MichaelY_HW5_Inference_Numerical

Michael Y.

March 24th, 2019

```
###setwd("c:/Users/Michael/DROPBOX/priv/CUNY/MSDS/201902-Spring/DATA606-Jason/Homework")
```

Homework - Chapter 5 - Inference for numerical data (pp.219-273)

Exercises: 5.6, 5.14, 5.20, 5.32, 5.48 (pp.257-273)

Datasets:

5.20 - hsb2

5.32 - epa2012

5.48 - gss2010

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 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.

```
p56_hi <- 77
p56_lo <- 65
p56_SampleMean <- (p56_hi+p56_lo)/2
cat("SampleMean: ", p56_SampleMean, "\n")
```

```
## SampleMean:  71
```

```
p56_MarginOfError <- (p56_hi-p56_lo)/2
cat("MarginOfError: ", p56_MarginOfError, "\n")
```

```
## MarginOfError:  6
```

```
p56_n <- 25

p56_df <- p56_n-1

p56_ConfidenceInterval <- 0.90
```

```r
p56_alpha <- 1 - p56_ConfidenceInterval
cat("alpha: ", p56_alpha, "\n")
```

```
## alpha:  0.1
```

```r
p56_t_percentile <- 1 - p56_alpha/2
cat("t_percentile: ", p56_t_percentile, "\n")
```

```
## t_percentile:  0.95
```

```r
p56_t_score <- qt(p56_t_percentile, p56_df)
cat("t_score: ", p56_t_score, "\n")
```

```
## t_score:  1.7108821
```

```r
p56_StandardError <- p56_MarginOfError / p56_t_score
cat("StandardError: ", p56_StandardError, "\n")
```

```
## StandardError:  3.5069629
```

```r
p56_SampleSTDev <- p56_StandardError * sqrt(p56_n)
cat("Sample Standard Deviation: ", p56_SampleSTDev, "\n")
```

```
## Sample Standard Deviation:  17.534815
```

Sample Mean: 71 .

Margin of Error: 6 .

Sample Standard Deviation: 17.53481456 .

#.

Exercise 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.

```r
p514_MarginOfError <- 25
p514_StandardDeviation <- 250
```

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

```r
p514_ConfidenceInterval <- 0.90
p514_alpha <- 1-p514_ConfidenceInterval
cat("alpha: ", p514_alpha, "\n")
```

```
## alpha:  0.1
```

```
p514_percentile <- 1 - p514_alpha/2
cat("norm_percentile: ", p514_percentile, "\n")
```

```
## norm_percentile:  0.95
```

```
p514_z_score <- qnorm(p514_percentile)
cat("Z-score: ", p514_z_score, "\n")
```

```
## Z-score:  1.6448536
```

```
p514_StandardError <- p514_MarginOfError / p514_z_score
cat("Standard Error: ", p514_StandardError, "\n")
```

```
## Standard Error:  15.198921
```

```
p514_SampleSize <- (p514_StandardDeviation / p514_StandardError)^2
### round up
p514_SampleSize <- ceiling(p514_SampleSize)
cat("SampleSize:  ", p514_SampleSize, "\n")
```

```
## SampleSize:   271
```

The required sample size is 271 .

(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.

The required sample size for Luke will be *larger* than that required for Raina.

Increasing the confidence interval causes the required percentile to move closer to 1 (or, zero) – i.e., we are farther out along the tails.

This translates to a larger Z-score (in absolute value terms.)

For a fixed Margin of Error, a larger Z-score translates to a smaller Standard Error.

$$MarginOfError = StandardError * Zscore \quad \Rightarrow \quad StandardError = \frac{MarginOfError}{Zscore}$$

As the sample size is proportional to the square of the standard deviation divided by the standard error, a smaller Standard Error translates into an increase in the required sample size.

$$StandardError = \frac{StandardDeviation}{\sqrt{SampleSize}} \quad \Rightarrow \quad SampleSize = \left(\frac{StandardDeviation}{StandardError}\right)^2$$

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

```
p514c_ConfidenceInterval <- 0.99
p514c_alpha <- 1-p514c_ConfidenceInterval
cat("alpha: ", p514c_alpha, "\n")
```

```
## alpha:  0.01
```

```
p514c_percentile <- 1 - p514c_alpha/2
cat("norm_percentile: ", p514c_percentile, "\n")
```

```
## norm_percentile:  0.995
```

```
p514c_z_score <- qnorm(p514c_percentile)
cat("Z-score: ", p514c_z_score, "\n")
```

```
## Z-score:  2.5758293
```

```
p514c_StandardError <- p514_MarginOfError / p514c_z_score
cat("Standard Error: ", p514c_StandardError, "\n")
```

```
## Standard Error:  9.7056121
```

```
p514c_SampleSize <- (p514_StandardDeviation / p514c_StandardError)^2
### round up
p514c_SampleSize <- ceiling(p514c_SampleSize)
cat("SampleSize:  ", p514c_SampleSize, "\n")
```

```
## SampleSize:   664
```

The required sample size for a 99% confidence interval is 664 .

#.

