Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

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.

→ Confidence interval for the mean claim on all policies is between **1127.554** and **1412.446**

$n = 36$ $\bar{x} = 1270$ $s = 421$

> tsum.test(n.x=36,mean.x=1270,s.x=421)

```
        One-sample t-Test

data:  Summarized x
t = 18.1, df = 35, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1127.554 1412.446
sample estimates:
mean of x
     1270
```

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.

→ Confidence interval for the mean claim on all policies is between **44.70213** and **49.29787**

$n = 49$ $\bar{x} = 47$ $s = 8$

> tsum.test(n.x=49,mean.x=47,s.x=8)

```
        One-sample t-Test

data:  Summarized x
t = 41.125, df = 48, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 44.70213 49.29787
sample estimates:
mean of x
       47
```

Karan A. Bhandarkar

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

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.

→

$n = 250$ $\quad\quad$ $\bar{x} = 1592$ $\quad\quad$ $s = 997$

$Z(\alpha/2) = Z(0.04) = 1.405072$

Confidence interval $= 1592 \pm 1.405072\ (997/250^{1/2})$
$= \mathbf{1503.4, 1680.6}$

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?

→ 95% confidence interval for the sleep variable for **men** is between **6.149609** and **10.314676**. 95% confidence interval for the sleep variable for **women** is between **7.097230** and **7.932182.**

```
> 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
```

Karan A. Bhandarkar

Yes, I am in the men's interval.
> 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
```

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?

→ 95% confidence interval for the pay for a haircut for men is between **15.97921** and **29.93307**.
95% confidence interval for the pay for a haircut for women is between **33.74673** and **69.64721**
Yes, the intervals appear very different.

> 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
```

> 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
```

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?

→ 95% confidence interval for the number of texts sent per day for men is between **26.90563** and **55.70151**.
No, I am not in this interval.
95% confidence interval for the number of texts sent per day for women is between **22.20156** and **48.82785**.
The intervals for men and women do differ but not that much.

> 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
```

> 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
```

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

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.

INDEPENDENT News Voices Sports Culture Indy/Life Video Daily Edition

News › World › Americas › US politics

Donald Trump would lose 2020 presidential election to Bernie Sanders or Joe Biden, new poll suggests

Senator Elizabeth Warren also beats incumbent in latest survey, as 53 per cent of voters still want Obama in the White House

Jon Sharman | Wednesday 19 July 2017 08:18 BST | 7 comments

Donald Trump would lose to Bernie Sanders and other Democrats in a 2020 Presidential election, according to a new poll.

The incumbent would lose by 13 points — 52 per cent to 39 per cent — and by an even greater margin against former Vice President Joe Biden, Public Policy Polling (PPP) found.

The Democrat-aligned firm tested voters' preferences for a string of potential opponents for Mr Trump.

READ MORE

Trump held secret hour-long meeting with Putin at G20

The billionaire would also lose to Senator Elizabeth Warren by a seven-point margin, it found in a survey of 836 registered voters.

Its poll had a margin of error of 3.4 per cent.

Mr Sanders, the Vermont senator who lost the Democratic presidential nomination for the 2016 election to Hillary Clinton, has been a fierce and outspoken critic of Mr Trump's administration and particularly its stance on healthcare.

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

→ 95% confidence interval for the number of texts sent per day for men is between **48.65%** and **55.42%**.
52% of the 836 sampled registered voters said yes

> 52 * 836 / 100
[1] 434.72
> binom.confint(435, 836)

	method	x	n	mean	lower	upper
1	agresti-coull	435	836	0.5203349	0.4864539	0.5540300
2	asymptotic	435	836	0.5203349	0.4864696	0.5542003
3	bayes	435	836	0.5203106	0.4864807	0.5541134
4	cloglog	435	836	0.5203349	0.4859164	0.5535873
5	exact	435	836	0.5203349	0.4858506	0.5546754
6	logit	435	836	0.5203349	0.4864280	0.5540555
7	probit	435	836	0.5203349	0.4864370	0.5540864
8	profile	435	836	0.5203349	0.4864462	0.5540992
9	lrt	435	836	0.5203349	0.4864557	0.5540883
10	prop.test	435	836	0.5203349	0.4858570	0.5546235
11	wilson	435	836	0.5203349	0.4864540	0.5540298

Interesting:
If we take the agresti-coull interval, it matches the margin of error exactly as that mentioned in the article.

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.

→ 95% confidence interval for the true proportion of those from their entire mailing list who may donate is between **0.04648758 and 0.04913242**
p = 4781 n = 100000
> binom.confint(4781,100000)

	method	x	n	mean	lower	upper
1	agresti-coull	4781	1e+05	0.04781000	0.04650475	0.04914999
2	asymptotic	4781	1e+05	0.04781000	0.04648758	0.04913242
3	bayes	4781	1e+05	0.04781452	0.04649467	0.04913945
4	cloglog	4781	1e+05	0.04781000	0.04649983	0.04914471
5	exact	4781	1e+05	0.04781000	0.04649571	0.04915060
6	logit	4781	1e+05	0.04781000	0.04650482	0.04914993
7	probit	4781	1e+05	0.04781000	0.04650215	0.04914713
8	profile	4781	1e+05	0.04781000	0.04649915	0.04914401
9	lrt	4781	1e+05	0.04781000	0.04649910	0.04914116
10	prop.test	4781	1e+05	0.04781000	0.04649993	0.04915494
11	wilson	4781	1e+05	0.04781000	0.04650486	0.04914988

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

→ The 5% rate is not plausible as it is more than the upper range for 95% confidence interval.

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.

→ 95% confidence interval for adults believing that cloning of humans should not be allowed is between **0.8710631 and 0.9095693**

p = 901   n = 1012

> binom.confint(901,1012)

