RMarkdown STAT 2218 Problem Set 1

Jack Ambery

Due Date 9/29/2023

For Problems 1 through 4 you will use the data set MedSchool.csv. Each problem is worth 10 points for a total of 40 points.

Data: MedSchool.csv used for Problems 1 to 4. Data collected from a random sample of students who applied to medical school in a given year. It contains the medical school admission status and information on GPA and standardized test scores for each subject.

* Subject: Code for different student subjects
* Acceptance: Indicator for whether student was accepted to med school or not: 1=accepted or 0=denied
* Sex: F=female or M=male
* SCIGPA: Grade Point Average in Science and Math classes
* GPA: Overall undergraduate grade point average
* Verbal: Verbal reasoning subscore on MCAT exam
* PhySci: Physical sciences subscore on MCAT exam
* Writing: Writing subscore on MCAT exam
* BioSci: Biological sciences subscore on MCAT exam
* MCAT: Score on the MCAT exam (sum of subscores above)
* NumApps: Number of medical schools applied to

# Input data file below.

Import data into RStudio using the base option, and then cut and paste YOUR file directory in the read.csv command below. It may be different from what I have listed below since your files are located in a different directory/folder. Then you can attach the file.

#Your file will be different! See instructions above.  
MedSchool <- read.csv("C:/Users/jacka/OneDrive - fairfield.edu/Fall 2023/Statistics II/datasets/MedSchool.csv")  
attach(MedSchool)

For Problems 1 - 4 run an appropriate hypothesis test. Use a significance level of 5%. Be sure to include:

* Your null and alternative hypotheses, defining your parameters,
* The type of test you used, along with your R code and output,
* A sentence summarizing your conclusion,
* If asked, comment on whether all assumptions of the test are met. Include evidence to support your claims.

# 1. Problem 1

Is there evidence that the average GPA of all students who did not get into medical school is less than the average GPA of all students accepted into medical school? Test using a significance level of 5%. Assess all conditions of the assumptions.

## Problem 1 R Code

#Insert R code here  
  
# This is to figure out whether to use a pooled or unpooled test, dependent on what the p-value results in.  
var.test(GPA~Acceptance)

##   
## F test to compare two variances  
##   
## data: GPA by Acceptance  
## F = 1.4335, num df = 24, denom df = 29, p-value = 0.3528  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.6655245 3.1788011  
## sample estimates:  
## ratio of variances   
## 1.433544

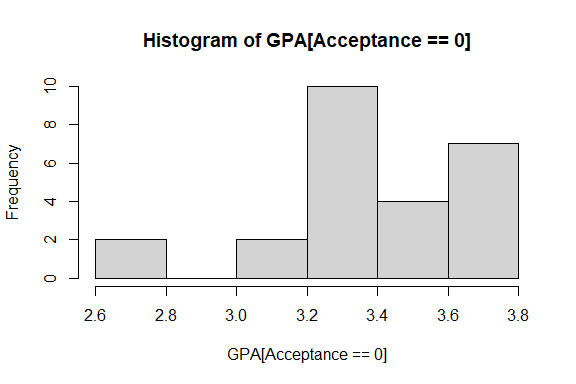
t.test(GPA~Acceptance, var.equal = TRUE, alternative = 'less')

##   
## Two Sample t-test  
##   
## data: GPA by Acceptance  
## t = -4.6775, df = 53, p-value = 1.021e-05  
## alternative hypothesis: true difference in means between group 0 and group 1 is less than 0  
## 95 percent confidence interval:  
## -Inf -0.1978508  
## sample estimates:  
## mean in group 0 mean in group 1   
## 3.385200 3.693333

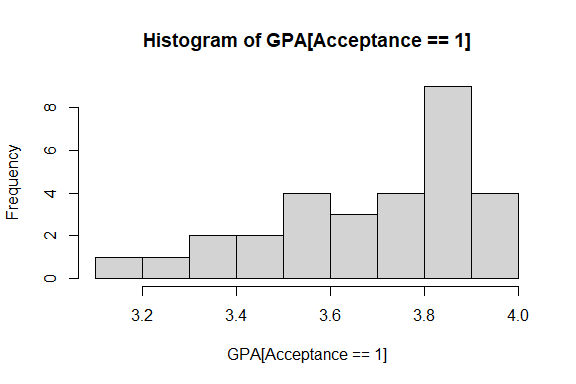
#Checking assumptions:  
  
tapply(GPA, Acceptance, shapiro.test)

