STA104-Project 2

June 1 2021

Class: STA104 Spring 2021

Professor: Maxime Pouokam

Members: Josemanuel Vega (#ID:916627124)
Kate Johnson (#ID:918200073)
Shih-Chi Chen (#ID:917995392)

Topic I: Question 2

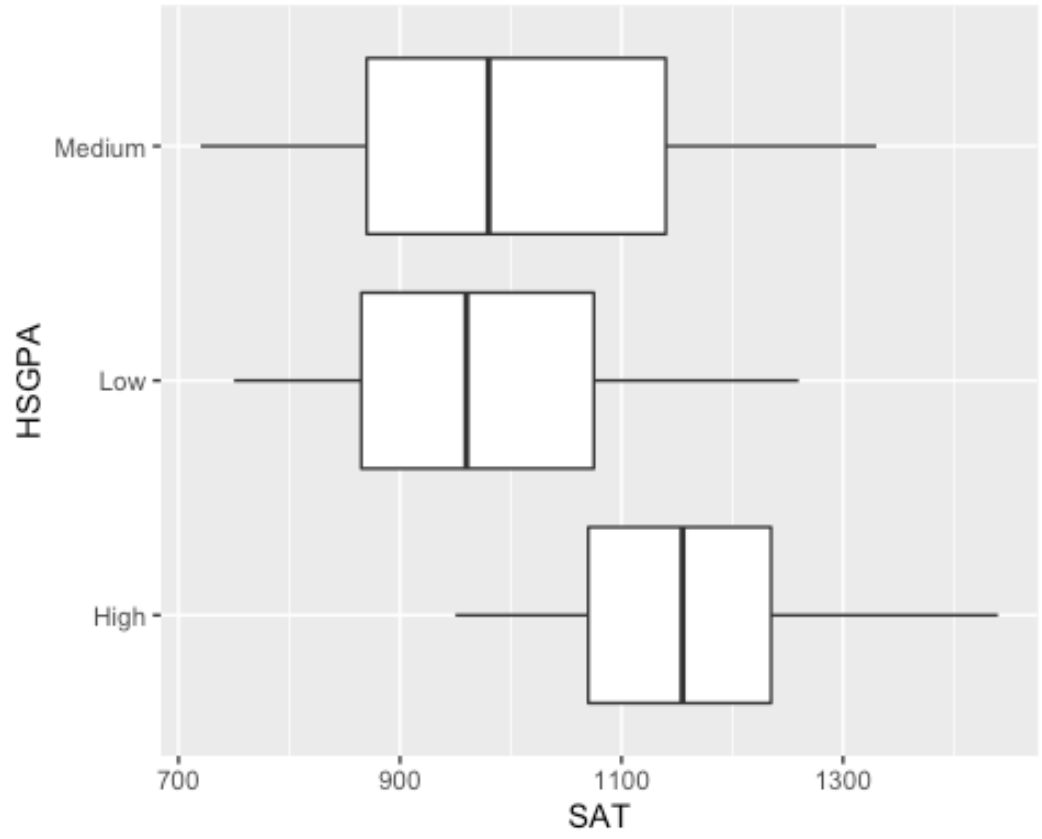
I. Introduction

This project aims at analyzing whether there are significant differences of SAT scores among GPA groups (Low, Medium, High). To conduct a reliable analysis, some boxplots of data are observed and checked with the assumptions at first. Then a decision for selecting appropriate tests is made. Afterwards, the pairwise differences are analyzed once significant differences of SAT scores among GPA groups exist. Finally, valuable conclusions are presented.

II. Summary of data and Diagnostics

- Boxplot

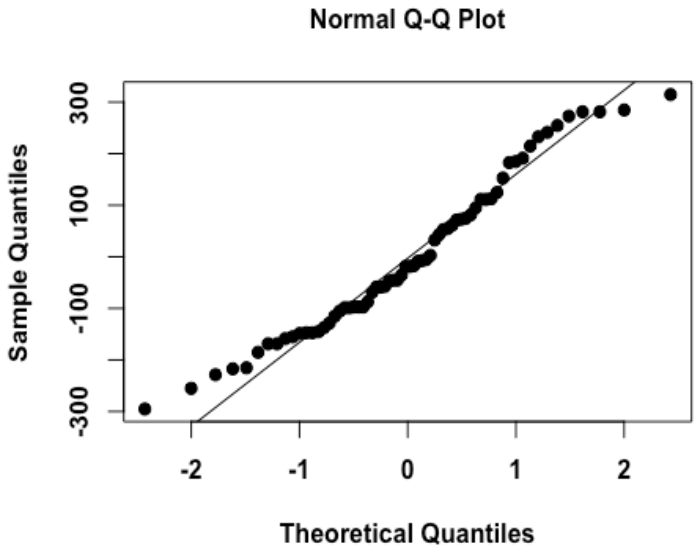
As shown in the boxplot, the maximum SAT scores is in High GPA group and the minimum SAT scores is in Medium GPA group. In addition, the medians of Medium and Low groups are not at the center in the box but close to the left, which means that the distributions of these two groups are right skewed.



