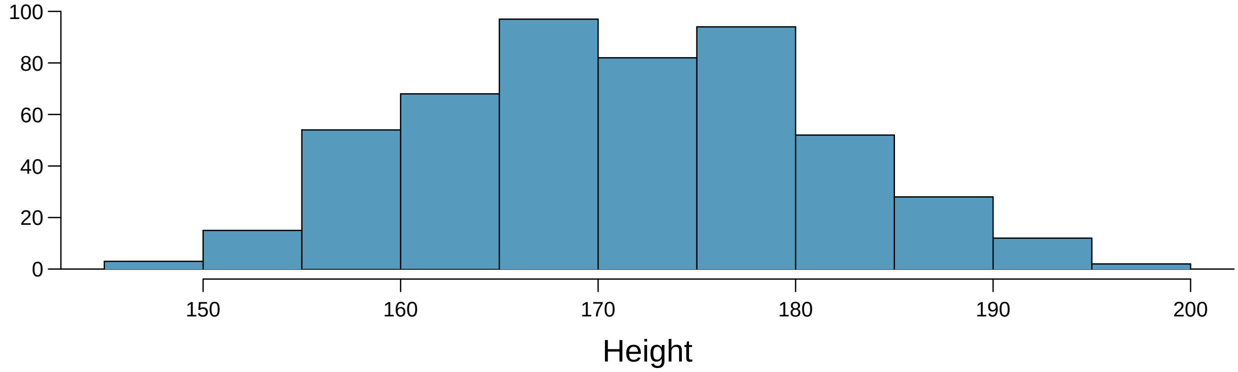
Chapter 5 - Foundations for Inference

Shana Green

DATA 606 - Homework 5

Due Date: 10/04/2020

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?

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

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

After running the summary of bdims$hgt, the point estimate of the average height is 171.1 and the median is 170.3.

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

```r
sd(bdims$hgt, na.rm = FALSE)
```

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

```r
IQR(bdims$hgt, na.rm = FALSE)
```

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

After running the sd and IQR (Interquartile Range) of bdims$hgt, the standard deviation of the heights of active individuals is 9.407205. The IQR is 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.

Let's calculate the Z-score for 180 cm:

```r
x <- 180
mu <- mean(bdims$hgt)
sd <- sd(bdims$hgt)

z <- (x - mu)/sd
paste0("The value of z is ", round(z,2), ".")
```

```
## [1] "The value of z is 0.94."
```

A person with a height of 180 cm is not unusually tall because it lies within 2 SD of the mean.

Let's calculate the Z-score for 155 cm:

```r
x <- 155

z <- (x - mu)/sd
paste0("The value of z is ", round(z,2), ".")
```

```
## [1] "The value of z is -1.72."
```

A person with a height of 155 cm would be considered unusually short because the values lie outside 2 SD of the mean.

(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.

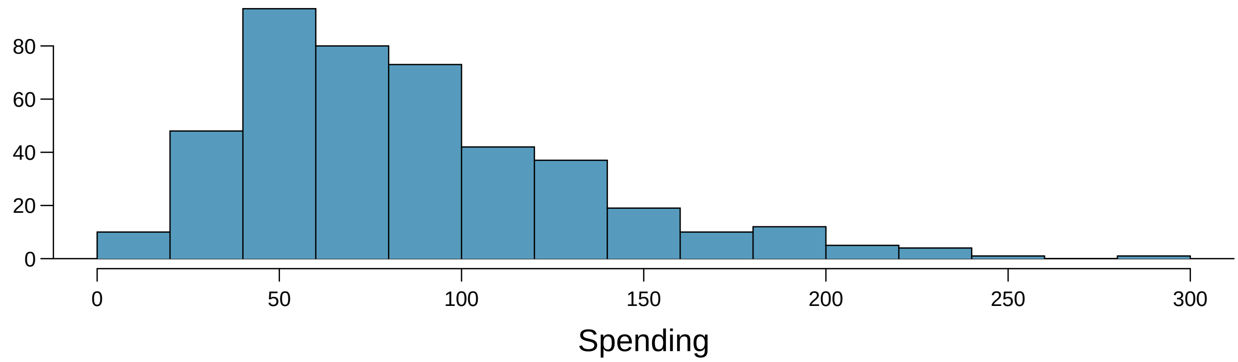
The mean and the standard deviation would be different due to randomness in the sampling.

