MichaelY_HW4_Inference

Michael Y.

October 21, 2018

```
setwd("c:/users/Michael/DROPBOX/priv/CUNY/MSDS/201809-Fall/DATA606_Jason/Homework")
```

Homework - Chapter 4 - Foundations for Inference (pp.168-218)

Exercises: 4.4, 4.14, 4.24, 4.26, 4.34, 4.40, 4.48 (pp.203-218)

Exercise 4.4 - Heights of adults.

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.

{G. Heinz et al. "Exploring relationships in body dimensions". In: Journal of Statistics Education 11.2 (2003).}

```
p44_Min = 147.2
p44_Q1  = 163.8
p44_Median = 170.3
p44_Mean = 171.1
p44_SD  = 9.4
p44_Q3  = 177.8
p44_Max = 198.1

p44_N = 507
```

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

The Mean is 171.1 .

What about the median?

The Median is 170.3 .

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

The Standard Deviation is 9.4 .

What about the IQR?

The IQR is Q3 - Q1 = 14 .

(c) *Is a person who is 1m 80cm (180 cm) tall considered unusually tall?*

```
p44_ZScore180 <- (180 - p44_Mean)/p44_SD
round(p44_ZScore180,4)
```

```
## [1] 0.9468
```

```
## .9468
p44_Q180 <- round(100*pnorm(p44_ZScore180),2)
p44_Q180
```

```
## [1] 82.81
```

Because the ZScore 0.9468085 is less than 1 standard deviation above the mean, the person is of above average height, but not unusually tall. The person is taller than 82.81 percent of the population.

And is a person who is 1m 55cm (155cm) considered unusually short?

```
p44_ZScore155 = (155 - p44_Mean)/p44_SD
round(p44_ZScore155,4)
```

```
## [1] -1.7128
```

```
## -1.7128
p44_Q155 <- round(100*pnorm(p44_ZScore155),2)
p44_Q155
```

```
## [1] 4.34
```

Because the ZScore -1.712766 is nearly 2 standard deviations below the mean, the person is indeed quite short. Only 4.34 percent of the population are shorter than this individual.

Explain your reasoning.

See above.

(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?*

No, I would not expect the mean and SD of another sample to match the ones given above.

Explain your reasoning.

Each sample is randomly drawn from a population, but the variation in the selection of each sample will cause variation in the sample mean for each sample.

(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 $SD\left(\bar{x}\right) = \frac{\sigma}{\sqrt{N}}$)

The measure is known as the Standard Error of the Sample Mean, and is computed using the above formula.

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

```
p44_SE = p44_SD / sqrt(p44_N)
p44_SE
```

```
## [1] 0.4174687
```

```
## 0.4175
```

The Standard Error of the mean is 0.4174687 .

#.

Exercise 4.14 - 4.14 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).

```
data(tgSpending)
str(tgSpending)
```

```
## 'data.frame':    436 obs. of  1 variable:
##  $ spending: num  52.8 187.6 63.9 195.6 120.8 ...
```