- Diagnostics

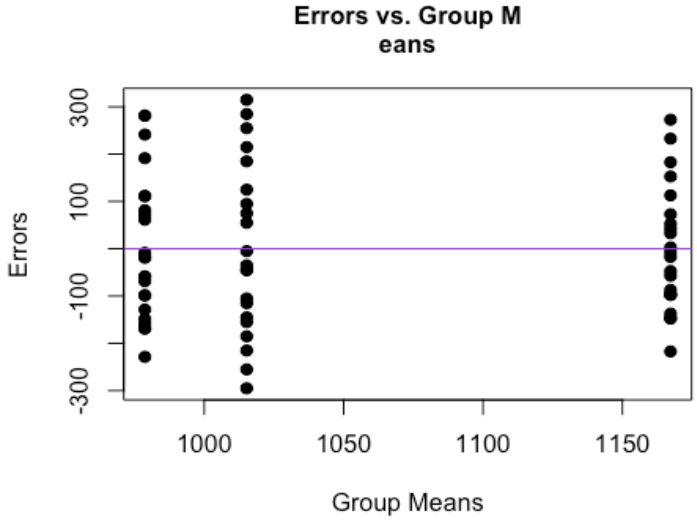
Check Normality: Q-Q Plot

As shown in the Q-Q Plot, some dots are not close to the line or on the line. Therefore, the assumption of normality is violated.



Check Equal Variance: Error vs. Group Means Plot

This plot shows that the vertical spreads of variance are approximately equal. Therefore, the equal variance assumption is hold.



Since the assumption of normality is violated and the sample size is large than 30, the chi-squared large sample approximation to Krustal-Wallis is appropriate.

III. Analysis

- Large sample approximation to Krustal-Wallis: (Use $\alpha = 0.05$)

State H_0 and H_a:

$H_0: F_H(X) = F_M(X) = F_L(X)$

H_a: At least one of the cumulative distributions is different among groups.

Test-statistic: KW=14.255

P-value: 0.0008026

Since p-value $< \alpha$, we would reject H_0 and conclude that at least one distribution is different from each other.

Since H_0 is rejected, at least one of the average SAT scores is different among GPA groups. Therefore, the issue that which groups are different is worth analyzing.

- Find the pairwise differences (Use $\alpha = 0.05$)

Since there are 3 groups (Low, Medium, High), totally 3 possible pairwise differences are considered. Besides, the asymptotic cut-off for Bonferroni, Tukey HSD is used to compare with $|R_i - R_j|$.

From the table, it shows that groups with a significant difference are High vs Low and High vs Medium because of $|R_i - R_j| \geq$ cutoff (BON and HSD).

	H vs L	H vs M	L vs M
Abs. Differences	20.66700	16.15043	4.516563
HSD	13.70043	14.01575	13.866152
BON	13.73675	14.05291	13.902908

IV. Interpretation

From the Krustal-Wallis test-statistic, it shows that H_0 would be rejected because of large test-statistic obtained. Furthermore, the conclusion is same as that of comparing p-value to α, which is due to p-value $< \alpha$ so that H_0 is rejected.

From the pairwise difference comparison, High vs Low and High vs Medium have significant difference because of $|R_i - R_j| \geq$ cutoff (BON and HSD). On the other hand, Low vs Medium do not have significant differences because of $|R_i - R_j| <$ cutoff.

V. Conclusion

First of all, the chi-squared large sample approximation to Krustal-Wallis is appropriate to use because there are more than 2 groups in this dataset, one of assumptions is violated, and sample size is large than 30.

Secondly, the Test-statistic and p-value of Krustal-Wallis suggest that there are significant differences of SAT scores among GPA groups. Therefore, finding which groups have significant differences is needed.

Finally, the asymptotic cut-off for Bonferroni and Tukey HSD is applied to compare the absolute values of different average of ranks between two groups. It shows that High vs Low and High vs Medium have significant differences on SAT scores.

