Homework 6: Confidence Intervals

NAME: Andres Castano Zuluaga

NETID: ac986

DUE DATE: October 25, 2016 by 1:00pm

For this homework, it will be helpful to have a copy of the knitted version of this document to answer the questions as much of it is written using mathematical notation that may be difficult to read when the document is not knitted.

Instructions

For this homework:

1. All calculations must be done within your document in code chunks. Provide all intermediate steps.
2. Incude any mathematical formulas you are using for a calculation. Surrounding mathematical expresses by dollar signs makes the math look nicer and lets you use a special syntax (called latex) that allows for Greek letters, fractions, etc. Note that this is not R code and therefore should not be put in a code chunk. You can put these immediately before the code chunk where you actually do the calculation.

Some Notation

Your solutions to the problems below must include the formula used for each calculation. To get you started, here is some mathematical expressions written in latex that you may find helpful when writing out the math in your answers. You can copy, paste, and edit these expressions as needed.

1. ($\bar{x}_n$ - $SE(\bar{X}_n) \times z_{\alpha/2}$, $\bar{x}_n$ + $SE(\bar{X}_n) \times z_{\alpha/2}$)
2. $n\hat{p}_n \geq 10$ and $n(1-\hat{p}_n) \geq 10$
3. ($\hat{p}_n$ - $SE(\hat{p}_n) \times z_{\alpha/2}$, $\hat{p}_n$ + $SE(\hat{p}_n) \times z_{\alpha/2}$)
4. ($\bar{x}_n$ - $SE(\bar{X}_n) \times t_{n-1,\alpha/2}$, $\bar{x}_n$ + $SE(\bar{X}_n) \times t_{n-1,\alpha/2}$)
5. $SE(\bar{X}_n) = \sigma/\sqrt{n}$
6. $s/\sqrt{n}$
7. $SE(\hat{p}_n)$ estimated by $\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$

In this homework we will practice creating confidence intervals for a population mean, μ.

Problem 1

A random sample of 110 lightning flashes in a certain region resulted in a sample average radar echo duration of .81 sec and a sample standard deviation of .34 sec.

a) What distribution should be used to determine the quantiles needed to calculate a 95% confidence interval for mean radar echo duration? Why?

In this example I consider that we should use the Standard Normal Distribution given that the Central Limit Theorem applies: with have a random sample (independent observations), n>30 and despite that we do not much about the skewness, its effects are less notorious when n is large (in our case a n of 110 seems reasonable). Also, it is important to note that with n>30 t_{n-1} is close to N(0,1) is reasonable to use the sample standard deviation (s_n) as the same that the population standard deviation (σ). This is important given that σ is unkown in our problem.

b) Calculate a 95% confidence interval for mean radar echo duration. Interpret this confidence interval in terms of the study.

The formula to get a $100(1-\alpha)\%$ confidence interval is:

($\bar{x}_n$ - $SE(\bar{X}_n)$ $\times$ $z_{\alpha/2}$, $\bar{x}_n$ + $SE(\bar{X}_n)$ $\times$ $z_{\alpha/2}$) = 1 - α. Given that we are interested in a 95% confidence interval, and that $SE(\bar{X}_n)$=$\sigma/\sqrt{n}$=$s/\sqrt{n}$, then the above expression becomes:

($\bar{x}_n$ - $s/\sqrt{n}$ $\times$ $z_{0.025}$, $\bar{x}_n$ + $s/\sqrt{n}$ $\times$ $z_{0.025}$) $\approx$ 95%

In R we get the lower and upper bounds for our 95% confidence interval as follows:

```r
n = 110
smean = 0.81
sd = 0.34
se = sd/sqrt(n)
quant = -qnorm(.025)
lower = smean - se*quant
upper = smean + se*quant
lower
```

```
## [1] 0.7464624
```

```r
upper
```

```
## [1] 0.8735376
```

Our 95% confidence interval is (0.746 , 0.8735). Which means that we are 95% confident that the population mean for radar echo duration is between 0.746 seconds and 0.8735 seconds.