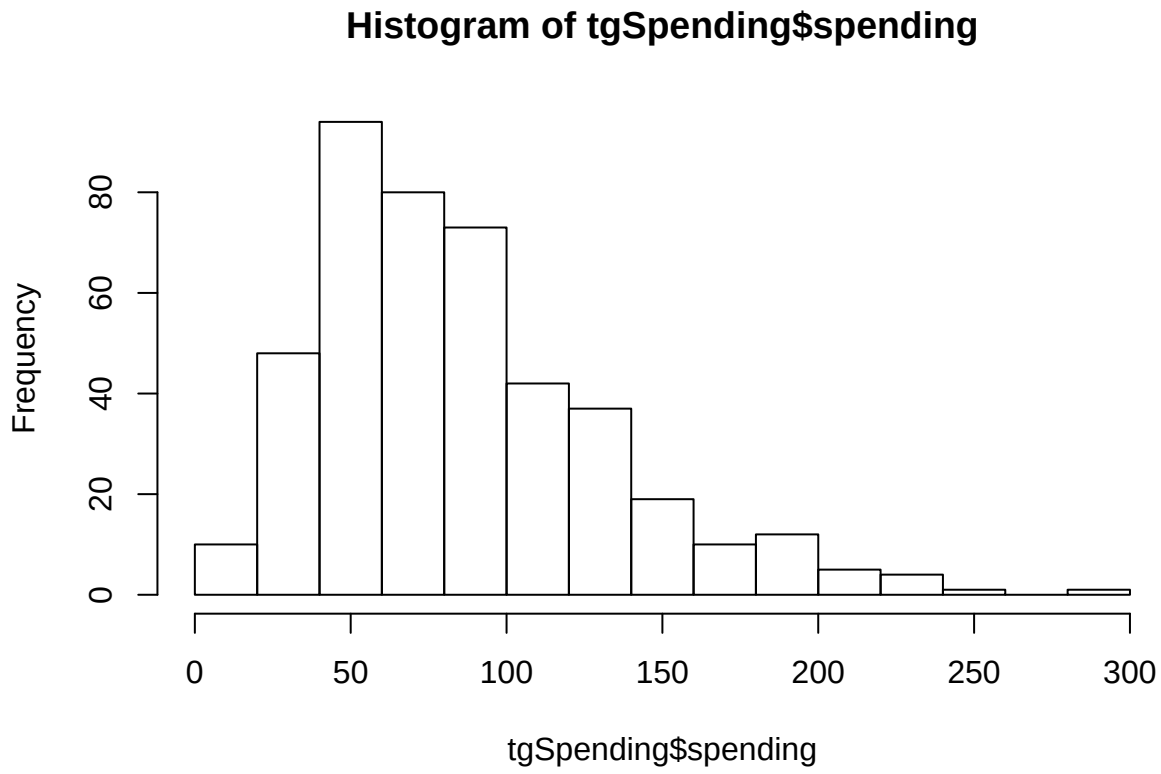
```
summary(tgSpending$spending)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.719  49.177  75.792  84.707 112.255 282.803
```

```
sd(tgSpending$spending)
```

```
## [1] 46.92851
```

```r
hist(tgSpending$spending)
```




```r
p414_mean <-  mean(tgSpending$spending)
p414_mean
```

```
## [1] 84.70677
```

```r
p414_SD <-  sd(tgSpending$spending)
p414_SD
```

```
## [1] 46.92851
```

```r
p414_N <-  length(tgSpending$spending)
p414_N
```

```
## [1] 436
```

```r
p414_SE <-  p414_SD / sqrt(p414_N)
p414_SE
```

```
## [1] 2.247468
```

```r
p414_lower95 <- p414_mean - 1.96 * p414_SE
p414_lower95
```

```
## [1] 80.30173
```

```r
p414_upper95 <- p414_mean + 1.96 * p414_SE
p414_upper95
```

```
## [1] 89.1118
```

```
p414_margin_of_error <- 1.96 * p414_SE
p414_margin_of_error
```

```
## [1] 4.405038
```

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 are 100% certain that the average spending of *these 436 American adults* is $84.71 .

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

FALSE

Although the distribution of spending in this sample happens to be right-skewed, by the Central Limit Theorem, the distribution of the sample means under repeated samples will be Normally distributed, so long as the samples are independent and identically distributed. The purpose of the confidence interval is to infer the population mean (i.e., average spending across all consumers during the 6-day observation period.)

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

FALSE

If repeated random samples of similar size (n=436) are taken from the population, we expect that the confidence interval, as constructed from the sample mean and standard deviation from each such sample, will cover the population mean in 95% of such samples.