Topic II: Question 1

I. Introduction

This project aims at analyzing non-parametrics test for independence upon is brought to us from a specific health club of a specific town. The details of the data brought to us is that it measures and accounts for the gender that was playing at the 3 of the sports played, and determine if the two categorical variables are independent of the other. Variable 1 will correspond to gender (either "M" of "F"), and variable 2 with sport (Handball-I, Racketball - II, Tennis - III). Afterwards, the pairwise differences are analyzed if Sport and Gender are not independent. Finally, meaningful conclusions are presented.

II. Summary of data

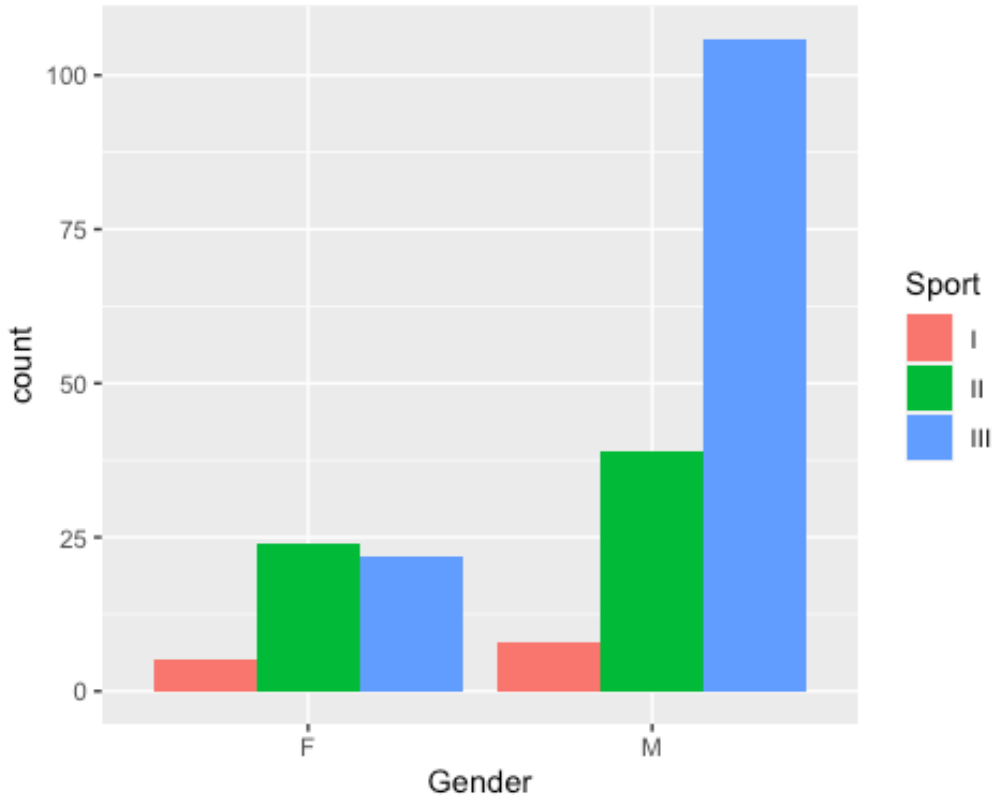
- Contingency table

From the contingency table, it shows that there is significant difference on Sport III(Tennis) between females and males. The amount of males is almost 5 times of that of females. Therefore, the dependency between Sport and Gender is expected.

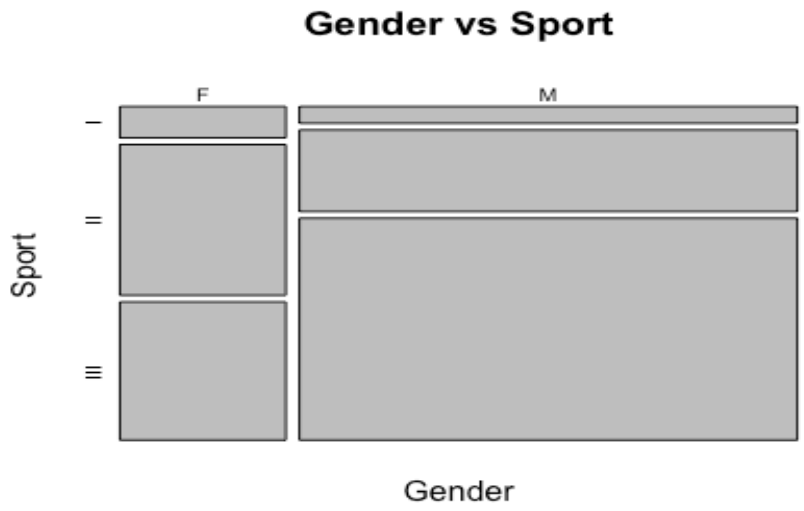
	Sport		
Gender	I (Handball)	II (Racquetball)	III (Tennis)
F	5	24	22
M	8	39	106

- Bar plot

From the bar plot, it shows that numbers of people who play Handball don't have large differences between two genders. However, numbers of people who play Tennis have large differences between two genders.