(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.

```r
n = 507 # physically active students
se <- sd / sqrt(n)
paste0("The Standard Error is used to quantify the variability. The standard error of mean is ", round(
```

```
## [1] "The Standard Error is used to quantify the variability. The standard error of mean 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 because the sample mean is always in the confidence interval. The 95% CI covers the population mean with a 95% probability.

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

This is FALSE because the CI may still be valid when the sample distribution is slightly skewed. The other conditions on the CI are met in this case.

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

This is FALSE because samples of different size may have different confidence intervals. It is true that the mean value of 95% of the random samples of size 436 lie within the confidence interval.

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

```r
mu<-mean(thanksgiving_spend$spending)
sd<-sd(thanksgiving_spend$spending)
l <- length(thanksgiving_spend$spending) - 1
error <- qt(.975,l)*sd/(sqrt(l))

neg<- mu - error
pos<- mu + error

neg
```

```
## [1] 80.28444
```

```
pos
```

```
## [1] 89.12909
```

This is TRUE.

(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.

(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 standard error is equal to $SD_x = \frac{\sigma}{\sqrt{n}}$. In order to get the CI to shrink to one third of what it is now, the sample size has to be 9 times bigger.

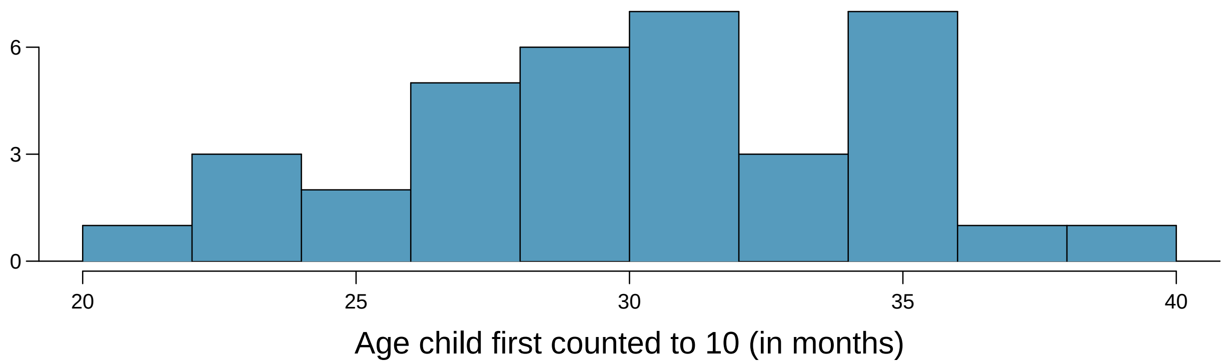
(g) The margin of error is 4.4.

```
paste0("The error is ",round(error,1), ".")
```

```
## [1] "The error is 4.4."
```

This is TRUE.

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?

In order for conditions to be met, we need to verify three things

Independent: 36 children are less than the 10 percent of the population of all children.

Randomization: The sample size is a random sample.

Normal: The sample size is larger than the mean of 30.69 and the central limit theorem tells us that the sampling distribution of the mean is approximately normal.**

(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.

The hypothesis is:

$$H_0 : \mu = 32$$
$$H_A : \mu < 32$$

The standard error is:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{4.31}{\sqrt{36}} = \frac{4.31}{6} \approx 0.7183$$

The z-value is:

$$z = \frac{\bar{x}-\mu}{\sigma\sqrt{n}} = \frac{30.69-32}{\frac{4.31}{\sqrt{36}}} \approx -1.82$$

**The standardized test score for 30.69:*