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

TRUE

The statistical inference from this model is that the population mean ("average spending of *all* American adults") [over the six-day period, which is presumed] falls within the 95% 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

To compute the 95% confidence interval we use

$$1.96 * SE\left(\bar{x}\right) = 1.96 * \frac{\sigma}{\sqrt{N}}$$

The critical value 1.96 comes from the following:

```
qnorm(0.975)
```

```
## [1] 1.959964
```

```
qnorm(0.025)
```

```
## [1] -1.959964
```

where we have a 2.5% tail on either side of the distribution, so 95% is in between.

If we only wanted a 90% confidence interval, then the critical value would be replaced by

```
qnorm(0.95)
```

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

```
qnorm(0.05)
```

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

1.6448536 instead of 1.959964 .

(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

The margin of error decreases with the reciprocal of the square root of the sample size.

Thus, to make the margin of error one-third of the present size, the required sample size would be *NINE* times the present size, rather than 3 times.

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

TRUE

The margin of error is defined as z * SE. Here z = 1.96 and SE = 2.2474681.

So, margin of error = 4.4050375 .

#.

Exercise 4.24 - 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.

{F.A. Graybill and H.K. Iyer. Regression Analysis: Concepts and Applications. Duxbury Press, 1994, pp. 511–516.}

```
data(gifted)
##str(gifted)
##summary(gifted)
count_age <- gifted$count
p424_summary <- summary(count_age)
p424_summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   28.00   31.00   30.69   34.25   39.00
```

```
p424_min_count_age  <- as.numeric(p424_summary["Min."])
p424_mean_count_age <- round(as.numeric(p424_summary["Mean"]),2)
p424_med_count_age  <- as.numeric(p424_summary["Median"])
p424_max_count_age  <- as.numeric(p424_summary["Max."])
p424_iqr_count_age  <- as.numeric(IQR(count_age))
cat(paste("Inter-Quartile Range of the sample: ",p424_iqr_count_age,"\n"))
```

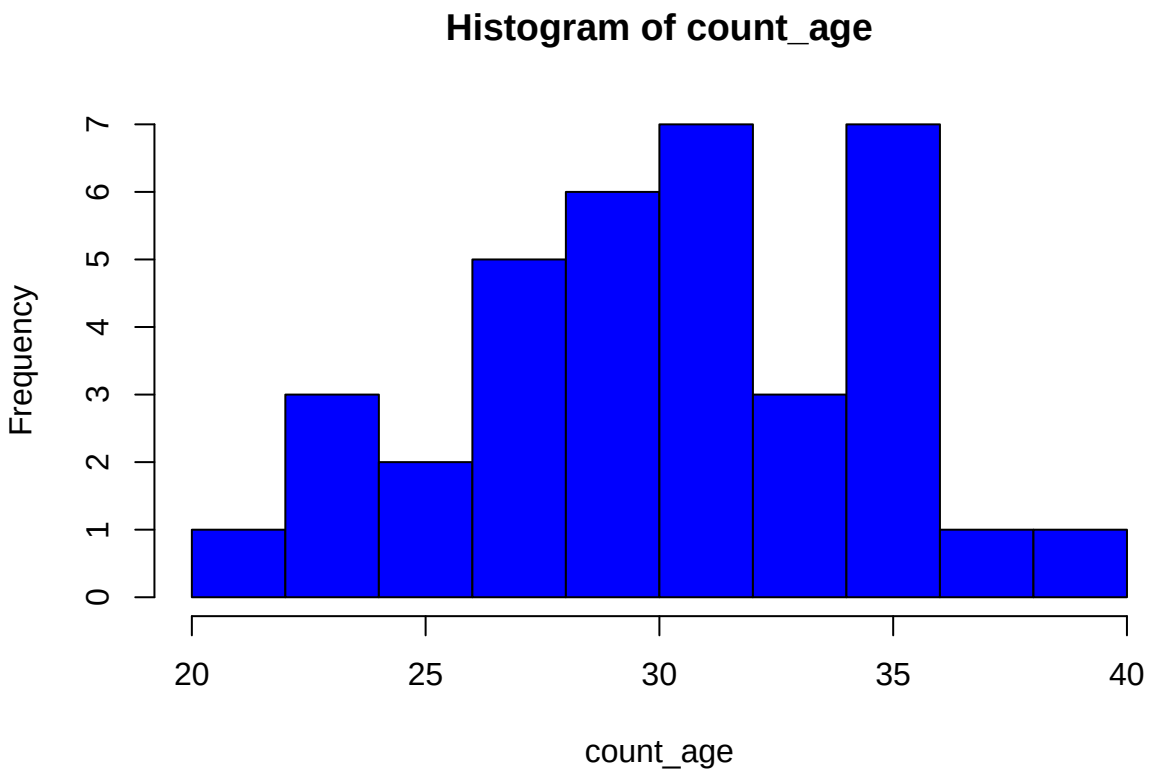
```
## Inter-Quartile Range of the sample:  6.25
```

```
p424_stdev_count_age <- round(as.numeric(sd(count_age)),2)
cat(paste("Standard Deviation of the sample: ",p424_stdev_count_age,"\n"))
```

```
## Standard Deviation of the sample:  4.31
```

```
p424_n <- length(count_age)
```

```
hist(count_age, col = "blue")
```



