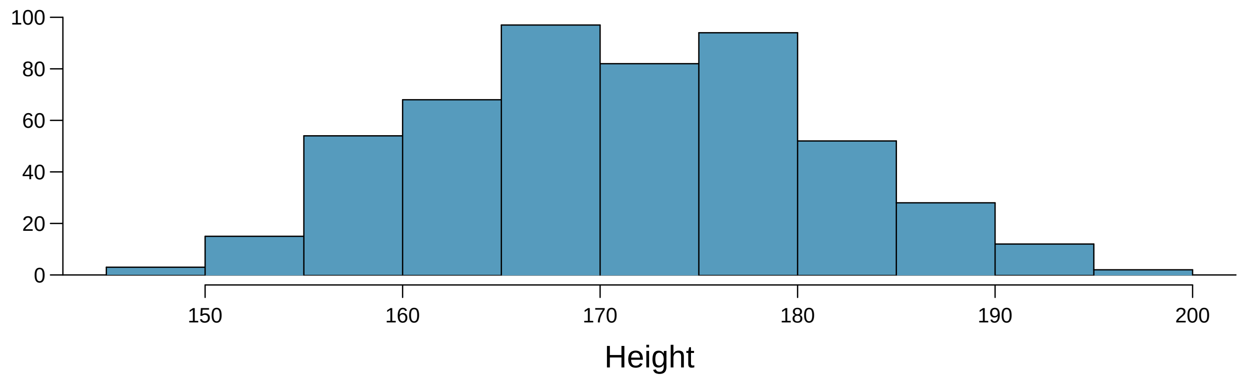
Chapter 5 - Foundations for Inference

Gabriel Campos

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?
Point estimates for *mean height*= 171.1 and *median height*=170.3

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

(b)

What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?
Point estimate for *standard deviation* = 9.407205 and for *IQR* = 14

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

```
## [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.
The range within 1 standard deviation of the mean is 161.6928 through 180.5072, therefore 180 is not unusually tall considering it is within that range. However 155cm, which is outside of 1 standard deviation from the mean can be considered unusually short.

(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.
The data collected showed the sample is not exactly normally distributed but nearly. This is supported, when noting the mean and median are not equivalent. Considering this, I would NOT expect the mean and standard deviation of the new sample to be identical, but it would be relatively close

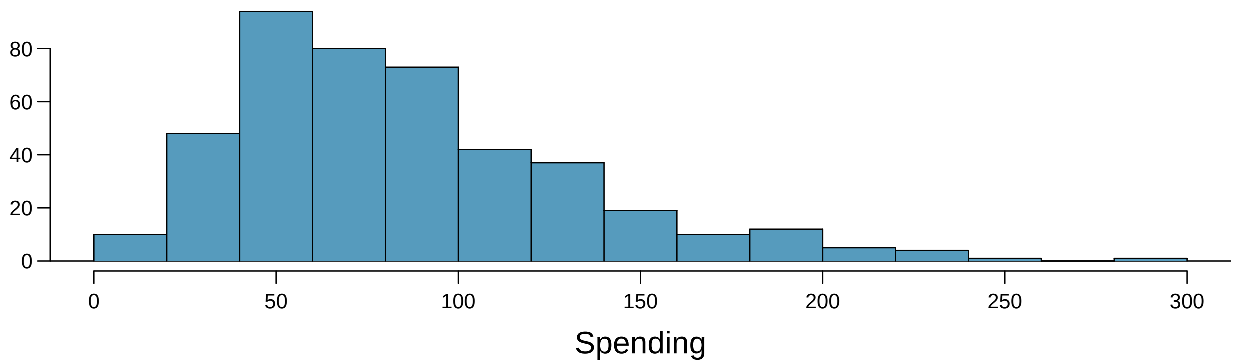
(e)

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.
The measure to quantify the variability is Standard Error. SE is represented by $SD_x = \frac{\sigma}{\sqrt{n}}$) where
σ = 9.407205 and n = 507