```
          method   x    n      mean     lower     upper
1  agresti-coull 901 1012 0.8903162 0.8695107 0.9081697
2     asymptotic 901 1012 0.8903162 0.8710631 0.9095693
3          bayes 901 1012 0.8899309 0.8704914 0.9089375
4        cloglog 901 1012 0.8903162 0.8694042 0.9080581
5          exact 901 1012 0.8903162 0.8694221 0.9089061
6          logit 901 1012 0.8903162 0.8695309 0.9081400
7         probit 901 1012 0.8903162 0.8698344 0.9083749
8        profile 901 1012 0.8903162 0.8700822 0.9085773
9            lrt 901 1012 0.8903162 0.8700823 0.9085780
10     prop.test 901 1012 0.8903162 0.8690371 0.9085679
11        wilson 901 1012 0.8903162 0.8695669 0.9081135
```

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

9) 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?

→ Proportions of sample reported drunk = x/n = 21/110 = 0.190 = 19%

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

→ We can not be certain based on this one sample but it does seem likely.

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

→ 95% confidence interval is between 0.1174638 and 0.2643543 (**11.74% to 26.4%**)

> binom.confint(21,110)

	method	x	n	mean	lower	upper
1	agresti-coull	21	110	0.1909091	0.1276771	0.2750010
2	asymptotic	21	110	0.1909091	0.1174638	0.2643543
3	bayes	21	110	0.1936937	0.1226980	0.2677990
4	cloglog	21	110	0.1909091	0.1238071	0.2691021
5	exact	21	110	0.1909091	0.1222334	0.2768977
6	logit	21	110	0.1909091	0.1279059	0.2751545
7	probit	21	110	0.1909091	0.1262268	0.2726917
8	profile	21	110	0.1909091	0.1250435	0.2710088
9	lrt	21	110	0.1909091	0.1250393	0.2710077
10	prop.test	21	110	0.1909091	0.1246374	0.2792968
11	wilson	21	110	0.1909091	0.1283941	0.2742840

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, the national average 30% is higher than the upper limit of 26.4% for 95% confidence interval.

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.

→ To keep the margin of error at most 5%, we need **384** students.
$n = (1.96)^2 * \hat{p}\,(1-\hat{p}\,) / (.05)^2 = (3.84)\,(0.5)\,(0.5) / (.05)^2 = 384$

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

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.

→ p = 40 n = 200 → 20%
> binom.confint(40,200)

	method	x	n	mean	lower	upper
1	agresti-coull	40	200	0.2000000	0.1501687	0.2611385
2	asymptotic	40	200	0.2000000	0.1445638	0.2554362
3	bayes	40	200	0.2014925	0.1472610	0.2573930
4	cloglog	40	200	0.2000000	0.1477966	0.2579957
5	exact	40	200	0.2000000	0.1468945	0.2622264
6	logit	40	200	0.2000000	0.1502336	0.2611850
7	probit	40	200	0.2000000	0.1492549	0.2599148
8	profile	40	200	0.2000000	0.1485812	0.2590552
9	lrt	40	200	0.2000000	0.1485884	0.2590550
10	prop.test	40	200	0.2000000	0.1482520	0.2635577
11	wilson	40	200	0.2000000	0.1504520	0.2608552

95% confidence interval: 14.46% to 25.54%
This is within 5.54%

$n = (1.96)^2(0.5)^2/e2$
$n*e^2 = (1.96)^2(0.5)^2$
$n_1*e_1^2 = n_2*e_2^2$
$200 * (0.0554)^2 = n2 * (0.03)^2$
$n2 = (0.613832) / (0.03)^2$
$n2 = 682.036$

The necessary sample is **683**

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

$\rightarrow n = (1.96)^2 * \hat{p}\,(1-\hat{p}\,) / (.03)^2$
$= (3.84)\,(0.5)\,(0.5) / (.03)^2$
$= 1066.66$

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

10) 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?

→ Margin of error = 24 - 22 / 2 = 1

$n = (1.96)^2\ \hat{p}\ (1-\hat{p}\ ) / (.01)^2$

$= (3.84)\ (0.5)\ (0.5) / (.01)^2$

$=$ **9600.00**

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).

→ Mean blood pressure is **133.9286** and confidence interval between **128.7077** and **139.1495**
> 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.

→ 95% confidence interval for the mean score is between **128.7** and **139.148.**

df = 14 -1 = 13
For 95% confidence, $t_{13} = 2.160369$
$133.9286 - 2.160369 * 9.04233 / (14)^{1/2}$, $133.9286 + 2.160369 * 9.04233/(14)^{1/2}$
= 133.9286 – 5.22, 133.9286 + 5.22
= 128.7, 139.148