(a) ***Are conditions for inference satisfied?***

The subjects were randomly selected, and we will assume that they are independent and identically distributed, as they are all age four. The sample size, while small, is larger than 30. The histogram does not exhibit severe skew, so the conditions for inference are met.

(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*** [sic] ***count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.***

$$H_0: \quad \mu_{age} = 32$$
$$H_A: \quad \mu_{age} < 32$$
$$\alpha = 0.10$$

```
### The significance level:
p424_alpha <- 0.10

### Determine the Critical Value for a one-tailed test at this significance level:
p424_criticalZ <- qnorm(p424_alpha)
p424_criticalZ
```

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

```
### The standard deviation of the sample:
p424_stdev_count_age
```

```
## [1] 4.31
```

```
### Determine the Standard Error of the Mean
p424_se <- p424_stdev_count_age / sqrt(p424_n)
p424_se
```

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

```
### Determine the Z-Score
p424_ZScore <- (p424_mean_count_age - 32) / p424_se
p424_ZScore
```

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

```
### Determine the p-value associated with this Z-Score
p424_pvalue <- pnorm(p424_ZScore)
p424_pvalue
```

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

We reject the null hypothesis, which stated that the age at which gifted children first count successfully to 10 is equal to the general average of 32 months, because the p-value of 0.0341013 is less than the designated significance level, alpha = 0.1 . Therefore we accept the alternative hypothesis, which stated that the age at which gifted children first count successfully to 10 is significantly less than 32 months. Additionally the Z-score of -1.8236659 is well below the critical value of -1.2815516 .

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

The extremely low p-value of 0.0341013 indicates that the the probability of the null hypothesis being valid is quite small.

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

```
p424_twotailZ <- qnorm(p=0.95)
p424_twotailZ
```

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

```
p424_lower <- p424_mean_count_age - p424_twotailZ * p424_se
p424_upper <- p424_mean_count_age + p424_twotailZ * p424_se
c(p424_lower,p424_upper)
```

```
## [1] 29.50845 31.87155
```

```
p424_lower_check <- qnorm(p=.05,mean = p424_mean_count_age,sd = p424_se, lower.tail=T)
p424_upper_check <- qnorm(p=.95,mean = p424_mean_count_age,sd = p424_se, lower.tail=T)
c(p424_lower_check,p424_upper_check)
```

```
## [1] 29.50845 31.87155
```