```
## [1] "ANSWER: 0.417788650505625"
```

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. The sample mean is always in the confidence interval. We are 95% confident that the average spending of the POPULATION falls between \$80.31 and \$89.11

(b)

This confidence interval is not valid since the distribution of spending in the sample is right skewed.
FALSE. Skewness would be a factor if the sample set were smaller. With a sample size of 436 or specifically 30 $\geq$ n where n=436 we can consider the results of this sample and ignore that it is right skewed.

(c)

95% of random samples have a sample mean between \$80.31 and \$89.11.
FALSE. The confidence interval addresses the population not sample mean.

(d)

We are 95% confident that the average spending of all American adults is between \$80.31 and \$89.11.
True. This is both the definition and purpose of a confidence interval of 95%.

(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. Confidence intervals become narrower as the confidence level decreases.

(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. Standard Error formula is $\frac{\sigma}{\sqrt{n}}$. In order to reduce the margin error to a third we would need n= n x 3^2 or just n to be 9 times greater.

(g)

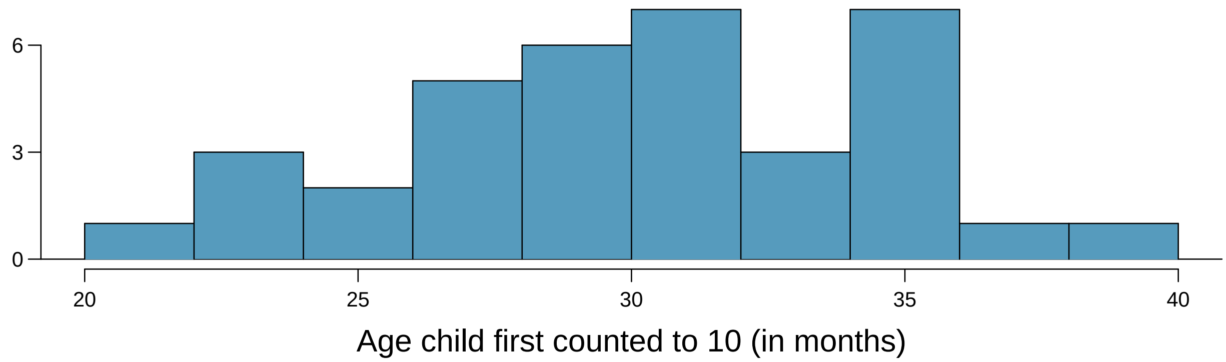
The margin of error is 4.4.

Margin of error can be calculated by subtracting the upper and lower confidence intervals and dividing by 2. The calculation verifies 4.4 is true.

```
## [1] 4.4
```

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?
Random sample with n $\geq$ 30 or 36 $\geq$ 30, each child is independent of each other, data is not heavily skewed, so yes conditions for inference is 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.

H_0: **32 months to count to 10 is the mean age**
H_A: **Mean does not equal 32 months**
Conditions: **Reject if p-value is less that 10%**

```
## [1] "Z-score equals: -1.82366589327146"
```

```
## [1] "P-value equals: 0.0341012909909363"
```

(c)

Interpret the p-value in context of the hypothesis test and the data.
P-value = 0.03410129 < .10 so we can reject the null hypothesis
The mean age for gifted children to learn to count to 10 is not 32 months

(d)

Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.
Z-score of 90% confidence is 1.645
intervals = $\frac{30.69 \pm 1.645 * 4.31}{\sqrt{(36)}}$

```
## [1] "upper interval is:31.8680666666667"
```

```
## [1] "lower interval is:29.5119333333333"
```

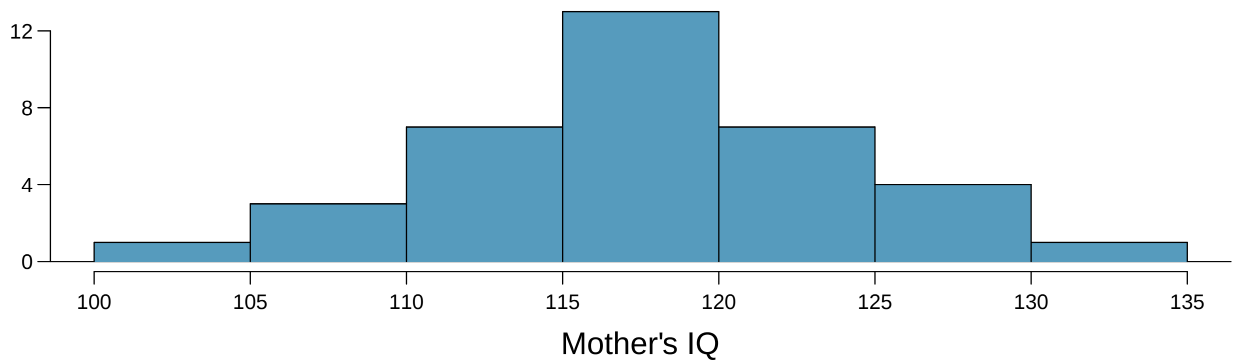
(e)

Do your results from the hypothesis test and the confidence interval agree? Explain.
Yes. 32 is outside of our CI, which supports the rejection of our null hypothesis

---

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 this data provides 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.

H_0*:* **mean IQ = 100 IQ**
H_A*:* **mean IQ $\neq$ 100 IQ**
Conditions: **Reject if p-value is less that 10%**
Z-score = $\frac{118.2-100}{SE}$
p-value = 1-pnorm(Z-score)

```
## [1] "We should reject hypothesis since the p-value is less than 0"
```

(b)

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