Exercise 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.

```
data(hsb2)
#str(hsb2)
#summary(hsb2)
satread <- hsb2$read
satwrite <- hsb2$write
satdiff <- satread - satwrite
summary(satread)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   44.00   50.00   52.23   60.00   76.00
```

```r
summary(satwrite)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  31.000  45.750  54.000  52.775  60.000  67.000
```

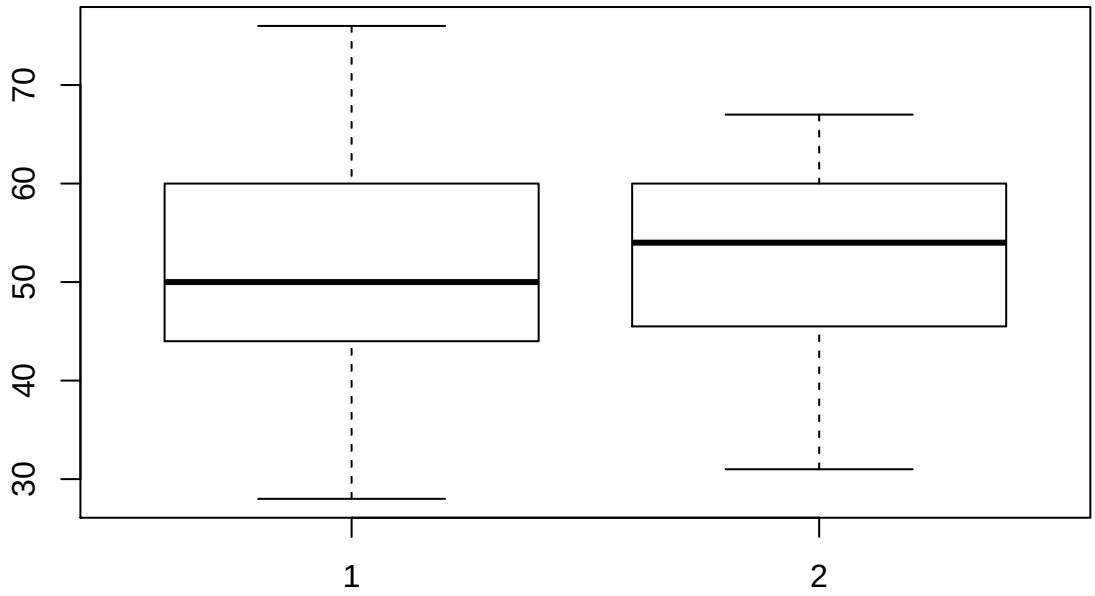
```r
summary(satdiff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -21.000  -7.000   0.000  -0.545   6.000  24.000
```

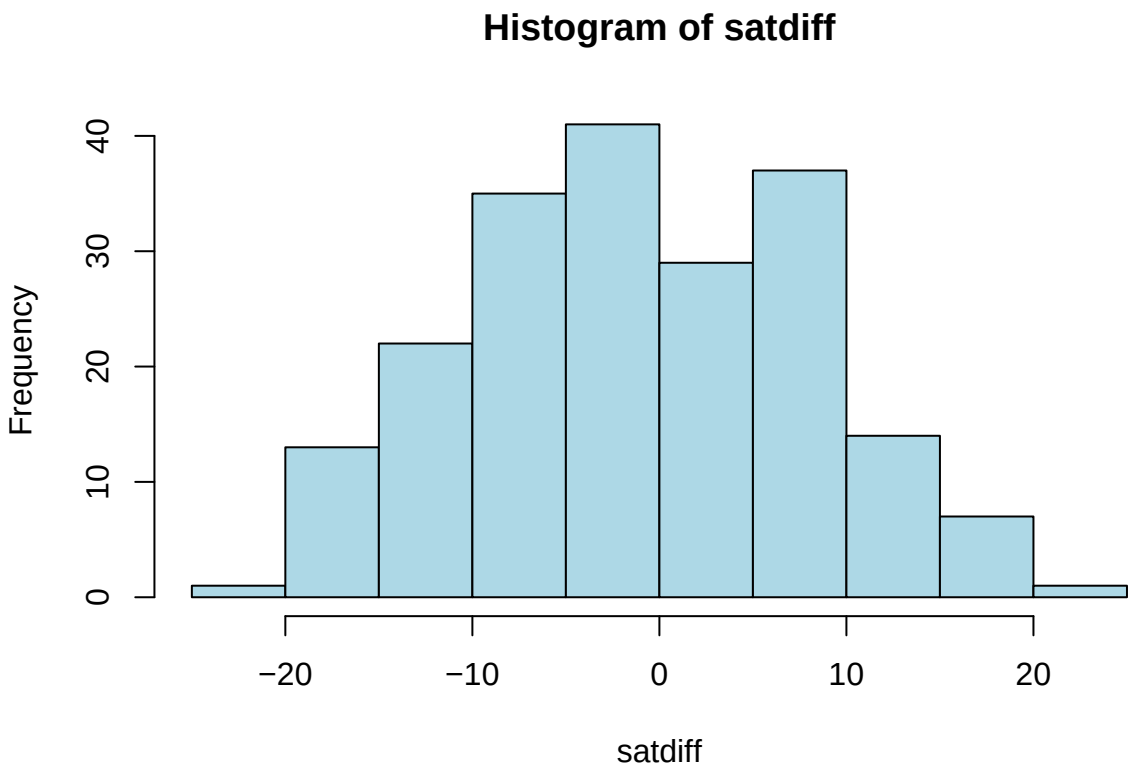
```r
cor.test(satread,satwrite)
```

```
##
##  Pearson's product-moment correlation
##
## data:  satread and satwrite
## t = 10.4652, df = 198, p-value < 0.000000000000000222
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.49938307 0.67927528
## sample estimates:
##       cor
## 0.59677648
```

```r
boxplot(satread, satwrite)
```




```r
hist(satdiff, col="lightblue")
```



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

While the above boxplots show that the median score of the writing exam is higher than the median score of the reading exam, it does not provide sufficient information regarding the average (i.e., the mean) scores.

The symmetry of the of the histogram suggests that the median and mean score difference are both centered at zero.

