STA 104 Project

Miguel Morales

Project #2

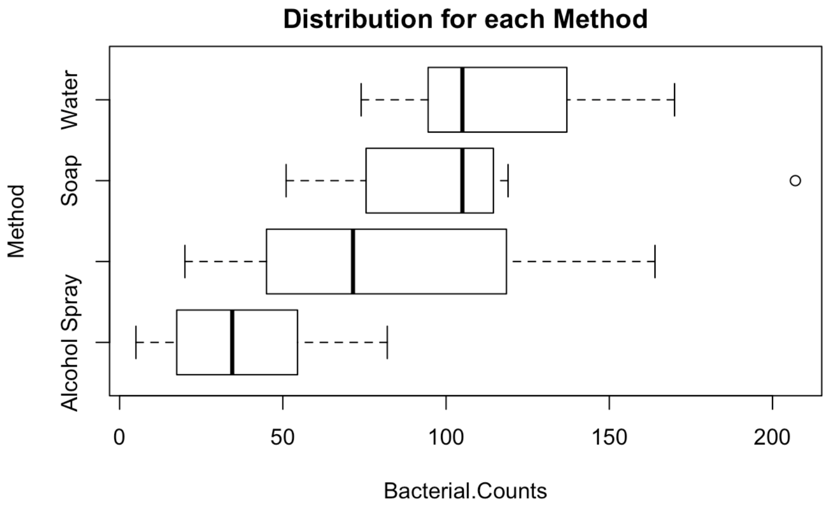
Topic 1

I. Introduction

With the case of the first topic, we will be exploring four different methods of washing hands and comparing how well they eliminate bacteria. This test was conducted by the subject testing four different methods of washing her hands: washing with water only, washing with regular soap, washing with antibacterial soap, and spraying hands with antibacterial spray. Each day one of these was chosen at random in the morning. After she placed her right hand on a sterile media plate designed to promote bacterial growth. Each plate was incubated for two days at 36 degrees celsius. This process took place over a month and the data was collected and given to us.

II. Summary

First we started by creating a boxplot to see how the different methods of washing hands compared in our dataset. We found that Alcohol Spray, Antibacterial Soap, and Water were skewed slightly to the right. From the boxplot we can also see that the Soap method contains an outlier.



```
## Bacterial.Counts    Method
## Min.   :  5.0        Length:27
## 1st Qu.: 51.0        Class :character
## Median : 82.0        Mode  :character
## Mean   : 84.7
##  3rd Qu.:109.0
##  Max.   :207.0
```

With the summary statistics above we can see a wide range of Bacterial Counts ranging from 5 up to 207. The 207 entry is obviously our outlier in the Soap method, but these summary statistics give us a good indication that most of the data is within the 1st and 3rd quartile.

Our goal is to determine if there is a difference in the amount of bacteria that is left between the different methods of hand washing. Therefore, our null hypothesis will be that bacteria amount on average differs significantly between the different methods of handwashing.

III. Analysis

We will begin conducting tests by using a large approximation Chi-Squared test using Kruskal-Wallis.

Our Test hypotheses are as follows:

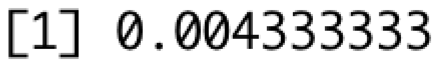
Ho: Fa(X) = Fb(X) = Fc(X) = Fd(X)
Ha: Fi(X) <= Fj(X) or Fi(X) >= Fj(X) for some i!= j

The alternative hypothesis means that at least one of the distribution of bacterial amounts for one of the methods is less or greater than a different method.

```
##  Asymptotic Kruskal-Wallis Test
##
## data:  Bacterial.Counts by
##   as.factor(Method) (Alcohol Spray, Antibacterial Soap, Soap, Water)
## chi-squared = 13.832, df = 3, p-value = 0.003144
```

Our Asymptotic KW tests returns us a p-value of 0.003144.

