Project I

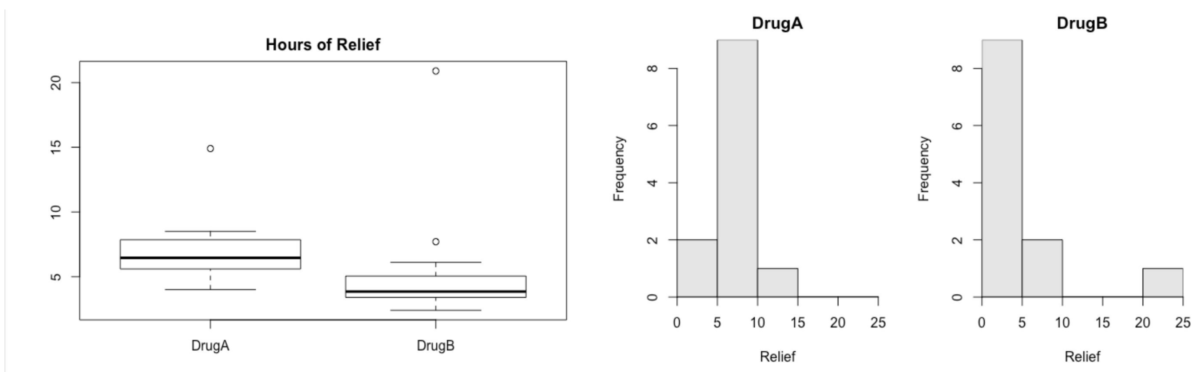
STA 104 Erin K. Melcon
Ruriko Imai

Topic II Question 1

Introduction:

The dataset Drug.csv contains a data of a study comparing the hours of pain relief for two common over-the-counter pain medications. There are two variables, Relief and Groups. The Relief variable describes the number of hours of relief provided by the Groups, either DrugA or DrugB. The two groups will be compared to observe if either DrugA or DrugB provides a longer time of relief from the pain. This might be of interest to consumers who buy those pain relief drugs in order to determine which one they would rather buy based on the effectiveness of the drug's pain relief.

Summary:



Looking at the box plot, the points outside the boxes represent that there are outliers present in the data. In the histogram, DrugB seems to have an outlier and the distributions for both drugs are not normally distributed.

The summary statistics of this data consists a total of 24 observations with 12 samples for each drug type. The overall mean and standard deviation are, respectively, 6.2417 and 4.0894. The average of DrugA is 6.9667 with a standard deviation of 2.8804. The average of DrugB is 5.5167 with a standard deviation of 5.0521.

The mean and standard deviations of each group also indicates that they might not be very similar, however it is difficult to infer that there is a significant difference based purely on the summary statistics. Therefore, we will analyze the data further using an appropriate statistical technique for this dataset.

Since the total sample size is less than 30 and there seems to be outliers present, we will use the Mann-Whitney, a non-parametric technique to analyze this dataset.

Analysis:

Consider Group 1 to be DrugA and Group 2 to be DrugB.
The hypothesis test is:
$H_0 : F_1(x) = F_2(x)$
$H_A : F_1(x) \le F_2(x)$

Therefore, the claim is that DrugA provides longer hours of pain relief than DrugB (i.e. $m_1 > m_2$).
In order to test the hypothesis, we need to find the Mann-Whitney test statistics and its corresponding p-value. The test statistic is 2.5481 and the p-value is 0.004664.

To further confirm the hypothesis test, here is the result for the Mann-Whitney confidence interval. The confidence interval is (0.4999269, 3.6000002).

Interpretation:

The interpretation of p-value is that if in reality the distributions of the relief period for the two drugs are the same, the probability of observing our data or more extreme is 0.004664.
Since the p-value is less than $\alpha = 0.05$ (or any given alpha level), we reject the null hypothesis.

The confidence interval does not include 0, therefore, it indicates that there is a significant difference between the two distributions of the drug's effectiveness. Since the upper and lower bounds of the interval are both positive, DrugA is bigger than DrugB (i.e. $m_1 - m_2 > 0$).