Therefore, the plots do not indicate a clear difference in the average reading and writing scores.

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

No, the scores of each student are not independent of each other, as a student who is good at writing is likely to be good at reading, and a student who is a poor reader is likely to be a poor writer. Indeed, the analysis above calculates a sample correlation of 0.5967765 , with a 95% confidence interval for the population correlation to be (0.4993831 , 0.6792753). If the scores were independent of each other, the correlation would be zero, as independence implies uncorrelatedness (but not vice-versa. . .)

(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?

Null Hypothesis: There is no difference in the average scores of students in the reading and writing exam.

In statistical notation:

$$H_0: \quad \mu_{read} = \mu_{write} \quad \Rightarrow \quad \mu_{read} - \mu_{write} = 0$$

where μ_{read} represents the average score of students on the reading exam and μ_{write} represents the average score of students on the writing exam.

Alternative Hypothesis: There is some difference in the average scores of students in the reading and writing exam:

$$H_0: \quad \mu_{read} \neq \mu_{write} \quad \Rightarrow \quad \mu_{read} - \mu_{write} \neq 0$$

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

(1) The distribution appears to be normal,

(2) The samples (of different students) are independent as we have randomly sampled 200 observations from a population which is certainly much larger than 2000, and

(3) Because the sample size is larger than 30, we can use the Normal distribution (rather than the Student-T) and do not need to be concerned with the modest amound of skew.

(e) The average observed difference in scores is

$$\bar{x}_{read-write} = -0.545$$

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?

```
p520_StandardDeviation <- 8.887
p520_DiffMean <- -0.545
p520_SampleSize <- 200

p520_StandardError <- p520_StandardDeviation / sqrt(p520_SampleSize)
cat("Standard Error: ",p520_StandardError ,"\n")
```

```
## Standard Error:  0.6284058
```

```
p520_Zscore <- p520_DiffMean / p520_StandardError
cat("Zscore: ", p520_Zscore, "\n")
```

```
## Zscore:  -0.86727399
```

```r
p520_pval <- 2*pnorm(p520_Zscore)
cat("pval: ", p520_pval)
```

```
## pval:  0.38579191
```

Because the pval=0.38579191 is so large, these data do not provide convincing evidence of a difference between the average scores on the two exams. Therefore we cannot reject the null hypothesis.

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

A Type 2 Error occurs when we fail to reject the null hypothesis when the alternative is actually true. In this context, we have concluded that there is no difference between the average scores on the two exams. We would have a type 2 error if there is indeed a significant difference, which may have occurred if we happen to have been unlucky with our drawn sample.

(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.

```r
p520_ConfidenceInterval= .9
p520_alpha <- 1-p520_ConfidenceInterval
cat("alpha: ", p520_alpha, "\n")
```

```
## alpha:  0.1
```

```r
p520_percentile <- 1 - p520_alpha/2
cat("norm_percentile: ", p520_percentile, "\n")
```

```
## norm_percentile:  0.95
```

```r
p520_p95_Zscore <- qnorm(p520_percentile)
cat("p95 Zscore :", p520_p95_Zscore, "\n")
```

```
## p95 Zscore : 1.6448536
```

```r
p520_p95_MarginOfError <- p520_p95_Zscore * p520_StandardError
cat("p95 Margin of Error: ",p520_p95_MarginOfError,  "\n")
```

```
## p95 Margin of Error:  1.0336356
```

```r
p520_lo <- p520_DiffMean - p520_p95_MarginOfError
p520_hi <- p520_DiffMean + p520_p95_MarginOfError
cat("90% Confidence Interval: (", p520_lo, ",",p520_hi, ")\n" )
```

```
## 90% Confidence Interval: ( -1.5786356 , 0.48863555 )
```

Yes, a confidence interval should include zero. Although no confidence interval has been specified, our sample mean is less than 1 standard deviation from zero. In the case of, say, a 90% confidence interval, the bounds would be (-1.57863555 , 0.48863555) , which includes zero. Indeed, we would have to weaken the confidence interval to 61% in order to fail to include zero.

#.

Exercise 5.32 Fuel efficiency 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 efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions manufactured in 2012.

```
data(epa2012)
#str(epa2012)
#summary(epa2012)
epa <- cbind(rownum=seq(nrow(epa2012)),epa2012)
bigepa <- spread(epa,key = transmission_desc, value=city_mpg)
MPGmanual <- bigepa[!is.na(bigepa$Manual),]$Manual
#summary(MPGmanual) ; sd(MPGmanual)
MPGautomatic <- bigepa[!is.na(bigepa$Automatic),]$Automatic
#summary(MPGautomatic) ; sd(MPGautomatic)
```

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 the conditions for inference are satisfied.

Null Hypothesis: There is no difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage.

In statistical notation:

$$H_0: \quad \mu_{automatic} = \mu_{manual} \quad \Rightarrow \quad \mu_{automatic} - \mu_{manual} = 0$$

where $\mu_{automatic}$ represents the average city MPG of cars with automatic transmissions and μ_{manual} represents the average city MPG of cars with manual transmissions.

