Stat 104: Quantitative Methods for Economics
Homework 6: Due Monday, October 23

1. An insurance company is interested in the average claim on its auto insurance policies. Using 36 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.

   1270±1.96*(421/sqrt(36))=(1132.47, 1407.53)

2. A random sample of the luggage of 49 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.

   47±1.96*(8/sqrt(49))=(44.76, 49.24)

3. A random sample of 250 credit card holders shows that the mean annual credit card debt for individual accounts is $1592 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.

   z-score for 4% in each tail is -1.75 and 1.75

   1592±1.75*997/sqrt(250)=(1481.65, 1702.35)

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

t.test(texts[female==1]) ## ci for just females
t.test(texts[female==0]) ## ci for just males
```

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

Female:

```
t.test(sleep[female==1])

   One Sample t-test

data:  sleep[female == 1]
t = 36.622, df = 33, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.097230 7.932182
sample estimates:
mean of x
 7.514706
```

Male:

```
t.test(sleep[female==0])

   One Sample t-test

data:  sleep[female == 0]
t = 7.9219, df = 55, p-value = 1.169e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  6.149609 10.314676
sample estimates:
mean of x
 8.232143
```

Student's answers will vary for the last part of the question.

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?

```
Male:
> t.test(haircut[female==0])

   One Sample t-test

data:  haircut[female == 0]
t = 6.5912, df = 56, p-value = 1.635e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 15.97921  29.93307
sample estimates:
mean of x
 22.95614

Female
> t.test(haircut[female==1])

   One Sample t-test
```

```
data:  haircut[female == 1]
t = 5.8664, df = 32, p-value = 1.599e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 33.74673  69.64721
sample estimates:
mean of x
 51.69697
```

The intervals are quite different. The intervals do not overlap and females pay more for haircut.

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?

Overall:

```
> t.test(texts)

        One Sample t-test

data:  texts
t = 7.6812, df = 89, p-value = 1.949e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 28.99792 49.23541
sample estimates:
mean of x
39.11667
```

Male:

```
> t.test(texts[female==0])

   One Sample t-test

data:  texts[female == 0]
t = 5.749, df = 55, p-value = 4.078e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 26.90563 55.70151
sample estimates:
mean of x
 41.30357
```

```
Female:

> t.test(texts[female==1])

   One Sample t-test

data:  texts[female == 1]
t = 5.4274, df = 33, p-value = 5.225e-06
```

```
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 22.20156 48.82785
sample estimates:
mean of x
 35.51471
```

The number of texts sent per day do not differ that much between males and females. Students' answer may vary.

5. In a newspaper or magazine of your choice (can be online), find a statistical study that contains an estimated population proportion. Include the article with your homework (via a cut and paste or copy of the article). Calculate a confidence interval for the sample proportion given.

Students' answers will vary. Check that student sets up the confidence interval for a proportion correctly.

6. 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.

      p.hat = 4781/100,000 = .04781
      95% CI = p.hat ±1.96* $\sqrt{\boldsymbol{p.hat} * (\mathbf{1} - \boldsymbol{p.hat})/\boldsymbol{n}}$
      = .04781 ±1.96*sqrt((0.04781*(1-0.04781))/100,000)
      = (0.046, 0.049)

      We are 95% confident that the true percentage of those from the mailing list who may donate is between 4.6% and 4.9%.

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

      Ho: p=0.05
      Ha: p not = 0.05
      Because 0.05 is outside the 95% confidence interval (even though it is quite close to the upper end of the confidence interval), the rate of 5% is not plausible.
      It is possible for answers to vary here if the upper bound was rounded to 0.05.

7. 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.

   P.hat = 901/1012 = 0.89
   CI = 0.89± 1.96*sqrt(0.89*0.11/1012) = (0.871, 0.909)

8. 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 = 0.19

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

      We cannot make a claim from just the sample proportion. We must find the 95% confidence interval.

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

      0.19±1.96*sqrt(0.19*(1-0.19)/110)=(0.12, 0.26)

   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, now we have reason to believe that the national level of 30% is not true of the middle school students in this city because the 95% confidence interval does not include 30%.

   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=$(1.96)^2$ * p.hat*(1-p.hat)/$(\text{margin of error})^2$

      n=$(1.96)^2(0.25)/(0.05)^2$ = 384.16 ~ 385

9. 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.

      We want the margin of error to be 0.03.

      n=$(1.96)^2$p.hat(1-p.hat)/(margin of error)^2

      n=$(1.96)^2(0.20)(1-0.20)/(0.03)^2$ = 682.95 ~ n must be at least 683

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

      Worst case scenario: p.hat=0.5
      n=$(1.96)^2$p.hat(1-p.hat)/(margin of error)^2

      n=$(1.96)^2(0.50)(1-0.50)/(0.03)^2$ = 1067.11 ~ n must be at least 1068

10. Obesity is defined as a body mass index (BMI) of 30 kg/m^2 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?

    p.hat = (0.22+0.24)/2 = 0.23

    0.24 = 0.23 + 1.96*sqrt(0.23*(1-0.23)/n)

    n = (0.23(1-0.23)/(0.24-0.23^2)/1.96)

    n = 6803.47 ~ 6804

11. When 14 different second-year medical students at Bellevue Hospital measured the blood pressure of the same person, they obtained the results listed below.

138 130 135 140 120 125 120 130 130 144 143 140 130 150

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).

