CUNY 606 Homework chapter 5

Valerie Briot

April 3, 2016

This document pertains to homeworks assigned for CUNY 606 Probability and Statistics on chapter 5: Inference for numerical data. The following homeworks have been assigned: 5.6, 5.14, 5.20, 5.32, 5.48

Exercise 5.6 Working backwards, Part II.

A 90% confidence interval for a population mean is (65,77). The population distribution is approximately normal and the population standard deviation is unkown. This confidence interval is based on a simple random sample of 25 observations.
Calculate the sample mean, the margin of error, and the sample standard deviation.

The mean is the mid-point of interval, hence $mu = \frac{77 - 66}{2} + 66 = 6 + 66 = 72$

me = 77 - 66 = 6

For a 90% interval calcualtion we would have zscore = 1.645

Confidence interval: point estimate +/ me = point estimate +/ zscore * SE

me = 1.645 * SE, SE = me/1.645 = 3.647

$SE = \frac{s}{\sqrt{n}}$,

hence, $s = SE \times \sqrt{n}$, s = 3.647 * 5 = 18.24

Exercise 5.14 SAT Scores

SAT scores of students at an Ivy League college are distributed with a standard deviation of 250 point. Two statistics students, Raina and Luke, want to estimate the average SAT score of students at this college as part of a class project. They want their marging of error to be no more than 25 points.

a) Raina wants to use a 90% confidence level interval. How large a sample should she collect? me = 25 s = 250

90% interval zscore = 1.645

me = zscore * SE

me = 1.645 * s/sqrt(n)

sqrt(n) = 1.645 * s/me = 1.645 * 250/25

n = (1.645 * 250/25)^2 = 270.6025, n = 271 (take next intger)

b) Luke wants to use a 99% confidence level interval. Without calculating the actual sample size determine whether his sample should be larger than Raina's and expliain your reasoning.

All things being equals, to increase the confience interval we need to widen the interval hence, the zscore is higher, actually zscore = 2.576. With a higher zscore with me and s being the same, n will have to be higher.

n = (2.576 * 250/25)^2 = 663.5776, n = 664

Exercise 5.20 High School and beyond, Part I

The National Center of Education Statistics conducted a survey of high school seniors, colllecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.

a) Is there a clear difference in the average reading and writng scores; From the box plot, it is clear that the median for writing score is higher than the one for the reading score. In addition the spread of the data is more important for the reading score. There does not seem to be any outliers for either distributions. They may be some right-skewedness to the reading distribution and possibly some left skewedness but the sampe size is large (200) and we can assume that the difference in median would also be reflected in a difference in average of reading scores and average in writing socre.

b) The reading and writing scores are paired.

c) H0 the average reading scores and writing scores are not different Ha the average reading scores and writing scores are different

d) The sample is randomly selected and is large (200) based on less that 10% of all senior in high school so we can assume independence.

e) $\overline{x_{read-write}} =$ -0.545 and s for difference = 8.887 points

SEdiff = sdiff/sqrt(200) = 8.887/14.14 = 0.6285

we now compute test statistic T-score, df = 200-1 = 199

T = (xdiff - 0)/SEdiff = (-0.545)/0.6285 = -0.8671 = -0.87

```
p_value = pt(-.87, df=199)
p_value * 2 #(2 tails)
```

```
## [1] 0.3853486
```

The p-value (0.385) > 0.05, hence there is no sufficient effidence to reject the H0 hypothesis.

f) we could have made a type 2 error, failed to reject the H0 hypothesis when Ha is actually true.

g) Yes, based on the results of the hypothesis test, we would expect a confidence interval for the average difference between reading and writing scores to include 0. Since we have failed to rejet the H0 hypothesis, we would expect that the confindence interval would span 0.

Exercise 5.32 Fuel efficiency of manual and automatic cars, Part 1.

Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manulactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.

H0: there is no difference in the average fuel efficiency between manual and automatic transmission in cars manufactured in 2012 h1: there is a difference in the average fuel efficiency between manual and automatic transmission in cars manufactured in 2012

we can assume that the conditions required for inference are met.

$T = \frac{x_m - x_a}{SE} = \frac{x_m - x_a}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_a^2}{n_m}}}$

```
xa <- 16.12
sa <- 3.58
na <- 26

xm <- 19.85
sm <- 4.51
nm <- 26

t_value <- (xm-xa)/(sqrt((sm^2/nm)+(sa^2/na)))
t_value
```

```
## [1] 3.30302
```

```
df <- 26-1

p_value <- dt(t_value, df)
p_value
```

```
## [1] 0.003564492
```

The p-value < 0.05 so we reject H0 hyptophese. There is a difference in average of fuel efficiency between manual and automatic transmission.

Exercise 5.48 Work hours and education

a) H0: any difference in the average hours worked between each group is due to chance Ha: the differences in the average hours worked between each group is not due to chance

b) There are 3 conditions to be checked for an ANOVA analysis: **Independence** we will assume that survey was sent ramdomly to residents, the number of respondants (1,172) is less than 10% of the population. Hence we have independence.

Approximately normal Each group have around 100 or above obervations. From the box plot, we may have outliers in many of the group distribution. It is very possible that we do not have normal distribution some of them (proably for Bachelor's group). However to proceed with the test we will assume near normal distribution.

Constance variance For the group Jr College, we can see from the box plot that this group has a greater variance but this may be the result of natural variation.

c)

```r
x <- 40.45
s <- 15.17
n <- 1172

x1 <- 38.67
s1 <- 15.81
n1 <- 121

x2 <- 39.6
s2 <- 14.97
n2 <- 543

x3 <- 41.39
s3 <- 18.1
n3 <- 97

x4 <- 42.55
s4 <- 13.62
n4 <- 253

x5 <- 40.85
s5 <- 15.51
n5 <- 155

k <- 5
dfg <- k-1

dfe <- n-k

ssg <- n1*(x1-x)^2 + n2*(x2-x)^2 + n3*(x3-x)^2 + n4*(x4-x)^2 + n5*(x5-x)^2
msg <- ssg/dfg

sse <- (n1-1)*s1^2 + (n2-1)*s2^2 + (n3-1)*s3^2 + (n4-1)*s4^2 + (n5-1)*s5^2
mse <- sse/dfe

F_value <- msg/mse

ssg
```

```
## [1] 2001.933
```

```r
sse
```

```
## [1] 266701.3
```

```r
msg
```

```
## [1] 500.4833
```

```r
mse
```

```
## [1] 228.5359
```

```
F_value
```

```
## [1] 2.189955
```