535 FINAL EXAM

Gustavo Esparza

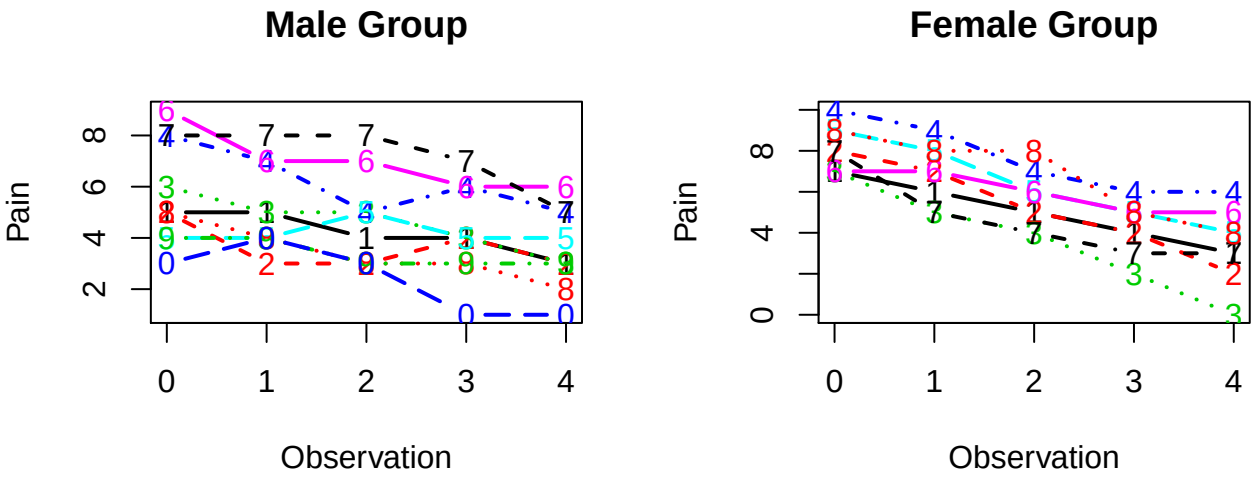
5/18/2019

1

A

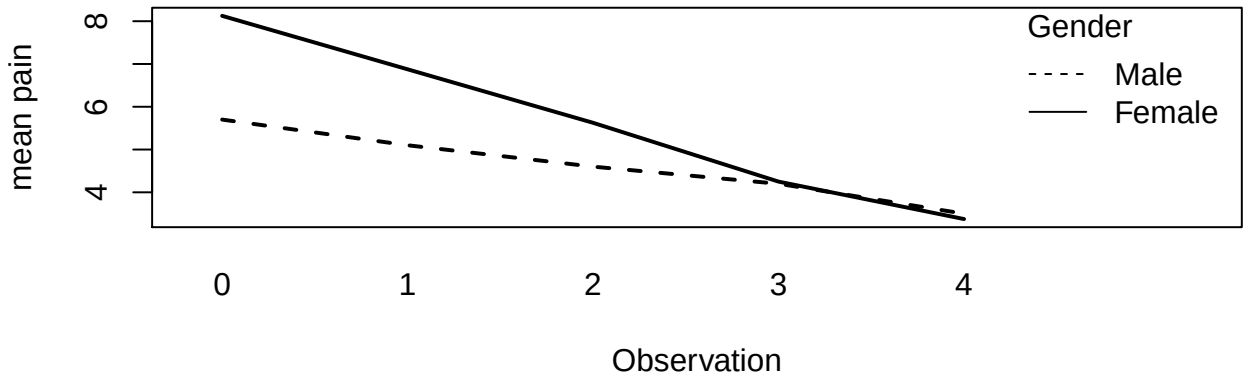
We are provided a dataset consisting of 18 patients suffering from Arthritis. We wish to quantify the effect of GNO44, a new supplement. Each patient ranked their pain on a scale from (0-10) every 3 months over the course of a year, for a total of 5 rankings.
Our first objective is to investigate whether there is evidence that GNO44 is helpful in reducing percieved pain amongst individuals suffering from Arthritis. Upon analyzing our dataset, it is important to first note that we are working with **Repeated Measures**. This means that we have data values that have observations measured for the same subjects over a speicific period of time.
When working with repeated measures, we will utilize an ANOVA test to see if GNO44 has an effect on the patients perception of time. The assumptions needed to proceed for this test is Independence and identical distribution.
Since our data is consisting of 5 groups of data (each group consisting of the pain rankings from each time period), our evidence of GNO44 being effective would be showing a difference between the groups. Conversely, a lack of evidence would be no difference found between the groups. In addition, we can also take the factor of **Gender** into consideration for the reduction in pain for patients.

Here is an initial plot that shows the pain measurements for the Male and Female groups over the period of five observations:

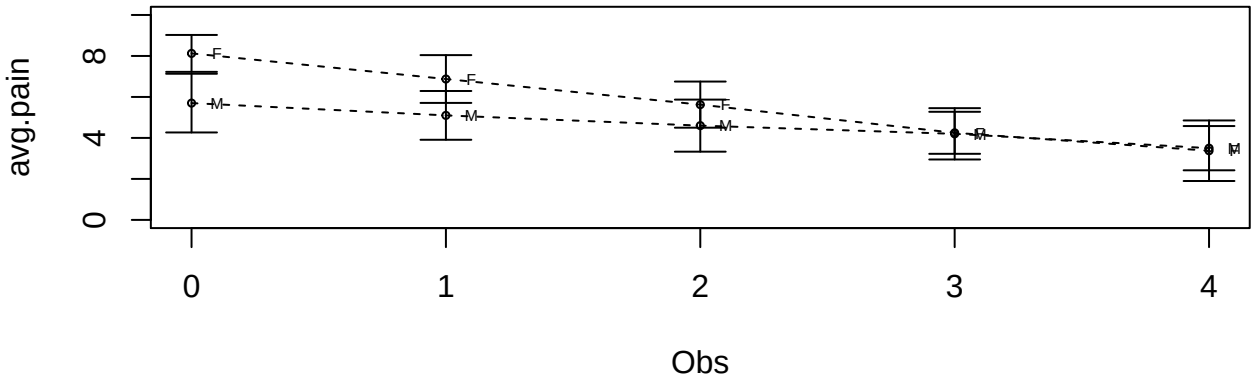


We can see that both groups do show a lower pain level over the span of the measurements, and we can also note that the Female group appears to have a more dramatic decline in pain.