Problem 2

Ten recently sold houses were randomly selected from Canton, NY. For each house, the sale price was recorded. The data are in the *HomesForSaleCanton.csv* file in the folder for homework 6 on blackboard. Assume the cost of homes in Canton, NY is normally distributed.

Initially we are going to read the data and calculate the sample mean and the sample standard deviation as follows:

```r
SaleCanton <- read.csv("HomesForSaleCanton.csv")
dim(SaleCanton)
```

```
## [1] 10  1
```

```r
mean_price=mean(SaleCanton$Price)
mean_price
```

```
## [1] 146.8
```

```r
sd_price=sd(SaleCanton$Price)
sd_price
```

```
## [1] 94.99801
```

a) What distribution should be used to determine the quantiles needed to calculate a 99% confidence interval for mean cost of a home in Canton, NY? Why?

The distribution that should be used is the t-student distribution for four reasons: observations are independent, the cost of homes in Canton (NY) is normally distributed, n<30, and the population variance (σ) is unknow.

b) Compute a 99% confidence interval for the average cost of a home in Canton, NY. Interpret this confidence interval in terms of the study. Use code chunks to compute all of the values needed for the confidence interval and to compute the lower and upper bounds of the confidence interval.

Our 99% interval take the following form:

$(\bar{x}_n$ - $s/\sqrt{n} \times t_{9,0.005}$, $\bar{x}_n + s/\sqrt{n} \times t_{9,0.005}) \approx$ 99%

```r
n=10
df=n-1
se_mean_prices = sd_price/sqrt(n)
quant = -qt(.005,df)
quant
```

```
## [1] 3.249836
```

```r
lower = mean_price - se_mean_prices*quant
upper = mean_price + se_mean_prices*quant
lower
```

```
## [1] 49.17166
```

```r
upper
```

```
## [1] 244.4283
```

Our 99% confidence interval is (49.17 , 244.42). Which means that we are 99% confident that the population mean for the prices of sold houses in Canton (NY) is between 49.17 and 244.42 (not units of measure specified in the problem).

c) How would this confidence interval change if the sample mean and variance are unchanged, but the sample size is 35?

Intuitively if n change from 10 to 35 the Standard error of the mean will be smaller given that the expression $s/\sqrt{n}$ decreases, and also, the t-value associate with a $p<0.005$ decreases. Ultimately this leads to an interval more precise given that lower bound is greater and the upper bound is lower. Numerically the new interval is:

```
n=35
df=n-1
se_mean_prices = sd_price/sqrt(n)
quant = -qt(.005,df)
quant
```

```
## [1] 2.728394
```

```
lower = mean_price - se_mean_prices*quant
upper = mean_price + se_mean_prices*quant
lower
```

```
## [1] 102.9885
```

```
upper
```

```
## [1] 190.6115
```

After the increase of n from 10 to 35 our 99% confidence interval is (102.98 , 190.65) which certainly seems more precise that the calculated for n=10. In this case, we are 99% confident that the population mean for the prices of sold houses in Canton (NY) is between 102.98 and 190.65 (not units of measure specified in the problem).

Problem 3

The data set *ICUAdmissions.csv* includes information on 200 randomly selected patients admitted to the Intensive Care Unit (ICU). One of the variables, *Status* indicates whether the patient lived (`Status=0`) or died (`Status = 1`).

a) Based on these data create a 99% confidence interval for the survival rate of ICU patients. Include all calculations in a code chunk and interpret this interval in the context of the study.

