HW4 - Confidence Intervals

Shyam BV

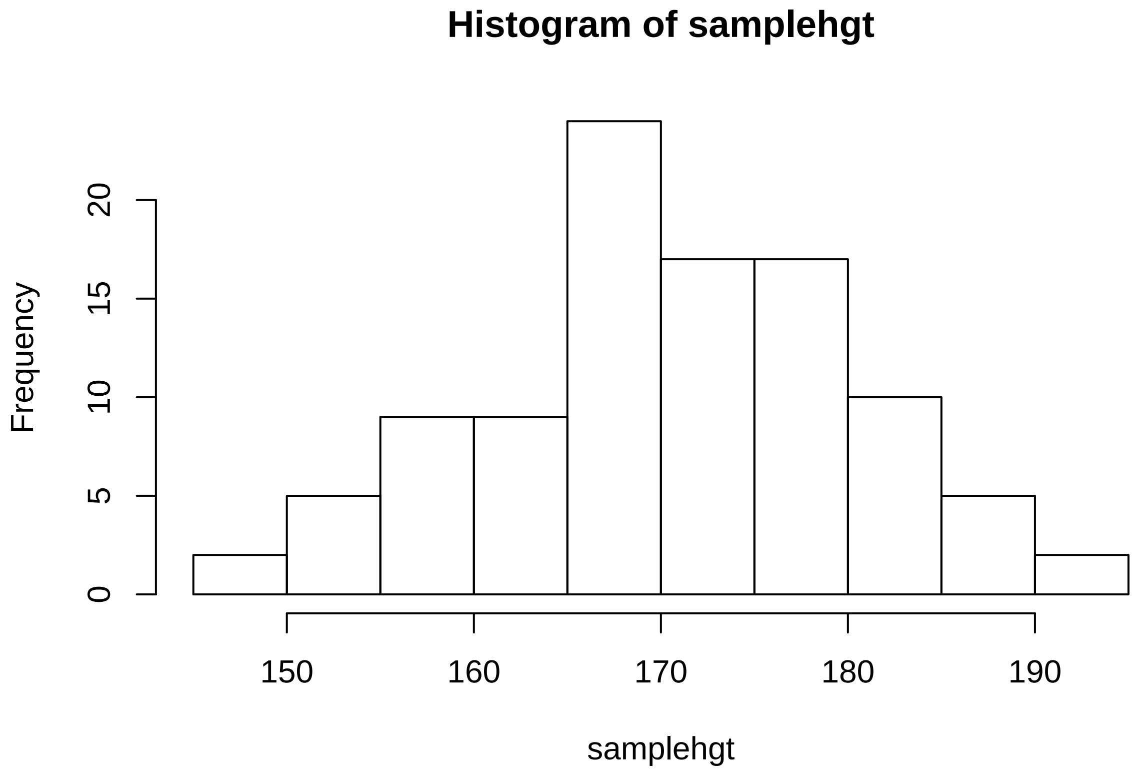
4.4 Heights of adults.

1. 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

```
hgtadults <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from
```

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

```
samplehgt <- sample(hgtadults$hgt,100)
hist(samplehgt)
```



Here we have taken a sample of 50 from the population. The point estimate for the average height is 171.143787. 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?

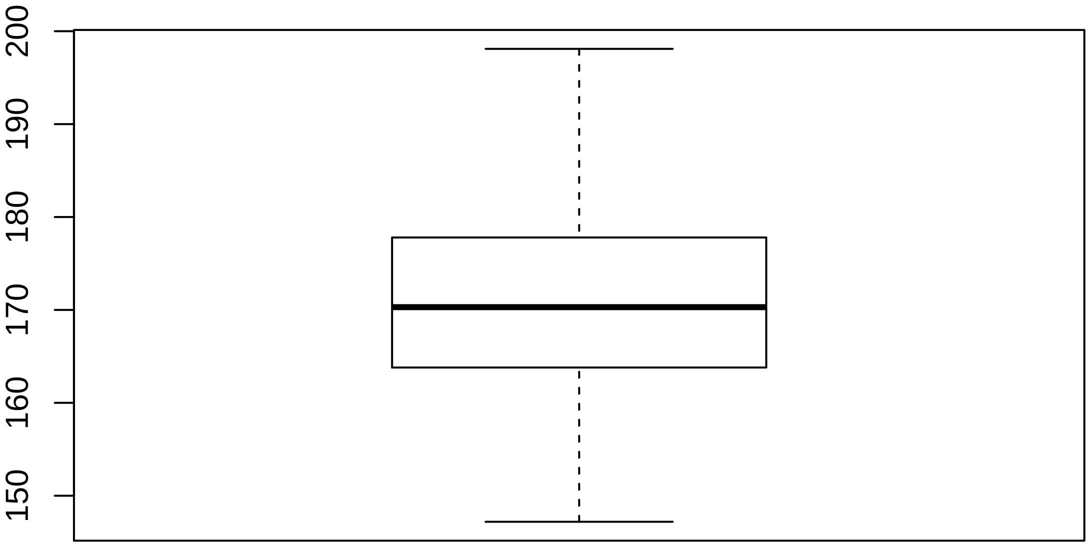
```
#Standard Deviation
sd(hgtadults$hgt)
```