Conclusion:

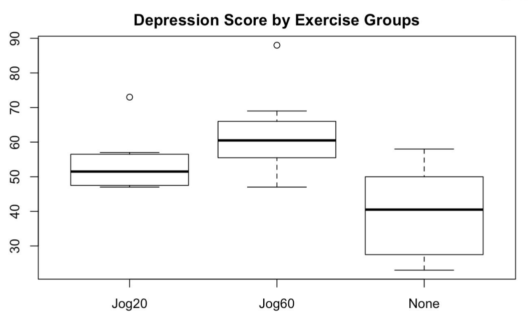
In conclusion, we reject the null hypothesis and support the claim that the distributions of the relief period offered by DrugA is significantly higher than DrugB. Furthermore, the confidence interval also confirms that DrugA offers longer period of relief (in hours) than DrugB. If sick/hurt patients are looking for over-the-counter medicine that relieves them from pain for a longer period, they should consider purchasing DrugA.

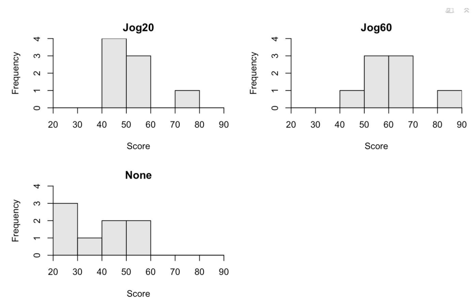
Topic III Question 1

Introduction:

The dataset Depress.csv consists of two columns, Score and Group. Score is the “Depression score” of the individuals, where 0 means they are completely and utterly depressed, and 100 means they are ecstatically happy. Group describes what exercise group individuals are in. The exercise groups are as follows, None (no regular exercise), Jog20 (20 minutes of jogging per day), and Jog60 (60 minutes of jogging per day). The three exercise groups will be compared to their depression scores in order to observe if exercise has any effect on depression. This is important to people who are depressed and are trying to find a way to lead a happier life.

Summary:





The box-plot shows outliers present for exercise groups Jog20 and Jog60. The group’s means also seems to be different. The histogram also presents outliers and the distributions of each exercise groups seems to be skewed.

Since there are outliers present and more than two groups to be compared, the best method to analyze the relationship between the two factors are using a non-parametric Kruskal-Wallis test. We cannot use the chi-squared Kruskal-Wallis test since the total sample size is less than 30. Kruskal Wallis test in general has higher power when data has outliers and observed distribution has “heavy tails” (i.e., highly skewed) for comparisons for two or more groups.

The summary statistics are as follows:

	Jog20	Jog60	None
Group Mean	53.875000	62.37500	39.6250
Group SD	8.724964	12.35126	12.8501
Rank Mean	12.687500	18.00000	6.8125
Sample Size	8.000000	8.00000	8.0000

SInce we are using the Kruskal-Wallis test, we will rank our data. In the summary statistics, the Rank means between the exercise groups seems to be different. Therefore, we will conduct a statistical analysis to determine if the distributions between the groups differ significantly, and also if the groups do differ significantly, which group is significant.

The standard deviation of the overall ranks (regardless of groups) is 49.8696.

Analysis:

Consider Group 1 to be Jog20 , Group 2 to be Jog60, and Group 3 to be None.
The hypothesis test is:
$H_0 : F_1(x) = F_2(x) = F_3(x)$
$H_A : \text{At least one of the } F_i(x) \text{ is different.}$

The observed Kruskal-Wallis test statistic is 27.3727.
Using the Kruskal-Wallis test with 2000 permutations gives a p-value of 0.0015.

Now, we would like to know which group means differ significantly since we reject the null hypothesis.

The permutation cutoff values for Bonferroni gives the following results:
The adjusted p-value for Bonferroni using α = 0.05 is 0.01667.
The p-value and the z test-statistics for Jog20 vs. Jog60, respectively, is 0.09713 and -1.6865.
The p-value and the z test statistics for Jog20 vs. None, respectively, is 0.06884 and 1.8406
The p-value and the z test statistics for Jog60 vs. None, respectively, is 0.001709 and 2.9428.

Interpretation:

The interpretation of p-value is that if in reality the distributions of the three exercise groups are the same, the probability of observing our data or more extreme is 0.0015.

Since the p-value is less than $\alpha = 0.05$ (or any given value of alpha), we reject the null hypothesis.

The interpretation of p-value is that if in fact the distributions of depressions scores are the same for all exercise groups, then the probability of observing our data or more extreme is 0.09713 for Jog20 vs. Jog60, 0.06884 for Jog20 vs. None, and 0.001709 for Jog60 vs. None.