Initially we are going to read the data and calculate the survival rate as survival rate = (patients with Status=0 / total patients)

```
ICU <- read.csv("ICUAdmissions.csv")
dim(ICU)
```

```
## [1] 200  21
```

```
table(ICU$Status)
```

```
##
##   0   1
## 160  40
```

```r
proportions=table(ICU$Status) / length(ICU$Status)
proportions
```

```
##
##   0   1
## 0.8 0.2
```

```r
survival_rate=proportions[1]
survival_rate
```

```
##   0
## 0.8
```

In this problem the survival rate is the proportion of patients (from the sample of 200) who lived after being admitted to the Intensive Care Unit. We are interested in the estimation of an interval for the true survival rate. By the Central Limit Theorem we know that if n is large enough and a random sample X_n is taken from a Bernoulli(p) distribution with mean=p, then:

$\bar{X}_n = \hat{p}_n \approx N(p, \sqrt{p(1-p)/n})$

The CLT applies if:

1) $np \geq 10$ and
2) $n(1-p) \geq 10$

Given that we do not know the true mean p, we are going to use the sample mean $\hat{p}_n$ as follows:

1) $n\hat{p}_n \geq 10$ and
2) $n(1-\hat{p}_n) \geq 10$

So, if (1) and (2) hold, quantiles from the standard normal distribution can be used to construct 100(1-α)% confidence intervals for p. in our case we know that $\hat{p}_n = 0.8$ and n=200, then:

1) $n\hat{p}_n = (200) * (0.8) = 160$
2) $n(1-\hat{p}_n) = (200) * (1-0.8) = 40$

We meet the conditios to use the quantiles of the normal standard distribution to construct a 99% confidence interval for the true proportion of patients who lived after being admitted to the ICU (p) :

($\hat{p}_n$ - $SE(\hat{p}_n) \times z_{0.005}$, $\hat{p}_n$ + $SE(\hat{p}_n) \times z_{0.005}$) $\approx$ 99%

with $SE(\hat{p}_n) = \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$

```r
n=200
p=0.8
SE_p=sqrt((p*(1-p))/n)
quant = -qnorm(.005)
lower = p - SE_p*quant
upper = p + SE_p*quant
lower
```

```
## [1] 0.7271445
```

```
upper
```

```
## [1] 0.8728555
```

Our 99% confidence interval is (0.726 , 0.893). Which means that we are 99% confident that the true proportion of patients who lived after being admitted to the ICU is between 0.726 and 0.893.

b) What assumptions are necessary for the confidence interval determined in (a) to be valid? Provide evidence that each of these assumptions is reasonable.

This question is answered above, we need to meet three requirements:

1. n large
2. $n\hat{p}_n \geq 10$
3. $n(1-\hat{p}_n) \geq 10$

In our case n=200, $n\hat{p}_n = 160$ and $n(1-\hat{p}_n) = 40$, so we meet the conditions to use que quatiles of the normal standard distribution and construct the 99% confidence interval for the true proportion of patients that lived after being admitted to the ICU (p).

Problem 4

A sample of 14 joint specimens of a particular type gave a sample mean proportional limit stress of 8.48 MPa. Assume proportional limit stress is approximately normally distributed. The variance of proportional limit stress is known to be $\sigma^2 = .6241$.

a) Calculate and interpret a 95% confidence interval for mean proportional limit stress.

Here we can treat the mean proportional limit stress as normal because the data distribution is itself normal and we know the population variance (σ^2). So, our interval of interest is:

($\bar{x}_n$ - $SE(\bar{X}_n) \times z_{0.025}$, $\bar{x}_n + SE(\bar{X}_n) \times z_{0.025}$) $\approx$ 95%

```
n = 14
sample_mean = 8.48
sample_sd = sqrt(0.6241)
SE_mean = sample_sd/sqrt(n)
SE_mean
```

```
## [1] 0.2111364
```

```
quant = -qnorm(.025)
quant
```

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

```
lower = sample_mean - SE_mean*quant
upper = sample_mean + SE_mean*quant
lower
```

```
## [1] 8.06618
```

```
upper
```

```
## [1] 8.89382
```

Our 95% confidence interval is (8.066 , 8.893). Which means that we are 95% confident that the true mean of the proportional limit stress of the specimens is between 8.06 MPa and 8.893 MPa.

b) Without recalculating this interval, explain how the width of this confidence interval would change if the confidence level was set to 90%.