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

```
#IQR
IQR(hgtadults$hgt)
```

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

```
boxplot(hgtadults$hgt)
```



```
summary(hgtadults$hgt)
```

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

Above are the standard deviation and IQR range from the samples. THis is the point estimate of the true population

(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

```
# Tall person

zscore_tall = (180-mean(hgtadults$hgt))/sd(hgtadults$hgt)

zscore_small = (155-mean(hgtadults$hgt))/sd(hgtadults$hgt)
```

180 cm is not unusually tall. Because it is within 1 SD from the mean. Similarlly, 155 is not too short. Because it is within 2 SD from 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.

```
mean(sample(hgtadults$hgt,100))
```

```
## [1] 170.286
```

If we run the above code multiple times, we can see that the mean changes each time. So he should not be suprised by the different mean. As long as it is in specified confidence interval, we should not be worried.

(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 ? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

```
samplehgt <- sample(hgtadults$hgt,100)

se_samplehgt <- sd(hgtadults$hgt)/sqrt(100)
```

The measure to quantify the variability of the estimate is called as standard error. It is also called as sampling distribution of sample mean.

4.14 Thanksgiving spending, Part I

2. 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. It should be said in a way that we are 95% confident that the interval will contain the population point estimate.

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

False. Although the distribution is right skewed, the sample size is large enough. As per central limit theorm, the sample will be normally distributed.

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

False. Confidence interval does not use sample mean.

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

True. We are 95% confident on the average spending of all american adults 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.

True. When the confidence interval reduces, the estimate will also reduce.

(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. To reduce the margin of error, we need to make sample^3(not multiply). Because SD/sqrt(n) is the formula. Although increasing the sample size will reduce the error, it cant reduce to 1/3rd.

4.24 Gifted children, Part I

3. 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

```r
gifted <-  read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from%2
```

(a) Are conditions for inference satisfied?

Yes. The conditions are satisfied. It has a minimum sample size above 30. Each child is independent and selection is random.

(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.

```r
qnorm(.10)
```

```
## [1] -1.281552
```

```
pnorm((30.69-32)/(4.31/sqrt(36)))
```

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

$H0:$ Months is greater or equal to 32. $H1:$ Months is less than 32.

The p-value is 0.03. It is less than the critical value of 0.05. So we reject the null hypohsis.

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

p-value is 0.03. Means there is only 3% probability that the months is greater than or equal to 32.

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

```
qnorm(.90)
```

```
## [1] 1.281552
```

```
30.69 - (1.28*(4.31/sqrt(36)))
```

```
## [1] 29.77053
```

```
30.69 + (1.28*(4.31/sqrt(36)))
```

```
## [1] 31.60947
```

Confidence interval of 90% is 29.77,31.61

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

Yes, it mathes the result. We rejected the hypothesis of greater than or equal to 32.

4.26 Gifted children, Part II

Exercise 4.24 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.

```
qnorm(.05)
```

```
## [1] -1.644854
```

```
#Z score
(118.2-100)/(6.5/sqrt(36))
```

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

```
1-pnorm((118.2-100)/(6.5/sqrt(36)))
```

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

$H0 = 100 \; H1 > 100$

As this is uppertail test, we need to subract from 1. The probability is very small. So we reject the null hypothesis.

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

```
118.2-(qnorm(0.05)*(6.5/sqrt(36)))
```

```
## [1] 119.9819
```

```
118.2+(qnorm(0.05)*(6.5/sqrt(36)))
```

```
## [1] 116.4181
```

The confidence interval for mothers IQ of gifted children are 116.41 to 119.98.

