Trends of Mortality in the United States

Joanna Jeon (jl7fb), Tommy Jun (tbj2cu), Jimin Lee (jl3jn), and Frank Woodling (fsw5vb)

I. Introduction

The purpose of this report is to analyze mortalities across the United States with respect to a number of causes. The data is extracted from the Center of Disease Control. A variety of nonparametric statistical tests will be performed on the data to draw conclusions on if there exists any significant relationships between not only mortality and life expectancy in the United States but also between states across various levels such as geographic location or economic status.

The methods applied in this analysis are divided into two distinct statistical tests: parametric and nonparametric. In general, nonparametric methods require minimal assumptions about the form of the distributions of the populations of interest. We make no assumptions about the probability distributions of the variables being assessed. These methods only assume that we have a continuous distribution, and the population distribution depends on the location or scale parameters. Parametric methods require more and strict assumptions. They require that the form of the population distribution be completely specified except for a finite number of parameters.

The four tests that are performed are: (a) the Wilcoxon Rank-Sum Test on general mortality between states at above and below average income levels, (b) the Kruskal-Wallis Test on heart disease rate between different geographic regions, (c) Pearson's correlation coefficient test between homicide and injury by firearms between states, and (d) Bootstrapping on influenza by taking samples from states and determining confidence intervals for the number of deaths and rate of deaths between states. Furthermore, the bootstrap methods will be compared to non-boostrap methods.

The reason why t-test cannot be used over nonparametric tests are because some of the basic assumptions of the t-test may not be met. These are: (1) the observations are independent or drawn randomly from an infinite sample, (2) The population is normally distributed, (3) the variances of the scores from each group are equivalent, and (4) there are no outliers. Generally, the normality assumption is not satisfied for many real-life cases. If we cannot assume the assumptions we mentioned above, nonparametric tests are appropriate to use.

II. Summary

A. Wilcoxon-Rank Sum

The 50 states (and Washington D.C.) are partitioned into two income groups which are greater or less than the national average. According to the data form the New Jersey Department of Labor and Workforce Development, District of Columbia, Connecticut, North Dakota, Massachusetts, New Jersey, Alaska, Wyoming, New Hampshire, Maryland, New York, Virginia, Washington, California, Nebraska, Minnesota, Colorado, Rhode Island, South Dakota, Pennsylvania, Illinois, Vermont, Delaware, Hawaii, Kansas, and Texas are categorized into the higher income group. Iowa, Oklahoma, Wisconsin, Florida, Louisiana, Ohio, Tennessee, Missouri, Maine, Nevada, Michigan, Montana, Oregon, Indiana, North Carolina, Georgia, Arizona, Arkansas, Alabama, Utah, Kentucky, Idaho, New Mexico, South Carolina, West Virginia, and Mississippi are grouped together into the lower income group. The summary statistics are as follows:

	High	Low
Min	3997	9511
Mean	52539.48	49365.62
Median	29630	43510
SD	63508.95	37413.69
Max	248400	181100

TABLE I

Summary Statistics for (a)

The corresponding histogram, quantile-quantile plot, and boxplot for the data are indicated on the next page.

As we can see from *Figure 1*, both of the higher income group and lower income group are skewed to the right and cannot assume that the data is normally distributed. This is further supported by the Q-Q plots in *Figure 2*. From *Figure 3*, we can see that the variations and means are not very different from each other.

B. Kruskal-Wallis Test

The 50 states (and Washington D.C.) are partitioned into four geographical regions, Northeast, Midwest, South, and West, based on the US Census Bureau's division assignments. *Table III* lists the states in each region. The summary statistics are as followed:

	Northeast	Midwest	South	West	All
Min.	179.6	107.9	152.0	96.0	96.0
1st Qu.	194.7	184.3	189.7	151.0	165.5
Median	207.4	192.9	218.2	161.2	194.7
Mean	207.8	194.9	213.5	155.5	193.4
3rd Qu.	224.1	227.8	249.3	166.0	224.4
Max	247.6	244.1	258.0	197.4	258.0

TABLE II

Summary Statistics for (b)

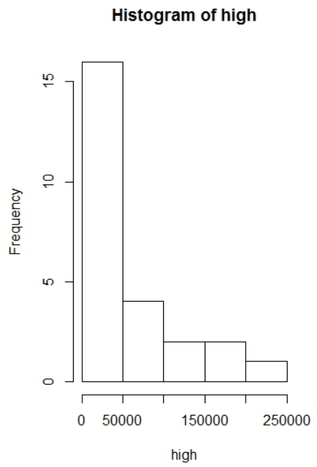

Fig. 1. Histogram of (a)