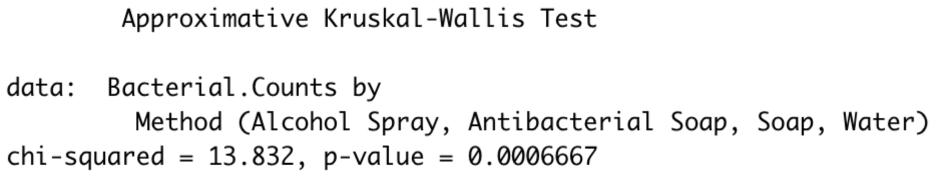
We also calculate the p-value using a permutation F-test , using R=3000 permutations:



```
[1] 0.004333333
```

This gave us a p-value of 0.004333333.

Then, we calculated the p-value using Kruskal-Wallis test, but this time using R=3000 permutations:



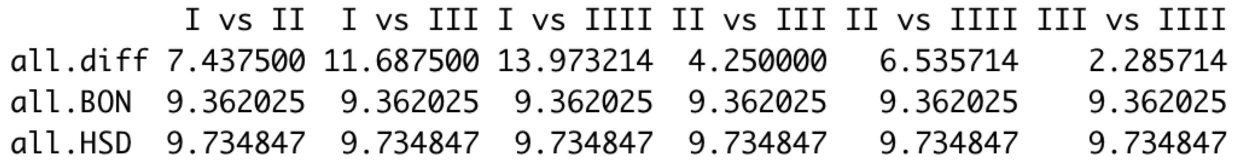
```
                Approximative Kruskal-Wallis Test

data:  Bacterial.Counts by
         Method (Alcohol Spray, Antibacterial Soap, Soap, Water)
chi-squared = 13.832, p-value = 0.0006667
```

Since, we already know that our data shows evidence of non-normality and clearly has an outlier, we should not use a parametric test. We should also note that the Kruskal-Wallis test will have more power than a regular permutation test due to the outlier shown in the boxplot. Kruskal-Wallis uses Ranks meaning that outliers will not affect it. Therefore, we will be using the Kruskal-Wallis test.

Lastly, since the p-value is 0.003144, this is lower than the alpha value we set earlier at 0.05. As a result, we can accept the null hypothesis that the means are different in each method.

Legend for pairwise comparisons: I= Water, II= Soap, III= Antibacterial Soap, IIII= Alcohol

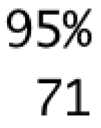
Now, to see which pairwise comparisons of the different hand washing methods significantly differ, we compare all combinations of two groups and their average ranks to calculated cutoffs. We use Tukey's HSD with an alpha of 0.05, but the non-paramatric version that uses ranks:



```
            I vs II  I vs III I vs IIII II vs III II vs IIII III vs IIII
 all.diff 7.437500 11.687500 13.973214  4.250000   6.535714    2.285714
 all.BON  9.362025  9.362025  9.362025  9.362025   9.362025    9.362025
 all.HSD  9.734847  9.734847  9.734847  9.734847   9.734847    9.734847
```

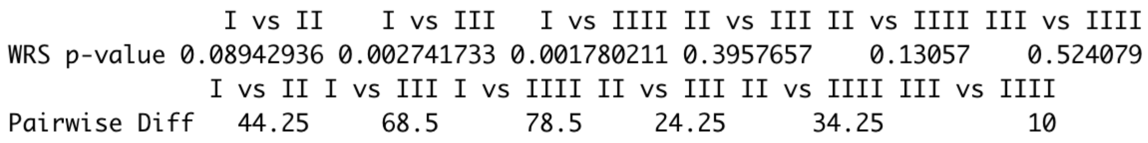
Looking at our pariwaise comparisons and cutoff, we find that groups I vs III and I vs IIII are siginfiicantly different.