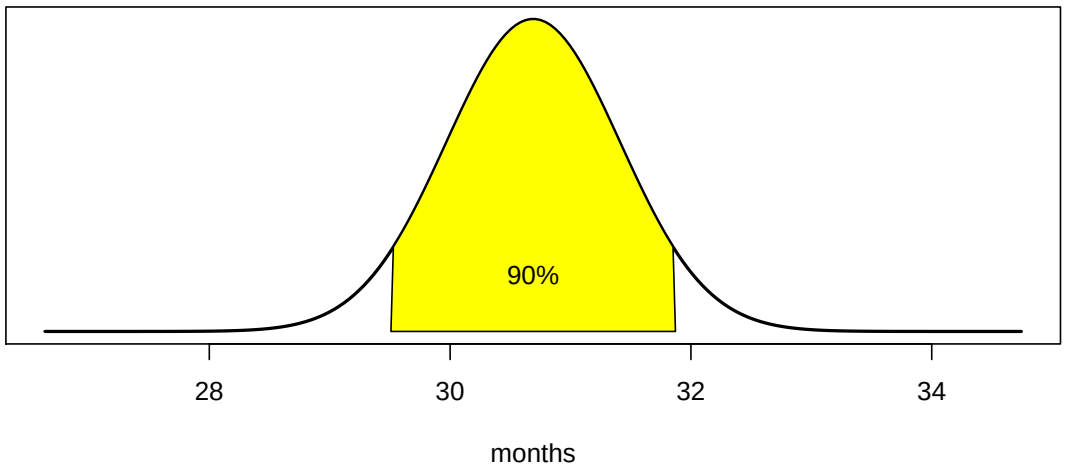
The 90% confidence interval for the average age at which gifted children first count successfully to 10 is (29.5084468 , 31.8715532) months.

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

The results are consistent, given that the value of 32 months is outside of the 10 percent confidence interval. However it is important to note that the significance level of alpha = 0.1 does not mean the same thing as a confidence interval of 90 percent.

Specifically, the hypothesis test employed a one-tail test to infer whether the counting age for gifted children was equal to 32 or less than 32. At 10 percent significance, this makes use of qnorm(.10) or qnorm(.90) to determine the critical value.

On the other hand, the 90 percent confidence interval employs a two-tail approach, which requires the use of qnorm(0.05) and qnorm(0.95) .



#.

Exercise 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.*****

```
data(gifted)
##str(gifted)
##summary(gifted)
iq <- gifted$motheriq
```

```r
p426_summary <- summary(iq)
p426_summary
```

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

```r
p426_min_iq  <- as.numeric(p426_summary["Min."])
### Round the below value to 1 decimal to match the figure shown in the text (118.2)
### Otherwise we would use the (more accurate) value of 118.17, but this would generate inconsistencies
p426_mean_iq <- round(as.numeric(p426_summary["Mean"]),1)
p426_med_iq  <- as.numeric(p426_summary["Median"])
p426_max_iq  <- as.numeric(p426_summary["Max."])
p426_iqr_iq  <- as.numeric(IQR(iq))
cat(paste("Inter-Quartile Range of the sample: ",p426_iqr_iq,"\n"))
```

