CWRU DSCI351-451: Homework 5 Inference

Roger H. French, JiQi Liu

26 October, 2018

Contents

Inference Guide

There is a useful Inference Cheat Sheet in your readings folder

- os2_extra_inference_guide.pdf

There is Hadley Wickham's book on ggplot in your readings textboosk folder

- Elegant Graphics for Data Analysis [@wickham_ggplot2:_2016]

5.0.0.1 1) Twitter users and News

5.0.0.1.1 Part I. (OIS 4.8)

A poll conducted in 2013 found that

- 52% of U.S. adult Twitter users get at least some news on Twitter.[@mitchell_twitter_2013]
- The standard error for this estimate was 2.4%,
  - and a normal distribution may be used to model the sample proportion.

Construct a 99% confidence interval for

- the fraction of U.S. adult Twitter users
- who get some news on Twitter,

and interpret the confidence interval in context.

```
#Confidence interval of percentage of US adult twitter users who get some news on twitter

#point estimate
pointestimate <- .52
#standard error
standarderror <- .024
#z score is calculated since it is normally distributed
zscore <- qnorm(.995)
```

```r
#lower limit of confidence interval
lowerlimit <- pointestimate - (zscore * standarderror)
#upper limit of confidence interval
upperlimit <- pointestimate + (zscore * standarderror)

print(lowerlimit)
```

```
## [1] 0.4581801
```

```r
print(upperlimit)
```

```
## [1] 0.5818199
```

Answer: We can be 99.5% sure that the percentage of US adult twitter users that get at least some news from twitter is between 45.8% & 58.2%.

5.0.0.1.2 Twitter users and news, Part II. (OIS 4.10)

Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(a) The data provide statistically significant evidence that

- more than half of U.S. adult Twitter users
  - get some news through Twitter.
- Use a significance level of $\alpha = 0.01$.

Answer: False. The lower limit of the confidence interval with a significance level of $\alpha = 0.01$ is 45.8% which means that there is a finite probability that less than half of US adult twitter users got some of their news through Twitter.

(b) Since the standard error is 2.4%,

- we can conclude that 97.6% of all U.S. adult Twitter users
  - were included in the study.

Answer: False. This is not what standard error means. Standard error is the sample deviation of the data within the sample. This implies that the average deviation of the sample data from the sample mean is 2.4%.

(c) If we want to reduce the standard error of the estimate,

- we should collect less data.

Answer: False. This is because the standard error is the standard deviation divided by the square root of the number of observations. Therefore, if we want to reduce the standard error of the estimate, we should be collecting more data.

(d) If we construct a 90% confidence interval

- for the percentage of U.S. adults Twitter users
  - who get some news through Twitter,
- this confidence interval will be wider
  - than a corresponding 99% confidence interval.

Answer: False. This is because the 90% confidence interval is narrower than the 99% confidence interval. This occurs because the higher the probability something exists within the confidence interval, the wider the range.

5.0.0.2 4.24 Gifted children,

5.0.0.2.1 Part I. (OIS 4.24)

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.[@graybill_regression_1994]

(a) Are conditions for inference satisfied?

Answer: No. This is because the standard deviation of the population is unknown.

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

Answer: The null hypothesis will be gifted children learn to read at or before 32 months of age while the alternative hypothesis is that gifted children learn to read before 32 months of age. Using the p-value approach we can reject the null hypothesis in favour of the alternative hypothesis since the p-value is less than the significance level.

```
#sample mean
samplemean <- 30.69
#sample standard deviation
sd <- 4.31
#sample size
samplesize <- 36
#popultation mean
populationmean <- 32
#significance level
significancelevel <- 0.1

#Conducting a t test
tstatistic <- (samplemean - populationmean)/(sd/sqrt(samplesize))
pvalue <- pnorm(tstatistic)
pvalue < significancelevel
```

```
## [1] TRUE
```

(c) Interpret the p-value in context

- of the hypothesis test
- and the data.

Answer: The p-value is the likelihood that the null hypothesis is true. Since the p-value is less than our significance level, we can reject the null hypothesis. If the p-value were more than our significance level,

however, we could not reject the null hypothesis.

(d) Calculate a 90% confidence interval

- for the average age at which gifted children
  - first count to 10 successfully.

Answer: (29.50834, 31.87166)

```
# z* and standard error calculations
standarderror <- sd/sqrt(samplesize)
zstar <- 1.645
# Confidence interval
lowerlimit <- samplemean - (zstar * standarderror)
upperlimit <- samplemean + (zstar * standarderror)
# Print lower and upper limits for confidence interval
print(lowerlimit)
```

```
## [1] 29.50834
```

```
print(upperlimit)
```

```
## [1] 31.87166
```

(e) Do your results from

- the hypothesis test and
- the confidence interval agree?