Alternative Hypothesis: There is some difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage:

$$H_0: \quad \mu_{automatic} \neq \mu_{manual} \quad \Rightarrow \quad \mu_{automatic} - \mu_{manual} \neq 0$$

```
p532_CityMPG <- matrix(c(16.12, 19.85, 3.58, 4.51, 26, 26), 3, 2, byrow = T, dimnames = list(c("Mean","S
p532_CityMPG
```

```
##      Automatic Manual
## Mean     16.12  19.85
## SD        3.58   4.51
## n        26.00  26.00
```

```
p532_mean_automatic  <- p532_CityMPG["Mean","Automatic"]
p532_sd_automatic    <- p532_CityMPG["SD","Automatic"]
p532_n_automatic     <- p532_CityMPG["n","Automatic"]
```

```r
p532_mean_manual     <- p532_CityMPG["Mean","Manual"]
p532_sd_manual       <- p532_CityMPG["SD","Manual"]
p532_n_manual        <- p532_CityMPG["n","Manual"]

p532_diff_mean_MPG   <- p532_mean_automatic - p532_mean_manual
cat("Difference in average MPG between Automatic vs. Manual: ", p532_diff_mean_MPG,"\n")
```

```
## Difference in average MPG between Automatic vs. Manual:  -3.73
```

```r
p532_StandardError   <- sqrt((p532_sd_automatic^2)/p532_n_automatic + (p532_sd_manual^2)/p532_n_manual)
cat("StandardError:", p532_StandardError, "\n")
```

```
## StandardError: 1.1292697
```

```r
p532_t_score         <- (p532_diff_mean_MPG - 0) / p532_StandardError
cat("T-score: ", p532_t_score, "\n")
```

```
## T-score:  -3.3030197
```

```r
p532_df <- min(p532_n_automatic-1, p532_n_manual-1)
cat("Degrees of Freedom: ", p532_df, "\n")
```

```
## Degrees of Freedom:  25
```

```r
p532_pval <- pt(q=p532_t_score,df = p532_df ) * 2
cat("pval: ", p532_pval, "\n")
```

```
## pval:  0.0028836148
```

Yes, these data provide strong evidence of a difference in fuel economy between automatic and manual transmission vehicles, based upon the City MPG sample given, as the p-value is 0.00288361 .

#.

Exercise 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.

```r
### The General Social Survey from NORC for 2010
gss2010 <- read.csv("gss2010.csv")
#summary(gss2010)
### The columns of interest are "hrs1" for hours worked and "degree" for highest degree.

### split out the hours from the individual column into 5 separate columns
HoursByDegree  <- gss2010 %>%
                  spread(key = degree,value = hrs1) %>%
```

```
                  select((names(table(gss2010$degree)))[c(5,3,4,1,2)])
### the above caused the original hrs1 column to be dropped.
###Replace it, but call it "Total"
HoursByDegree$TOTAL <- gss2010$hrs1

###Drop all rows where the number of hours was NA -- this reduces the table from 2044 rows to 1172
HoursByDegree <- HoursByDegree[!is.na(HoursByDegree$TOTAL),]

###compute the means, sd, and count for each column
means <- apply(X = HoursByDegree, 2, mean, na.rm=T)
stdevs <- apply(X = HoursByDegree, 2, sd, na.rm=T)
counts <- unlist(lapply(apply(X = HoursByDegree, 2,na.omit),length))

### reproduce the chart that appears in the textbook on page 272
Educ.Attainment <- rbind(means,stdevs,counts)
Educ.Attainment
```

```
##        LT HIGH SCHOOL HIGH SCHOOL JUNIOR COLLEGE   BACHELOR   GRADUATE
## means       38.669421   39.597070      41.391753  42.549407  40.845161
## stdevs      15.814228   14.971248      18.103612  13.617308  15.505400
## counts     121.000000  546.000000      97.000000 253.000000 155.000000
##              TOTAL
## means    40.452218
## stdevs   15.167393
## counts 1172.000000
```

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

Null Hypothesis: There is no difference between the average number of hours worked across each of the five groups -i.e., the average number of hours worked is the same across all groups:

$$H_0: \quad \mu_{lths} = \mu_{hs} = \mu_{jc} = \mu_{bach} = \mu_{grad}$$

where

μ_{lths} represents the average hours worked by the members of the group with less than high school education,
μ_{hs} represents the average hours worked by the members of the group with a high school diploma,
μ_{jc} represents the average hours worked by the members of the group with a junior college degree,
μ_{bach} represents the average hours worked by the members of the group with a bachelor's degree, and
μ_{grad} represents the average hours worked by the members of the group with a graduate degree.

Alternative Hypothesis: The average number of hours worked across each of the five groups differs.

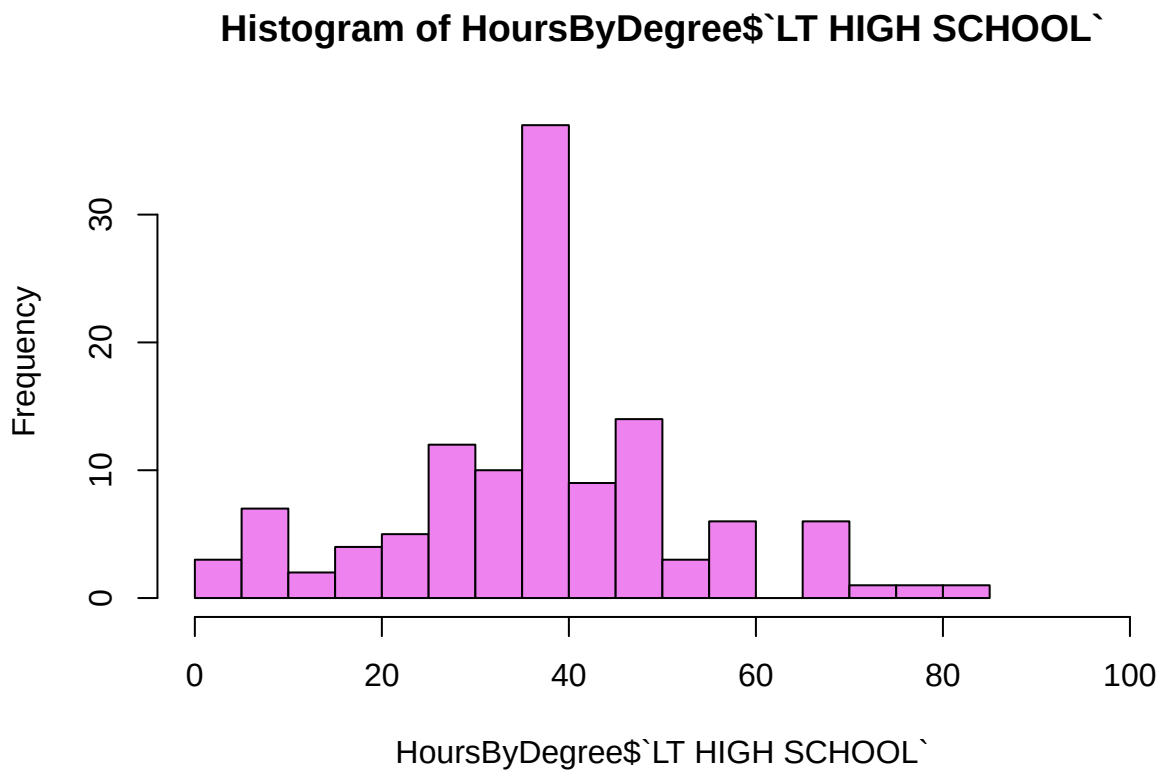
(b) Check conditions and describe any assumptions you must make to proceed with the test.