We also find the permutation based cutoff for Tukeys HSD at an alpha of 0.05.



```
95%
 71
```

We conduct two sample tests using the Wilcoxon-Mann-Whitney test to find more p-values



```
               I vs II    I vs III   I vs IIII II vs III II vs IIII III vs IIII
WRS p-value 0.08942936 0.002741733 0.001780211 0.3957657    0.13057    0.524079
              I vs II I vs III I vs IIII II vs III II vs IIII III vs IIII
Pairwise Diff   44.25     68.5      78.5     24.25      34.25          10
```

From this we can see that using Bonferroni's significance level of alpha/6, we see that groups I vs III and I vs IIII are significantly different, the same findings to our comparisons before. Using the permutation based Tukeys cutoff of 71, we see that group I vs IIII are significantly different.

IV. Interpretation

With the p-value at 0.003144 and the alpha value at 0.05, we reject the null hypothesis that the distribution of the different methods is the same and conclude that at least one of the distributions of the different methods are not equal. Additionally we choose to use the same Kruskal Wallis test but this time permutation the data 3,000 times. We ran this test because the dataset as a whole was rather small so we ran the permutation test to see how the results from earlier would change. This permutation version of the KW test gave us a p-value of 0.0006667, meaning that it reinforces our rejection of the Null Hypothesis.

To find out which groups significantly differ, we calculated the average ranks of the different methods and compared them to Tukey's HSD cutoff using an alpha of 0.05. When comparing the differences to the cutoff, we find that groups I vs III and I vs IIII have siginficantly different ranks. We again did this using the Wilcoxon-Mann-Whitney test and also concluded that the same groups were significantly different. When using Tukey's permutation cutoff, we found only groups I vs IIII differ significantly.

V. Conclusion

From this dataset we aimed to see if there was any difference in the 4 washing methods that are given to us in this dataset. The Kruskal-Wallis test and its permutation version, using R=3000, were used to find the p-value and we concluded that at least one of the methods distributions was different. To find out which methods were different, we compared the methods difference to Tukey's HSD at an alpha of 0.05. We found out that the average rank difference between Water and Antibacterial Soap as well as Water and Alcohol have significantly different ranks. However, these differences needed anova assumptions of independent random errors with the same variance for all groups and normality. Something that we can disprove by seeing the outlier and skewed data in our boxplot. Also, using Ranks does put an upper limit to how violated the assumptions can be. We also conducted two sample tests and used Tukey's permutation cutoff. We compared the two sample tests to Bonferroni's significance level and also found the groups of Water and Antibacterial Soap as well as Water and Alcohol are different. Tukey's permutation showed us that only Water and Alcohol were significantly different, but

Water and Antibacterial Soap just shy of meeting the cutoff. Based on our results we would expect that washing hands with alcohol reduces the amount of bacteria left on someone's hands. Antibacterial soap also does a better job at reducing bacteria count but not as well as alcohol, since Tukey’s is less conservative.

Topic II

I. Introduction

For this analysis, we are using a dataset called "Mind.csv". This dataset contains two columns. Column 1 is the two categories of treatment given to a subject with a particular, unknown to us, mental disorder (Medication, Therapy) and Column 2 is the ordinal scale of improvement shown by the patient after a period of 6 months using the treatment (None, Mild, Moderate, Major). We will be studying the relationship between the variables by answering the following key questions:

Question #1: What direction of nonparametric tests will we use to analyze this data?
Question #2: Are the two variables independent or dependent?
Question #3: If the variables are dependent, which categories of the two variables are dependent?

To answer question one, we should simply take a look at our data set and see if the columns are numerical or categorical data and understand any information that was given about this dataset. .

To answer question two, we will use permutation tests for contingency tables to test the relationship between the two variables.

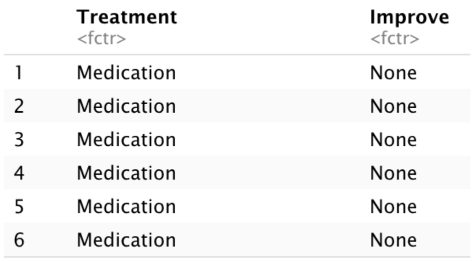
To answer question three, we will look at the differences in patient improvement between Medication and Therapy and analyze these differences using a cutoff.

