Homework 4

Erik Lee

Mon Jul 23 20:38:38 2018

Problem 1

An insurance company is interested in the average claim on its auto insurance policies. Using 40 randomly selected claims, it finds the mean claim to be $1,270 with a standard deviation of $421. Construct a 95 percent confidence interval for the mean claim on all policies.

95% CI = [1140,1400]

```
# Using the 95% CI equation for population mean:
# n = 40, x_bar = 1270, s.d.=421
# lower bound:
1270-1.96*(421/sqrt(40))
```

```
## [1] 1140
```

```
# upper bound:
1270+1.96*(421/sqrt(40))
```

```
## [1] 1400
```

Problem 2

A random sample of the luggage of 53 passengers of Jet Blue finds that the mean weight of the luggage is 47 pounds with a standard deviation of 8 pounds. Construct a 95 percent confidence interval for the mean weight of Jet Blue Airlines luggage.

95% CI = [44.85,49.15]

```
# n = 53, x_bar = 47, s.d. = 8
# lower bound:
47-1.96*(8/sqrt(53))
```

```
## [1] 44.85
```

```
# upper bound:
47+1.96*(8/sqrt(53))
```

```
## [1] 49.15
```

Problem 3

A random sample of 250 credit card holders shows that the mean annual credit card debt for individual accounts is $1600 with a standard deviation of $997. Use this information to construct a 92% (yes that is not a typo) confidence interval for the mean annual credit card debt for the population of all accounts.

92% CI = [1490,1710]

```
# alpha/2 = (1-0.92)/2 = 0.08/2 = 0.04
# z_alpha/2 = z_0.04 = 1.75 (z score table)
# n = 250, x_bar = 1600, s.d. = 997
# lower bound:
1600-1.75*(997/sqrt(250))
```

```
## [1] 1490
```

```
# upper bound:
1600+1.75*(997/sqrt(250))
```

```
## [1] 1710
```

Problem 4

For this problem we are going to use class survey data from a previous offering of Stat 111. Enter the following commands into R:

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/stat111_survey.csv")

weight=mydata$weight
female=mydata$female
sleep=mydata$sleep
haircut=mydata$haircut
texts=mydata$texts
```

Note that we can find confidence intervals in R using this data as follows. For number of texts someone sends a day:

```
t.test(texts) ## ci for everyone
```

```
##
##  One Sample t-test
##
## data:  texts
## t = 7.7, df = 89, p-value = 2e-11
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  29.00 49.24
## sample estimates:
## mean of x
##     39.12
```

```
t.test(texts[female==1]) ## ci for just females
```

```
##
##  One Sample t-test
##
## data:  texts[female == 1]
## t = 5.4, df = 33, p-value = 5e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  22.20 48.83
## sample estimates:
## mean of x
##     35.51
```

```
t.test(texts[female==0]) ## ci for just males
```

```
##
##  One Sample t-test
##
## data:  texts[female == 0]
## t = 5.7, df = 55, p-value = 4e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  26.91 55.70
## sample estimates:
## mean of x
##      41.3
```

a) Find a 95% confidence interval for the sleep variable for men and women separately. Compare the results. Are you inside your respective interval?

```
# 95% CI for men's sleep
t.test(sleep[female==0])
```

```
##
##  One Sample t-test
##
## data:  sleep[female == 0]
## t = 7.9, df = 55, p-value = 1e-10
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   6.15 10.31
## sample estimates:
## mean of x
##     8.232
```

```
# 95% CI for women's sleep
t.test(sleep[female==1])
```

```
##
##  One Sample t-test
##
## data:  sleep[female == 1]
## t = 37, df = 33, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  7.097 7.932
## sample estimates:
## mean of x
##     7.515
```

CI for Men's Sleep = [6.15,10.31] CI for Women's Sleep = [7.097,7.932]

By subsetting the data for the sleep for men and women and applying the t.test() function to each, we find the 95% CI for men and women. When looking at the sleep for men (female==0), the 95% CI is between 6.15 and 10.31 hours. I would say I fit into this confidence interval. On week days I average about 7 hours of sleep and 9 for weekends. Both these values fit in the CI range.

b) The variable haircut is what do you usually pay for a haircut. Find a 95% confidence interval for this variable for men and women separately. Do the intervals appear that different?