According to the textbook:

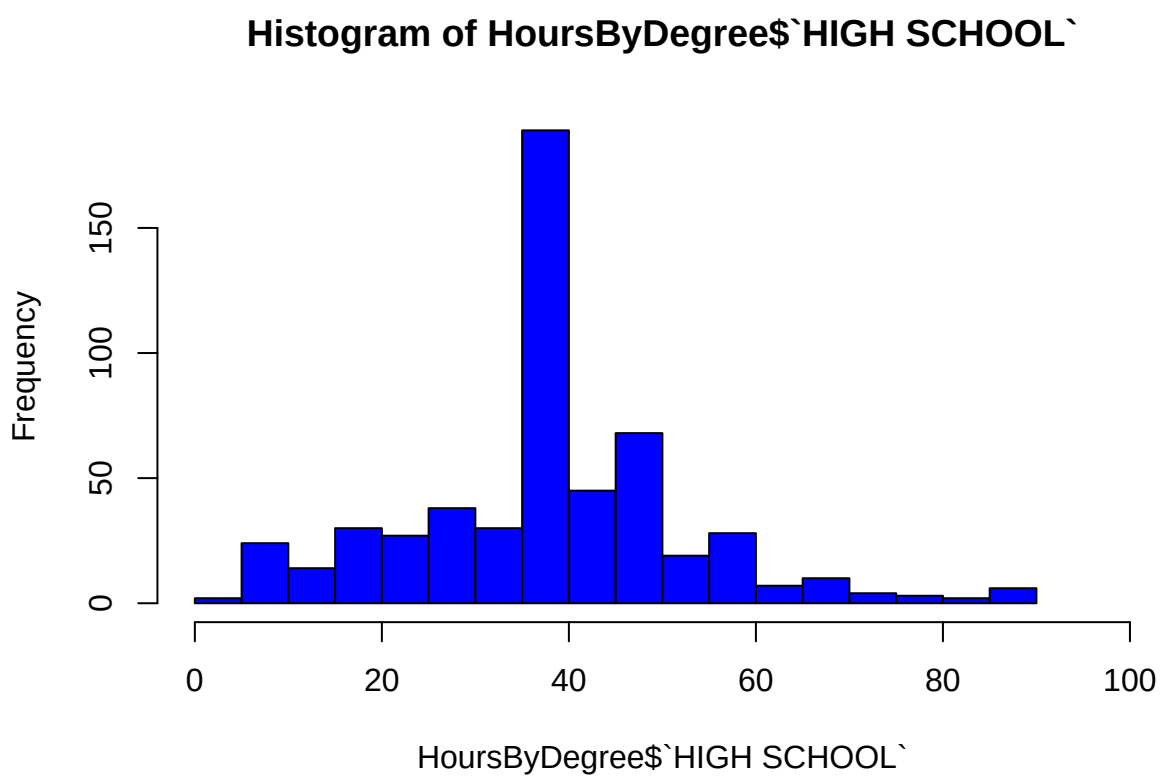
Generally we must check three conditions on the data before performing ANOVA: • the observations are independent within and across groups, • the data within each group are nearly normal, and • the variability across the groups is about equal.

(1) We can assume that the observations are independent as the NORC makes a great effort to randomly select a broadly distributed group for their biannual surveys. However, it is noteworthy that data on hours is available for only 57% of the cases (1172 out of 2044), forcing the elimination of 872 observations. Either these individuals are out-of-the-workforce (e.g., unemployed, retired, etc.), or we need to assume that the number of hours worked for these individuals is proportional to the figures for those who responded.