No knowledge about the distribution of the data was given and assumptions needed for parametric tests were not given, therefore we will be conducting these tests using nonparametric techniques.

II. Summary of Data

First, we will be taking a look at the head of our dataset. This will provide us with information to answer question #1, showing the types of variables that we will be analyzing. We also create a contingency table to view the different categories for each variable.

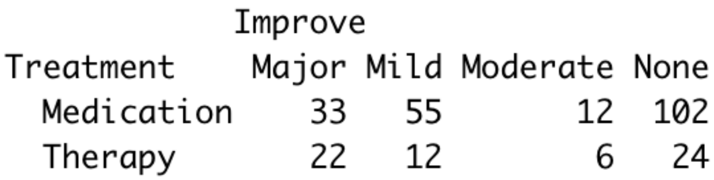
Head of dataset:



	Treatment <fctr>	Improve <fctr>
1	Medication	None
2	Medication	None
3	Medication	None
4	Medication	None
5	Medication	None
6	Medication	None

Here, we see that our data set contains two columns, the first named Treatment and the second named Improve. From the head of the data, we can see that the two variables are categorical, meaning that the nonparametric tests we will be using will have to be compatible with categorical data, not numerical.

Contingency Table of dataset:



```
            Improve
Treatment    Major Mild Moderate None
  Medication    33   55       12  102
  Therapy       22   12        6   24
```

To further elaborate on the variables, the type of Treatment is either Medication or Therapy, and Improve is using an ordinal scale of patient improvement using major, mild, moderate, and none. We can also see the number of instances for each combination of categories from the sample.

III. Analysis

$\phi(x)$

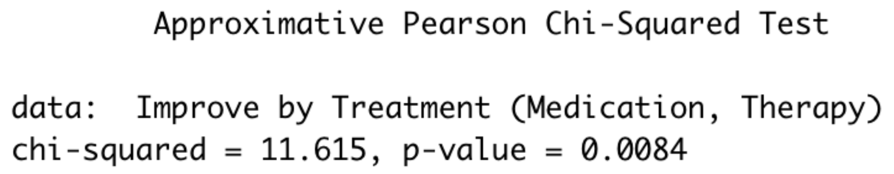
From our summary, we saw that our data uses categorical variables and we are not given any parametric assumptions of our dataset.

Therefore, for a contingency table, we can check if the variables are independent or dependent using a Chi-squared test. However, we will be using the nonparametric version of this test which is the chi-squared permutation test. Since we want to know about the independence of the two variables our alternative hypothesis will be if the variables Treatment and Improve are dependent.

We will use the following hypotheses:

Ho : The Improvement that the patient showed and the Treatment that the patient was given are independent.
Ha : At least one category from patients Improvements and Treatment are dependent.

Chi-squared permutation test results:



```
                Approximative Pearson Chi-Squared Test

data:  Improve by Treatment (Medication, Therapy)
chi-squared = 11.615, p-value = 0.0084
```

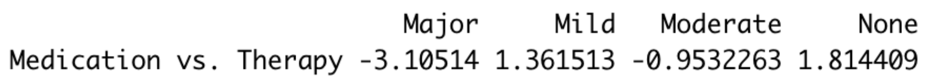
For this chi-squared permutation test, we used 5000 permutations. Our chi-squared permutation test value is 11.615 and our p-value is 0.0084. We will reject Ho and conclude that at least one category from patients Improvements and Treatment is dependent

Now, we want to find out which combination or combinations of the different catagories between the variables are dependent. For this we will calculate the Z values for comparing the row values with each column. Then we will compare them to Tukeys inspired cuttoff. If any cutoff is greater that Tukey's, we will know that it is statistically significant.