## $`0`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.92748, p-value = 0.07602  
##   
##   
## $`1`  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.92006, p-value = 0.02692

hist(GPA[Acceptance == 0])



hist(GPA[Acceptance == 1])



## Problem 1 Analysis

State your hypotheses (defining your parameter(s)); the test used; the test statistic, degree(s) of freedom, and p-value; write your conclusion and explain whether all of the assumptions of the test are met. Refer to the relevant R output above.

#Insert your analysis here

= 5%

µ1 = average GPA of all students who did not get accepted into medical school

µ2 = average GPA of all students who did get into medical school

H0: µ1 = µ2

Ha: µ1 != µ2

Using var.test, I figured out that:

p-value = 0.353

Since this p-value is greater than , we will use a pooled t-test.

Then from the results of the pooled t-test, I found:

t\* = -4.68 p-value = 1.02E-5

Therefore at a 5% confidence level, we have sufficient evidence to show that the average GPA of all students who were not accepted into med school was lower than the average GPA of all the students who did get accepted into med school.

Checking the assumptions:

1. Two independent populations and two unknown means µ1, µ2 - This is true, the two groups of students are indpendent from eachother and the means are not explicitly given in the problem.
2. Assumed that the two population standard deviations are equal (σ1 = σ2) - We proved this to be true using var.test
3. Samples from each population are randomly selected - This was said to be true in the problem.
4. n1 and n2 < 10% of their respective population sizes - It is safe to assume the 55 students in this test make up less than 10% of all medical school applicants in a given year.
5. Distributions of each population are approximately normal:

To test if these populations are normally distributed, we can use the Shapiro-Wilk test. For this Shapiro-Wilk test:

H0: Population is normally distributed

Ha: Population is not normally distributed.

and we will use a significance level of = 5%.

t = -4.6775 df = 53 p-value = 1.021e-05

Since the p-value is significantly less than 0.05, we can accept that these populations are normally distributed.

We can also look at the histograms for each group above. These show us that the populations are approximately normal as well since they are not severely skewed.

At the 5% significance level, there is evidence to support that the mean of the GPAs of students accepted to medical school is greater than the average GPAs of students who were not accepted into medical school. The assumptions of the test were all met as well.

# 2. Problem 2

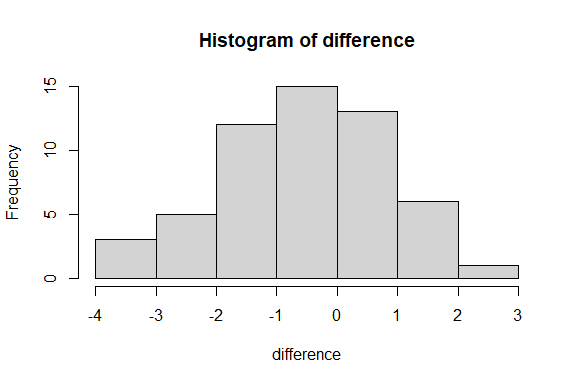
Is there evidence of a difference in the average score on the Physical Sciences portion of the MCAT exam and the average score on the Biological Sciences portion of the MCAT exam? Test using a significance level of 5%. Also check the normality assumption of the test.

## Problem 2 R Code

#Insert R code here  
  
# This creates the new array of mu1 - mu2  
difference = MedSchool$PhySci - MedSchool$BioSci  
  
# mu is 0 here because we are assuming each value in our new column is 0 (the scores are equal)  
t.test(difference, mu = 0, alternative = "two.sided")

##   
## One Sample t-test  
##   
## data: difference  
## t = -0.375, df = 54, p-value = 0.7091  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -0.4615523 0.3160978  
## sample estimates:  
## mean of x   
## -0.07272727

# Checking the assumptions  
  
hist(difference)



## Problem 2 Analysis

State your hypotheses (defining your parameter(s)); the test used; the test statistic, degree(s) of freedom, and p-value; write your conclusion and explain whether all of the assumptions of the test are met. Refer to the relevant R output above.

