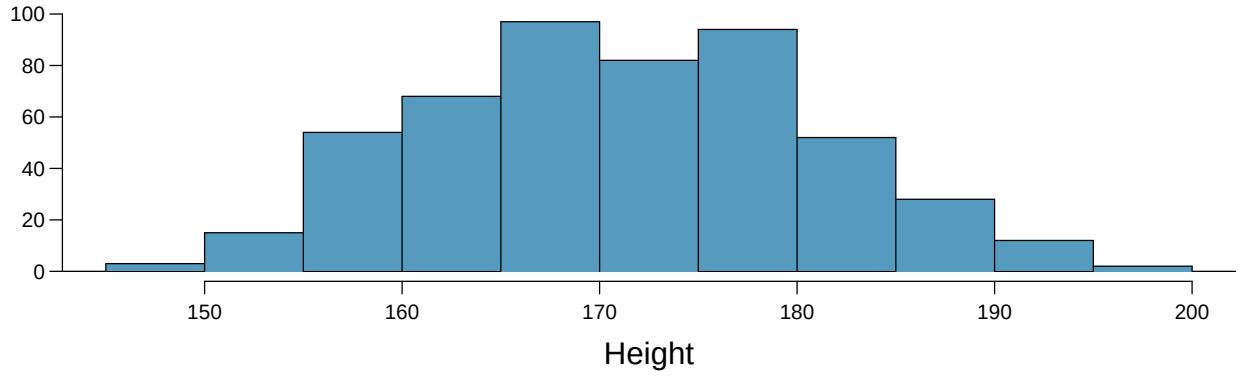
Chapter 5 - Foundations for Inference

David Simbandumwe

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.




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

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

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

- **mean 171.14**
- **median 170.3**

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR? -**SD 9.4** -**IQR 14**

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

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

```
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.

```r
iqr_h <- IQR(bdims$hgt)
q1 <- quantile(bdims$hgt, .25)
q3 <- quantile(bdims$hgt, .75)

c(q1 - 1.5 * iqr_h,q3 + 1.5 * iqr_h)
```

```
##   25%   75%
## 142.8 198.8
```

-180cm is within the range and is not an outlier -155cm is within the range and is not an outlier

(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.

-no there will likely be some variation from the random sampling of a population

(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.

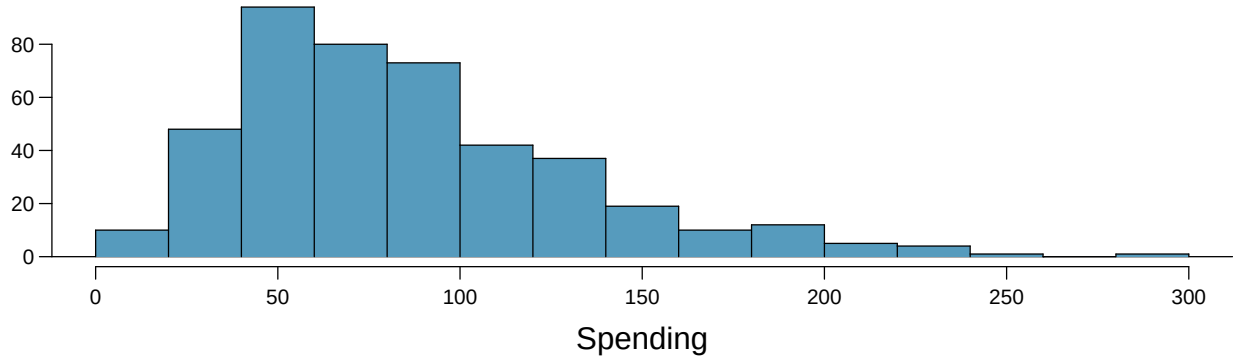
-standard deviation -9.4

