Chapter\_5homework

Priya Shaji

3/23/2019

## Inference for Numerical Data

### Ques\_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 unknown. 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.

Answer 5.6

1. Sample Mean

Since we know that the sample mean is (x2+x1)/2 where the confidence interval is (x1,x2)

n <- 25  
x1 <- 65  
x2 <- 77  
SMean <- (x2 + x1) / 2  
SMean

## [1] 71

The sample mean is 71.

1. Margin of Error

Since we know that the margin of error is (x2−x1)/2 where the confidence interval is (x1,x2)

n <- 25  
x1 <- 65  
x2 <- 77  
  
ME <- (x2 - x1) / 2  
ME

## [1] 6

The margin of error is ME = 6.

1. Sample standard deviation

To calculate the sample standard deviation we use ME=t∗⋅SE by using the qt() function and df = 25 - 1.

df <- 25 - 1  
p <- 0.9  
p\_2tails <- p + (1 - p)/2  
  
t\_val <- qt(p\_2tails, df)  
# Since ME = t \* SE  
SE <- ME / t\_val  
  
# Since SE = sd/sqrt(n)  
sd <- SE \* sqrt(n)  
sd

## [1] 17.53481

The standard deviation is sd = 17.5348146.

### Ques\_5.14

SAT scores. SAT scores of students at an Ivy League college are distributed with a standard deviation of 250 points. 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 margin of error to be no more than 25 points.

Answer 5.14

1. Raina wants to use a 90% confidence interval. How large a sample should she collect?

Answer a)

For this, I will use as follows: ME=z⋅SE and since SE=sdn√

we have as follows: ME=z⋅sdn√ at the end we obtain: MEz=sdn√

n=(z⋅sdME)2

z <- 1.65 # due to 90% Confidence interval  
ME <- 25  
sd <- 250  
  
n <- ((z \* sd) / ME ) ^ 2  
n

## [1] 272.25

The sample size should be 273 students.

1. Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina’s, and explain your reasoning.

Answer b)

Luke’s sample should be larger since it will require a higher z number multiplied by the standard deviation and then squared.

1. Calculate the minimum required sample size for Luke.

Answer c)

z <- 2.575 # due to 99% Confidence interval  
ME <- 25  
sd <- 250  
  
n <- ((z \* sd) / ME ) ^ 2  
n

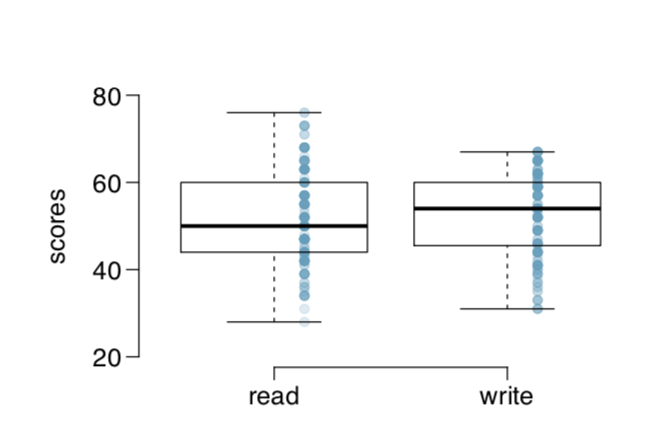
## [1] 663.0625

The sample size should be 664 students.

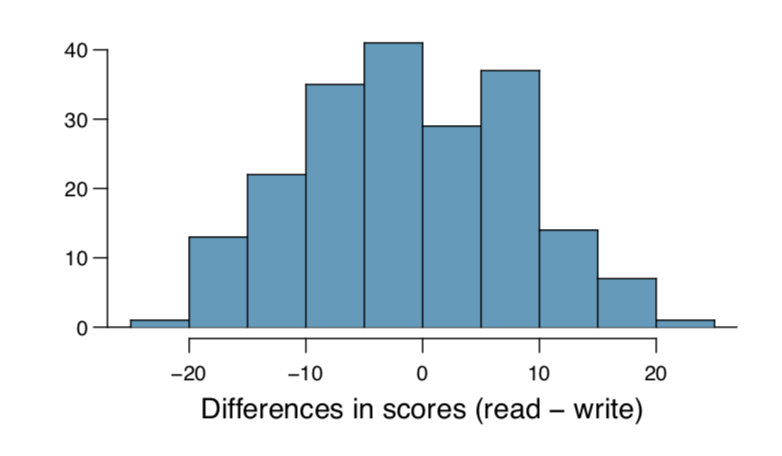
### Ques\_5.20

High School and Beyond, Part I. The National Center of Education Statistics conducted a survey of high school seniors, collecting 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.