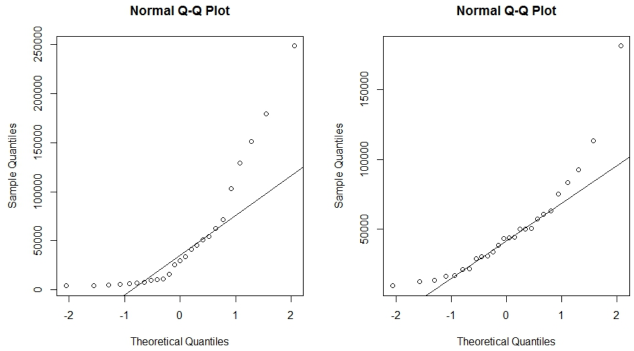

Fig. 2. Q-Q Plot of (a)

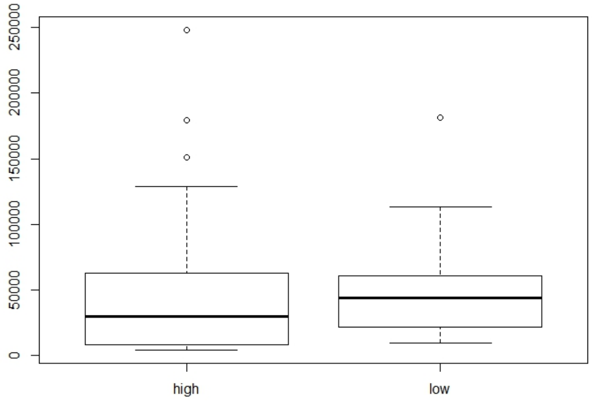

Fig. 3. Boxplot of (a)

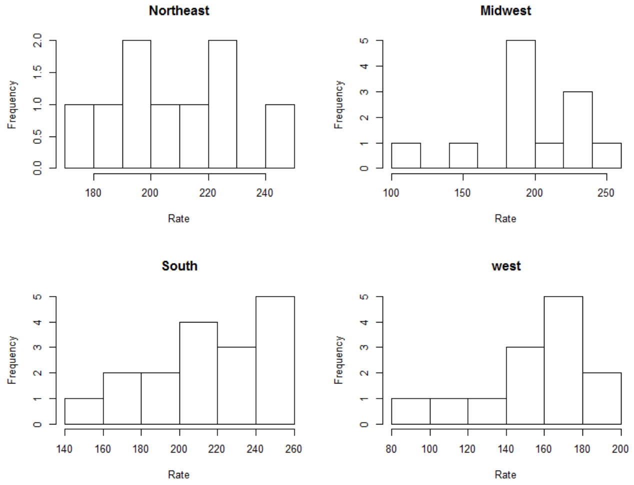

Fig. 4. Histogram of (b)

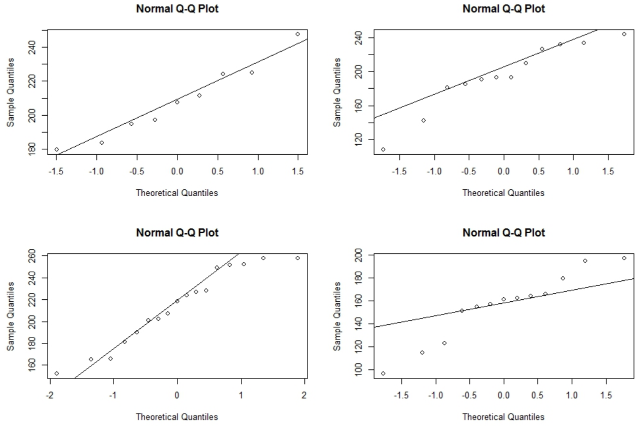

Fig. 5. Q-Q Plot of (b)

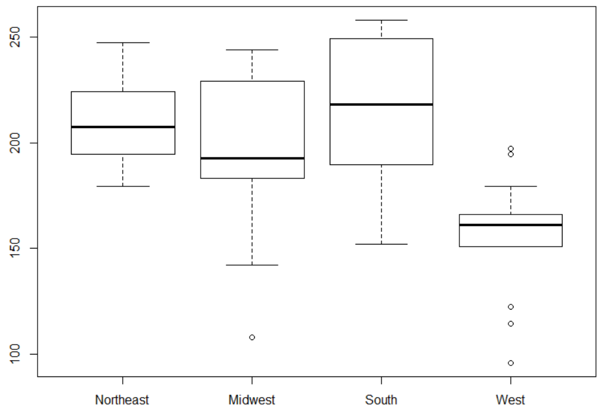

Fig. 6. Boxplot of (b)