```r
sd(bdims$hgt) / sqrt(507)
```

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

The standard deviation of the sample is 0.4178

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 100% of the average spend 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. **-false - the sample is greater that 30 and large enough to model a close to normal standard distribution**

(c) 95% of random samples have a sample mean between $80.31 and $89.11. **-false - we cannot assume that exactly 95% of the samples will have the random samples**

(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11. **-true - based on the definition of 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. **-true - the more confidence we want in the accuracy of the estimate for the population the larger the confidence interval**

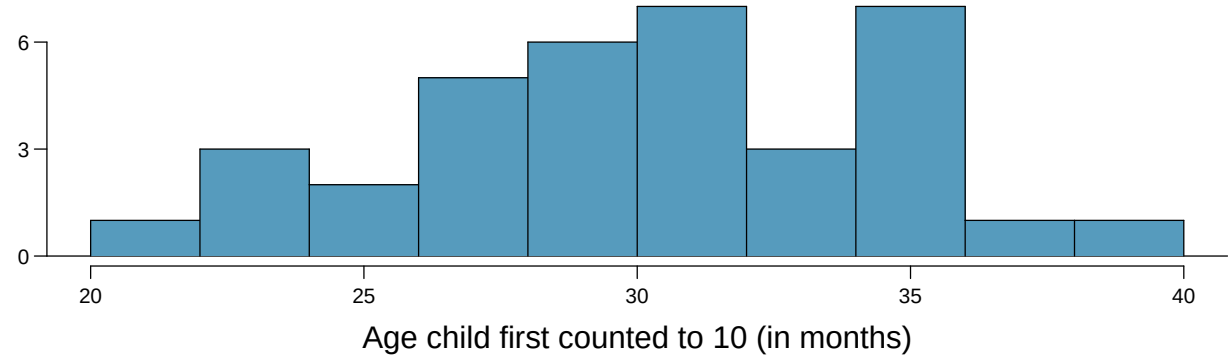
(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger. **-false - we would need to increase our sample size by 9**

(g) The margin of error is 4.4. **-true**

```
(89.11 - 80.31)/2
```

```
## [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?

- **yes - randomized method was used to gather sample, observations are independent, distribution is not overly sckewed**

(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.

- **H0 - average age of gifted children counting to 10 is the same / more then the population in general**
- **HA - average age of gifted children counting to 10 is less than the population in general**
- **since 32 is not within the 90% confidence interval we will reject the null hypothesis in favor of the alternate hypothesis**

```
t.test(gifted$count, alternative = "less", mu = 32, conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  gifted$count
## t = -1.8154, df = 35, p-value = 0.03902
## alternative hypothesis: true mean is less than 32
## 90 percent confidence interval:
##     -Inf 31.6338
## sample estimates:
## mean of x
##  30.69444
```

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

- **The p value of 0.039 is less than 0.10 so we do not have enough evidence to support the null hypothesis**

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

- **The 90% confidence interval for the average age when a gifted child counts to 10 is between 29.51 and 31.88**

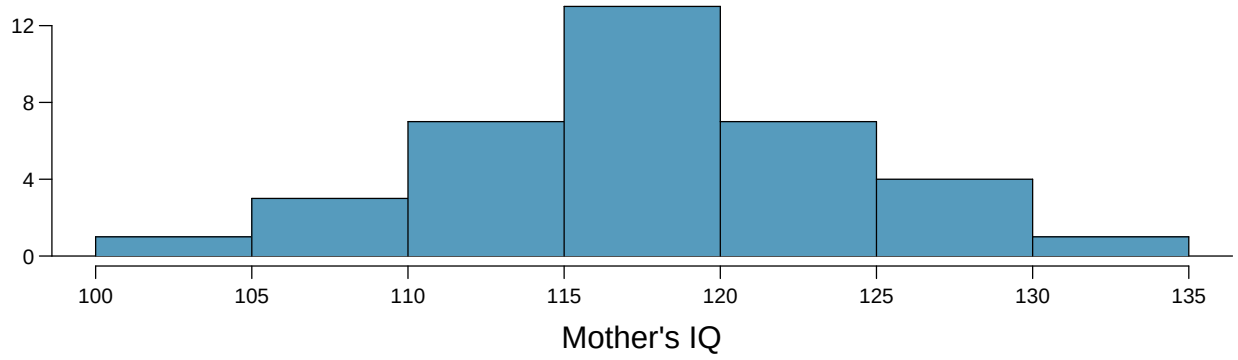
```
sig_level <- .9
z_score <- qnorm((1-sig_level)/2)*-1

 (g.ci <- c(mean(gifted$count) - z_score * sd(gifted$count) / sqrt(length(gifted$count)),
            mean(gifted$count) + z_score * sd(gifted$count) / sqrt(length(gifted$count))))
```

```
## [1] 29.51155 31.87734
```

(e) Do your results from the hypothesis test and the confidence interval agree? Explain. -**yes they both agree**

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.

- **H0 - average IQ of mother of gifted children is not different than the average IQ of the population 100**
- **HA - average IQ of mother of gifted children is different than the average IQ of the population**
- **The p-value is much lower than the significance level of .1 so there is not enough evidence to support the null hypothesis. Therefore we will reject the null hypotheis in favor of the alternate**

```
t.test(gifted$motheriq, alternative = "two.sided", mu = 100, conf.level = 0.90)
```

```
##
##  One Sample t-test
##
## data:  gifted$motheriq
## t = 16.756, df = 35, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 100
## 90 percent confidence interval:
##  116.3349 119.9984
## sample estimates:
## mean of x
##  118.1667
```

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

```r
sig_level <- .9
z_score <- qnorm((1-sig_level)/2)*-1

 (g.ci <- c(mean(gifted$motheriq) - z_score * sd(gifted$motheriq) / sqrt(length(gifted$motheriq)),
            mean(gifted$motheriq) + z_score * sd(gifted$motheriq) / sqrt(length(gifted$motheriq))))
```

```
## [1] 116.3834 119.9499
```

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

-yes they agree - The general population mean of 100 is not withing the 90% confidence interval so we will reject the null hypothesis in favor of the alternative 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.