```
> t.test(mydata)
    One Sample t-test

    data:  mydata
    t = 55.419, df = 13, p-value < 2.2e-16
    alternative hypothesis: true mean is not equal to 0
    95 percent confidence interval:
     128.7077          139.1495
    sample estimates:
    mean of x
     133.9286
```

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.

```
> mean(mydata)
[1] 133.9286
> sd(mydata)
[1] 9.04233

> qt(.975,13)

[1] 2.160369
```

mean = 133.93, sd=9.04
95%CI=(133.93-2.16*9.04/sqrt(14), 133.93+2.16*9.04/sqrt(14))
=(128.71, 139.15)

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=(133.93-1.96*9.04/sqrt(14), 133.93+1.96*9.04/sqrt(14))
=(129.19, 138.67)

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

The intervals from Part (a) and (b) are the same. This means that R uses the t-distribution by default when calculating confidence intervals. The interval calculated with the normal distribution is narrower than the interval calculated with the t-distribution.

12. 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. We are 95% confident that the true mean is in (33.8, 39.0). However, the true mean may not be included. Answers may vary.

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

95%*1000=950 interval

14. Suppose we know that "a 95% confidence interval for the mean age of all U.S. millionaires is from 54.3 years to 62.8 years." Decide which of the following sentences provide a correct interpretation of the statement in quotes. Justify your answer.

   a. Ninety-five percent of all U.S. millionaires are between the ages of 54.3 years and 62.8 years.
   b. There is a 95% chance that the mean age of all U.S. millionaires is between 54.3 years and 62.8 years.
   c. **We can be 95% confident that the mean age of all U.S. millionaires is between 54.3 years and 62.8 years.**
   d. The probability is 0.95 that the mean age of all U.S. millionaires is between 54.3 years and 62.8 years.

15. Why is a sample proportion generally used to estimate a population proportion instead of obtaining the population proportion directly?

The population proportion is usually not available to us. Especially when the population is large, we cannot collect data from the entire population. Student's answer may vary.

16. 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 because we are seeing if the average is greater than 28 miles and only looking at the right end of the tail.

   b. Write the null and alternative hypothesis.

      Ho: average gas mileage = 28 mph
      Ha: average gas mileage > 28 mph

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

      If the CEO made a Type I error, he would conclude that the average mpg is greater than 28 mpg when it actually is not, and would thus buy the new device.

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

      If the CEO made a Type II error, he would conclude that the average mpg is 28 mpg and fail to reject the null when the alternative is true. He would not buy the device when he should because it does increase gas mileage.

17. The real estate industry claims that it is the best and most effective system to market residential real estate. A survey of randomly selected home sellers in Illinois found that a 95% confidence interval for the proportion of homes that are sold by a real estate agent is 69% to 81%. Interpret the interval in this context.

a. In 95% of the years, between 69% and 81% of homes in Illinois are sold by a real estate agent.
b. 95% of all random samples of home sellers in Illinois will show that between 69% and 81% of homes are sold by a real estate agent.
c. If you sell a home in Illinois, you have a 75% ± 6% chance of using a real estate agent.
d. We are 95% confident that between 69% and 81% of homes in this survey are sold by a real estate agent.
e. We are 95% confident, based on this sample, that between 69% and 81% of all homes in Illinois are sold by a real estate agent.

18. Each of the following paragraphs calls for a statistical test about a population mean. 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.

The null hypothesis is that the average diameter of the spindle in the sample is 5 mm. The alternative hypothesis is that the average diameter of the spindles in the sample is not 5mm.
Ho:μ = 5mm, Ha:μ not = 5mm

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.

The null hypothesis is that the mean household income of mall shoppers = $42,500. The alternative hypothesis is that the mean household income of mall shoppers>$42,500.
Ho: μ = $42,500, Ha: μ > $42,500

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.

The null hypothesis is that the mean commuting distances for workers in this city is 15 miles. The alternative hypothesis is that the mean commuting distance for workers in this city is greater than 15 miles.
Ho: μ = 15 miles, Ha: μ >15 miles

19. 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.

Ho: μ = $250

Ha: μ not = $250

n = 100

s = $95.23

xbar = $234.85

CI = (234.85-1.96*95.23/10, 234.85+1.96*95.23/10)

CI = (216.184, 253.515)

Because 250 lies within the 95% confidence interval, we fail to reject our null hypothesis. Therefore, we do not have sufficient evidence to suggest that the mean donation from the charity was not $250.

20. 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

Ho: p= .4, Ha: p not = .4, n = 500, p.hat = .34

Significance level .05
95% Confidence Interval = (0.34 ± 1.96*sqrt(0.34*(1-0.34)/500))= (.298, .382)

At the 5% significance level, because our confidence interval does not include 0.4, we have evidence to reject the null hypothesis. We have evidence to suggest that the approval rating this year is different from the 40% last year.