```r
# Men's haircut cost CI
t.test(haircut[female==0])
```

```
##
##  One Sample t-test
##
## data:  haircut[female == 0]
## t = 6.6, df = 56, p-value = 2e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  15.98 29.93
## sample estimates:
## mean of x
##     22.96
```

```r
# Women's haircut cost CI
t.test(haircut[female==1])
```

```
##
##  One Sample t-test
##
## data:  haircut[female == 1]
## t = 5.9, df = 32, p-value = 2e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  33.75 69.65
## sample estimates:
## mean of x
##      51.7
```

CI for Men's Haircuts = [15.98,29.93] CI for Women's Haircuts = [33.75,69.65]

Yes the Confidence Intervals for men and women's hair cut costs are different. For each 95% CI, men (female==0) have an interval between \$15.98 and \$29.93. Women (female==1) have an interval between \$33.75 and \$69.65. Women have an interval of almost double the cost as compared to men.

c) Find a 95% confidence interval for the variable texts, the number of texts you send per day. Are you inside this interval? Do the separate intervals for men and women differ that much?

```r
# 95% CI for men's texting
t.test(texts[female==0])
```

```
##
##  One Sample t-test
##
## data:  texts[female == 0]
## t = 5.7, df = 55, p-value = 4e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  26.91 55.70
## sample estimates:
## mean of x
##      41.3
```

```r
# 95% CI for women's texting
t.test(texts[female==1])
```

```
##
```

```
##  One Sample t-test
##
## data:  texts[female == 1]
## t = 5.4, df = 33, p-value = 5e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  22.20 48.83
## sample estimates:
## mean of x
##     35.51
```

CI for Men's Texting = [26.91,55.70] CI for Women's Texting = [22.20,48.83]

Comparing to the 95% CI for men, I send an average of 15 texts per day. This does not fall into the CI for men between 26.91 and 55.70 texts.

The intervals for men and women are faily similar. Men have an interval between 26.91 and 55.70. Women have an interval between 22.20 and 48.83. The interval is a bit wider for men, but the ranges are close, men's CI have a range of 28.79 and women's CI have a range of 26.63.

Problem 5

The paralyzed Veterans of America is a philanthropic organization that relies on contributions. They send free mailing labels and greeting cards to potential donors on their list and ask for voluntary contribution. To test a new campaign they recently sent letters to a random sample of 100,000 potential donors and received 4781 donations.

a) Give a 95% confidence interval for the true proportion of those from their entire mailing list who may donate.

95% CI = [0.04648,0.04912]

```
# sucess = 4781, n = 100,000, p_hat = 4781/100000 = 0.0478
# lower bound:
0.0478-1.96*sqrt((0.0478*(1-0.0478))/100000)
```

```
## [1] 0.04648
```

```
# upper bound:
0.0478+1.96*sqrt((0.0478*(1-0.0478))/100000)
```

```
## [1] 0.04912
```

b) A staff member thinks that the true rate is 5%. Given the confidence interval you found, do you find that percentage plausible?

No, a 5% true rate is not plausible based on the 95% confidence interval found in Part A. The 95% confidence interval is between 4.648% and 4.912%. A 5% true return falls outside the CI so the likelihood given the sample statistics does not hold.

Problem 6

A recent Gallup poll consisted of 1012 randomly selected adults who were asked whether "cloning of humans should or should not be allowed." Results showed that 901 of those surveyed indicated that cloning should not be allowed. Construct a 95% confidence interval estimate of the proportion of adults believing that cloning of humans should not be allowed.

95% CI = [0.971,0.9096]

```r
# n = 1012, sucess = 901, p_hat = 901/1012 = 0.8903
# lower bound =
0.8903-1.96*sqrt((0.8903*(1-0.8903))/1012)
```

```
## [1] 0.871
```

```r
# upper bound =
0.8903+1.96*sqrt((0.8903*(1-0.8903))/1012)
```

```
## [1] 0.9096
```

Problem 7

A national health organization warns that 30% of middle school students nationwide have been drunk. Concerned, a local health agency randomly and anonymously surveys 110 of the middle 1212 middle school students in its city. Only 21 of them report having been drunk.

a) What proportion of the sample reported having been drunk?

p_hat = 21/110 = 0.1909

b) Does this mean that this city's youth are not drinking as much as the national data would indicate?

Not necessarily. This is just one survey of 110 students from the middle school. Even though the percentage of students reported drunk of 19.09% is much lower than the reported 30%, there is no saying what the percentage would be for subsequent samples of middle school students nor what the actual percentage is for the whole school, unless a census was taken. It may be the case that this particular random sample of students has a lower percentage or this may accurately reflects the actual percentage for the school. More data and evidence must be collected to better compare this city's youth with the national average.

c) Create a 95% confidence interval for the proportion of the city's middle school students who have been drunk.

95% CI = [0.1175,0.2643]