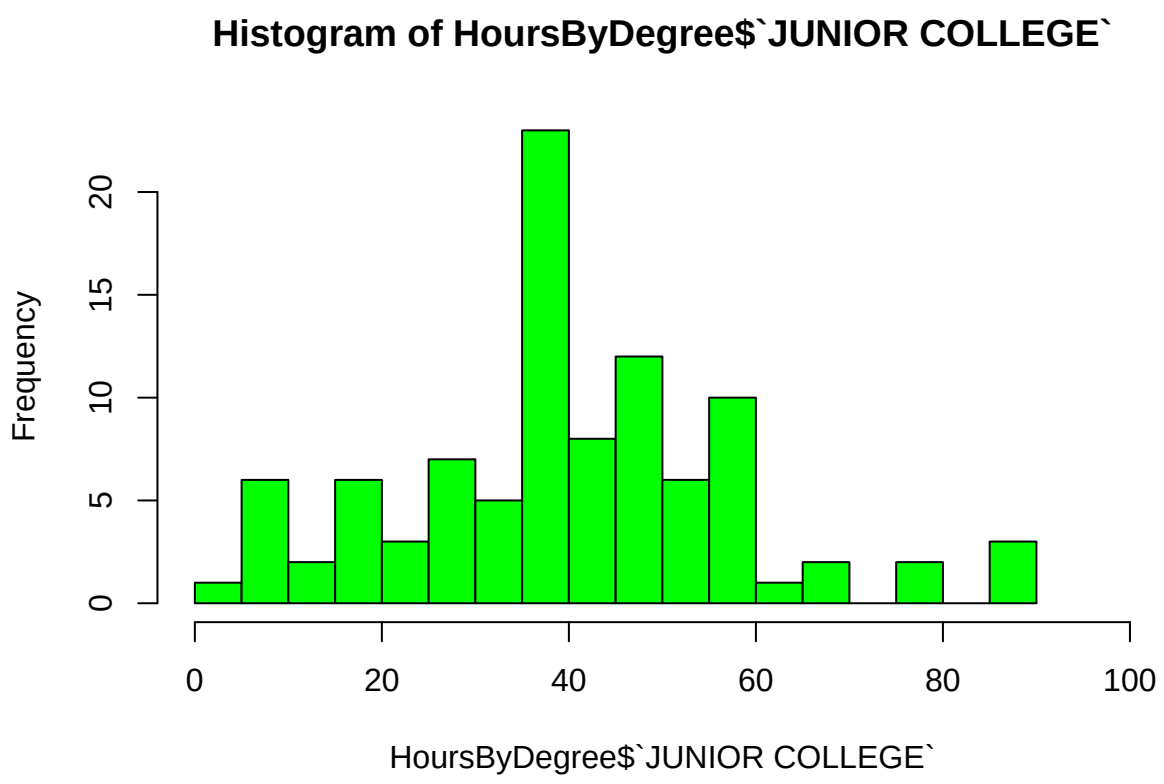
```
hist(HoursByDegree$`LT HIGH SCHOOL`,breaks = 20,xlim = c(0,100),col="violet")
```



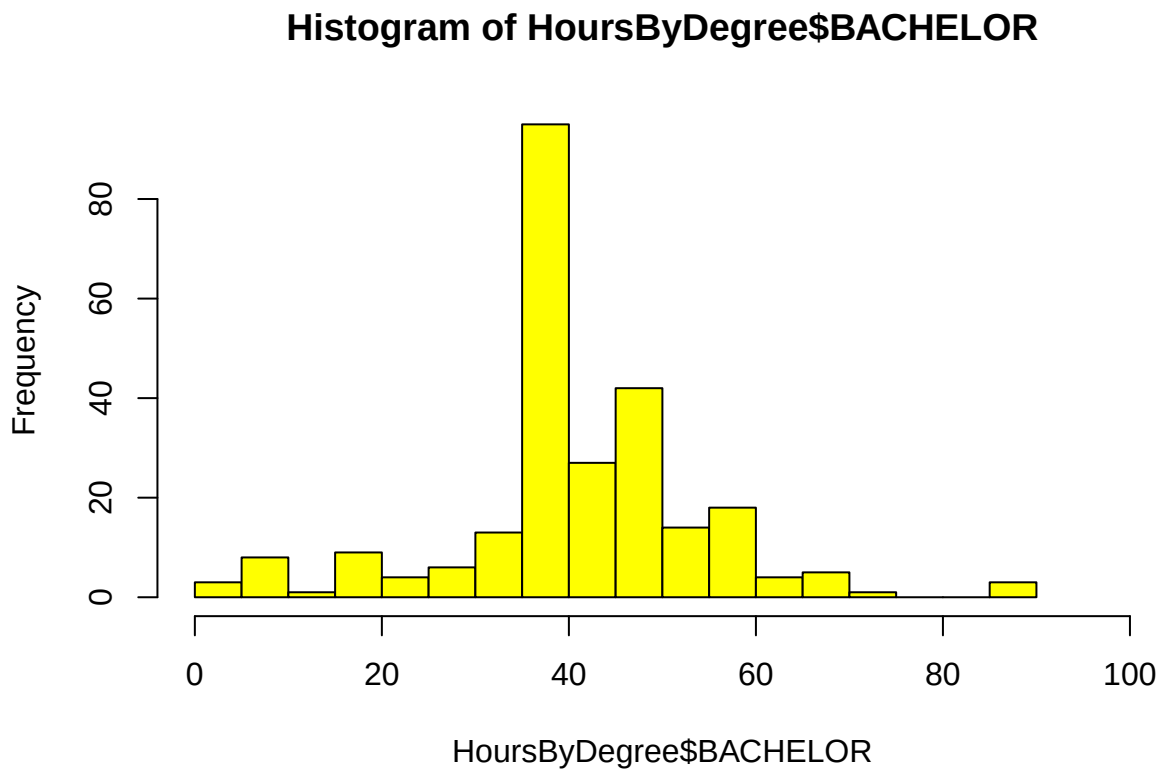

```
hist(HoursByDegree$`HIGH SCHOOL`,breaks = 20,xlim = c(0,100),col="blue")
```