If the confidence interval is set to a 90% the quantile of the normal standard distribution will be in absolute value lower, then the lower bound of the interval is going to increase and upper bound of the interval is going to decrease (compared to the situation with 95% confidence), which ultimately leads to a narrower interval of confidence interval. In common words we are less sure about if our interval will contain the true population mean compared to the case with 95% confidence.

Problem 5

United Airlines Flight 179 is non-stop from Boston's logan airport to San Francisco International Airport. An important factor in scheduling flights is the actual airborne flying time from takeoff to touchdown. In the data file, *FlightData.csv*, is the airborne time in minutes for 3 dates each month in the year 2010 for this flight and one other flight. The data of interest can be found in the column named `Flight179`. Create a 95% confidence interval for the mean airborne time for this flight based on the data and interpret it in terms of the study. Also, give support for how you chose the quantiles to calculate this interval.

Initially we are going to read the data:

```
flight <- read.csv("FlightData.csv")
dim(flight)
```

```
## [1] 36  3
```

```
library(Hmisc)
```

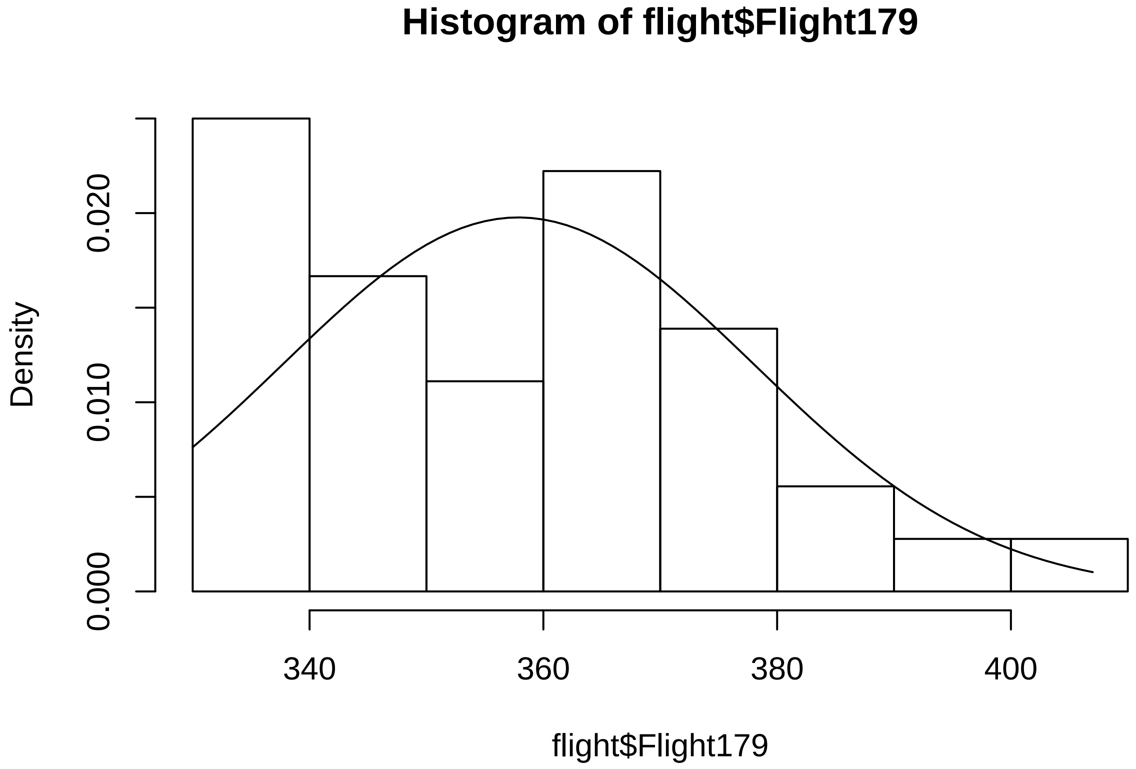
```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```r
describe(flight)
```

```
## flight
##
##  3  Variables      36  Observations
## ---------------------------------------------------------------------------
## Date
##       n missing  unique
##      36       0      36
##
## lowest : 01/05/2010 01/15/2010 01/25/2010 02/05/2010 02/15/2010
## highest: 11/15/2010 11/25/2010 12/05/2010 12/15/2010 12/25/2010
## ---------------------------------------------------------------------------
## Flight179
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      36       0      28       1   357.9   330.0   331.0   341.5   358.5
##     .75     .90     .95
##   370.2   380.5   391.2
##
## lowest : 330 332 334 335 339, highest: 374 387 390 395 407
## ---------------------------------------------------------------------------
## Flight180
##       n missing  unique    Info    Mean     .05     .10     .25     .50
##      36       0      30       1   301.5   281.8   285.5   289.8   301.0
##     .75     .90     .95
##   312.5   322.0   324.2
##
## lowest : 280 281 282 285 286, highest: 318 322 323 328 329
## ---------------------------------------------------------------------------
```