Since the p-value for Jog60 vs. None is less than the adjusted p-value, we reject the null hypothesis indicating that the distributions are same for all groups.

Conclusion:

Since p-value was less than $\alpha = 0.05$ and we rejected the null hypothesis, we conclude that there is a significant differences in the distributions of the three exercise groups and their corresponding depression scores.

After rejecting the null hypothesis, we tested to see which group means differed significantly. By comparing the rank differences and the critical cutoff values for Bonferroni, we were able to deduce that there is a significant difference between Jog60 and None. Therefore, we can say that not exercising at all vs. jogging for 60 minutes a day may have a significant affect on one’s depression score.

The benefit to using the permutation based cutoffs is that they do not rely on an asymptotic distribution. Thus, if sample size is small or the distribution is very highly skewed, you will get better performance from the permutation based cutoffs.

The Bonferroni cutoff is used in this situation since the g (number of groups to be compared with one another) is between 3 and 10. If g was larger than 10, using Tukey cutoffs will be a better choice and if g is smaller than 3 Fisher's LSD should be used.

Appendix:

```
library(coin)
drug = read.csv("~/downloads/Drug.csv")
depress = read.csv("~/downloads/Depress.csv")
depress$Group
drug$Groups
head(drug)
head(depress)
boxplot(Relief ~ Groups, data = drug, main="Hours of Relief")
library(FSA)
hist(Relief ~ Groups, data = drug)
dim(drug)
table(drug$Groups)
meanOverall = mean(drug$Relief) #overall mean
meanA = mean(drug$Relief[1:12]) #mean of GroupA
meanB = mean(drug$Relief[13:24]) #mean of GroupB
sdOverall = sd(drug$Relief)
sdA = sd(drug$Relief[1:12])
sdB = sd(drug$Relief[13:24])
wilcox_test(Relief ~ Groups, drug, distribution = "exact",alternative = "greater")
the.test = wilcox_test(Relief ~ Groups, data = drug, conf.level = 0.95, conf.int = TRUE)
the.CI = as.numeric(confint(the.test)$conf.int)
the.CI
boxplot(Score ~ Group, data = depress, main="Depression Score by Exercise Groups")
hist(Score ~ Group, data = depress)
table(depress$Group)
dim(depress)
table(depress$Score)
depress$Rank = rank(depress$Score, ties = "average")
Group.order = aggregate(Score ~ Group, data = depress, mean)$Group
Xi = aggregate(Score ~ Group, data = depress, mean)$Score
si = aggregate(Score ~ Group, data = depress, sd)$Score
Ri = aggregate(Rank ~ Group, data = depress, mean)$Rank
ni = aggregate(Score ~ Group, data = depress, length)$Score
results = rbind(Xi,si,Ri,ni)
rownames(results) = c("Group Mean","Group SD","Rank Mean","Sample Size")
colnames(results) = as.character(Group.order)
results
SR.2 = var(depress$Rank)
N = 24
KW.OBS = 1/SR.2*sum(ni*(Ri - (N+1)/2)^2) #Note, this assumes you calculate ni and Ri above
R = 2000
many.perms.KW = sapply(1:R,function(i){
  permuted.data = depress #So we don't overwrite the original data
  permuted.data$Group = sample(permuted.data$Group, nrow(permuted.data), replace = FALSE)
#Permuting the groups
  SR.2 = var(permuted.data$Rank)
```

```
  ni = aggregate(Rank ~ Group, data = permuted.data,length)$Rank
  Ri = aggregate(Rank ~ Group, data = permuted.data,mean)$Rank
  KW.i= 1/SR.2*sum(ni*(Ri - (N+1)/2)^2)
  return(KW.i)
})
p.value = mean(many.perms.KW > KW.OBS)
split.groups = split(depress,depress$Group) #Makes a list of K groups (in alphabetical order)
IvsII = rbind(split.groups[[1]],split.groups[[2]]) #Binds 1 and 2 back togeather
IvsIII = rbind(split.groups[[1]],split.groups[[3]]) #Binds 1 and 3 back togeather
IIvsIII = rbind(split.groups[[2]],split.groups[[3]]) #Binds 2 and 3 back togeather
library(coin)
wilcox_test(Score ~ Group, data = IvsII,distribution = "exact")
wilcox_test(Score ~ Group, data = IvsIII,distribution = "exact")
wilcox_test(Score ~ Group, data = IIvsIII,distribution = "exact")
0.05/3
```