```r
# n = 110, sucess = 21, p_hat = 0.1909
# lower bound:
0.1909-1.96*sqrt((0.1909*(1-0.1909))/110)
```

```
## [1] 0.1175
```

```r
# upper bound:
0.1909+1.96*sqrt((0.1909*(1-0.1909))/110)
```

```
## [1] 0.2643
```

d) Is there any reason to believe that the national level of 30% is not true of the middle school students in the city?

Yes, based on the 95% confidence interval for this random sample, the percentage of middle school students who have been drunk falls between 11.75% and 26.43%. The 30% figure is not within this interval and is high compared to the upper bound of the confidence interval.

It is important to notice that the national level of 30% represents the larger population of middle school students. However, based on this survey of 110 from a specific city's middle school of 1212 students, the 30% is not represented from the data's confidence interval. And while no other evidence/data supports the city school matching the national average, we can see that this middle school has a lower average and confidence interval indicates a lower percentage of students reported drunk.

e) To keep the margin of error at most 5%, how many middle school students do we need to survey? Assume we have no prior idea what the true proportion is.

n = 237.3 students

```
# n = (1.96^2)*(0.1909)*(1-0.1909)/e^2, e = 0.05
((1.96^2)*(0.1909)*(1-0.1909))/(0.05^2)
```

```
## [1] 237.3
```

Problem 8

A researcher wishes to be 95% confident that her estimate of the true proportion of individuals who travel overseas is within 3% of the true proportion.

a) Find the sample necessary if, in a prior study, a sample of 200 people showed that 40 traveled overseas last year.

n = 683 individuals

p = 40/200 = 0.2 z_alpha/2 = 1.96 for 95% CI

```
#sample necessary n=((1.96^2)(0.2)(1-0.2))/(e^2), e=0.03
((1.96^2)*(0.2)*(1-0.2))/(0.03^2)
```

```
## [1] 683
```

b) If no estimate of the sample proportion is available, how large should the sample be?

n = 1067 individuals

This is an example of the worse-case scenario for p_hat, where p_hat=0.5.

```
# n = (1.96^2)*(0.5)*(0.5)/e^2, e = 0.03, p_hat=0.5
(1.96^2)*(0.5)*(0.5)/(0.03^2)
```

```
## [1] 1067
```

Problem 9

Obesity is defined as a body mass index (BMI) of 30 kg/m2 or more. A 95% confidence interval for the percentage of U.S. adults aged 20 years and over who were obese was found to be 22% to 24%. What was the sample size?

n = sample size = 68.03 people

95% CI = [0.22,0.24] p_hat = 0.23 e = error = 0.1

```
# p_hat=0.23, e=0.1
# n = ((1.96^2)*(p)*(1-p))/(e^2)
# n = ((1.96^2)*(0.23)*(1-0.23))/(0.1^2)
((1.96^2)*(0.23)*(1-0.23))/(0.1^2)
```

```
## [1] 68.03
```

Problem 10

When 14 different second-year medical students at Bellevue Hospital measured the blood pressure of the same person, they obtained the results listed below. You can read this data into R by entering the command:

```
mydata=c(138, 130, 135, 140, 120, 125, 120, 130, 130, 144, 143, 140, 130, 150)
```

a. Using R, find the 95% confidence interval for the mean blood pressure (use the t.test command).

95% CI = [128.7,139.1]