Furthermore, here is an interactive plot that shows the average pain levels for each group:



We can also add Confidence Intervals to our average pain levels for each gender over the span of the trial period. This gives a more in depth on the percieved pain for all patients within each gender group:



Essentially, our ANOVA will test to see if the differences between the groups is due to an effect that the treatment is having, rather than coincidence or random chance. Here is the hypothesis that we will be working with:

$H_o : Group_1 = Group_2 = Group_3 = Group_4 = Group_5$ vs. H_a : At least two groups are different

The resulting test-statistics is an F value that provides a corresponding p-value. Here are the p-values found for the data provided.

```
Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
gender     1  23.58   23.58   2.104  0.166
Residuals 16 179.28   11.21

Error: Within
                   Df Sum Sq Mean Sq F value   Pr(>F)
Observation         4 125.07  31.267   64.10  < 2e-16 ***
gender:Observation  4  21.31   5.328   10.92 8.19e-07 ***
Residuals          64  31.22   0.488
```

First, we can see that the p-value for the Gender factor is large enough that we will Fail to Reject our Null Hypothesis when considering Gender alone. Essentially, we can conclude that Gender alone does not have an effect on the pain felt by patients suffering from Arthritis.

Then, we can see that the following two p-values for our Observation factors and the interaction of Observations and Gender are both quite small. Thus, we can conclude that GNO44 does reduce the pain ratings as the

observations are measured throughout the year. The two p-values inform us that GNO44 reduces pain when considering the entire group AND when considering gender of each patient for pain perception. Thus, we can conclude that GNO44 has a very strong effect in reducing pain for patients suffering from Arthritis.

B

There are a few changes to improve this study that come to mind when reviewing the general setup.

First, we should note that there are only 18 observations. This seems like far too small of a sample size to really measure the impact of GNO44 on patients suffering from Arthritis. Similar to the minimal sample size, we may also consider increased the number of measurements taken for each observation. Perhaps rather than every three months, a more frequent measurement interval(such as monthly or biweekly) would strengthen the results of our tests.

Going along with how frequent pain is being measured, we should also consider what we know about the patients when beginning the test trials. We are simply given the gender of the patient, and surely this is too little information to really understand the impact that a drug has on percieved pain. Factors such as age can also contribute a great deal of information to this study.

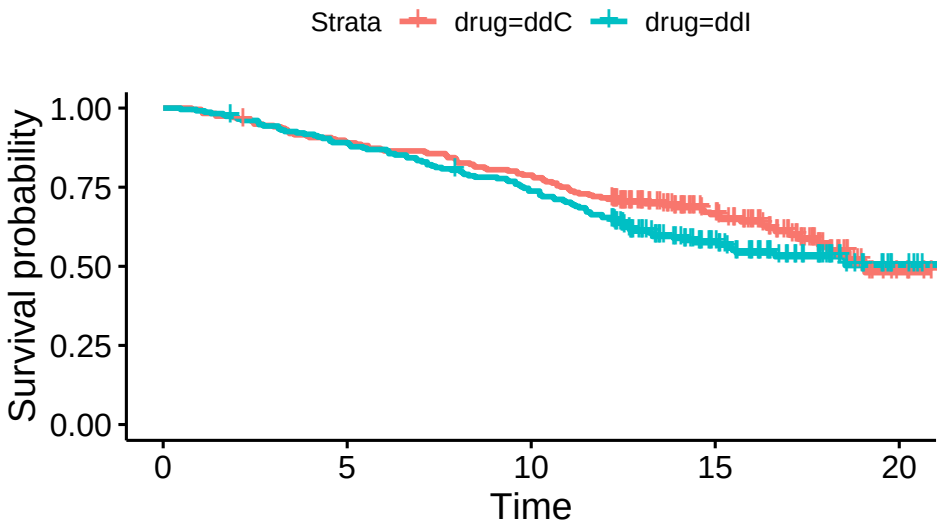
Beyond the size and details of the subject group, we must also consider having a control group that is given a common Arthritis medication so that the effect of GNO44 can be measured against another set of results. This would be ideal for any statistical study that is holistic in researching the effects of any experiement/treatment.

2

We are presented with a dataset consisting of 3 variables: Time (until death or censorship), Death (1 = death, 0 = censored), and CD4 (initial white blood cell count). The interest of this dataset is reporting the effects of Zalcitabine (ddC) vs. Didanosine (ddI) in terms of prolonging life for a patient diagnosed with AIDS. We will be using the data provided to interpret the relative benefit of one drug over the other.

We will begin our analysis with a graph that displays the cummulative survival probabilties for each group. Utilizing the **Survival** package, we will begin with a survival rate of 1 and modify that value by the instantaneous survival rate at each time, t. Thus, the Survival Probability Plot provides a useful display of the two survival rates, over t months, for a patient with AIDS.

Here is the plot of the survival rates for both drugs, ddC and ddI:



Although we can see the Survival Rate for patients using ddC is consistently higher for patients using ddI, we can not say for certain whether this difference is significant in determining prolonging death for patients.

Now having an idea of what the survival rates are for our two groups, we can perform a **Log-Rank Hypothesis test** (Mantel–Haenszel) for a difference in the two groups.

If S_{ddx} is defined as the survival rate for each of our two drug choices, our hypothesis for this test is as follows:

$H_o : S_{ddC} = S_{ddI}$ vs $H_a : S_{ddC} \neq S_{ddI}$

For this test, our test statistic is a Chi-Square statistic that uses Expected and Observed values from the Kaplan-Meier table that is in turn used to create the previously seen Survival Curves. Here is the basic structure of our test-statistic:

$$\chi^2 = \sum \frac{(O_{jt} - E_{jt})^2}{V}$$

Now, having established our Hypothesis and test statistic, we have the following result:

```
## P-value for test in difference between Survival Rates of ddC and ddI =  0.1502432
```

When considering a relatively small (<.10) significance level, our resulting p-value is large enough to fail to reject our null hypothesis. Thus, when considering ddC and ddI patients as two seperate groups, we do not have enough evidence to infer that there is any significant benefit of one drug over the other in terms of prolonging death for AIDS patients.