```
mu <- round(mean(gifted$count),2)
n <-  36
sd <- round(sd(gifted$count),2)
x <- 32
se <- sd / sqrt(n)
z <- round((mu - x)/se, 2)

z
```

```
## [1] -1.82
```

```
p_value<-round(pnorm(z, mean = 0, sd = 1),4)
p_value
```

```
## [1] 0.0344
```

Since

$$P < 0.10 \rightarrow Reject\ H_0$$

there is enough evidence to support the claim that the average age at which children first count to 10 successfully is less than the general average of 32 months.

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

From part(b):

```
p_value<-round(pnorm(z, mean = 0, sd = 1),4)
p_value
```

```
## [1] 0.0344
```

```
mean(gifted$count)
```

```
## [1] 30.69444
```

**If the population had a mean of 32 months, there would be a 3.44% chance of obtaining a sample of 36 children, since the mean is 30.69.

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

```
CL<-c(0.90) # 90% confidence interval
p<-1-(1-CL)/2
p
```

```
## [1] 0.95
```

```r
multiplier<-round(qnorm(p, mean=0,sd=1),3)
multiplier
```

```
## [1] 1.645
```

```r
se <- round((sd / sqrt(n)),4) # Standard Error
se
```

```
## [1] 0.7183
```

```r
neg<-round(mu - (multiplier*se),4)
pos<-round(mu + (multiplier*se),4)

pos
```

```
## [1] 31.8716
```

```r
neg
```

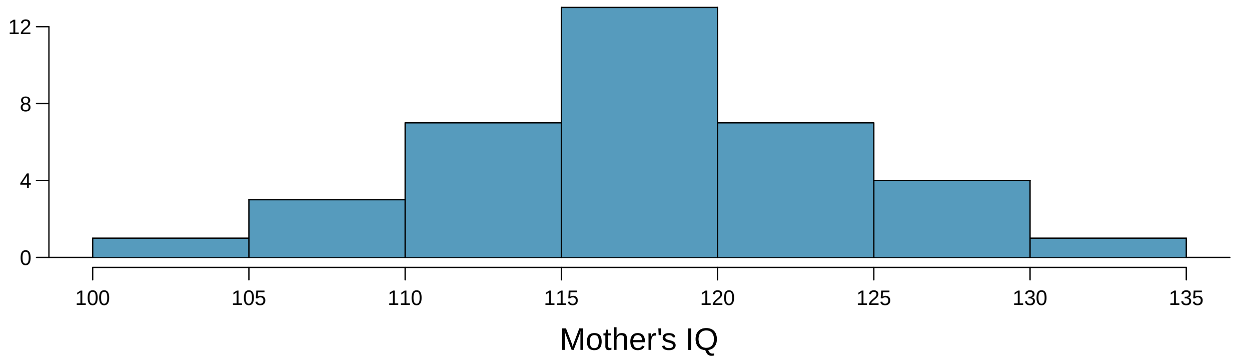
```
## [1] 29.5084
```

The 90% confidence interval is (29.5084, 31.8716)

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

Yes it does. The confidence interval does not contain 32, so both tests reject the null hypothesis in favor of the alternative 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 the se 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.

The hypothesis is:

$$H_0 : \mu = 100$$
$$H_A : \mu \neq 100$$

The standard error is:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{6.5}{\sqrt{36}} = \frac{6.5}{6} \approx 1.0833$$

The z-value is:

$$z = \frac{\bar{x}-\mu}{\sigma\sqrt{n}} = \frac{118.2-100}{\frac{6.5}{\sqrt{36}}} \approx 16.80$$

Since

$$P < 0.10 \rightarrow Reject\ H_0$$

there is enough evidence to support the claim that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100.

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

```
summary(gifted$motheriq)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   101.0   113.8   118.0   118.2   122.2   131.0
```

```
n<-100
mu <- round(mean(gifted$motheriq),1)
sd <- round(sd(gifted$motheriq),1)
CL<-c(0.90) # 90% confidence interval
p<-1-(1-CL)/2
multiplier<-round(qnorm(p, mean=0,sd=1),3)
se <- round((sd / sqrt(n)),4) # Standard Error

neg<-round(mu - (multiplier*se),4)
pos<-round(mu + (multiplier*se),4)

pos
```

```
## [1] 119.2692
```

```
neg
```

```
## [1] 117.1308
```

The 90% confidence interval is (117.1308, 119.2692).

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

Yes it does, because the CI does not include 100. Therefore, it comes to the same conclusion as the hypothesis test.

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 is best described as a probability distribution of a statistic, that’s obtained through a large number of samples drawn from a specific population. The sampling distribution of the mean takes the mean of each sample. When the sample size increases, the shape normalizes and has a smaller spread.

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?