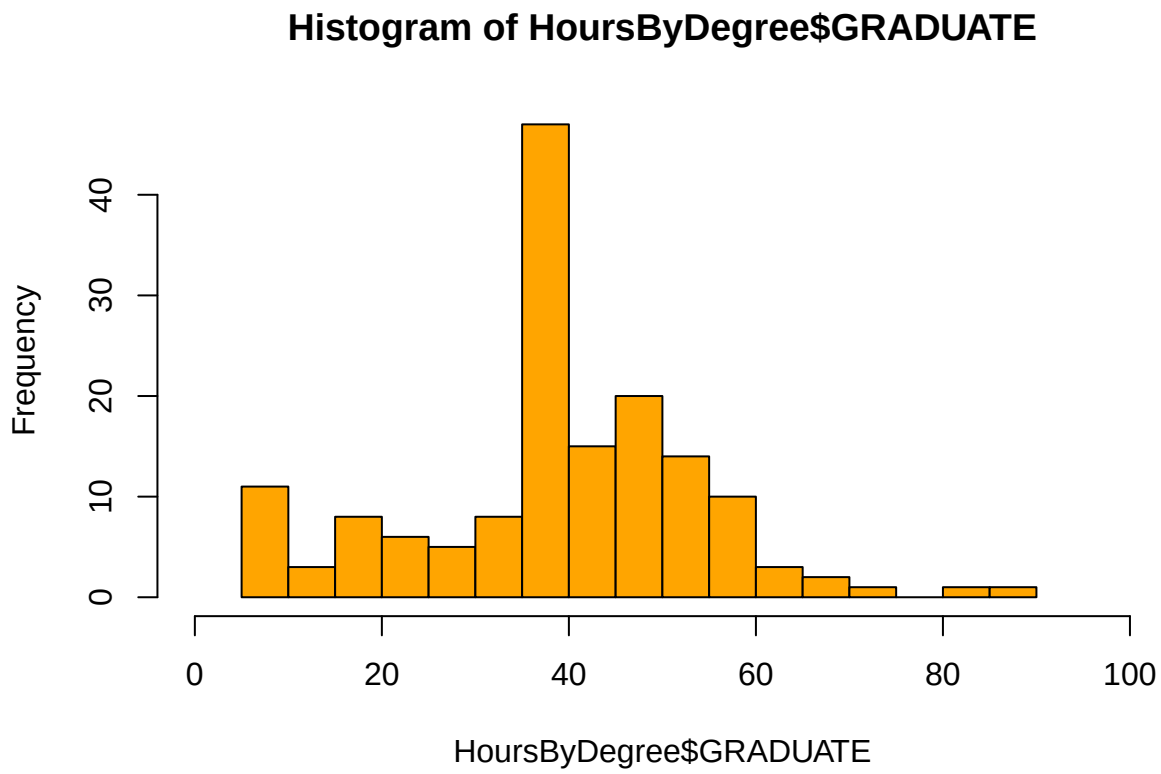

```
hist(HoursByDegree$`JUNIOR COLLEGE`,breaks = 20,xlim = c(0,100),col="green")
```



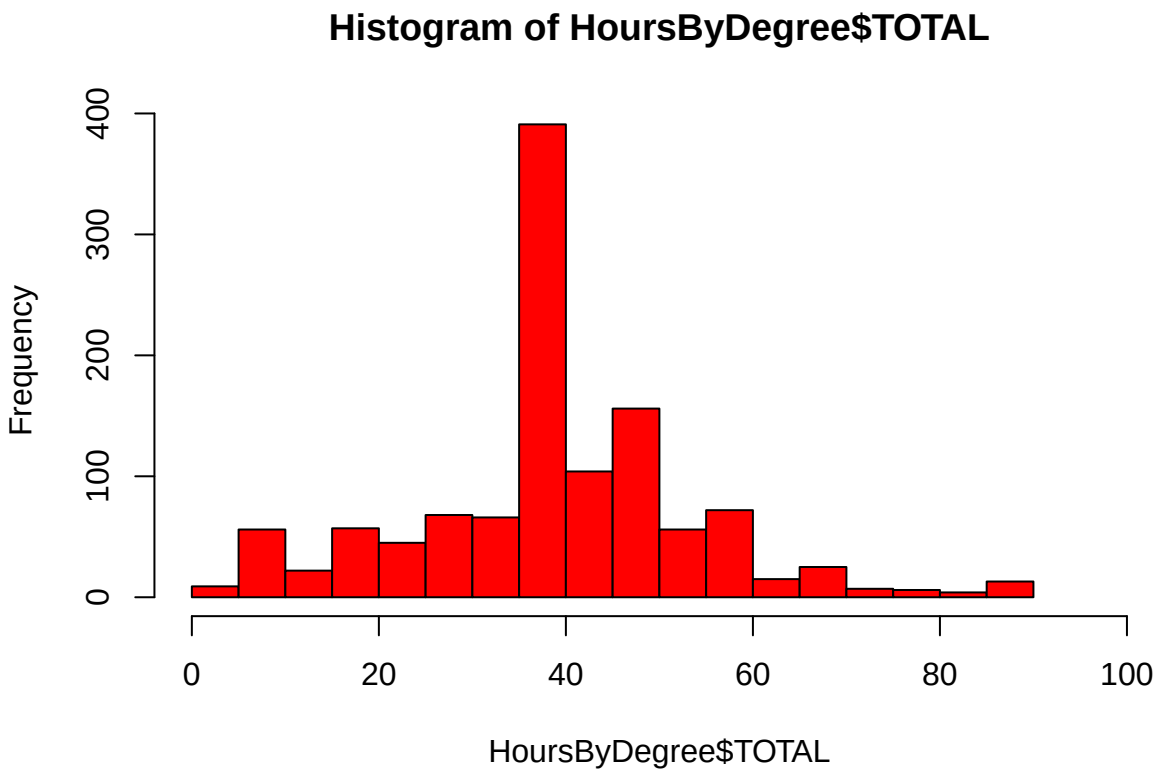

```
hist(HoursByDegree$BACHELOR,breaks = 20,xlim = c(0,100),col="yellow")
```




```
hist(HoursByDegree$GRADUATE,breaks = 20,xlim = c(0,100),col="orange")
```




```
hist(HoursByDegree$TOTAL,breaks = 20,xlim = c(0,100),col="red")
```