#Insert your analysis here

= 5%

µ1 = average score on the Physical Sciences portion

µ2 = average score on the Biological Sciences portion

H0: µ1 = µ2

Ha: µ1 != µ2

Since our data is paired in this problem, I first made a new column of data called “difference” which was defined as µ1 - µ2. I then ran a one sample t-test using “difference.”

The t-test gave me:

t = -0.375 df = 54 p-value = 0.71

Since the p-value was significantly larger than , we can conclude that there is not enough evidence to support Ha.

Therefore, at the 5% level, there is no evidence to support that there was a difference between the scores of the Physical Science Section and Biological Section of the MCAT among the sample of recent medical school applicants.

Checking the assumptions:

1. The two populations/samples are dependent - This is true because scores from each individual test are compared, it would not make sense to compare one student’s physics score to a different student’s biology score.
2. Within the population, the subjects are chosen independently. - This is true. The problem states students were chosen randomly and each student and their data are independent of all others.
3. n < 10% of the population size - It is safe to assume that 55 students is less than 10% of all medical school applicants of a given year.
4. The differences have an approximately normal distribution. Normality can be check by creating a histogram of the differences. Looking at the histogram generated above, since n is large (n > 30) and the graph is not severely skewed, it is safe to assume the differences are normally distributed.

At the 5% significance level, there is not evidence that the mean scores of the Physical Sciences section of the MCAT from the first population is different than the mean of the scores of the Biological Sciences section from the second population [dbar = -0.07, Sd = 1.44, n = 55, t\* = -0.375, df = 54, p-value = 0.71]. The assumptions of the test were met.

# 3. Problem 3

Is there a difference in the average number of medical schools students applied to based on the sex? Test at the 5% level and assess the normality condition(s) of the assumptions.

## Problem 3 R code

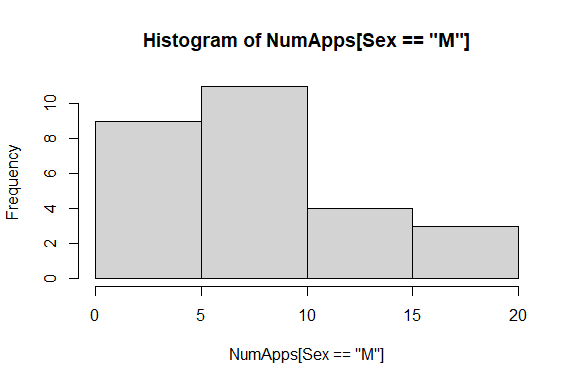
#Insert code here  
  
# var.test to ensure equal/unequal variance between male and female.  
var.test(NumApps ~ Sex)

##   
## F test to compare two variances  
##   
## data: NumApps by Sex  
## F = 1.1798, num df = 27, denom df = 26, p-value = 0.6754  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.5402426 2.5618922  
## sample estimates:  
## ratio of variances   
## 1.179843

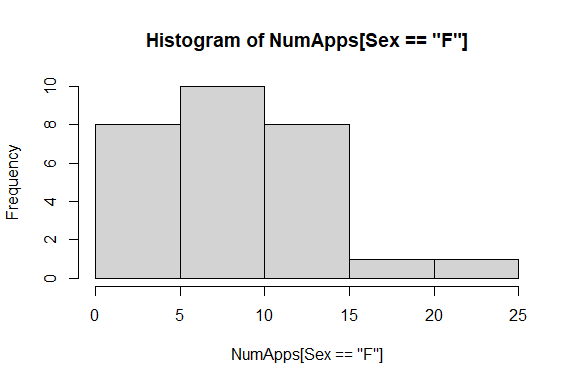
# pooled t-test  
t.test(NumApps ~ Sex, mu = 0, var.equal = TRUE, alternative = "two.sided")

##   
## Two Sample t-test  
##   
## data: NumApps by Sex  
## t = 0.61114, df = 53, p-value = 0.5437  
## alternative hypothesis: true difference in means between group F and group M is not equal to 0  
## 95 percent confidence interval:  
## -1.795997 3.370071  
## sample estimates:  
## mean in group F mean in group M   
## 8.750000 7.962963

# checking the assumptions  
#histogram of male applications  
hist(NumApps[Sex == 'M'])



#histogram of female applications  
hist(NumApps[Sex == 'F'])



## Problem 3 Analysis

State your hypotheses (defining your parameter(s)); the test used; the test statistic, degree(s) of freedom, and p-value; write your conclusion and explain whether the normality assumption(s) of the test are met. Refer to the relevant R output above.