knitr::include\_graphics("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/homeworks/homework\_5/Screen Shot 2019-03-24 at 4.13.23 PM.png")



knitr::include\_graphics("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/homeworks/homework\_5/Screen Shot 2019-03-24 at 4.13.32 PM.png")



Answer 5.20

1. Is there a clear difference in the average reading and writing scores?

Answer a)

Clear difference is not visible in the average of the reading and writing scores. The difference distribution is fairly normal around the zero difference, though it seems to be a slight skew to the right.

1. Are the reading and writing scores of each student independent of each other?

Answer b)

The scores are independent of each student but not of each score, that is reading and writing scores are not independent of each other for each student.

1. Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

Answer c)

Since the question is referring for the difference in the average score of students, and not referring to the average difference in scores. The hypotheses could be as follows:

H\_0: The difference of average in between reading and writing equal zero. That is: μr−μw=0

H\_A: The difference of average in between reading and writing does NOT equal zero. That is: μr−μw≠0

1. Check the conditions required to complete this test.

Answer d)

Independence of observations: The difference histogram suggested the data are paired. If paired, then they wouldn’t be independent.

Observations come from nearly normal distribution: The box plot provided in the text suggests the data are reasonably normally distributed and no outliers exist.

1. The average observed difference in scores is x ̄read

Answer e)

The hypotheses for the average difference test are:

H0

The difference of average scores is equal to zero. That is: μdiff=0

HA

The difference of average scores is NOT equal to zero. That is: μdiff≠0

The paired data is presumably from less than 10% of the population of senior high schoolers, and from a simple random sample. We noted that the differences are nearly normally distributed, so the conditions are met in order to apply the t-distribution.

sd\_Diff <- 8.887  
mu\_Dif <- -0.545  
n <- 200  
  
SE\_Diff <- sd\_Diff / sqrt(n)  
  
# Compute T statistic  
t\_value <- (mu\_Dif - 0) / SE\_Diff  
  
df <- n - 1  
  
p <- pt(t\_value, df = df)  
p

## [1] 0.1934182

Since the p-value is not less that 0.05, this implies that there is not convincing evidence of a difference in student’s reading and writing exam scores maintaining our NULL hypothesis.

The above conclusion need to be analyzed with further detail since the data need to be independent and currently is not.

1. What type of error might we have made? Explain what the error means in the context of the application.

Answer f)

Type I error: Incorrectly reject the null hypothesis.

Type II error: Incorrectly reject the alternative hypothesis.

In the case, we may have made a type II error by rejecting the alternative hypothesis HA. that is, we might have wrongly concluded that there is not a difference in the average student reading and writing exam scores.

1. Based on the results of this hypothesis test, would you expect a confidence interval for the average di↵erence between the reading and writing scores to include 0? Explain your reasoning.

Answer g)

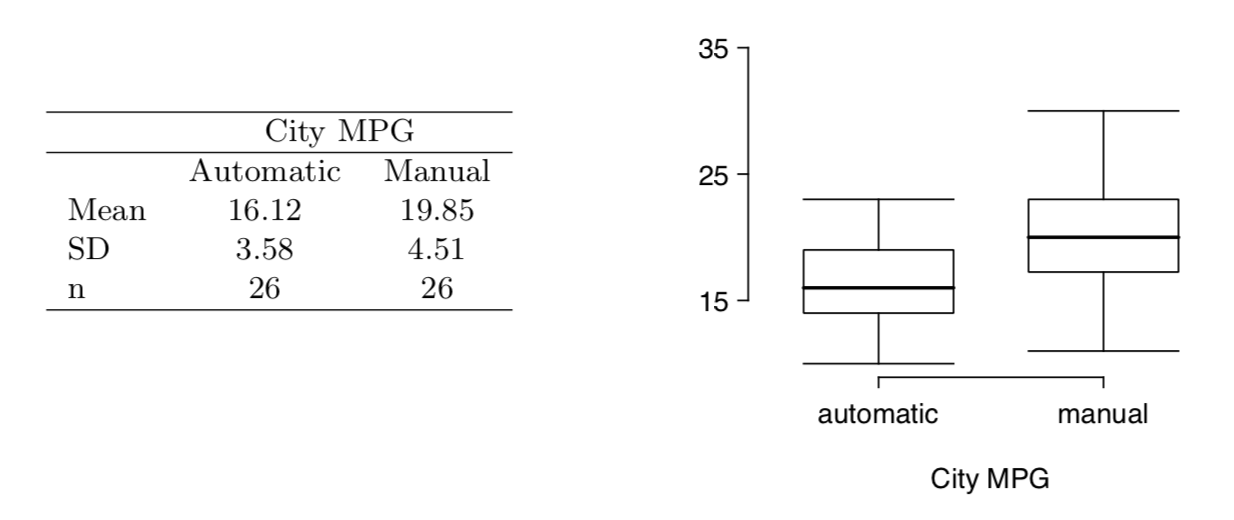
Yes,there should be a confidence interval for the average difference between reading and writing scores to include 0.