(2) From the above histograms, the data within each group somewhat resembles a normal distribution, except there are a very large number of cases (335 out of 1172, or 28 percent) where the individual reports 40 hours, which is understandable as that is considered a "full-time" workweek in the US.

(3) The standard deviations are all in the same "ballpark", ranging from a low of 13.62 to a high of 18.1 .

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

```
## create an empty grid
p548_grid <- data.frame(matrix(rep(NA,3*5),3,5,
                               dimnames = list(c("degree","Residuals","Total"),
                                               c("Df","SumSq","MeanSq","Fvalue","Pr(>F)")))
)
## correct the column header
colnames(p548_grid)[5]<-"Pr(>F)"
## populate the 3 given values
p548_MSG <- p548_grid["degree","MeanSq"] <- 501.54

p548_pval <- p548_grid["degree","Pr(>F)"] <- .0682

p548_SSE <- p548_grid["Residuals","SumSq"] <- 267382
```

```
## display the starting grid
p548_grid
```

```
##           Df  SumSq MeanSq Fvalue Pr(>F)
## degree    NA     NA 501.54     NA 0.0682
## Residuals NA 267382     NA     NA     NA
## Total     NA     NA     NA     NA     NA
```

Find the missing values:

```
p548_n <- Educ.Attainment["counts","TOTAL"]
cat("Total (n): ", p548_n, "\n")
```

```
## Total (n):  1172
```

```
p548_k <- 5
cat("Groups (k): ", p548_k, "\n")
```

```
## Groups (k):  5
```

```
### Compute the degrees of Freedom column
p548_dfg <- p548_k-1
cat("Degree-DF (dfg): ", p548_dfg, "\n")
```

```
## Degree-DF (dfg):  4
```

```
p548_dfe <- p548_n - p548_k
cat("Residuals-DF (dfe): ", p548_dfe, "\n")
```

```
## Residuals-DF (dfe):  1167
```

```
p548_dft <- p548_dfg + p548_dfe
cat("Total-Df (dft)", p548_dft, "\n" )
```

```
## Total-Df (dft) 1171
```

```
### Compute the sum-of-squares column
p548_SSG <- p548_MSG * p548_dfg
cat("degree-SumSq (SSG): ", p548_SSG, "\n")
```

```
## degree-SumSq (SSG):  2006.16
```

```
p548_SST <- p548_SSG + p548_SSE
cat("Total-SumSq (SST): ", p548_SST, "\n")
```

```
## Total-SumSq (SST):  269388.16
```

```
### Compute the Mean Squares column
p548_MSE <- p548_SSE / p548_dfe
cat("Residuals-MeanSq (MSE) :", p548_MSE, "\n")
```

```
## Residuals-MeanSq (MSE) : 229.11911
```

```
### Compute the F statistic
p548_F <- p548_MSG / p548_MSE
cat("F-statistic: ", p548_F, "\n")
```

```
## F-statistic:  2.1889925
```

Put the computed values in the grid:

```
p548_grid["degree","Df"]        <- p548_dfg
p548_grid["Residuals","Df"]     <- p548_dfe
p548_grid["Total","Df"]         <- p548_dft
p548_grid["degree","SumSq"]     <- p548_SSG
p548_grid["Total","SumSq"]      <- p548_SST
p548_grid["Residuals","MeanSq"] <- p548_MSE
p548_grid["degree","Fvalue"]    <- p548_F

p548_grid
```

```
##             Df     SumSq    MeanSq    Fvalue Pr(>F)
## degree       4   2006.16 501.54000 2.1889925 0.0682
## Residuals 1167 267382.00 229.11911        NA     NA
## Total     1171 269388.16        NA        NA     NA
```

Check the F statistic vs. the given pval:

```
cat("Given pval:      ", p548_pval, "\n")
```

```
## Given pval:       0.0682
```

```
p548_calculated_pval <- pf(p548_F,p548_dfg,p548_dfe,lower.tail = F)
cat("Calculated pval: ", p548_calculated_pval, "\n")
```

```
## Calculated pval:  0.068193249
```

Check results vs. actual ANOVA calculations on the full GSS data set:

```
p548_linearmodel <- lm(formula = gss2010$hrs1 ~ gss2010$degree)
p548_anovatable <- anova(p548_linearmodel)
p548_actual_ANOVA <- rbind(data.frame(p548_anovatable),Total=c(sum(p548_anovatable$Df), sum(p548_anovata
p548_actual_ANOVA
```

```
##                  Df      Sum.Sq   Mean.Sq   F.value      Pr..F.
## gss2010$degree    4   2006.1624 501.54059 2.1889937 0.068193109
## Residuals      1167 267382.1619 229.11925        NA          NA
## Total          1171 269388.3242        NA        NA          NA
```

(d) What is the conclusion of the test?

Fail to reject the null because the p-value is $0.068 > 0.05$.

There is insufficient evidence to support the alternative hypothesis, i.e., that the average number of hours worked differs across educational attainment levels.