Considering WBC

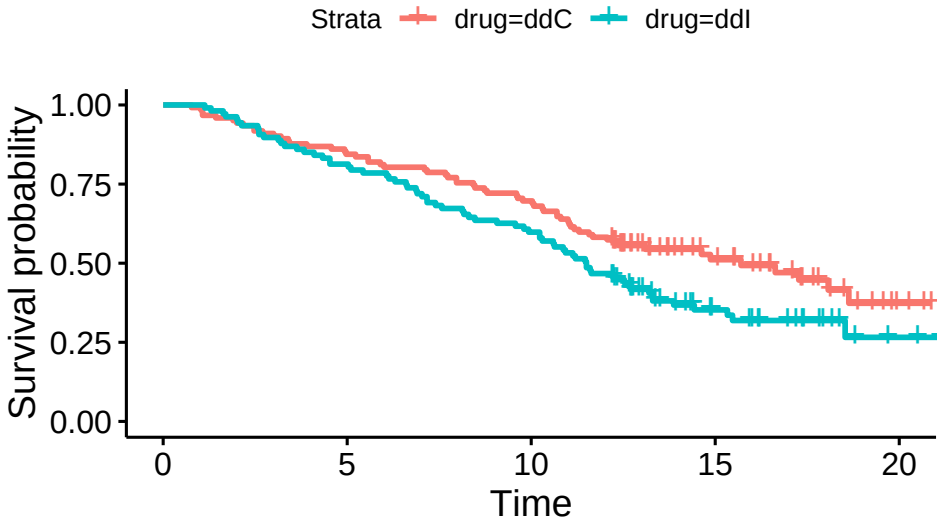
Although our last hypothesis test did not provide any statistically significant conclusions regarding one drug having a greater benefit than the other, it is important to take notice of the variable CD4: White Blood Cell Count at initial observation. After a quick literature review, we have the understanding that a higher CD4 count implies a stronger immune system. Considering this information, perhaps a difference in medication would have a larger impact on patients who had a significantly lower CD4 count upon initial observation.

NOTE: It is understood that the **Cox Hazards Model** could also be used to determine the effects of the drugs while also considering wbc count as a potential confounding factor. For this problem, as well as problem 5, we will partition our additional factor to get another perspective using a grouping method with our Survival Analysis. This methodology was discussed and accepted during Office Hours.

We will choose to parition our groups by dividing WBC count by the median value. Thus, we have a group of patients with a count less than 6 and the other group having a count that is greater than or equal to 6.

Low WBC

Here are the results for the low CD4 group (the lower quarter consists of any patient with a CD4 count less than 6):



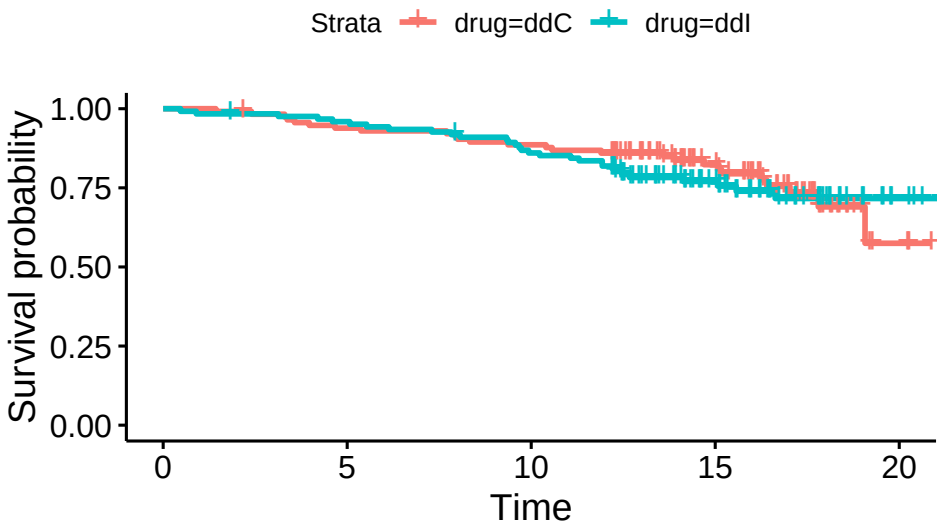
We can see that our survival curves appear to have greater distance than when considering the entire dataset. This is a good sign for our complain, so we will compute our test and compare the p-value.

```
## P-value for test in difference between Low CD4 Survival Rates of ddC and ddI =  0.03697045
```

Our new p-value is smaller than our original value, so we have greater evidence to support the claim that there is a greater benefit in ddC for patients with a lower CD4 count.

High WBC

Here are the results for the high CD4 group (greater than or equal to 6):




```
## P-value for test in difference between High CD4 Survival Rates of ddC and ddI =  0.6692187
```