When the confidence interval include 0 for this kind of hypothesis test, it indicates that the difference is not in one side or another.

### Ques\_5.32

Fuel efficiency of manual and automatic cars, Part I. Each year the US Environ- mental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel e

knitr::include\_graphics("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/homeworks/homework\_5/Screen Shot 2019-03-24 at 4.39.55 PM.png")



Answer 5.32

The hypotheses for this test are as follows:

H0

The difference of average miles is equal to zero. That is: μdiff=0

HA

The difference of average miles is NOT equal to zero. That is: μdiff≠0

From the text we have as follows:

n <- 26  
# Automatic  
mu\_a <- 16.12  
sd\_a <- 3.58  
# Manual  
mu\_m <- 19.85  
sd\_m <- 4.51  
  
# difference in sample means  
mu\_Diff <- mu\_a - mu\_m  
  
# standard error of this point estimate  
SE\_Diff <- ( (sd\_a ^ 2 / n) + ( sd\_m ^ 2 / n) ) ^ 0.5  
  
t\_val <- (mu\_Diff - 0) / SE\_Diff  
df <- n - 1  
p <- pt(t\_val, df = df)  
p

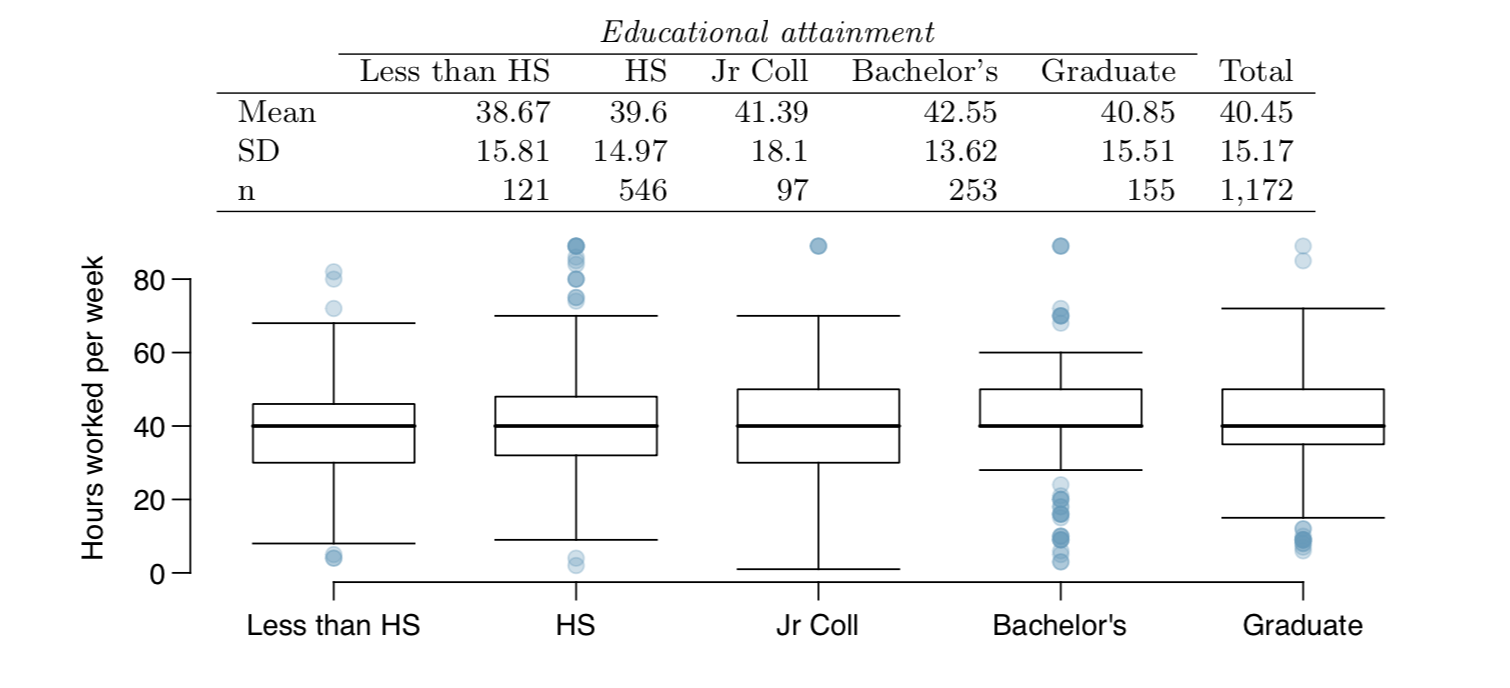
## [1] 0.001441807

Since the p-value is less than 0.05, we reject the null hypothesis H0 and conclude that there is strong evidence of a difference in fuel efficiency between manual and automatic transmissions.

### Ques\_5.48

Work hours and education. The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents.47 Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

knitr::include\_graphics("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/homeworks/homework\_5/Screen Shot 2019-03-24 at 4.44.53 PM.png")



1. Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

Answer a)

The hypotheses for this ANOVA test follow:

H0

The difference of ALL averages is equal. That is: μl=μh=μj=μb=μg

HA

There is one average that is NOT equal to the other ones.