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

Yes. The results from hypothesis test and confidence interval match with each other. In the hypothesis test, we rejected that the mothers IQ is equal to 100. Confidence interval shows that their IQ is 116.41 to 119.98.

4.34 CLT.

5. 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 which is formed with the sample means. We fetch a sample from a population distribution, if the population is not heavily skewed, then the sample distribution of size 30 will be a normal distribution.

If we do n number of samples from the population, the sampling distribution of sample means will be normally distributed.

The shape will be perfect bell curve which is unimodel and the spread is lesser than the original population.

4.40 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?

```
mean=9000
x=10500
sd=1000

1-pnorm((x-mean)/sd)
```

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

The probability of the randomly chosen light bulb lasts more than 10500 is 6.68%

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

```
samplelight <- rnorm(15,9000,1000)
mean(samplelight)
```

```
## [1] 9027.47
```

```
sd(samplelight)
```

```
## [1] 800.3846
```

The sample mean distribution of the 15 light bulbs is around `rmean(samplelight)`. This might vary depending on the samples selected.

As the sample size is 15(less than the standard 30), we might need to use t-distribution to calculate the population distribution.

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

```
#t distribution
(mean(samplelight)-10500)/((sd(samplelight))/sqrt(15))
```

```
## [1] -7.125429
```

```
1-pt((10500-9000)/((sd(samplelight))/sqrt(15)),14)
```

```
## [1] 2.084771e-06
```

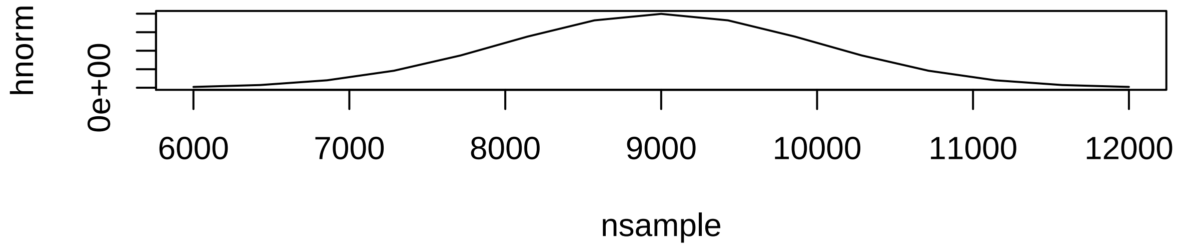
There is 4.964283e-06 ~=0 % probability of choosing a bulb which is more than 10,500.

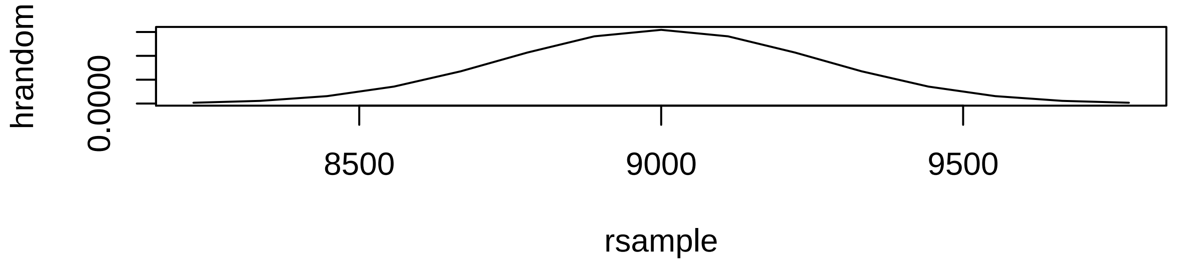
(d) Sketch the two distributions (population and sampling) on the same scale.

```
nsample <- seq(mean - (3 * 1000), mean + (3 * 1000), length=15)
rsample<- seq(mean - (3 * (1000/sqrt(15))), mean + (3 * (1000/sqrt(15))), length=15)
hnorm <- dnorm(nsample,mean,1000)
hrandom<- dnorm(rsample,mean,(1000/sqrt(15)))

par(mfrow = c(2, 1))

plot(nsample,hnorm,type="l")
plot(rsample,hrandom,type="l")
```





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

```
1-pnorm((x-mean)/sd)
```

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

```
1-pt((10500-9000)/((sd(samplelight))/sqrt(15)),14)
```

```
## [1] 2.084771e-06
```

We can estimate the probability, but the skewness can't be determined.

4.48 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 sample size increases. The standard error decreases, the z-score increases, and therefore the p value will decrease.