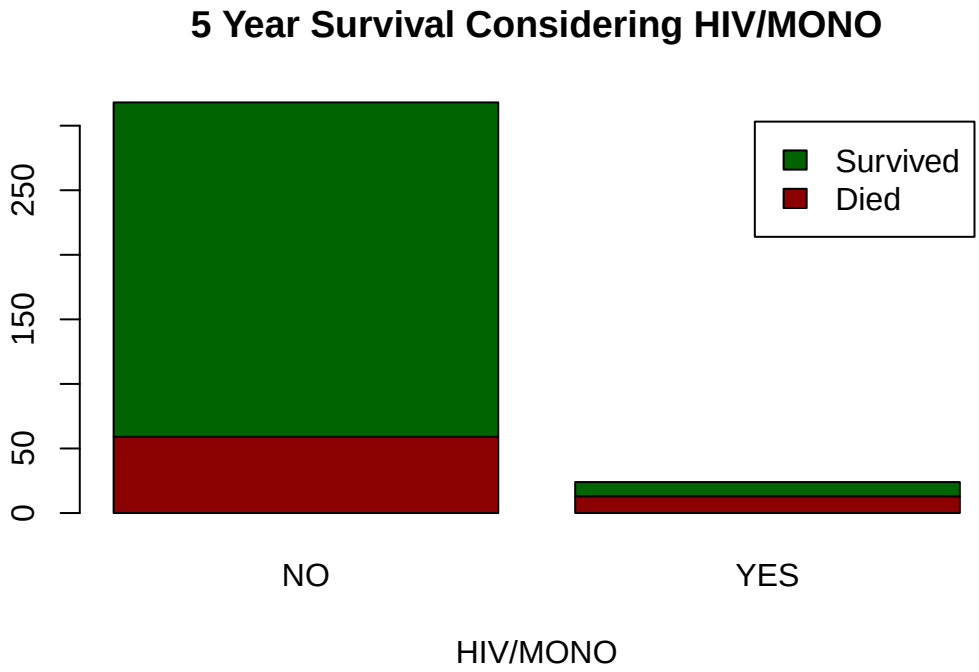
We can see that we do not have evidence for showing a difference between the two drug treatments when applied to a group of patients with a relatively high White Blood Cell count.

Thus, we have shown that ddC does not have an overall benefit for patients as opposed to ddI for the overall population of patients. However, we have seen that the relatively lower WBC group has a greater benefit when using ddC as opposed to ddI.

3

A

We wish to assess the impact of having previously had Mono/HIV as a risk factor for death prior to five years of Hodgkin's diagnosis. A nice initial interpretation for this would be a simple bar plot that shows the proportions for both groups (Yes/No to MONO/HIV)



Although there are not an equal amount of subjects in each category, we can see that the proportion of patients that did NOT have HIV/MONO and SURVIVED after 5 years is quite large. As for the group that did have HIV/Mono, there seems to be an equal split for survival. This is a good first impression of the effect that HIV/Mono can have on survival, but we really need to look at our Survival variable in a more holistic sense that considers all variables.

We can see that our dataset consists of several predictors (both continuous and categorical) and a binary response. Thus, the course of action that is appropriate for determining any inference regarding impact or risk of a variable will be a **Logistic regression Model**

Our Model is defined as follows:
Alive ~ Stage + Age + RBC + Gender + HIV.Mono + WBC

Before implementing our Logistic regression model we should always analyze the residual plots to ensure there are no extreme violations. Upon inspection, we can see that although all residuals are not perfect, there is no clear violation that would affect our ability to make any ifference with our model. Thus, we can continue with our model intact.

Now, we can use our Logistic model to assess the impact of having Mono/HIV as a risk factor for death prior to five years of being diagnosed with Hodgkin's Disease. While holding all other predictors constant, we can use the odds ratio as a point estimate for the impact that having Mono/HIV has on Dying within 5 years of diagnosis. Thus, we have the following result:

```
##
## Change in Odds From a One Unit Increase in HIVMONO =  3.99934
```

We can also construct a Confidence Interval, using our regression coeffecient and corresponding standard error, that can giver greater insight to the impact of having Mono/HIV on our response of Dying within 5 years of diagnosis:

```
##
## Confidence Interval = [ 1.264992  ,  12.64413 ]
```

We can see that our confidence interval is always above one, so we have confidence that having HIV/Mono does increase the risk of dying within 5 years of being diagnosed, in the context of all other predictors.

B

We wish to determine how much more at risk an individual diagnosed with stage IV Hodgkins disease as opposed to stage I, in the context of the other potential confounding variables.

Again, we will be utilizing the coeffecient and standard error provided in our model summary to display a point estimate and a confidence interval for our odds ratio. Here are the results:

```
##
## Change in Odds From having stage IV Hodgkins as opposed to stage I =  6.05278
```

```
##
## Confidence Interval for change in odds from having stage IV = [ 2.092801  ,  17.5058 ]
```

We can see that our confidence interval contains values greater than one, so we are confident that a patient with stage IV Hodgkins is more likely to die within 5 years of being diagnosed as opposed to a patient with stage I Hodgkins. We can also note that the odds can be as high as 17 times more likely to die within 5 years, so the severity of Hodgkins is very impactful.

C

We will be utilising the predict function to find a probability of dying within 5 years of being diagnosed and having the characteristics of the individual described. Here is the result:
Note: We will be taking the complement of the predicted value to provide an estimate of survival rather than dying.

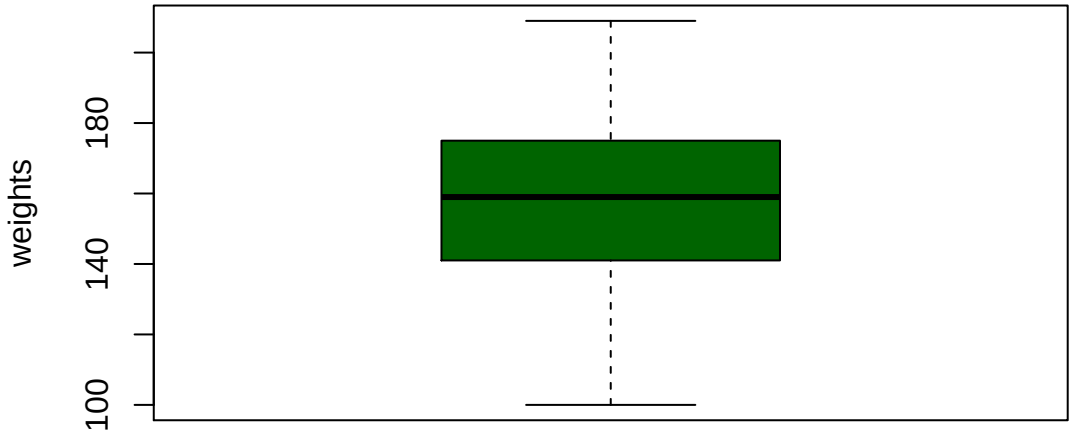
```
## Given the characteristics provided, this individual has a 49.54523
## percent chance of being alive 5 years from the point of diagnosis.
```

Thus, we can claim that this individual has about a 50% chance of surviving 5 years from their time of diagnosis. Given their characteristics, I would also detail the risk that HIV/MONO has on their chances of survival so that their chances of survival are not reduced any further.