```r
n<-10500
mu<-9000
sd<-1000

p<-1-pnorm(n,mu,sd)
paste0("The probability that a randomly chosen light bulb lasts more than 10,500 hours is ",(round(p,3)
```

```
## [1] "The probability that a randomly chosen light bulb lasts more than 10,500 hours is 0.067."
```

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

$$x_1, x_2, ..., x_{15} \sim N(9000, 1000^2)$$
$$\bar{x} \sim N(9000, \frac{1000^2}{15}) \text{ or } N(9000, \frac{1000^2}{15})$$

The distribution of the mean lifespan of 15 light bulbs is normal.

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

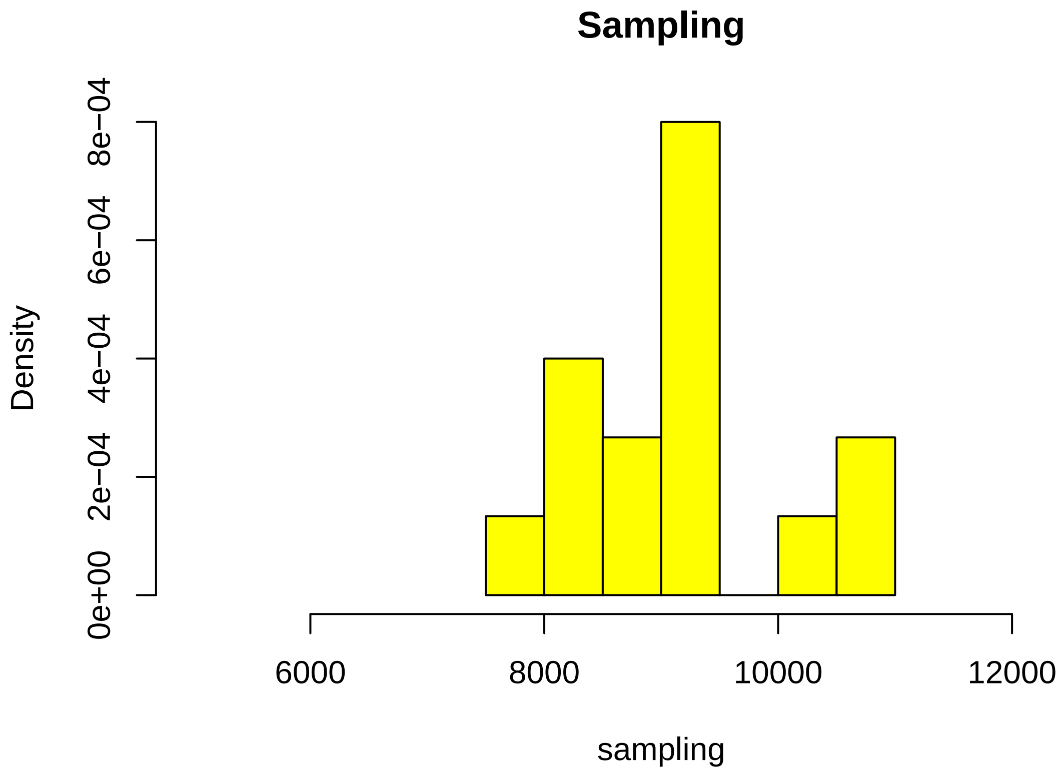
```r
x<-15
se <- round((sd / sqrt(n)),4)
z <- (n - mu)/se
p <- 1 - pnorm(z)
p
```

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

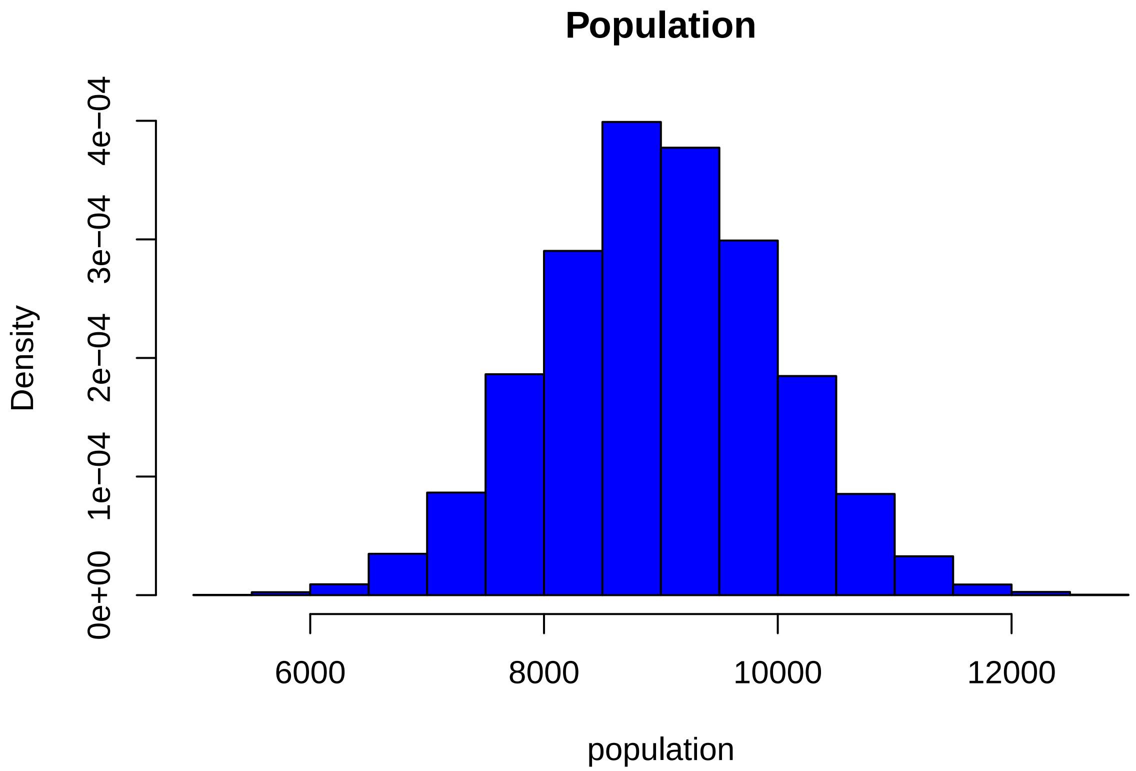
The probability is zero.

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