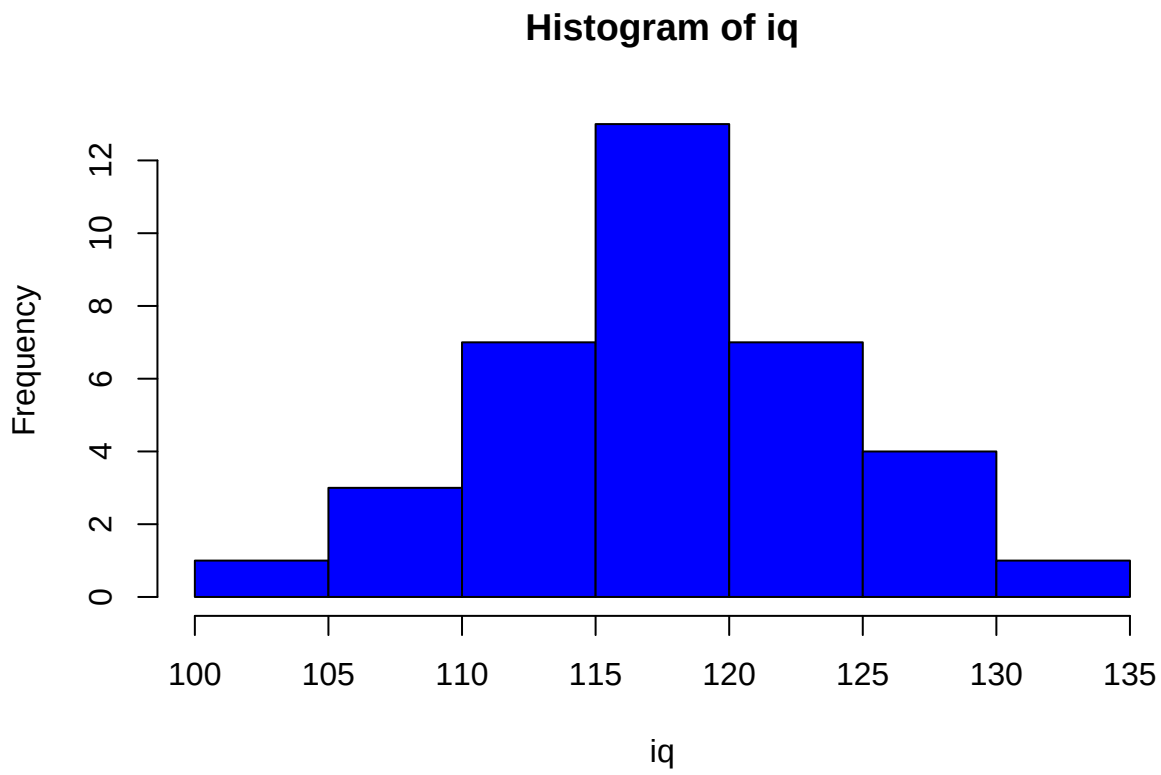
```
## Inter-Quartile Range of the sample:  8.5
```

```r
p426_stdev_iq <- round(as.numeric(sd(iq)),2)
cat(paste("Standard Deviation of the sample: ",p426_stdev_iq,"\n"))
```

```
## Standard Deviation of the sample:  6.5
```

```r
p426_n <- length(iq)
```

```r
hist(iq, col = "blue")
```



(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.***

$$H_0: \quad \mu_{iq} = 100$$
$$H_A: \quad \mu_{iq} \neq 100$$
$$\alpha = 0.10$$

```
### The significance level:
p426_alpha <- 0.10

### Determine the Critical Value for a TWO-tailed test at this significance level:
p426_criticalZ <- qnorm(1-p426_alpha/2)
p426_criticalZ
```

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

```
### The standard deviation of the sample:
p426_stdev_iq
```

```
## [1] 6.5
```

```
### Determine the Standard Error of the Mean
p426_se <- p426_stdev_iq / sqrt(p426_n)
p426_se
```

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

```
### Determine the Z-Score
p426_ZScore <- (p426_mean_iq-100) / p426_se
p426_ZScore
```

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

```
### Determine the p-value associated with this Z-Score
p426_pvalue <- 1-pnorm(p426_ZScore)
p426_pvalue
```

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

We reject the null hypothesis, which stated that the average IQ of mothers of gifted children is equal to the general average IQ of 100, because the p-value of 0 is less than the designated significance level, alpha/2 = 0.05 . (In contrast to the previous problem, here we need to use a two-tailed test because this question stated that the IQ was "different than" rather than "greater than". This requires that we cut alpha in half.)

Therefore we accept the alternative hypothesis, which stated that the average IQ of mothers of gifted children is significantly different from the general average IQ of 100. Additionally, the Z-score of 16.8 is well above the critical value of 1.6448536 .