Explain.

Answer: Yes, because the p value is much lower than the significance level and the null hypothesis is rejected. The population mean does not lie in the confidence interval for the sample mean, so it provides for statistically signficant evidence.

5.0.0.2.2 Part II. (OIS 4.26)

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.

(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.

Answer: The null hypothesis in this case is that the average IQ of mothers of gifted children is equal to the population average while the alternative hypothesis states it is different. Using the p-value approach we can reject the null hypothesis in favour of the alternative hypothesis since the p-value is less than the significance level.

```
#sample mean
samplemean <- 118.2
```

```r
#sample standard deviation
sd <- 6.5
#sample size
samplesize <- 36
#popultation mean
populationmean <- 100
#significance level
significancelevel <- 0.1

#Conducting a t test
tstatistic <- (samplemean - populationmean)/(sd/sqrt(samplesize))
pvalue <- pnorm(tstatistic, lower.tail = FALSE)
pvalue < significancelevel
```

```
## [1] TRUE
```

(b) Calculate a 90% confidence interval

- for the average IQ of mothers of gifted children.

Answer: (116.4179,119.9821)

```r
# z* and standard error calculations
standarderror <- sd/sqrt(samplesize)
zstar <- 1.645
# Confidence interval
lowerlimit <- samplemean - (zstar * standarderror)
upperlimit <- samplemean + (zstar * standarderror)
# Print lower and upper limits for confidence interval
print(lowerlimit)
```

```
## [1] 116.4179
```

```r
print(upperlimit)
```

```
## [1] 119.9821
```

(c) Do your results from

- the hypothesis test
- and the confidence interval agree?

Explain.

Answer: The significance level of 0.1 is more than the p-value and so rejects the null hypothesis while the confidence interval does not include 100 so that also states that with 90% confidence 100 is not part of the interval. Therefore, our confidence interval and our hypothesis test agree.

5.0.0.3 Spray Paint (OIS 4.42)

Suppose the area that can be painted using a single can of spray paint

- is slightly variable
- and follows a nearly normal distribution
  - with a mean of 25 square feet
  - and a standard deviation of 3 square feet.

(a) What is the probability that

- the area covered by a can of spray paint

- is more than 27 square feet?

Answer: 0.252

```
#Calculating the probability that the area covered by a can of spray paint is more than 27 square feet
pnorm(27, mean = 25, sd = 3, lower.tail = FALSE)
```

```
## [1] 0.2524925
```

(b) Suppose you want to spray paint

- an area of 540 square feet
- using 20 cans of spray paint.

On average, how many square feet

- must each can be able to cover
- to spray paint all 540 square feet?

Answer: 27

```
540/20
```

```
## [1] 27
```

(c) What is the probability

- that you can cover a 540 square feet area
- using 20 cans of spray paint?

Answer: 0.0429%

```
# Sample size
samplesize <- 20
# Standard deviation
sd<- 3
# Sample mean
samplemean <- 25
# Standard error
standarderror <- sd/sqrt(samplemean)
# Z-score
zscore <- (27 - samplemean)/standarderror
pnorm(zscore, lower.tail = FALSE)
```

```
## [1] 0.0004290603
```

(d) If the area covered by a can of spray paint

- had a slightly skewed distribution,
- could you still calculate the probabilities in parts (a) and (c)
  - using the normal distribution?

Answer: Yes. This is because a slight skew will not change our observations much in part (a) and (c) since the sample population is greater than 10.

5.0.0.4 Fuel efficiency of Prius. (OIS 5.8)

Fueleconomy.gov,

- the official US government source
  - for fuel economy information,
- allows users to share gas mileage information on their vehicles.

The histogram below shows

- the distribution of gas mileage in miles per gallon (MPG)
  - from 14 users who drive a 2012 Toyota Prius.
- The sample mean is 53.3 MPG
  - and the standard deviation is 5.2 MPG.

Note that these data are user estimates

- and since the source data cannot be verified,
- the accuracy of these estimates are not guaranteed.[@noauthor_gas_nodate]

(a) We would like to use these data to evaluate

- the average gas mileage of all 2012 Prius drivers.

Do you think this is reasonable?

- Why or why not?

Answer: Yes. This is becausealthough the data has extreme values, the sample data are approximately normally distributed and we have an unbiased estimate of the population mean and the population standard deviation.

(b) The EPA claims that a 2012 Prius gets 50 MPG

- (city and highway mileage combined).

Do these data provide strong evidence against this estimate

- for drivers who participate on fueleconomy.gov?
- Note any assumptions you must make as you proceed with the test.