4

A

Here is the boxplot of our Weight distribution for the 472 randomly selected students:



Distribution of Weights

For the weights provided, we have the following 5-Number Summary that reflects our boxplot:

```
## Min =  100
## Q1  =  141
## Med =  159
## Q3  =  175
## Max =  209
```

We can see that the boxplot has no outliers which can be confirmed by the lower and upper fence values:

```
## lower fence =  90
## upper fence  =  226
```

B

Our objective is to create a 95% Confidence Interval for the boxplot that we previously created. Thus, we must create a 95% Confidence interval for each of the values in our previous 5 number summary. That is, a confidence interval for the minimum value, first quarter, median, third quarter and maximum value.

All we are provided with in this data set are the 472 values without any information regarding normality for the given population. Specifically, we cannot quite apply any CLT procedures for this dataset. Thus, our alternative option is to bootstrap our 5 confidence intervals.

The procedure for the bootstrapping will be as follows:
1.) Take N samples of our original data (with replacement).
2.) For each sample, compute boxplot values (min,Q1,Med,Q3,Max) and store them in 5 vectors of size N.
3.) Create a credible 95% confidence interval for each of these 5 vectors of our bootstrapped boxplot values.

Here is our result:

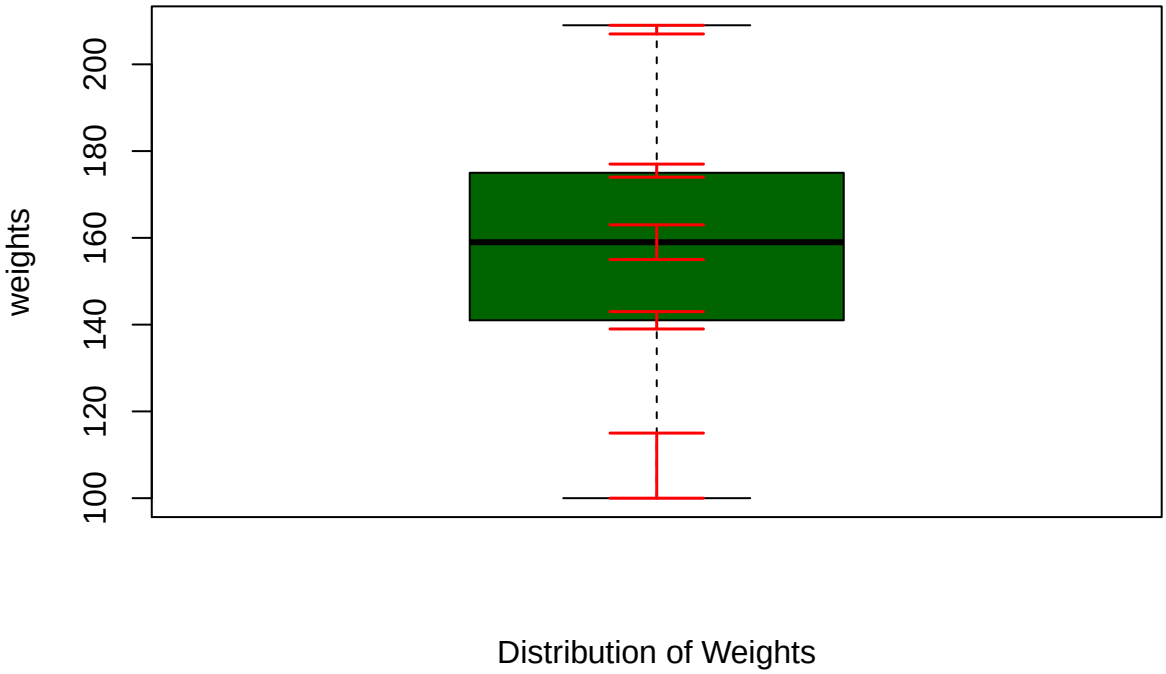
95% confidence intervals for each of the 5 values in our boxplot:

```
## Confidence Interval for MIN = [  100 , 115 ]
## Confidence Interval for  Q1 = [  139 , 143 ]
## Confidence Interval for MED = [  155 , 163 ]
## Confidence Interval for  Q3 = [  174 , 177 ]
## Confidence Interval for MAX = [  207 , 209 ]
```

In addition, we can also compute a 95% confidence interval for the lower and upper fence values that categorize outliers:

```
## Confidence Interval for lower fence = [  83.5 , 95 ]
## Confidence Interval for upper fence = [  222 , 231.5 ]
```

Here is the original boxplot of the distribution of weights, with the confidence bands added:

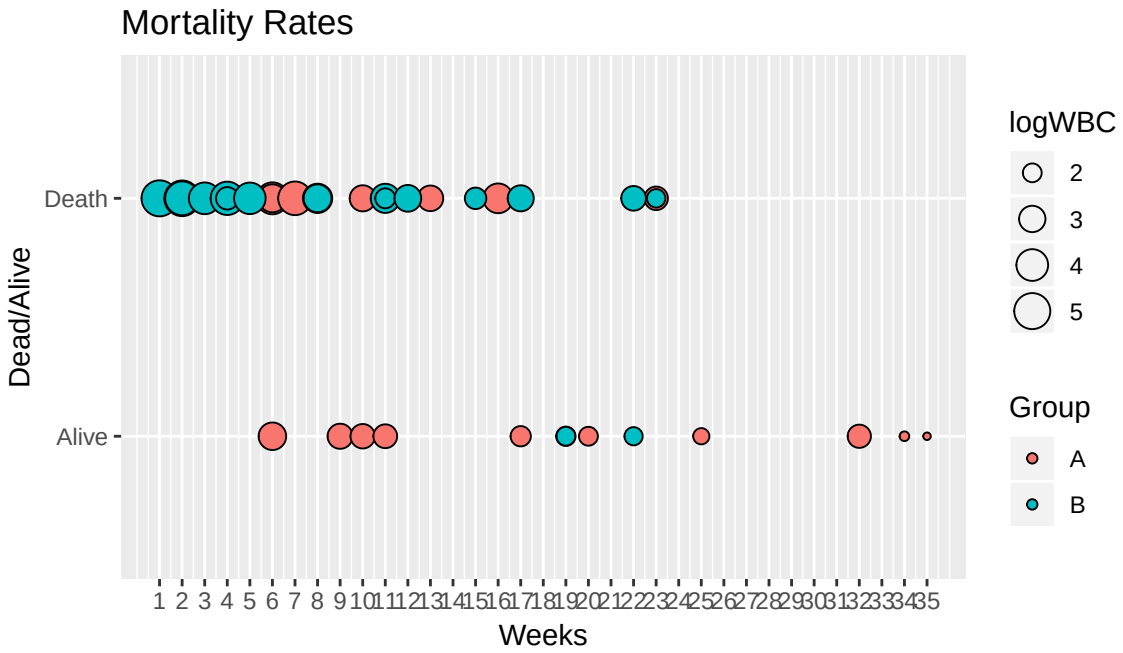


Thus, we have been succesful in creating confidence intervals for our boxplot. Furthermore, we could create a confidence interval (via bootstrapping) for any desired statistic of our data.

5

We have been provided with time to event data for 44 patients who were diagnosed with Leukemia. The data consists of 4 variables: Group,Result,Time, and log white blood cell count. Our objective is to determine whether there is a benefit to being placed in the treatment group as opposed to the control group.

First and foremost, we will take a look at a graph for the two treatment groups while considering mortality and white blood cell count:



Here, it is very evident that the control group had more instances of death than the treatment group. We can also make note of the log white blood cell count and its contribution to survival,

Thus, we have intuition to suggest that the treatment group is indeed more effecient than our control group. To verify our claim with statistical significance, we will compute a Survival Analysis on the Mortality Rates for the Control and Treatment groups.

Thus, we have the following hypothesis statement: If ID is defined as the Incident Density for each group then we have

$H_o : ID_A = ID_B$ vs $H_a : ID_A < ID_B$

Our test statistic is defined as a z-score, so we are able to utilize the normal density to compute our p-value. Here is the result from our test:

```
## P-value for Survival Analysis test =  0.000720004
```

Our resulting p-value is significantly small, so we are able to conclude that there is a benefit to being placed in the treatment group opposed to the control group.

Considering White Blood Cell Count

Once again, we should utilize any aspect of the data provided to gain deeper insight into how the treatments benefit our patients. Previously, we conducted our Survival Analysis without regarding the White Blood Cell

Count. Thus, we will partition our data into two subsets of low and high WBC count. Considering the range of our WBC values, we will create a threshold of a median value of 2.765.

NOTE: It is understood that the **Cox Hazards Model** could also be used to determine the effects of the drugs while also considering wbc count as a potential confounding factor. For this problem, as well as problem 5, we will partition our additional factor to get another perspective using a grouping method.

LOW WBC

```
## P-value for Survival Analysis test of low WBC =  0.01329341
```

We can see that the low WBC group does show evidence to suggest that the treatment medication is more beneficial than the control medication.

HIGH WBC

Here are the results for the higher White Blood Cell count group:

```
## P-value for Survival Analysis test =  0.1193075
```

We can see that the significance level for the high white blood cell group is not as low as the previous two considerations. Thus, we are not as confident that the treatment medication is more beneficial than the control treatment for this subgroup of the dataset.

APPENDIX

```
#LIBRARIES
library(MASS)
library(readr)
library(survival)
library(survminer)
library(dplyr)
library(car)
library(ggplot2)
```

PROBLEM 1

SETTING UP DATA:

```
arthritis= read.csv("/Users/gustavo/Desktop/MATH 535/#FINAL/Arthritis.csv", header=TRUE)
```

```
#FORMAT DATA
score_0=arthritis$Initial
score_3months=arthritis$X3mo
score_6months=arthritis$X6mo
score_9months=arthritis$X9mo
score_12months = arthritis$X12mo

arthritis1 = data.frame(O1 = score_0, O2 = score_3months, O3 = score_6months,
                        O4 = score_9months, O5 = score_12months,
           Treatment = as.factor(c(rep("initial",18),rep("Three",18),rep("Six",18),
                                   rep("Nine",18),rep("Twelve",18))))
```

```
#Creating new datset for use in ANOVA
Observation = as.factor(rep(0:4,18))

pain = rep(0,18*5)
for(i in 1:18){
    for(j in 1:5)
    pain[5*(i-1)+j] = arthritis1[i,j]}

subject = as.factor(rep(1:18,rep(5,18)))
gender1 = as.factor(c(rep(1:10,rep(5,10)),rep(1:8,rep(5,8))))
gender = as.factor(c(rep("Male",50),rep("Female",40)))
arthritis.RM = data.frame(subject=subject,gender=gender,Observation=Observation,pain=pain)
```

PLOTS OF DATA:

```
matplot(0:4,t(arthritis1[1:10,1:5]),xlab="Observation",
        ylab="Pain",main="Male Group",type="b",lwd = rep(2,10))
```

```
matplot(0:4,t(arthritis1[11:18,1:5]),xlab="Observation",
        ylab="Pain",main="Female Group",type="b",lwd = rep(2,10))
```

```
interaction.plot(Observation,gender,pain,lty=1:3,lwd=2,
                 ylab="mean pain",xlab = "Observation",trace.label = "Gender")
```