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

```
p426_twotailZ <- qnorm(p=0.95)
p426_twotailZ
```

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

```
p426_lower <- p426_mean_iq - p426_twotailZ * p426_se
p426_upper <- p426_mean_iq + p426_twotailZ * p426_se
c(p426_lower,p426_upper)
```

```
## [1] 116.4181 119.9819
```

```
p426_lower_check <- qnorm(p=.05,mean = p426_mean_iq,sd = p426_se, lower.tail=T)
p426_upper_check <- qnorm(p=.95,mean = p426_mean_iq,sd = p426_se, lower.tail=T)
c(p426_lower_check,p426_upper_check)
```

```
## [1] 116.4181 119.9819
```

The 90 percent confidence interval for Mothers' IQ is (116.4180752 , 119.9819248) .

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

The results concur, in that the value associated with the Null Hypothesis (IQ = 100) is well outside the 90 percent confidence interval associated with the sample.

#.

Exercise 4.34 - 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.

The sampling distribution of the mean represents the distribution of the point estimates (here, the sample averages) based on samples of a fixed size from a certain population. The distribution of the sample mean is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes. Also it can be a problem if the distribution is extremely skewed.

As the sample size increases:

The shape of the sampling distribution more closely resembles the normal distribution.

The center of the sampling distribution converges onto the population mean.

The spread of the sampling distribution decreases according to the square root of the sample size, so quadrupling the sample size will cause the standard error to become half as wide.

#.

Exercise 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.

```
p440_mu <- 9000
p440_sigma <- 1000
```

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

```
p440a_probability <- pnorm(q=10500,mean=p440_mu,sd = p440_sigma, lower.tail = F)
p440a_probability
```

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

The probability is 0.0668072 .

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

The distribution of the average lifespan of 15 lightbulbs would be Normal with mean = 9000 and standard deviation = $\frac{1000}{\sqrt{15}}$

i.e., $N\left(\mu = 9000, \quad \sigma = \frac{1000}{\sqrt{15}}\right)$

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

```
p440c_prob1 <-  pnorm(q = 10500, mean = p440_mu, sd = p440_sigma/sqrt(15), lower.tail = F)
p440c_prob1
```

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

```
p440c_prob2 <- pnorm((10500 - p440_mu)/(p440_sigma/sqrt(15)),lower.tail = F)
p440c_prob2
```