- **sample distribution of the mean - if we take random independent samples the distribution of the means for all the samples is called the sample distribution of the mean**
- **shape - closer to the shape of a normal distribution**
- **center - becomes taller as the frequency of values close to the true population mean increase)**
- **spread - becomes narrower**

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
```

-6.7%

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

```
mean = 9000
sd = 1000
n = 15
se = sd / sqrt(n)
```

- **near normal population distribution $N(9000, 1000^2)$ and the SE=258.2**

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

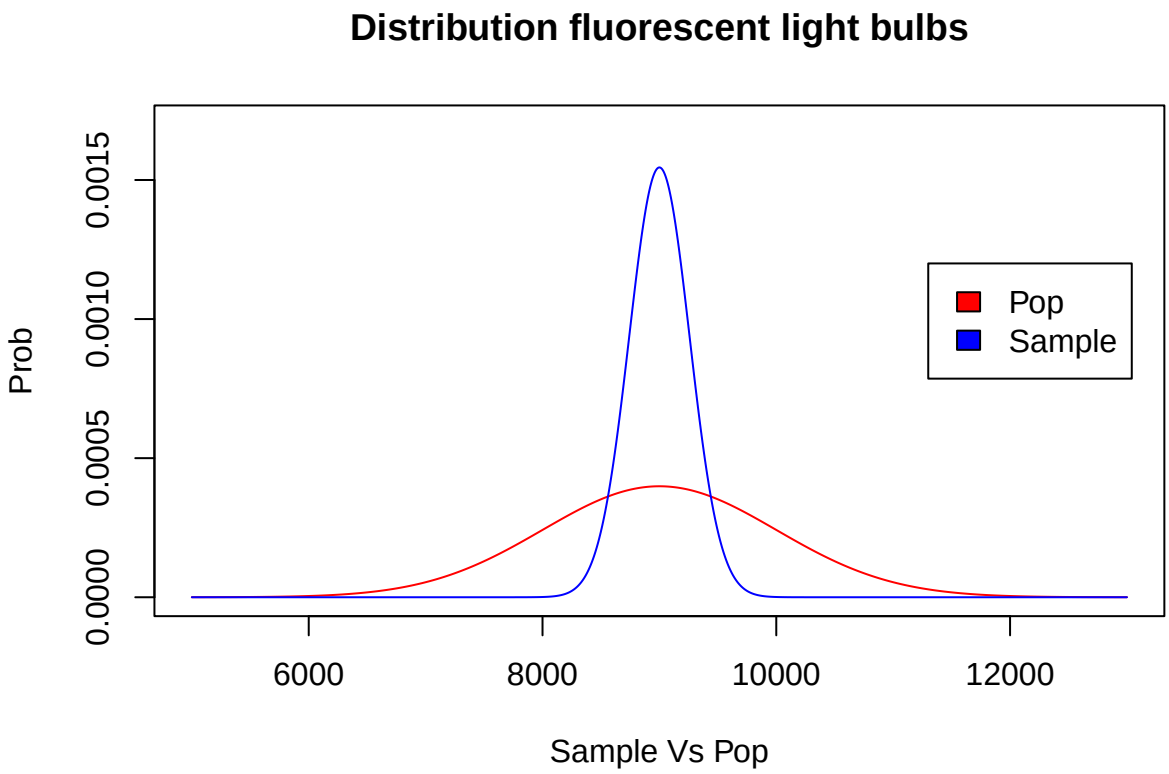
```
((1 - pnorm(q=10500, mean=9000, sd=se)))
```

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

- **approximately to zero**

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

```
sd_p <- 1000
sd_s <- 258.2
x <- 5000:13000
dist_pop <- dnorm(x, mean = 9000, sd = sd_p)
dist_sample <- dnorm(x, mean = 9000, sd = sd_s)
plot(x, dist_pop, type = "l", main = "Distribution fluorescent light bulbs",
     xlab = "Sample Vs Pop", ylab = "Prob", col = "red", ylim = c(0, 0.0017))
lines(x, dist_sample, col = "blue")
legend(11300, 0.0012, legend = c("Pop", "Sample"), fill = c("red", "blue"))
```



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

- **The z values were based on a normal distribution and the sample size is small. You would require 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.

- **As n increases the p-value to get smaller.**