1. Check conditions and describe any assumptions you must make to proceed with the test.

Answer b)

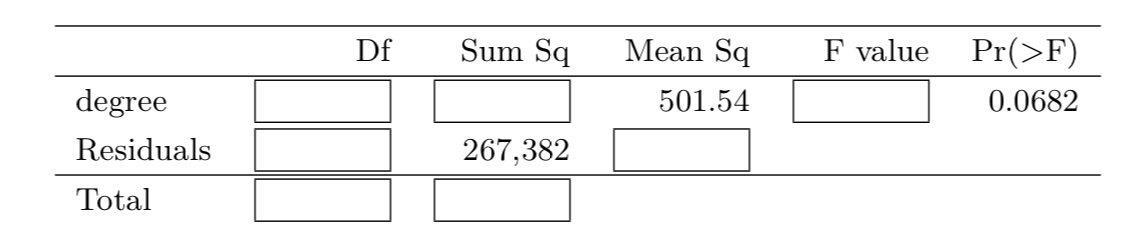
The observations are independent within and across groups: I will assume independence within and across the groups based on the nature of the provided data.

The data within each group are nearly normal: The box plots do not support nearly normal data within each group. Each group has outliers some groups seem to follow a normal distribution.

The variability across the groups is about equal: There seems to be a similarity of variability in between some of the groups just by observing the standard deviations.

1. Below is part of the output associated with this test. Fill in the empty cells.

knitr::include\_graphics("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/homeworks/homework\_5/Screen Shot 2019-03-24 at 4.46.54 PM.png")



Answer c)

mu <- c(38.67, 39.6, 41.39, 42.55, 40.85)  
sd <- c(15.81, 14.97, 18.1, 13.62, 15.51)  
n <- c(121, 546, 97, 253, 155)  
data\_table <- data.frame (mu, sd, n)

n <- sum(data\_table$n)  
k <- length(data\_table$mu)  
  
# Finding degrees of freedom  
df <- k - 1  
dfResidual <- n - k  
  
# Using the qf function on the Pr(>F) to get the F-statistic:  
  
Prf <- 0.0682  
F\_statistic <- qf( 1 - Prf, df , dfResidual)  
  
# F-statistic = MSG/MSE  
  
MSG <- 501.54  
MSE <- MSG / F\_statistic  
  
# MSG = 1 / df \* SSG  
  
SSG <- df \* MSG  
SSE <- 267382  
  
# SST = SSG + SSE, and df\_Total = df + dfResidual  
  
SST <- SSG + SSE  
dft <- df + dfResidual

Df Sum Sq Mean Sq F value Pr(>F)

degree 4 2006.16 501.54 2.188984 0.0682 Residuals 1167 267,382 229.12  
Total 1171 269388.16

1. What is the conclusion of the test?

Since the p-value = 0.0682 is greater than 0.05, We conclude that there is not a significant difference between the groups and the null hypothesis does not get rejected.

Answer d)

Answer 5.48