Inference for numerical data

Shyam BV

October 30, 2016

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.

```
#For any confidence interval, the center point is the mean of two values
samplesize = 25
samplemean <- (65+77)/2

#Calculate t-score

me <- 77-samplemean

tscore <- qt(.05,24)*-1

s <- (me/tscore)*(sqrt(samplesize))
```

sample mean = 71 Marigin of error = 6 sample SD = 17.5348146

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.

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

```
s <- 250
me <- 25
z_score <- qnorm(0.95)

((z_score*s)/me)^2
```

```
## [1] 270.5543
```

```
#total Sample size required is at-least 271 students
```

(b) 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.

To determine the sample size, we need more samples. Below are the reasons

1. The confidence level increases, but the margin of error should be less than 25.
2. If we assume that 99% of samples are in that specific interval, then the smaller sample size will not be sufficient. So we need a larger number of samples,
3. Mathematically, ME = z-score*(sd/sqrt(n)). If n increases, SE decreases and ME deceases.

(c) Calculate the minimum required sample size for Luke.

```r
s <- 250
me <- 25
z_score <- qnorm(0.995)

((z_score*s)/me)^2
```

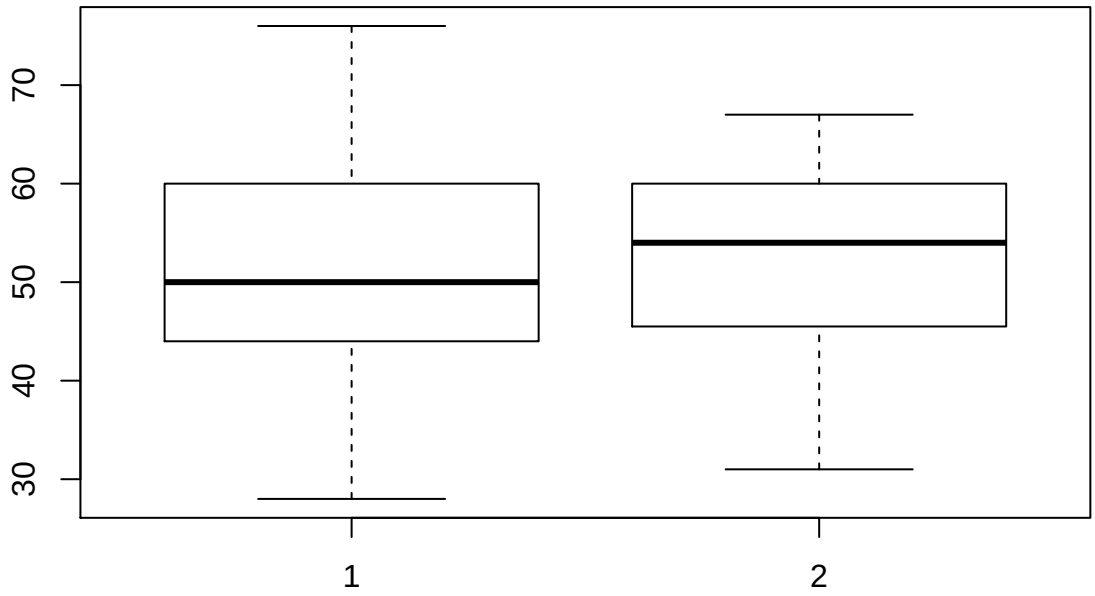
```
## [1] 663.4897
```

5.20 High School and Beyond,

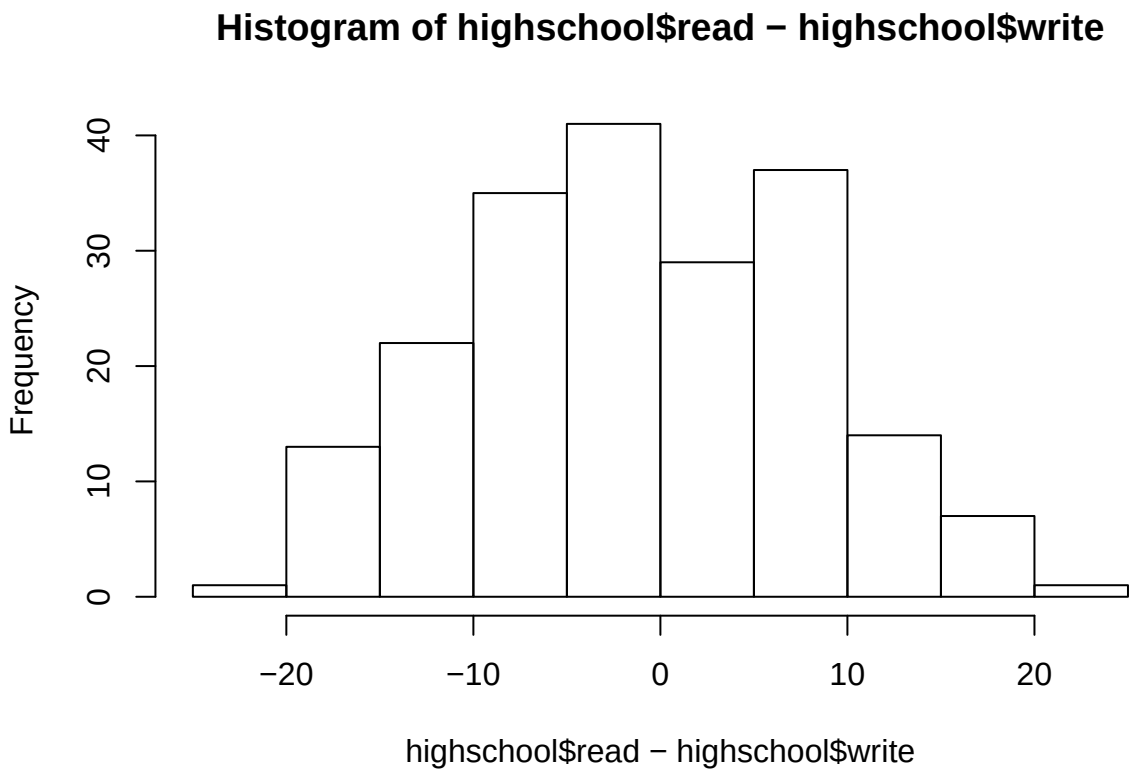
Part I. The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as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

```r
highschool <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20fro

boxplot(highschool$read,highschool$write)
```



```r
hist(highschool$read-highschool$write)
```



(a) Is there a clear difference in the average reading and writing scores?

```r
mean(highschool$read)-mean(highschool$write)
```

```
## [1] -0.545
```

Above values shows there is almost no difference between the reading and writing scores.