```
## [1] "Lower CI:  116.417916666667"
```

```
## [1] "Upper CI: 119.982083333333"
```

(c)

Do your results from the hypothesis test and the confidence interval agree? Explain.
100 does not reside in the intervals so they do not contradict each other.

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 distribution of the mean from various samples (say size n) out of a population. Symmetry, median and mean come closer to a normal distribution as you increase size n.

CFLB

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.

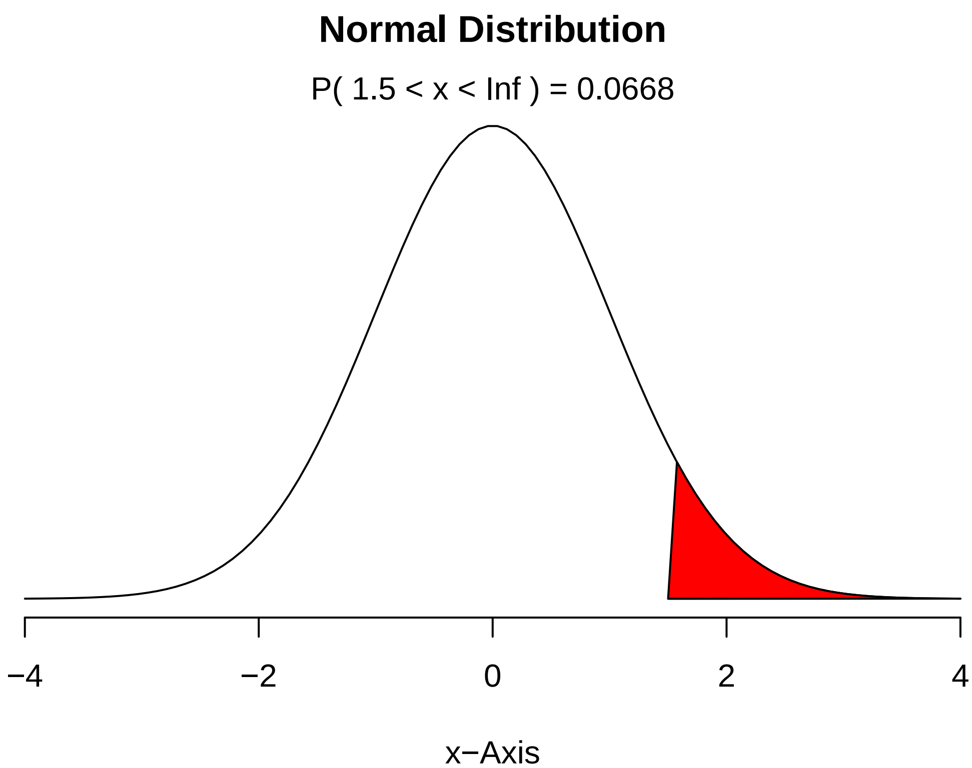
(a)

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

```
## [1] "Z-score: 1.5"
```

```
## [1] "Probability equals to 6.68%"
```

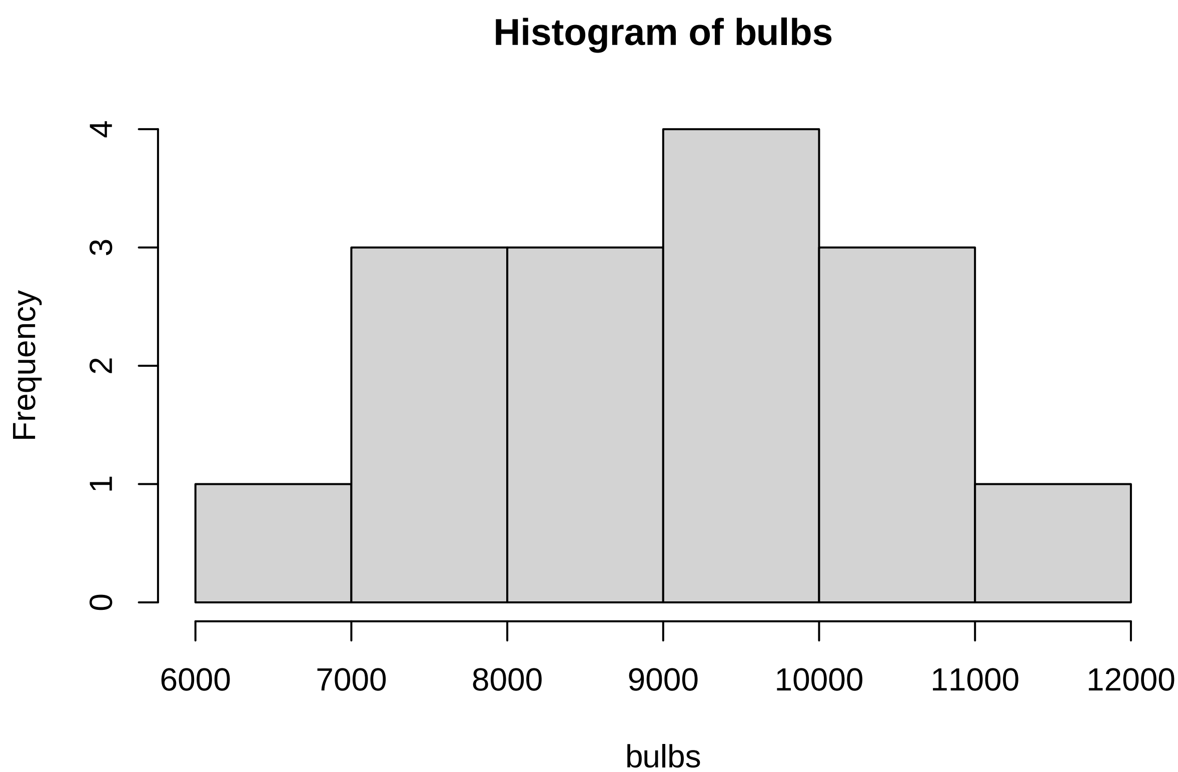
Added the normalPlot to verify further



(b)

Describe the distribution of the mean lifespan of 15 light bulbs.
Sample size is too small and it is skewing the data as seen by the histogram.This data should be *'nearly'* normal as per the description, but the size makes it difficult to verify that.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6866    8014    9220    9091    9981   11009
```