PLOTS WITH CONFIDENCE INTERVALS:

```
avg.pain = c(apply(arthritis1[1:10,1:5],2,mean),apply(arthritis1[11:18,1:5],2,mean))
sd.pain = c(apply(arthritis1[1:10,1:5],2,sd),apply(arthritis1[11:18,1:5],2,sd))
se.mean = sd.pain/c(rep(sqrt(10),5),rep(sqrt(8),5))
t.score = c(rep(qt(.975,9),10),rep(qt(.975,7),10))
low.bound = avg.pain-t.score*se.mean
high.bound = avg.pain+t.score*se.mean
Obs = rep(0:4,2)
Treatment = rep(c("M","F"),rep(5,2))

plot(Obs,avg.pain,cex=.5,ylim = c(0,10))
points(Obs+.1,avg.pain,pch=Treatment,cex=.5)
lines(Obs[1:5],avg.pain[1:5],lty=2)
lines(Obs[6:10],avg.pain[6:10],lty=2)

for(i in 1:10){
    lines(c(Obs[i]-.1,Obs[i]+.1),c(high.bound[i],high.bound[i]))
    lines(c(Obs[i]-.1,Obs[i]+.1),c(low.bound[i],low.bound[i]))
    lines(c(Obs[i],Obs[i]),c(low.bound[i],high.bound[i]))}
```

ANALYSIS OF VARIANCE TEST

PROBLEM 2

```
aids= read.csv("/Users/gustavo/Desktop/MATH 535/#FINAL/aids.csv", header=TRUE)
```

Here is the plot of the survival rates for both drugs, ddC and ddI:

```
aids_obj = Surv(time = aids$Time, event = aids$death)
aids_fit = survfit(aids_obj ~ drug, data = aids)

ggsurvplot(aids_fit, data = aids, pval = F)
```

LOG RANK TEST:

```
aids_log_rank = survdiff(Surv(Time, death)~drug, data=aids)
X2 = aids_log_rank$chisq
pval= 1-pchisq(X2,1)
cat("P-value for test in difference between Survival Rates of ddC and ddI = ",pval,"")
```

PLOTS AND DATA FOR PARTITIONED WBC:

```
aids_low=subset(aids,CD4<6)
aids_low_obj = Surv(time = aids_low$Time, event = aids_low$death)
aids_low_fit = survfit(aids_low_obj ~ drug, data = aids_low)
ggsurvplot(aids_low_fit, data = aids_low)
```

```
aids_log_rank = survdiff(Surv(Time, death)~drug, data=aids_low)
X2 = aids_log_rank$chisq
pval= 1-pchisq(X2,1)

cat("P-value for test in difference between Low CD4
    Survival Rates of ddC and ddI = ",pval,"")
```

```
aids_high=subset(aids,CD4>=6)
```

```
aids_high_obj = Surv(time = aids_high$Time, event = aids_high$death)
aids_high_fit = survfit(aids_high_obj ~ drug, data = aids_high)

ggsurvplot(aids_high_fit, data = aids_high, pval = T)
```

```
aids_log_rank = survdiff(Surv(Time, death)~drug, data=aids_high)
X2 = aids_log_rank$chisq
pval= 1-pchisq(X2,1)

cat("P-value for test in difference between High CD4 Survival
    Rates of ddC and ddI = ",pval,"")
```

Problem 3

LOAD DATA:

```
hodgkins = read.csv("/Users/gustavo/Desktop/MATH 535/#FINAL/Hodgkins.csv",
                    header = T,stringsAsFactors =T)
attach(hodgkins)
hodgkins$alive= as.character(alive5yr)

hodgkins[hodgkins$alive=="Y",8]=0
hodgkins[hodgkins$alive=="N",8]=1
hodgkins$alive=as.integer(hodgkins$alive)
```

PLOT OF HIV/MONO VS SURVIVAL:

```
hodgkins2 = read.csv("/Users/gustavo/Desktop/MATH 535/#FINAL/Hodgkins.csv",
                     header = T,stringsAsFactors =F)

survive=hodgkins2$alive5yr
survive[survive=="Y"]="Survived"
survive[survive=="N"]="Died"
hm=hodgkins2$HIV.mono
hm[hm=="Y"]="YES"
hm[hm=="N"]="NO"

counts = table(survive,hm)
barplot(counts, main="5 Year Survival Considering HIV/MONO",
  xlab="HIV/MONO", col=c("dark red","dark green"),
  legend = rownames(counts))
```

LOGISTIC MODEL:

```
model=glm(alive~stage+age+rbc+gender+HIV.mono+wbc,family = "binomial",data = hodgkins)
```

```
#Check residuals
summary(model)
residualPlots(model)
```

IMPACT OF HIV/MONO:

```
beta_hm=model$coefficients[8]
ste_hm=summary(model)$coefficients[8,2]
OR_hm = exp(beta_hm)
#cat("\nChange in Odds From a One Unit Increase in HIVMONO = ",OR_hm,"")
```

CONFIDENCE INTERVAL:

```
#Confidence interval
z_star = 1.96
L = exp(beta_hm - z_star*ste_hm)
U = exp(beta_hm + z_star*ste_hm)
#cat("\nConfidence Interval = [",L," , ",U,"]")
```

ODDS FOR HODGKINS STAGE:

```
beta_IV= model$coefficients[4]
ste_IV= summary(model)$coefficients[4,2]
OR_IV = exp(beta_IV)

#Confidence interval
z_star = 1.96
L = exp(beta_IV - z_star*ste_IV)
U = exp(beta_IV + z_star*ste_IV)
#cat("\nChange in Odds From having stage IV Hodgkins as opposed to stage I = ",OR_IV,"")
#cat("\n")
#cat("\nConfidence Interval for change in odds from having stage IV = [",L," , ",U,"]")
```

CASE PREDICTION:

```
Case1 = predict(model,newdata=data.frame(stage="II",age=33,gender="M",rbc=4.3,wbc=12000,HIV.mono="N"),
cat("Given the characteristics provided, this individual has a",Case1*100,
    "\npercent chance of being alive 5 years from the point of diagnosis.")
```

PROBLEM 4

LOAD DATA:

```
weights = read.csv("/Users/gustavo/Desktop/MATH 535/#FINAL/Weight.csv", header=TRUE)
weights2=weights$weight

min=min(weights2)
Q1=sort(weights2)[472*.25]
med=median(weights2)
Q3=sort(weights2)[472*.75]
max=max(weights2)
```

BOXPLOT:

```
boxplot(weights,col = "dark green",xlab="Distribution of Weights",ylab="weights")
```

```
cat("Min = ",min,
    "\nQ1  = ",Q1,
    "\nMed = ",med,
    "\nQ3  = ",Q3,
    "\nMax = ",max,"")
```