```
>  qt(.975,13)
[1] 2.160369
> sd(mydata)
[1] 9.04233
```

(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% confidence interval for the mean score is between **129.192** and **138.66**

= $133.9286 - 1.96 * 9.04233 / (14)^{1/2}$, $133.9286 + 1.96 * 9.04233 / (14)^{1/2}$
= 133.9286 – 4.736, 133.9286 + 4.736
= 129.192, 138.66

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

→ Confidence interval in **a and b are same**. They are calculated using R vs by hand.
Confidence interval in **a and b using the t-distribution is bigger than c** as we have less than 30 number number of observations.

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're 95% confident that the population mean lies in the confidence interval but it could lie outside this range as well.

13) 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 intervals would actually contain μ.

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

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.

→ (c) Confidence interval tells us how confident we are of the population mean being inside the confidence interval and not the probability.

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

→ In many real world scenarios, the population size is too large to obtain descriptive statistics. For such cases, we estimate a plausible sized population proportion instead.

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

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.

→ It is a one-tailed test since they testing is for 'equal versus greater than' and does not consider less than.

b) Write the null and alternative hypothesis.

→ Null hypothesis: H0: Average <= 28
Alternative hypothesis: Ha: Average > 28

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

→ Type I error is reject the null hypothesis when the null hypothesis is true. In this case CEO will buy 100,000 devices for its new line of cars even when the car average is not greater than 28.

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

→ Type II error is failure to reject the null hypothesis when the alternative hypothesis is true. In this case CEO will not buy 100,000 devices for its new line of cars even when the car average is greater than 28.

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.

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

18) 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.

→ Null hypothesis: H0: diameter = 5
Alternative hypothesis: Ha: Average ≠ 5

(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.

→ Null hypothesis: H0: mean income > 42,500
Alternative hypothesis: Ha: mean income <= 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.

→ Null hypothesis: H0: commuting distance > 15
Alternative hypothesis: Ha: commuting distance <= 15

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.

→ Fail to reject that charity organization's claims that the average donation from contributors to the charity is not $250.00. **The confidence interval 215.9543, 253.7457 contains 250.00**

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

        One-sample t-Test

data:  Summarized x
t = 24.661, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 215.9543 253.7457
sample estimates:
mean of x
   234.85
```

Karan A. Bhandarkar

Homework 6
STAT 104 - Introduction to Quantitative Methods for Economics

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

→ Fail to reject that your candidate's approval rating has changed from the previous dismal 40% after a major policy announcement. **The confidence interval 0.2702986, 0.4170611 contains .40**

```
> prop.test(57.8,170)
```

```
	1-sample proportions test with continuity correction

data:  57.8 out of 170, null probability 0.5
X-squared = 16.774, df = 1, p-value = 4.211e-05
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.2702986 0.4170611
sample estimates:
   p
0.34
```