The following are the hypotheses for the pairwise comparisons:

Ho : P(improve is Major/Mild/Moderate/None| treatment is Medication) = P(improve is Major/Mild/Moderate/None| treatment is Therapy)

Ha : at least one of the following is true:
P(improve is Major|treatment is Medication) != P(improve is Major|treatment is Therapy)
P(improve is Mild|treatment is Medication) != P(improve is Mild|treatment is Therapy)
P(improve is Moderate|treatment is Medication) != P(improve is Moderate|treatment is Therapy)
P(improve is None|treatment is Medication) != P(improve is None|treatment is Therapy)



```
                          Major     Mild   Moderate     None
Medication vs. Therapy -3.10514 1.361513 -0.9532263 1.814409
```

From our calculations we see that the Z values between the treatments and improve are -3.10514 for major, 1.361513 for mild, -0.9532263 for moderate, and 1.814409 for None. Now we compare the the calculated z values for our conditional probabilities to Tukey's inspired cutoff, to see if any conditional probability is statistically significant.

```
[1] 2.473236
```

Our calculation of Tukey's inspired cutoff results in a cutoff of 2.473236. We will use this to compare with our calculated Z values.

IV. Interpretation

Our chi-squared permutation test value is 11.615 and our p-value is 0.0084. Using an alpha of 0.05, since the p-value from our chi-squared permutation test is less than our alpha, we reject Ho. There is sufficient evidence to conclude that at least one category from the patients Improve and the given Treatment is dependent. Now we interpret the Z values to find the dependence. Based on our calculation for Tukey's inspired cutoff of 2.473236 and comparing them to our conditional cutoffs, we can see that there is a dependence between the treatments and the Major improve category. The cutoff is larger than our Tukey's cutoff, meaning that it is statistically significant. This dependence supports our chi-squared permutation alternative hypotheses. Furthermore since the cutoff is negative for Major patient improvement, it suggests Therapy provides significantly more **major** patient improvement.

V. Conclusion

This analysis was conducted to see if there was a dependence between the different types of treatment and the extent of a patients improvement. Through a chi-squared permutation test we uncovered that at least one of the combinations of the two variables was dependent. By calculating the differences between the conditional probabilities of the data and comparing them to Tukey's cutoff we discovered that the z-score for Medication vs Therapy is negative for Major patient improvement, suggesting that there is a greater probability that Therapy will provide a patient with major improvement, than with medication.