(b) Are the reading and writing scores of each student independent of each other?

No. The reading and writing scores of each student are not independent. Because reading and writing is produced by a same student. So it is not independent.

(c) 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?

H_0: Average scores of students in reading and writing exam are equal H_A: Averag scores of students in reading and writing exam are different.

(d) Check the conditions required to complete this test.

1. Independance: Since the students are sampled randomly and it is less than 10% of actual high school students, we can assume that the reading and writing scores of each student are independent of another
2. As the sample size is more than 30, any reasonalbe skew in the sample is accepted.

(e) The average observed difference in scores is x read-write = -0.545, and the standard deviation of the differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams?

```r
#calculating z score and p value
zscore <-(-0.545-0)/(8.887/sqrt(200))
pnorm(zscore)*2
```

```
## [1] 0.3857919
```

This shows that we cannot reject null hypotheses. Reading and writing scores are equal(not different).

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

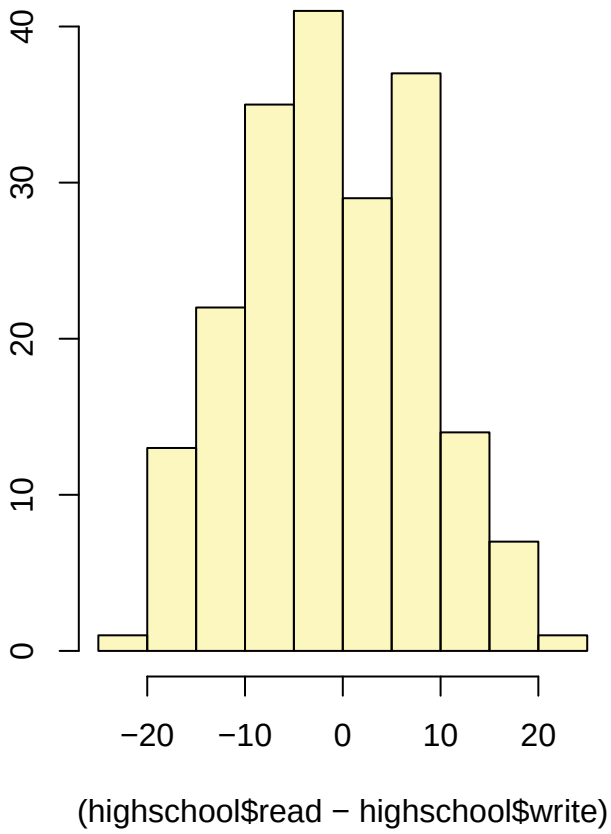
We might have made Type II error. Type II error means, in actual scenario, the null hypotheses should have been rejected.

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

Yes. The confidence interval for average difference should include 0.

```r
inference(y=(highschool$read-highschool$write),est = "mean",conflevel = 0.95,null = 0,alternative = "tw
```

```
## Single mean
## Summary statistics:
```




```
## mean = -0.545 ;  sd = 8.8867 ;  n = 200
## Standard error = 0.6284
## 95 % Confidence interval = ( -1.7766 , 0.6866 )
```

This confidence interval has 0. So we cannot reject null hypotheses.

