in the validity of the confidence internal.

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

False, 95% only applies to the confidence interval of the random sample of 436 American adults and not any random samples.

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

False, the 95% confident interval only applies to the 436 American adults, not all American adults.

(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, if we do not need to be as sure about our estimate then we can have a narrower 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.

False, since are looking into the margin of error, to get a third of what it is now we would need a sample 9 times larger.

(g) The margin of error is 4.4.

True, when calculated, the margin error is 4.4.

```r
#Margin of error
me <- (89.11 - 80.31 ) / 2
me
```

```
## [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.

Age child first counted to 10 (in months)
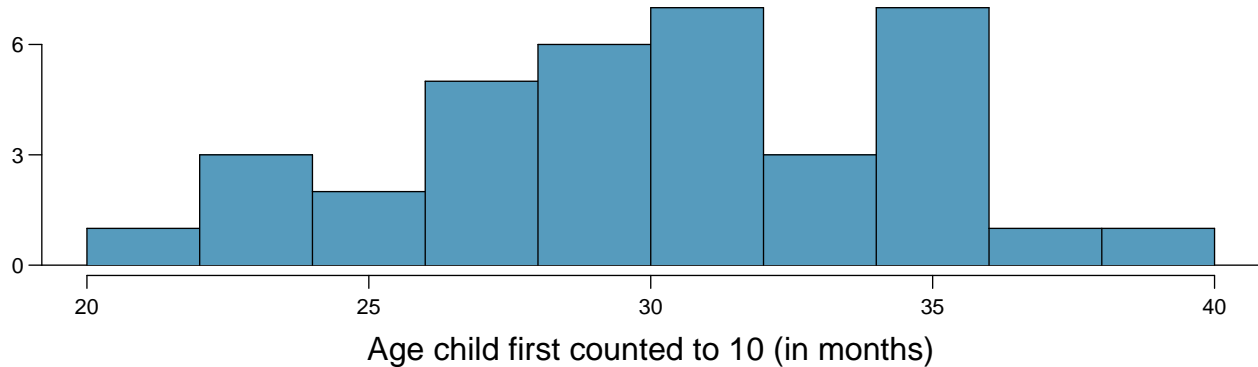
n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

Yes, given that the sample was taken from schools in a large city, then the conditions for inference are 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.

Ho: The average age of a gifted child to successfully count to 10 is 32 months.

Ha: The average age of a gifted child to successfully count to 10 is less than 32 months.

```
mean_32 <- 30.69
sd_32 <- 4.31
size_32 <- 36
population_32 <- 32

#Find standard error
#standard error = standard deviation / sqrt (size)
se_32 <- sd_32 / sqrt(size_32)
se_32
```

```
## [1] 0.7183333
```

```
#Find z-score
#z-score = (mean - population) / standard error
zs_32 <- (mean_32 - population_32) / se_32
zs_32
```

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

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

We can reject our null hypothesis since the value of p is lower than .10

```
#Find p-value
pvalue_32 <- round((2*pnorm(zs_32, mean = 0, sd = 1)), 2)
pvalue_32
```

```
## [1] 0.07
```

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

A 90% confidence interval for the average age at which gifted children first count to 10 successfully is (20.86, 23.23).

```
#90% confidence interval
lower <- (mean_32 - 1.645) * se_32
lower
```

```
## [1] 20.86399
```

```
upper <- (mean_32 + 1.645) * se_32
upper
```

```
## [1] 23.22731
```

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

Since both lower and upper 90% confidence interval are less than 32, then the hypothesis test do agree with the 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.
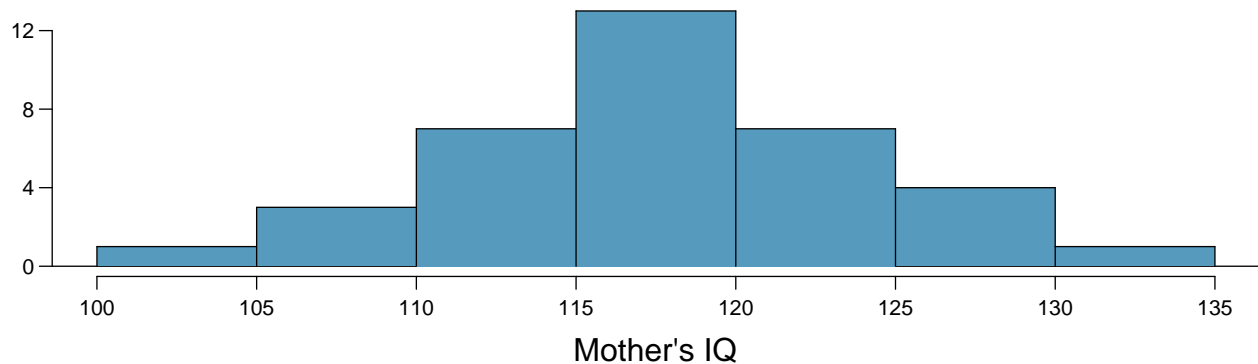
n	36
min	101
mean	118.2
sd	6.5
max	131

(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.

Ho: The average IQ of a mother of a gifted child is 100.

Ha: The average IQ of a mother of a gifted child is not equal to 100.