```
t.test(mydata)
```

b. By hand, and using the t distribution, find the 95% confidence interval for the mean score. You can use the summary statistics from R. In R, the command qt(.975,df) will calculate the appropriate t cut-off value, where df=n-1.

95% CI = [128.7,139.1]

mean=133.9 sd=9.042

df = 14-1 = 13 t_0.975 = 2.160

Lower Bound: 133.9-2.160*(9.042/(sqrt(14))) = 128.7

Upper Bound: 133.9+2.160*(9.042/(sqrt(14))) = 139.1

c. By hand, and using the normal distribution, find the 95% confidence interval for the mean score (i.e. use "1.96"). You can use the summary statistics from R.

95% CI = [129.2,138.6]

mean=133.9 sd=9.042

z_alpha/2 = 1.96

Lower Bound: 133.9-1.96*(9.042/(sqrt(14))) = 129.2

Upper Bound: 133.9+1.96*(9.042/(sqrt(14))) = 138.6

d. Discuss the difference between (a) and (b) and (c).

Part A and B have the exact same confidence interval between 128.7 and 139.1. This makes sense because both apply the t-test/t-distribution where the df=13 and t-value=2.160. Part C a confidence interval between 129.2 and 138.6.

Part C has a narrower confidence interval compared to the previous two intervals. Since the error, e, is being multiplied by 1.96, the interval is smaller. This normal distribution accounts for the scenario where the sample size n is large, greater than 30. But since the sample size is 14, it is not appropriate to use a normal distribution to find the 95% confidence interval.

Part A and B, using the t-test, has a wider confidence interval. This accounts for having such a small sample size to test from. In this case, since the sample size is 14, it is more approporiate to use the t-test results to provide a better confidence interval for this distribution.

Problem 11

Answer true or false to the following statement and give a reason for your answer: If a 95% confidence interval for a population mean, μ, is from 33.8 to 39.0, the mean of the population must lie somewhere between 33.8 and 39.0.

False. It is not guaranteed that this particular sample mean and confidence interval falls within the group of 95% of confidence intervals containing the population mean. All we know is that this particular interval has a range between 33.8 and 39.0. This either could be an one of the 95% of intervals containing the population mean or it could be part of the 5% of intervals without the population mean. Since we are not sure about this particular confidence interval's relation or the population mean value, we cannot assume that this statment is true.

Problem 12

If you obtained one thousand 95% confidence intervals for a population mean, μ, roughly how many of the intervals would actually contain μ?

Roughly 950 of the 1000 confidence intervals contain mu.

Problem 13

1) A worker at a car manufacturer invented a new device that he believes will increase gas mileage. The current car averages 28 miles per hour. The CEO decides to put the new device on 100 of its vehicles and measure the average from that sample. If the average gas mileage from the 100 cars is significantly greater than the current average of 28, the CEO will buy 100,000 devices for its new line of cars.

a) Is this a one or two tailed test? Explain.

This is a one-tailed test. The scenario looks at the average (mean) gas milage of a sample set of 100 cars. The CEO wants to know if adding a new device increases gas mileage. Seeing if the mean mileage (mu) is greater than the current mean (mu=28) is an example of a right-tailed test.

b) Write the null and alternative hypothesis.

H_o: mu = 28 H_a: mu > 28 One (right) tailed test

Null Hypothesis (H_o) = average gas mileage for new cars does not change with the new device

Alternative Hypothesis (H_a) = average mileage for new cars increases with the new device to a value greater than 28

c) In this context, what would happen if the CEO made a Type I error?

Type I is when the null hypothesis is rejected, but it is true. In this case based on the sample data, the CEO buys and installs the 100,000 devices for the new cars, when in fact these do not increase the average gas mileage of the new cars.

d) In this context, what would happen if the CEO made a Type II error?

Type II is when the the null hypothesis is accepted, but the alternative hypothesis is true. In this case based on the sample data, the CEO passes on buying the 100,00 devices for new cars, but mileage is significantly greater when these devices were installed in new cars.

Problem 14

Each of the following paragraphs calls for a statistical test about a population mean μ. State the null hypothesis Ho and the alternative hypothesis Ha in each case.

a. The diameter of a spindle in a small motor is supposed to be 5 mm. If the spindle is either too small or too large, the motor will not work properly. The manufacturer measures the diameter in a sample of motors to determine whether the mean diameter has moved away from the target.

H_o: mu = 5mm H_a: mu != 5mm

Null Hypothesis (H_o) = mean diameter of the small motor spindle has not moved away from the target

Alternative Hypothesis (H_a) = mean diameter of the small motor spindle has moved, either increased or decreased, away from the target

b. Census Bureau data show that the mean household income in the area served by a shopping mall is $42,500 per year. A market research firm questions shoppers at the mall. The researchers suspect the mean household income of mall shoppers is higher than that of the general population.

H_o: mu = $42,500 H_a: mu > $42,500

Null Hypothesis (H_o) = the mean household income of mall shoppers is the same as the mean of the general population at $42,500

Alternative Hypothesis (H_a) = the mean houshold income of mall shoppers is higher than that of the general population

c. A study in 2002 established the mean commuting distance for workers in a certain city to be 15 miles. Because of the westward spread of the city, it is hypothesized that the current mean commuting distance exceeds 15 miles. A traffic engineer wishes to test the hypothesis that the mean commuting distance for workers in this city is greater than 15 miles.

H_o: mu = 15 miles H_a: mu > 15 miles

Null Hypothesis (H_o) = the mean commuting distance for workers in that city has not changed at 15 miles

Alternative Hypothesis (H_a) = the mean commuting distance for workers in that city is greater than 15 miles

---

Problem 15

The fundraising officer for a charity organization claims the average donation from contributors to the charity is $250.00. To test the claim, a random sample of 100 donations is obtained. The sample yielded a sample mean of $234.85 and sample standard deviation of $95.23. State and run the appropriate hypothesis test using the confidence interval approach. Clearly state your conclusion.

Null Hypothesis (H_o) is that $250.00 is a plausible value for the average donation to this charity. By constructing a 95% confidence interval, we can determine if the value for the null hypothesis, in this case $250.00 as the average, falls withing the confidence interval. If this is true, then we can accept the null hypothesis. If this is false, we reject the null hypothesis. The alternative hypothesis (H_a) is that $250.00 is not a plausible value for the average donation to this charity.

H_o: mu = $250.00 H_a: mu != $250.00