5.32 Fuel effeciency of manual and automatic cars,

Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel effciency (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 effciency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied

Conditions for inference

1. Since the cars are randomly sampled and less than 10% of all cars, we can assume that fuel effeciency between manual and automatic transmissions are different for each car.
2. We have more than 30 samples. So any slight skewness in data should be fine.

```
carsfuel <- read.csv("https://raw.githubusercontent.com/jbryer/DATA606Fall2016/master/Data/Data%20from%
```

Hypothesyes conditions

H_0: Fuel effeciency of manual and automatic cars are equal. H_1: Fuel effeciency of manual and automatic cars are different.

Calculation from dataset

```
#Mean, SD, n

by(carsfuel$comb_mpg,carsfuel$transmission_desc,mean)
```

```
## carsfuel$transmission_desc: Automatic
## [1] 20.36879
## ------------------------------------------------------------
## carsfuel$transmission_desc: Manual
## [1] 23.64286
```

```
by(carsfuel$comb_mpg,carsfuel$transmission_desc,sd)
```

```
## carsfuel$transmission_desc: Automatic
## [1] 8.6233
## ------------------------------------------------------------
## carsfuel$transmission_desc: Manual
## [1] 4.911105
```

```
by(carsfuel$comb_mpg,carsfuel$transmission_desc,length)
```

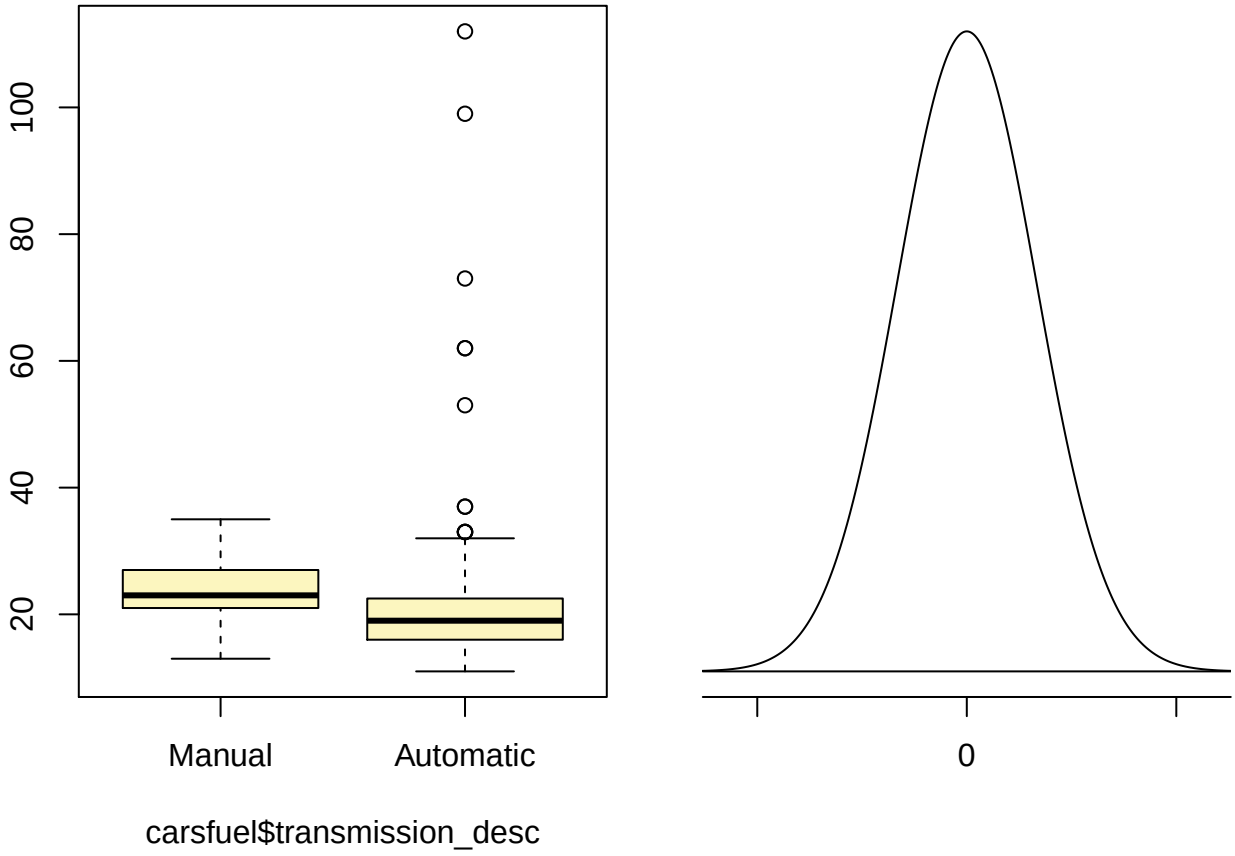
```
## carsfuel$transmission_desc: Automatic
## [1] 423
## ------------------------------------------------------------
## carsfuel$transmission_desc: Manual
## [1] 252
```

```
inference(y=carsfuel$comb_mpg,x=carsfuel$transmission_desc,est = "mean",order = c("Manual","Automatic")
```

```
## Response variable: numerical, Explanatory variable: categorical
## Difference between two means
## Summary statistics:
## n_Manual = 252, mean_Manual = 23.6429, sd_Manual = 4.9111
## n_Automatic = 423, mean_Automatic = 20.3688, sd_Automatic = 8.6233
```

```
## Observed difference between means (Manual-Automatic) = 3.2741
##
```