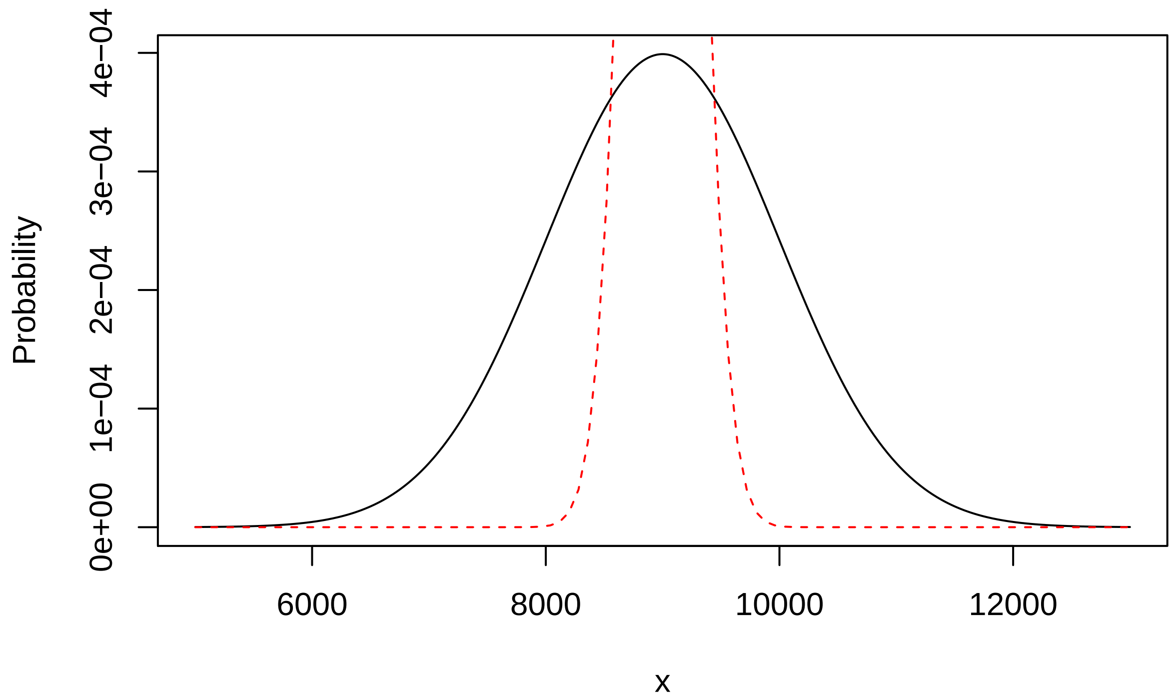
(c)

What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?
Z-score=$\frac{10{,}500-9{,}000}{\frac{1000}{sqrt(15)}}$ **P-value:**

```
## [1] "Probability equals to 0.999999996866548 so basically 0% chance"
```

(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?
For parts (a) and (c) without a nearly normal distribution, we cannot make estimates and (c) specifically needs a larger sample size.

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 impact n has on the calculation for SE $\frac{sd}{\sqrt(n)}$, means Z-score $\frac{x-mean}{SE}$ is impacted because of the standard error. With that said increasing the increase of n from 50 to 500 will increase the Z-score and therefore increase the p-value.