- Mosaic Plot



From the Mosaic plot, it shows that most males play tennis (III), some play racketball (II), and few play handball(I). On the other hand, numbers of female play tennis (III) and racketball (II) are approximately same, and few play handball(I).

III. Analysis

- Comparing conditional probabilities test (Use $= 0.05$)

State the null and alternative hypothesis in terms of conditional probabilities:

H_0: P (Sport | Female) = P (Sport | Male)

H_a: P (Sport | Female) ≠ P (Sport | Male) for at least some level of Sport

Test statistic $\chi^2_{S,OBS} = 11.18498$

Use R to find the permutation based p-value, based on R = 3000 random shufflings:

Permutation based p-value=0.004

We would reject H_0 without having to choose a significance level because this p value is much smaller than any α that we care about, and conclude that there is difference in Sport subject by gender (They are not independent).

- Pairwise comparisons

Since there is difference in sport subject by gender, pairwise comparisons are needed.

Z values for comparing the row values of each column (the sport of the subject given the gender (Female, Male)) are shown below:

Sport	I (Handball)	II (Racquetball)	III (Tennis)
F vs M	1.158415	2.887278	−3.344280

From these Z values, the largest difference exists in Sport III (Tennis) for Female vs. Male because the largest absolute value of Z values obtained.

Tukey inspired cutoff value (using $\alpha = 0.05$):

Tukey inspired cutoff=2.340996

There are two significant differences, which are Sport II (Racquetball) for Female vs. Male and Sport III (Tennis) for Female vs. Male because absolute value of z-scores are above the Tukey cutoff.

The difference P (Sport | F) - P (Sport | M) is shown below:

Sport	I (Handball)	II (Racquetball)	III (Tennis)
P (Sport | F) - P (Sport | M)	0.04575163	0.21568627	-0.26143791

For Sport II (Racquetball), it suggests that the proportions of males and females who play racquetball are different, with female tending to play racquetball more often.

For Sport III (Tennis), it suggests that the proportions of males and females who play tennis are different, with male tending to play tennis more often.

IV. Interpretation

From the comparing conditional probabilities test, H_0 is rejected and it is concluded that there is difference in Sport subject by gender.

From the pairwise comparisons, Z values show that most differences are in Sport III (Tennis) for Female vs. Male. Afterwards, the comparisons between the Tukey inspired cutoff and Z values shows that Sport II (Racquetball) for Female vs. Male and Sport III (Tennis) for Female vs. Male have significant differences.

From the pairwise comparisons, it shows that Sport I dose not have significant differences. Accordingly, only the differences of conditional probabilities for Sport II and Sport III are analyzed.

For Sport II (Racquetball), since the positive probability is obtained, it suggests that females are more often to play racquetball than males.

For Sport III (Tennis), since the negative probability is obtained, it suggests that males are more often to play tennis than females.

V. Conclusion

First of all, the comparing conditional probabilities test is appropriate to use because the conditional probabilities of sport in given different gender are focused on.

Secondly, the Test-statistic and p-value of comparing conditional probabilities test suggest that Sport and Gender are not independent. Therefore, finding which groups have significant differences is needed.

Thirdly, the comparison between Z values and Tukey inspired cutoff shows that two significant differences exist, which are Sport II (Racquetball) for Female vs. Male and Sport III (Tennis) for Female vs. Male.

Finally, looking at actual difference of these two group and conclude that the proportion of males tending to play tennis more often than that of females. On the other hands, the proportion of females tending to play racquetball more often than that of males.