B.) BOOTSTRAPPING

```
N=10000

#Create Vectors
```

```
bs_min = rep(0,N)
bs_q1 = rep(0,N)
bs_med = rep(0,N)
bs_q3 = rep(0,N)
bs_max = rep(0,N)

#Fill in 5 number summary vectors
for(i in 1:N){
bs_weights = sample(weights[,1],472,replace = T)

bs_min[i] = min(bs_weights)
bs_q1[i] = sort(bs_weights)[472*.25]
bs_med[i] = median(bs_weights)
bs_q3[i] = sort(bs_weights)[472*.75]
bs_max[i] = max(bs_weights)}
```

95% confidence intervals for each of the 5 values in our boxplot:

```
minL = sort(bs_min)[N*0.025]
minU = sort(bs_min)[N*0.975]

Q1L = sort(bs_q1)[N*0.025]
Q1U = sort(bs_q1)[N*0.975]

medL = sort(bs_med)[N*0.025]
medU = sort(bs_med)[N*0.975]

Q3L = sort(bs_q3)[N*0.025]
Q3U = sort(bs_q3)[N*0.975]

maxL = sort(bs_max)[N*0.025]
maxU = sort(bs_max)[N*0.975]

cat("Confidence Interval for MIN = [ ",minL,",",minU,"]",
    "\nConfidence Interval for  Q1 = [ ",Q1L,",",Q1U,"]",
    "\nConfidence Interval for MED = [ ",medL,",",medU,"]",
    "\nConfidence Interval for  Q3 = [ ",Q3L,",",Q3U,"]",
    "\nConfidence Interval for MAX = [ ",maxL,",",maxU,"]")
```

Here is the original boxplot of the distribution of weights, with the confidence bands added:

```
boxplot(weights,col = "dark green",xlab="Distribution of Weights",ylab="weights")

#MIN
lines(c(0.95,1.05),c(minL,minL),col="red",lwd=1.5)
lines(c(0.95,1.05),c(minU,minU),col="red",lwd=1.5)
lines(c(1,1),c(minL,minU),col="red",lwd=1.5)

#Q1
lines(c(0.95,1.05),c(Q1L,Q1L),col="red",lwd=1.5)
lines(c(0.95,1.05),c(Q1U,Q1U),col="red",lwd=1.5)
lines(c(1,1),c(Q1L,Q1U),col="red",lwd=1.5)

#MED
lines(c(0.95,1.05),c(medL,medL),col="red",lwd=1.5)
```

```
lines(c(0.95,1.05),c(medU,medU),col="red",lwd=1.5)
lines(c(1,1),c(medL,medU),col="red",lwd=1.5)

#Q3
lines(c(0.95,1.05),c(Q3L,Q3L),col="red",lwd=1.5)
lines(c(0.95,1.05),c(Q3U,Q3U),col="red",lwd=1.5)
lines(c(1,1),c(Q3L,Q3U),col="red",lwd=1.5)

#MAX
lines(c(0.95,1.05),c(maxL,maxL),col="red",lwd=1.5)
lines(c(0.95,1.05),c(maxU,maxU),col="red",lwd=1.5)
lines(c(1,1),c(maxL,maxU),col="red",lwd=1.5)
```

PROBLEM 5

LOAD DATA:

```
leukemia = read.csv("/Users/gustavo/Desktop/MATH 535/#FINAL/Leukemia.csv", header=TRUE)
attach(leukemia)
A =subset(leukemia, Group=="A")
B= subset(leukemia, Group=="B")

d1= as.numeric(table(A$Result)[2])
t1=sum(A$Time)
d2= as.numeric(table(B$Result)[2])
t2=sum(B$Time)
```

PLOT OF DATA:

```
ggplot(leukemia, aes(x = Time, y = Result, size = logWBC, fill = Group)) +
        geom_point(shape = 21) +
        ggtitle("Mortality Rates ") +
        labs(x = "Weeks", y = "Dead/Alive") +
        scale_x_continuous(breaks = seq(1, 35, 1))
```

INCIDENT DENSITY TEST:

```
IDhat.male = d1/t1
IDhat.female = d2/t2
V = (d1+d2)*(t1/(t1+t2))*(t2/(t1+t2))
E = (d1+d2)*(t1/(t1+t2))
Z = (abs(d1-E) - .5)/sqrt(V)

pvalue = 2*pnorm(-abs(Z))
cat("P-value for Survival Analysis test = ",pvalue,"")
```

CONSIDERING WHITE BLOOD CELL:

LOW:

```
leukemia_l = subset(leukemia,logWBC < 2.765)
A =subset(leukemia_l, Group=="A")
B= subset(leukemia_l, Group=="B")

d1= as.numeric(table(A$Result)[2])
t1=sum(A$Time)
```

```
d2= as.numeric(table(B$Result)[2])
t2=sum(B$Time)
```

```
IDhat.male = d1/t1
IDhat.female = d2/t2
V = (d1+d2)*(t1/(t1+t2))*(t2/(t1+t2))
E = (d1+d2)*(t1/(t1+t2))
Z = (abs(d1-E) - .5)/sqrt(V)

pvalue = pnorm(-abs(Z))
cat("P-value for Survival Analysis test of low WBC = ",pvalue,"")
```

HIGH:

```
leukemia_h = subset(leukemia,logWBC > 2.765)
A =subset(leukemia_h, Group=="A")
B= subset(leukemia_h, Group=="B")

d1= as.numeric(table(A$Result)[2])
t1=sum(A$Time)
d2= as.numeric(table(B$Result)[2])
t2=sum(B$Time)
```

```
IDhat.male = d1/t1
IDhat.female = d2/t2
V = (d1+d2)*(t1/(t1+t2))*(t2/(t1+t2))
E = (d1+d2)*(t1/(t1+t2))
Z = (abs(d1-E) - .5)/sqrt(V)

pvalue = pnorm(-abs(Z))
cat("P-value for Survival Analysis test = ",pvalue,"")
```