Before deciding which quantiles use is important to know how the actual airbone time is distributed:

```r
sd_179=sd(flight$Flight179)
mean_179=mean(flight$Flight179)
hist(flight$Flight179, breaks=10, freq=FALSE)
xvalues = seq(330,407,1)
yvalues = dnorm(xvalues,mean_179, sd_179)
lines(xvalues,yvalues)
```



As we can observed the actual airborne time does not seem to follow a normal distribution even with a n=36. However, the skewness is not as bad as it can be. Also is important to note that the observed distribution of the actual airborne time has heavier tails compared to the theoretical normal distribution. Given this description it seems reasonable to use the the quantiles of the standard normal distribution for two reasons:

1.) Despite that we do not know if the database used is based on a Simple Random Sampling (so we can not assure the independence of the observations), the histogram shows evidence that the data is not extremely skewed. So we can be reasonable confident that if n increases the sample mean of the actual airborne time is going to follow a normal distribution.

2.) In this problem we do not know the population variance of the actual airborne times (σ^2) but given that n>30 we can use its sample counterpart (S^2) to estimate the Standard deviation of the mean with no much difference in the case of the normal standard 95% confidence interval, compared to the 95% t-student interval used when σ^2 is unknow.

After this explanations, our 95% confidence interval is:

$(\bar{x}_n$ - $s/\sqrt{n} \times z_{0.025}$, $\bar{x}_n + s/\sqrt{n} \times z_{0.025}) \approx 95\%$

```
n=36
sd_179=sd(flight$Flight179)
mean_179=mean(flight$Flight179)
SE_mean = sd_179/sqrt(n)
quant = -qnorm(.025)
lower = mean_179 - SE_mean*quant
upper = mean_179 + SE_mean*quant
lower
```

```
## [1] 351.2698
```

```
upper
```

```
## [1] 364.4525
```

Our 95% confidence interval is (351.27 , 364.45). Which means that we are 95% confident that the true mean of actual airborne times is between 351.27 minutes and 364.45 minutes.

As an additional exercise we have calculated the same interval based on the quantiles of the t-student distribution as follows:

```
n=36
sd_179=sd(flight$Flight179)
mean_179=mean(flight$Flight179)
SE_mean = sd_179/sqrt(n)
quant = -qt(.025, 36)
lower = mean_179 - SE_mean*quant
upper = mean_179 + SE_mean*quant
lower
```

```
## [1] 351.0406
```

```
upper
```

```
## [1] 364.6816
```

As we can see, there is not huge differences in the lower and upper bounds of the 95% confidence intervals based on the quantiles of the standard normal distribution and the quantiles of t-student distribution.