#Insert your analysis here

= 5%

µ1 = average number of applications submitted by male students

σ1 = standard deviation of the number of applications submitted by male students

µ2 = average number of applications submitted by female students

σ2 = standard deviation of the number of applications submitted by female students

Since these samples are independent, I used var.test to determine if the variances between the two samples are equal or unequal. For this test:

H0: σ1 = σ2

Ha: σ1 != σ2

var.test gave me:

F = 1.18 p-value = 0.68

Since the p-value is greater than , it is likely that σ1 = σ2 and H0 is true. Therefore we will use a pooled t-test.

For the pooled t-test:

H0: µ1 = µ2

Ha: µ1 != µ2

Running the t.test above gave:

t = 0.61 df = 53 p-value = 0.54

Since the p-value is greater than , 5%, there is not enough evidence to support Ha.

Checking the assumptions:

1. Two independent populations with unknown means µ1, µ2 -
2. Assumed that the two population standard deviation are equal(σ1 = σ2) - this was proved to be true earlier using var.test
3. Samples from each population are randomly selected. - This is true, stated in the problem.
4. n1 and n2 < 10% of their respective population sizes - It is safe to assume that the 55 male and female students make up less than 10% of male and female students who apply to medical school in a given year.
5. Distributions of each populations are approximately normal

This can be checked by creating histograms for each the male students and female students. The two histograms above appear to both be skewed to the right. Since both n1 and n2 are less than 30, but the histograms are skewed, it is not safe to assume each sample is normally distributed.

At the 5% significance level, there is not evidence that the mean number of applications submitted my male students from the first population is different that the mean number of applications submitted by female students from the second population [xbar1 = 7.96, S1 = 4.57, n1 = 27, xbar2 = 8.75, S2 = 4.96, n2 = 28, t\* = 0.611(pooled), df = 53, p-value = 0.543]. The assumptions were not all met since there was not enough evidence that either population was normally distributed.

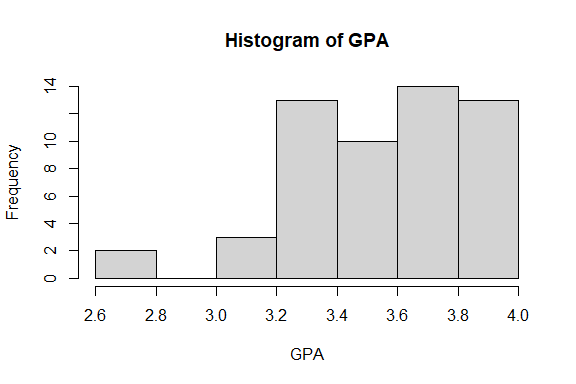
# 4 Problem 4

## Part a.

Assess whether GPA is likely to have come from a normal distribution. Provide evidence to support your claim, using two different criteria.

### Problem 4a R Code

#Insert your code here  
  
# GPA histogram  
hist(GPA)



# Shapiro-Wilk test:  
shapiro.test(GPA)

##   
## Shapiro-Wilk normality test  
##   
## data: GPA  
## W = 0.95019, p-value = 0.02344

### Problem 4a Analysis

#Insert your analysis here

The first way we can assess normality of the GPAs is by creating a histogram. Because the histogram above is not symmetric or bell-shaped, it is likely that GPA is not normally distributed.

Another way we can assess normality is by running the Shapiro-Wilk test on the dataset. For this test:

H0: Population is normally distributed

Ha: Population is not normally distributed.

and we will use a significance level of = 5%.

According to the Shapiro-Wilk test run above, it is clear that the data set is not normally distributed because the p-value is 0.023 which is significantly lower than which is 0.05. This means that we reject the assumption that H0 is true.