```r
x<-15
set.seed(123)
sampling <- rnorm(x, mean = mu, sd = sd)
population <- rnorm(n, mean = mu, sd = sd)
hist(sampling, freq = FALSE, xlim=c(5000,13000),col="yellow",main="Sampling")
```



```
hist(population, freq = FALSE,xlim=c(5000,13000),col="blue",main="Population")
```



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

In part a), it was not normal and in c), the CLT says that the means are close to normally distributed, even if the original observations are not. But here, we only have n=15.

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.

So we have two values $n_1 = 50$ and $n_2 = 500$**. Since** σ **is not provided, we know the standard error equation:**

$$SE = \frac{\sigma}{\sqrt{n}}$$
$$SE_1 = \frac{\sigma}{\sqrt{n_1}} > SE_2 = \frac{\sigma}{\sqrt{n_2}}$$

In this case, we know that the larger the denominator, the smaller the fraction, so: $SE_1 > SE_2$.

If we calculate the z-value:

$$Z = \frac{\mu - x}{SE}$$
$$Z_1 = \frac{\mu - x}{SE_1} < Z_2 = \frac{\mu - x}{SE_2}$$

the p value will always change if the sample size changes. For the case above, the p-value will decrease in this equation: $1 - pnorm(Z, mean = 0, sd = 1)$