Answer: Assuming we have an independent normally distributed random sample without a strong skew, we can make our null hypothesis that the 2012 Prius gets 50 MPG while our alternative hypothesis is that the MPG value of the Prius is not equal to 50. We can reject the null hypothesis since p is less than the significance level.

```
#sample mean
samplemean <- 53.3
#sample standard deviation
sd <- 5.2
#sample size
samplesize <- 14
#popultation mean
populationmean <- 50
#significance level
significancelevel <- 0.05
#degrees of freedom
degreesoffreedom <- samplesize-1

#Conducting a t test with a two sided p value
tstatistic <- (samplemean - populationmean)/(sd/sqrt(samplesize))
pvalue <- 2*pt(-abs(tstatistic), df = degreesoffreedom)
pvalue < significancelevel
```

```
## [1] TRUE
```

(c) Calculate a 95% confidence interval

- for the average gas mileage of a 2012 Prius
- by drivers who participate on fueleconomy.gov.

Answer: (50.29812,56.30188)

```
# z* and standard error calculations
standarderror <- sd/sqrt(samplesize)
tstar <- 2.16
# Confidence interval
lowerlimit <- samplemean - (tstar * standarderror)
upperlimit <- samplemean + (tstar * standarderror)
# Print lower and upper limits for confidence interval
print(lowerlimit)
```

```
## [1] 50.29812
```

```
print(upperlimit)
```

```
## [1] 56.30188
```

5.0.0.5 Diamonds

5.0.0.5.1 Diamonds Part I. (OIS 5.28)

Prices of diamonds are determined by what is known as the 4 Cs:

- cut,
- clarity,
- color,
- and carat weight.

The prices of diamonds go up

- as the carat weight increases,
- but the increase is not smooth.

For example, the difference between the size

- of a 0.99 carat diamond and
  - a 1 carat diamond is undetectable to the naked human eye,
- but the price of a 1 carat diamond tends to be much higher
  - than the price of a 0.99 diamond.

In this question we use two random samples of diamonds,

- 0.99 carats and 1 carat,
- each sample of size 23,

and compare the average prices of the diamonds.

In order to be able to compare equivalent units,

-we first divide the price for each diamond - by 100 times its weight in carats.

That is, for a 0.99 carat diamond, we divide the price by 99.

For a 1 carat diamond, we divide the price by 100.

The distributions and some sample statistics are shown below.[@wickham_ggplot2:_2016]

(a) Conduct a hypothesis test to evaluate

- if there is a difference between the average standardized prices
- of 0.99 and 1 carat diamonds.

Make sure to

- state your hypotheses clearly,
- check relevant conditions,
- and interpret your results in context of the data.

Answer:

The null hypothesis states that there is no difference between the standardized prices for a 0.99 and 1 carat diamonds while the alternative hypothesis states that there is a difference between the standard prices for 0.99 and 1 carat diamonds.

Assumptions include that the distributions are not skewed and the sample size is 10% of the whole population. We can reeject the p-value because as shown below it is less than the significance level.

```
# Significance level
significancelevel <- 0.05
# Population mean
populationmean <- 0

## For 0.99 carats
# Point estimate
mean99 <- 44.51
# Standard deviation
sd99 <- 13.32
# Sample size
size99 <- 23

##For 1 carat
# Point estimate
mean1 <- 56.81
# Standard deviation
sd1 <- 16.13
# Sample size
size1 <- 23

# Standard error
standarderror <- sqrt(((sd99^2)/size99)+((sd1^2)/size1))

# t statistic
tstatistic <- ((mean1 - mean99) - populationmean)/standarderror

# Degrees of freedom (used an online calculator)
degreesoffreedom <- 42.4808

# p-value (two-tailed)
pvalue <- 2 * pt(-abs(tstatistic), df = degreesoffreedom)

pvalue < significancelevel
```

```
## [1] TRUE
```

5.0.0.5.2 Diamonds Part II. (OIS 5.30)

We discussed diamond prices

- (standardized by weight)
- for diamonds with weights 0.99 carats and 1 carat.

See the table for summary statistics,

- and then construct a 95% confidence interval
- for the average difference
  - between the standardized prices of 0.99 and 1 carat diamonds.

You may assume the conditions for inference are met.

Answer: (2.878324,21.72168)

```
# t-score
tstar <- 2.16
# Confidence interval
lowerlimit <- (mean1 - mean99) - tstar * standarderror
upperlimit <- (mean1 - mean99) + tstar * standarderror
print(upperlimit)
```

```
## [1] 21.72168
```

```
print(lowerlimit)
```

```
## [1] 2.878324
```

5.0.0.6 Links

http://www.r-project.org

http://rmarkdown.rstudio.com/

https://www.openintro.org/stat/textbook.php?stat_book=os

5.0.0.7 References