```
mean_moms <- 118.2
sd_moms <- 6.5
size_moms <- 36
population_moms <- 100

#Find standard error
#standard error = standard deviation / sqrt (size)
se_moms <- sd_moms / sqrt(size_moms)
se_moms
```

```
## [1] 1.083333
```

```
#Find z-score
#z-score = (mean - population) / standard error
zs_moms <- (mean_moms - population_moms) / se_moms
zs_moms
```

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

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

The 90% confidence interval for the average IQ of mothers of gifted children is (126.27, 129.83).

```
#Find p-value
pvalue_moms <- round((2*pnorm(zs_moms, mean = 0, sd = 1)), 2)
pvalue_moms
```

```
## [1] 2
```

```
#90% confidence interval
lower <- (mean_moms - 1.645) * se_moms
lower
```

```
## [1] 126.2679
```

```
upper <- (mean_moms + 1.645) * se_moms
upper
```

```
## [1] 129.8321
```

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

No, my hypothesis test and the confidence interval do not agree. The results for the confidence interval are above the hypothesis testing the average IQ of mothers of gifted children to be 100. Therefore we can reject our null hypothesis and accept the alternate hypothesis.

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.

Sampling distribution of the mean is the probability distribution that are obtained from a large sample size within a population. As the sample size increases, the sample distribution gets closer to the normal distribution. The shape is more symmetrical, the spread becomes narrower as well as the center.

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?

The probability that a randomly chosen light bulb lasts more than 10,500 hours is 6.68%

```
#Probability that a randomly chosen light bulb last more than 10,500 hrs
mean_lb <- 9000
sd_lb <- 1000

P <- 1 - pnorm(10500, 9000, 1000)
P
```

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

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

Based on the given data we know that the distribution is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours. With this small sample size of 15, the distribution can vary in shape.

(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 3.133e-09 or almost zero.

```
size_lb <- 15
x <- 10500

#Standard error
sample15 <- sd_lb / sqrt(size_lb)
sample15
```
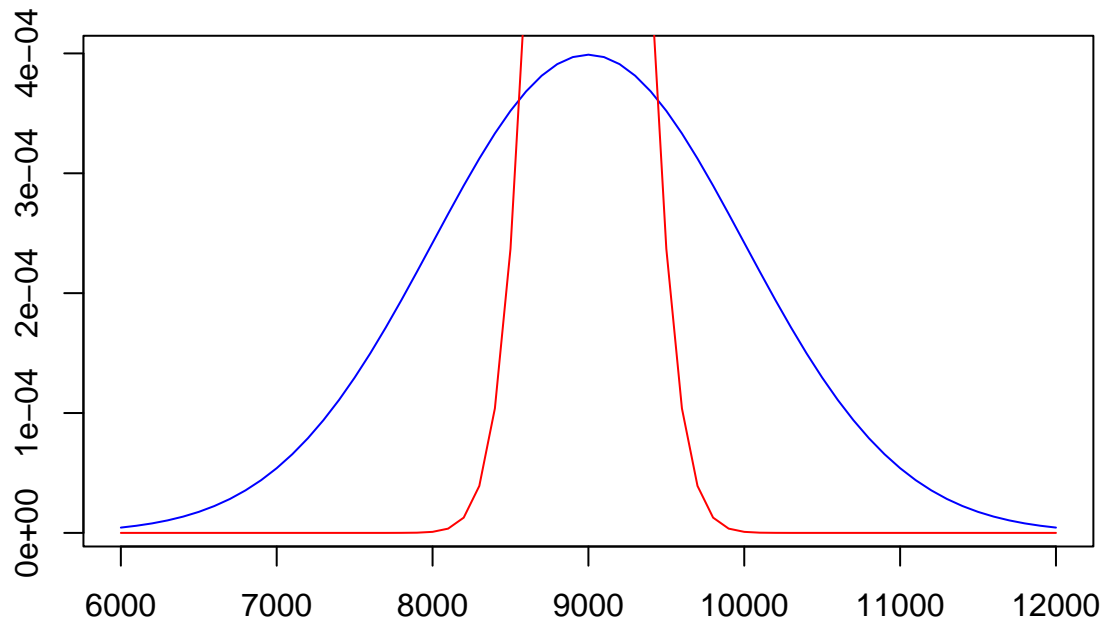
```
## [1] 258.1989
```

```
#Probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10500 hours
P15 <- 1 - pnorm(x, mean_lb, sample15)
P15
```

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

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

```
s <- seq(6000, 12000, 100)

plot(s, dnorm(s, mean_lb, sd_lb), type = "l", xlab = 'Population vs Sampling Lightbulbs', ylab = '', co
lines(s, dnorm(s, mean_lb, sample15), col = "red")
```

Population vs Sampling Lightbulbs

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

We would not be able to estimate the probabilities from parts (a) and (c) if the the lifespans of light bulbs had a skewed distribution because the sample size is not big enough.

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 as the same size increases as seen below:

```
sd <-   1
mean <- 0

#standard error = standard deviation / sqrt (size)
se50 <- sd / sqrt(50)
se50
```

```
## [1] 0.1414214
```

```
se500<- sd / sqrt(500)
se500
```

```
## [1] 0.04472136
```

```
#Z-score of 50
#z-score = (mean - population) / standard error
zscore50 <- (mean - 50) / se50
zscore50
```

```
## [1] -353.5534
```

```
#Z-score of 500
#z-score = (mean - population) / standard error
zscore500 <- (mean - 500) / se500
zscore500
```

```
## [1] -11180.34
```