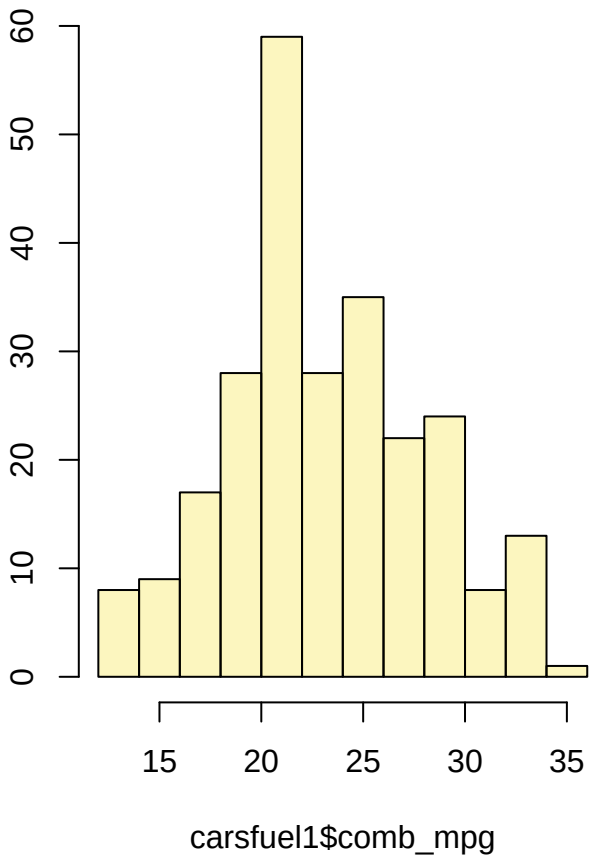
```
## H0: mu_Manual - mu_Automatic = 0
## HA: mu_Manual - mu_Automatic != 0
## Standard error = 0.521
## Test statistic: Z =  6.283
## p-value =  0
```




```
carsfuel1 <- filter(carsfuel,transmission_desc=="Manual")

#Confidence Interval of manual cars
inference(y=carsfuel1$comb_mpg,est = "mean",null = 0,alternative = "twosided",type = "ci",method = "the
```

```
## Single mean
## Summary statistics:
```




```
## mean = 23.6429 ;  sd = 4.9111 ;  n = 252
## Standard error = 0.3094
## 99 % Confidence interval = ( 22.846 , 24.4397 )
```

5.48 Work hours and education

The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents. 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.

Statistic	< HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1172

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

H_0: Mean work hours is same for all the residents with different education levels H_1: Mean work hours is different for at-least one education level compared to the other education levels among residents.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

1. Observations within the group and between the group are independent.
2. Sample random sample is less than 10% of the population.

3. Observations within each group should be nearly normal.
4. Variability across the groups should be about equal

(c) Below is part of the output associated with this test. Fill in the empty cells.

df_G Row1,Column1: 4 df_E Row2,Column1: 1167 df_T Row3,Column1: 1171

SSG Row1,Column2:

```
dfg = 4
dfe = 1167
dft=1171

ssg <- (121*(38.67-40.45)^2)+(546*(39.6-40.45)^2)+(97*(41.39-40.45)^2)+(253*(42.55-40.45)^2)+(155*(40.85
```

SSG Row3,Column2: 2004.101+267382

MSG Row1,Column3: 2004.101/4 = 501.0253 MSE Row2,Column3: 267382/1167 = 229.1191

F Score: 501.0253/229.1191

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
degree	**4**	**2004.10**	501.54	**2.188984**	0.0682
Residuals	**1167**	267,382	**229.12**		
Total	**1171**	**269386.1**			

```
pf(2.186,4,1167,lower.tail = FALSE)
```

```
## [1] 0.06852388
```

(d) What is the conclusion of the test?

As the p value is little greater than 5%, hence we failed to reject the null hypothesis. The work hours may be equal for different education level. We need more data to reject null hypothesis.