R Appendix

Topic I:

```
library(coin)
bacteria <- read.csv("/Users/PavanJohal/Desktop/bacterial.csv")
data <- bacteria
newdata <- data[order(data$Method),]

library(ggplot2)
boxplot(Bacterial.Counts ~ Method, data = newdata, main = "Distribution for each Method",
horizontal = TRUE)
summary(newdata)
the.test = kruskal_test(Bacterial.Counts ~ as.factor(Method), data = bacteria)
The.test

F.obs = summary(lm(Bacterial.Counts ~ Method, bacterial))$fstatistic["value"]
R =3000
R.perms = sapply(1:R,function(i){
  the.data = bacterial
  the.data$Method = sample(the.data$Method,length(the.data$Method),replace = FALSE)
  FR = summary(lm(Bacterial.Counts ~ Method, the.data))$fstatistic["value"]
  return(FR)
  })
p.value = mean(R.perms >= F.obs)
p.value

approxkw<-kruskal_test(Bacterial.Counts ~ Method, data = bacterial, distribution =
approximate(nresample = 3000))
Approxkw

bacterial$Rank = rank(bacterial$Bacterial.Counts, ties = "average")
```

```
Group.order = aggregate(Bacterial.Counts ~ Method, data = bacterial, mean)$Method
Ri = aggregate(Rank ~ Method, data = bacterial, mean)$Rank
#Ri
SR.2 = var(bacterial$Rank)
#SR.2

all.diff = as.numeric(dist(Ri,method = "manhattan"))
names(all.diff) = c("I vs II","I vs III","I vs IIII","II vs III","II vs IIII", "III vs IIII")

all.diff

SR.2 = var(bacterial$Rank)
K = length(unique(bacterial$Method))
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

alpha =0.05
R = 3000
R.perms = sapply(1:R,function(i){
 permute.data =  bacteria
 permute.data$Method = sample(permute.data$Method,length(permute.data$Method),replace =
FALSE)
 Ri = aggregate(Bacterial.Counts ~ Method, data = permute.data, mean)$Bacterial.Counts
 all.diff = as.numeric(dist(Ri,method = "manhattan"))
 max.diff = max(all.diff)
 return(max.diff)
})
```

```
tukey.cutoff = quantile(R.perms,1-0.05)
tukey.cutoff

tuksplit.groups = split(bacterial,bacterial$Method) #Makes a list of K groups (in alphabetical order)
one.two = rbind(tuksplit.groups[[1]],tuksplit.groups[[2]]) #Binds 1 and 2
one.three = rbind(tuksplit.groups[[1]],tuksplit.groups[[3]]) #Binds 1 and 3
one.four = rbind(tuksplit.groups[[1]],tuksplit.groups[[4]]) #Binds 1 and 4
two.three = rbind(tuksplit.groups[[2]],tuksplit.groups[[3]]) #Binds 2 and 3
two.four = rbind(tuksplit.groups[[2]],tuksplit.groups[[4]]) #Binds 2 and 4
three.four = rbind(tuksplit.groups[[3]],tuksplit.groups[[4]]) #Binds 3 and 4

pval12 = pvalue(wilcox_test(Bacterial.Counts ~ Method, data = one.two))
pval13 = pvalue(wilcox_test(Bacterial.Counts ~ Method, data = one.three))
pval14 = pvalue(wilcox_test(Bacterial.Counts ~ Method, data = one.four))
pval23 = pvalue(wilcox_test(Bacterial.Counts ~ Method, data = two.three))
pval24 = pvalue(wilcox_test(Bacterial.Counts ~ Method, data = two.four))
pval34 = pvalue(wilcox_test(Bacterial.Counts ~ Method, data = three.four))
comp.names  = c("I vs II","I vs III","I vs IIII","II vs III","II vs IIII", "III vs IIII")
all.pvalues = c(pval12,pval13,pval14,pval23,pval24,pval34)
names(all.pvalues) = comp.names

Xb = aggregate(Bacterial.Counts ~ Method,bacterial, mean)$Bacterial.Counts

all.diff = as.numeric(dist(Xb,method = "manhattan"))
names(all.diff) = comp.names


all.pvals = matrix(all.pvalues, nrow = 1)
rownames(all.pvals) = "WRS p-value"
colnames(all.pvals) = comp.names
all.pvals

all.diff = matrix(all.diff, nrow = 1)
rownames(all.diff) = "Pairwise Diff"
colnames(all.diff) = comp.names
all.diff
```

Topic II:

```
library(coin)
mind<- read.csv("~/Downloads/Mind.csv")
head(mind)

table(mind)

chisq_test(Improve~Treatment, mind, distribution = approximate(nresample = 5000))

n = sum(table(mind))
ni. = rowSums(table(mind))
n.j = colSums(table(mind))
all.pjG1 = table(mind)[1,]/ni.[1] #all conditional probabilites for row 1
all.pjG2= table(mind)[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])) #The
z-test-statistics
R <- 5000
r.perms.cutoff = sapply(1:R,function(i){
  perm.data = mind
  perm.data$Treatment = sample(perm.data$Treatment,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)))
cutoff.q
all.Zij = matrix(all.Zij,nrow=  1)
colnames(all.Zij) = c("Major","Mild","Moderate","None")
rownames(all.Zij) = c("Medication vs. Therapy")
all.Zij

cutoff.q = as.numeric(quantile(r.perms.cutoff,(1-alpha)))
```

cutoff.q