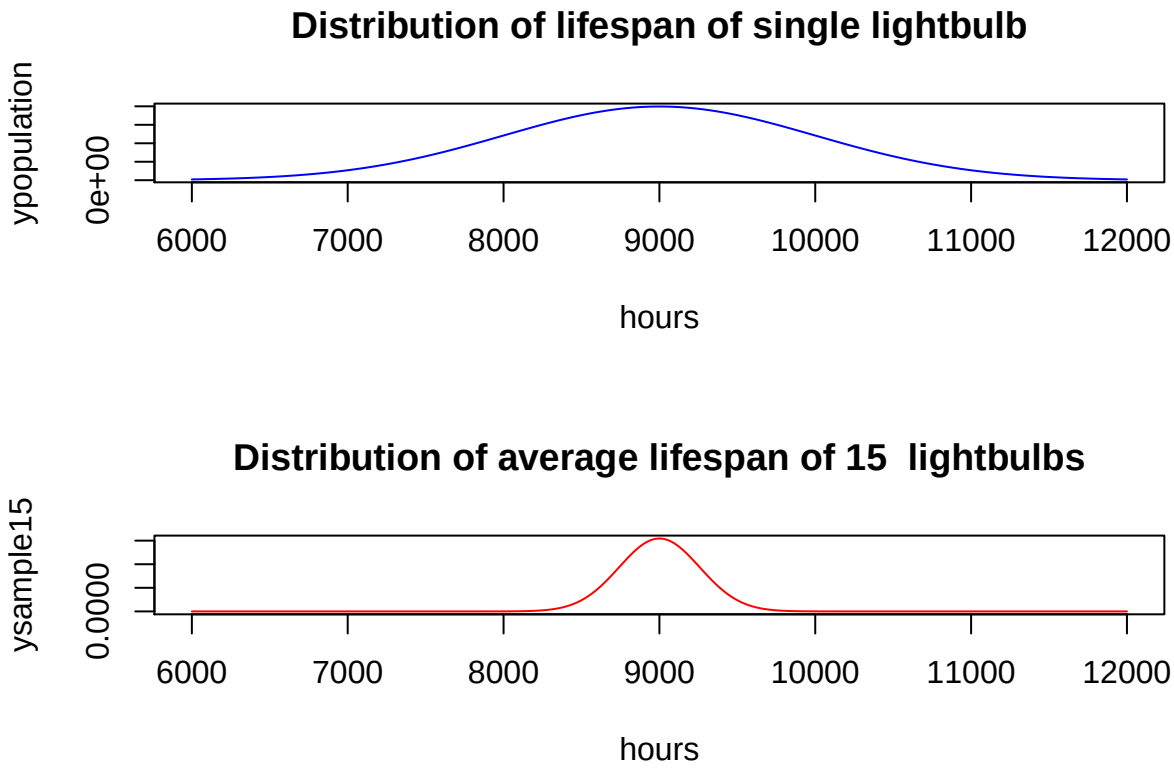
```
## [1] 3.133452e-09
```

The probability is extremely small: 3.133452e-09 .

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

```
xmin <- (p440_mu - 3*p440_sigma)
xmax <- (p440_mu + 3*p440_sigma)
xrange <- xmin:xmax
ypopulation <- dnorm(x = xrange, mean = p440_mu, sd = p440_sigma)
```

```r
ysample15 <- dnorm(x = xrange, mean = p440_mu, sd = p440_sigma/sqrt(15))
par(mfrow = c(2, 1))
plot(xrange,ypopulation, type = "l", col="blue", main = "Distribution of lifespan of single lightbulb",
plot(xrange, ysample15, type="l", col="red", main = "Distribution of average lifespan of 15  lightbulbs
```



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

The analysis for part (a) requires that the distribution be Normal, and a Normal distribution has zero skew. If the Normal distribution is not appropriate then the estimate would not be valid.

The analysis for part (c) does not require that the underlying distribution be Normal, but it requires that the sample size be "large enough" to avoid the effects of skew in the underlying distribution. This is generally considered to be of size N = 30 or more. Because we are only averaging the lifespans of 15 lightbulbs, the Central Limit Theorem may not hold, resulting in estimates which are not valid.

#.

Exercise 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. Under the initial (incorrecly-formulated test) the standard error was calculated based upon the incorrect assumption that n = 50. This calculation resulted in a p-value of 0.08. For a given significance level such as alpha = 0.05, this p-value would have appeared to be too large to reject the null hypothesis in favor of the alternative. However, this incorrect p-value would have stemmed from a smaller-than expected Z-score, which resulted from a larger-than-expected Standard Error (SE).

As the Standard Error decreases with the reciprocal of the square root of the increase in the sample size, increasing the sample size by a factor of 10 would cause the SE to decrease by a factor of sqrt(10) = 3.162 .

Correcting the p-value by a factor of 10 will cause the SE to become smaller by a factor of sqrt(10) which will cause the Z-score to become larger by a like proportion. This will result in a substantially smaller p-value. Indeed, the corrected p-value may be close to zero, resulting in the rejection of the Null hypothesis in favor of the alternative.

```
bad_pvalue <- 0.08
bad_ZScore <- qnorm(1-bad_pvalue)
bad_ZScore
```

```
## [1] 1.405072
```

```
### Because the SE will decrease by  factor of sqrt(10), the ZScore will increase by such a factor
good_ZScore <- sqrt(10) * bad_ZScore
good_ZScore
```

```
## [1] 4.443226
```

```
good_pvalue <- 1- pnorm(good_ZScore)
good_pvalue
```

```
## [1] 4.430991e-06
```