```r
# Calculating a 95% confidence interval by hand
# n=100, x=234.85, sd=95.23
# lower bound
234.85-1.96*(95.23/sqrt(100))
```

```
## [1] 216.2
```

```r
# upper bound
234.85+1.96*(95.23/sqrt(100))
```

```
## [1] 253.5
```

```r
# Creating a confidence interval with BSDA package
# n=100, x=234.85, sd=95.23
# must install.packages("BSDA")
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```r
tsum.test(n.x=100, mean.x=234.85, s.x=95.23)
```

```
## Warning in tsum.test(n.x = 100, mean.x = 234.85, s.x = 95.23): argument
## 'var.equal' ignored for one-sample test.
```

```
##
##  One-sample t-Test
##
## data:  Summarized x
## t = 25, df = 99, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  216.0 253.7
## sample estimates:
## mean of x
##     234.8
```

95% CI = [216.0,253.7]

By calculating a 95% confidence interval, both by hand and using the tsum.test() function, we find that $250.00 falls within the confidence interval between $216 and $253. Based on this finding, we do not reject the null hypothesis as the the value for the average since this value is found within the confidence interval making the value plausible. So the conclusion is that $250.00 is a plausible value for the average donation from contributors to a charity.

Problem 16

You want to test whether your candidate's approval rating has changed from the previous dismal 40% after a major policy announcement. You run a survey and 170 out of a random sample of 500 voters approve of your candidate. ($\hat{p} = 34\%$). Construct a hypothesis test using a two sided confidence interval to test if the approval rating is now different from 40%. Clearly state your conclusion

The Null Hypothesis is that it is not plausible the 40% approval rating has not changed after a major policy announcement. The Alternative Hypothesis is that it its plausible the 40% approval rating changed after a major policy announcement. Using a two sided confidence interval, we can test if the value for the null hypothesis falls within a 95% confidence interval. If this is true, we accept the null hypothesis, and if this is false, we reject the null hypothesis. Since this confidence itnerval is two sided, the change could either be an increase or decrease in rating percentage for the alternative hypothesis.

H_o: p = 0.40 H_a: p != 0.40

```r
# Calculating a 95% confidence interval by hand
# n=500, p=0.34
# lower bound
0.34-1.96*sqrt((0.34)*(1-0.34)/500)
```

```
## [1] 0.2985
```

```
# upper bound
0.34+1.96*sqrt((0.34)*(1-0.34)/500)
```

```
## [1] 0.3815
```

```
# Creating a confidence interval prop.test() function
# x=170, n=500
prop.test(170,500)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  170 out of 500, null probability 0.5
## X-squared = 51, df = 1, p-value = 1e-12
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2989 0.3836
## sample estimates:
##    p
## 0.34
```

95% CI = [0.2989,0.3836]

Based on the results of calculating CI and prop.test() function, the value of 0.40 does not fall within the range of the 95% confidence interval between 0.29 and 0.38. From this, we can conclude that 0.40 or 40% is not a plausible value within the confidence interval and we reject the null hypothesis because there is not enough evidence to support it, based on this sample data. As a result, we can accept the alternative hypothesis saying it is plausible the approval rating has changed after a major policy announcement. Based on the range of the confidence interval, the lower and upper bound are below 40%, indicating that this policy change may have lowered the approval rating.