#Appendix Code

```r
#topic1
#Boxplot
Academic <- read.csv("~/Downloads/Academic.csv")
library(ggplot2)
ggplot(data = Academic, mapping = aes(x = HSGPA, y = SAT)) +
    geom_boxplot() +
    coord_flip()
sat_lm <- lm(SAT ~ HSGPA, data = Academic)
#QQ plot
qqnorm(sat_lm$residuals,pch = 19,font = 2,font.lab =2,cex =1,cex.lab=1, cex.a
xis=1, cex.main=1, cex.sub=1)
qqline(sat_lm$residuals)
#Errors vs. Group Means plot
plot(sat_lm$fitted.values, sat_lm$residuals, main = "Errors vs. Group M
eans",xlab = "Group Means",ylab = "Errors",pch = 19,font = 1,font.lab =1,cex=
1,cex.lab=1, cex.axis=1, cex.main=1, cex.sub=1)
abline(h = 0,col = "purple")
#test statistic and p values
Academic$Rank = rank(Academic$SAT, ties = "average")
N = nrow(Academic)
K = length(unique(Academic$HSGPA))
Ri = aggregate(Rank ~ HSGPA, data = Academic, mean)$Rank
ni = aggregate(SAT ~ HSGPA, data = Academic, length)$SAT
SR.2 = var(Academic$Rank)
KW.OBS = 1/SR.2*sum(ni*(Ri - (N+1)/2)^2)
p.value = pchisq(KW.OBS, df = K-1,lower.tail = FALSE)
#all diff for rank
all.diff = as.numeric(dist(Ri,method = "manhattan"))
names(all.diff) = c("H vs L","H vs M","L vs M")
all.diff
#cutoff
K = length(unique(Academic$HSGPA))
alpha = 0.05
g = K*(K-1)/2
BON12 = qnorm(1-alpha/(2*g))*sqrt(SR.2*(1/ni[1] + 1/ni[2]))
BON13 = qnorm(1-alpha/(2*g))*sqrt(SR.2*(1/ni[1] + 1/ni[3]))
BON23 = qnorm(1-alpha/(2*g))*sqrt(SR.2*(1/ni[2] + 1/ni[3]))
all.BON = c(BON12, BON13, BON23)
HSD12 = qtukey(1-alpha,K,N-K)*sqrt((SR.2/2)*(1/ni[1] + 1/ni[2]))
HSD13 = qtukey(1-alpha,K,N-K)*sqrt((SR.2/2)*(1/ni[1] + 1/ni[3]))
HSD23 = qtukey(1-alpha,K,N-K)*sqrt((SR.2/2)*(1/ni[2] + 1/ni[3]))
all.HSD = c(HSD12,HSD13,HSD23)
all.crits = rbind(all.diff, all.BON,all.HSD)
all.crits

##topic2
#contingency table
library(readxl)
```

```r
play <- read_excel("~/Downloads/play.xlt")
two.way = table(play)
two.way
#Bar graph
ggplot(play, aes(x = Gender, fill = Sport)) + geom_bar(position = "dodge")

#Mosaic Plot
mosaicplot(two.way, main = "Gender vs Sport")

#eij and chi-square test statistic
the.test = chisq.test(two.way,correct = FALSE)
eij = the.test$expected
chi.sq.obs = as.numeric(the.test$statistic)
chi.sq.obs
#Permutation P-value
R = 3000
r.perms = sapply(1:R,function(i){
perm.data = play
perm.data$Sport = sample(perm.data$Sport,nrow(perm.data),replace = FALSE)
chi.sq.i = chisq.test(table(perm.data),correct = FALSE)$stat
return(chi.sq.i)
})
perm.pval = mean(r.perms >= chi.sq.obs)
#z-test-statistics
n = sum(two.way)
ni. = rowSums(two.way)
n.j = colSums(two.way)
all.pjG1 = two.way[1,]/ni.[1] #all conditional probabilites for row 1
all.pjG2= two.way[2,]/ni.[2] #all conditional probabilites for row 2
all.pbar = n.j/n #all probabilities regardless of group
all.Zij = c(all.pjG1 - all.pjG2)/sqrt(all.pbar*(1-all.pbar)*(1/ni.[1] + 1/ni.
[2]))

#cut off
r.perms.cutoff = sapply(1:R,function(i){
perm.data = play
perm.data$Gender = sample(perm.data$Gender,nrow(perm.data),replace = FALSE)
row.sum = rowSums(table(perm.data))
col.sum = colSums(table(perm.data))
all.pji = table(perm.data)[1,]/row.sum[1]
all.pji.= table(perm.data)[2,]/row.sum[2]
all.pbar = col.sum/sum(row.sum)
all.Zij = c(all.pji - all.pji.)/sqrt(all.pbar*(1-all.pbar)*(1/row.sum[1] + 1/
row.sum[2]))
Q.r = max(abs(all.Zij))
return(Q.r)
})
alpha = 0.05
cutoff.q = as.numeric(quantile(r.perms.cutoff,(1-alpha)))
#proportion
```

```
row.sum = rowSums(table(play))
all.pji = table(play)[1,]/row.sum[1]
all.pji.= table(play)[2,]/row.sum[2]
all_proportion=all.pji-all.pji.
all_proportion
```