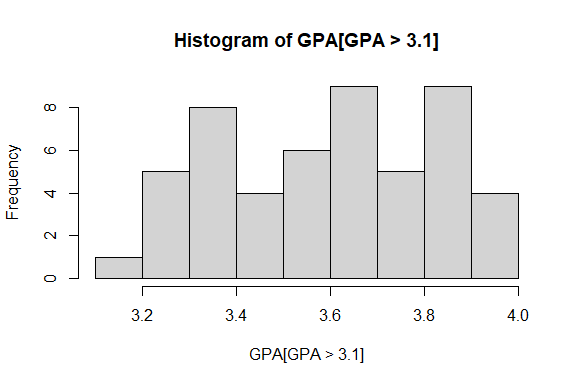
Using two different methods, we have confirmed that GPA is not normally distributed

## Part b.

If needed, transform the GPA data so that the transformed data likely comes from a normal distribution. Provide evidence to support that your transformation works. Use two different criteria.

### Problem 4b R Code

#Insert your code here  
  
# transformed GPA histogram  
hist(GPA[GPA > 3.1])



# Shapiro-Wilk test:  
shapiro.test(GPA[GPA > 3.1])

##   
## Shapiro-Wilk normality test  
##   
## data: GPA[GPA > 3.1]  
## W = 0.96049, p-value = 0.08756

### Problem 4b Analysis

#Insert analysis here

In order to transform the data to likely be more normally distributed, I eliminated the outliers and only included GPAs above a 3.1.

The first way we can assess normality of the transformed GPAs is by creating a histogram. Because the histogram above is approximately symmetric and bell-shaped, it is likely that the transformed GPA is normally distributed.

Another way we can assess normality is by running the Shapiro-Wilk test on the dataset. For this test:

H0: Population is normally distributed

Ha: Population is not normally distributed.

and we will use a significance level of = 5%.

According to the Shapiro-Wilk test run above, it is clear that the data set is normally distributed because the p-value is 0.088 which is greater than which is 0.05. This means that we do not reject the assumption that H0 is true.

Using two different methods, we have confirmed that transformed GPA is normally distributed.

## Part c.

Test if the average of all transformed GPAs is more than 3.4. Use a significance level of 5%.

### Problem 4c R Code

#Insert code here  
  
t.test(GPA[GPA > 3.1], mu = 3.4, alternative = "greater")

##   
## One Sample t-test  
##   
## data: GPA[GPA > 3.1]  
## t = 6.4206, df = 50, p-value = 2.419e-08  
## alternative hypothesis: true mean is greater than 3.4  
## 95 percent confidence interval:  
## 3.550404 Inf  
## sample estimates:  
## mean of x   
## 3.603529

### Problem 4c Analysis

State your hypotheses (defining your parameter(s)); the test used; the test statistic, degree(s) of freedom, and p-value; write your conclusion. Refer to the relevant R output above.

#Insert analysis here

Using a one-sample upper-tail t-test, we can assess if the average of all transformed GPAs is more than 3.4. For this test:

= 0.05

µ = the average GPA of the transformed GPA dataset

µ0 = 3.4

H0: µ = µ0

Ha: µ > µ0

Running t.test (shown above) results in:

t\* = 6.42 df = 50 p-value = 2.42e-08

Considering the results above, since the p-value is 2.42e-08 which is significantly smaller , we can make the conclusion that there is enough evidence to support Ha to be true.

Checking the assumptions:

1. Normal population with unknown µ0 - we proved above that the transformed GPAs were normally distributed and we were never given a value for µ0 therefore it is unknown.
2. Assumed that population standard deviation is not known - this value is never given, therefore it is unknown to us.
3. Samples from each population are randomly selected - this is true based on information given in the original problem.
4. n < 10% of the population size - it is same to assume that 51 students make up for less than 10% of students applying to medical school on any given year.

At the 5% significance level, there is evidence to support that the mean of the transformed GPAs is greater than the value µ0, 3.4 [xbar = 3.6, n = 51, S = 0.226, t\* = 6.42, df = 50, p-value = 2.42e-08]. The assumptions of the test were all met as well.

End of Problem Set