Northeast	Midwest	South	West
Connecticut	Indiana	Delaware	Arizona
Maine	Illinois	D.C	Colorado
Massachusetts	Michigan	Florida	Idaho
New Hampshire	Ohio	Georgia	New Mexico
Rhode Island	Wisconsin	Maryland	Montana
Vermont	Iowa	North Carolina	Utah
New Jersey	Kansas	South Carolina	Nevada
New York	Minnesota	Virginia	Wyoming
Pennsylvania	Missouri	West Virginia	
	Nebraska	Alabama	
	North Dakota	Kentucky	
	South Dakota	Mississippi	
		Tennessee	
		Arkansas	
		Louisiana	
		Oklahoma	
		Texas	

TABLE III

State Divisions for (b)

From this table, we can see right away that the western region not only contains the minimum observation, but it is lower than overall in every statistic. The northeast and southern regions seem to be higher on average, and the Mid-western region is very close to the overall averages. These observations are noted in the corresponding histograms, Q-Q plots, and boxplots as indicated in the previous page. None of the histograms seem to be normally distributed, the corresponding Q-Q plots suggest the same. Three of them (all but the Northeast) are clearly left-skewed. This suggests that while some regions have similar rates of death amongst states, there are also nearby states with very different rates. The fact that the data is not normal means that nonparametric tests will be our best option.

C. Pearson's

The Spearman rank correlation test method is used to determine the strength of the relationship between rate of homicide per state and the rate of injury by firearm. The data collected is the total number and rate or mortalities (per person, per 100,000). The non parametric approach to testing the relationship between two variables, using a calculated correlation coefficient, will allow us to determine the extent to which injury per firearm is related to, and hence can predict the rate of homicide. There are a total of 49 samples for each dataset from each state, omitting US States and territories with insufficient data. The following are the relevant summary statistics for the number of deaths due to firearms and homicides in the US:

	Firearms	Homicide
Min.	2.79	1.7000
Median	11.58	4.992
Mean	11.70	5.000
SD	4.09	2.52
Max	19.60	13.900

TABLE IV

Summary Statistics for (c)

A scatterplot is used to determine if there is any noticeable trend between the two variables. Initially there does seem to be a positive correlation. A histogram and quantile-quantile plot is used to determine normality.

D. Bootstrap

The bootstrap methods will be used to determine which States in the United States are the most prone to the impacts of influenza. In particular, data is collected for the total number and rate of mortalities (persons per 100,000) due to influenza in each State. Nonparametric bootstrap methods are applied with many iterations to determine point estimates and corresponding confidence intervals for both statistics. The summary statistics are as followed:

	Number	Rate
Min.	66	9.00
1st Qu.	340	15.40
Median	761	18.50
Mean	1117	18.53
3rd Qu.	1399	21.15
Max	6551	32.50

TABLE V

Summary Statistics for (d)

Histograms and Q-Q plots are constructed to determine the normality of the data. The distribution of Number of Deaths is not normally distributed while the distribution of Rate of Deaths is normally distributed. A parametric bootstrap method can be applied to the Rate of Deaths; however, for the purposes of this report a nonparametric bootstrap method will be applied to both statistics. Nonparametric methods are used because no assumptions are made on the underlying distributions from which the data is drawn. The cost of this flexibility is a decreased statistical power; however, it is necessary for sets of data where normality cannot be assumed. Two other assumptions must also be met: (1) the observations are independent and (2) they are randomly selected.

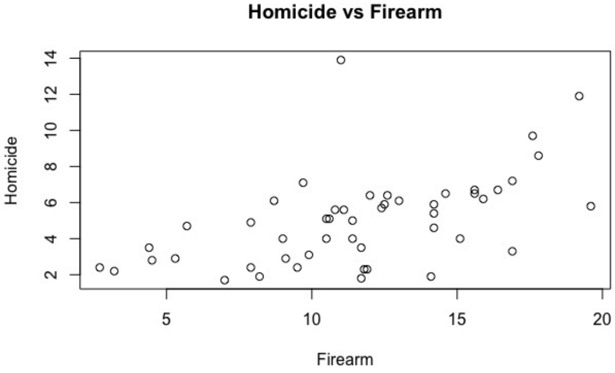

Fig. 7. Scatterplot of (c)

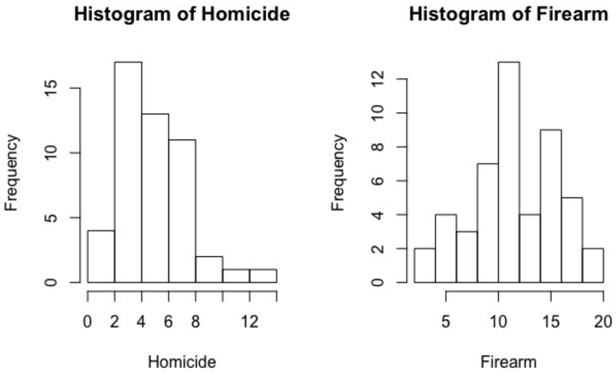

Fig. 8. Q-Q Plot of (c)

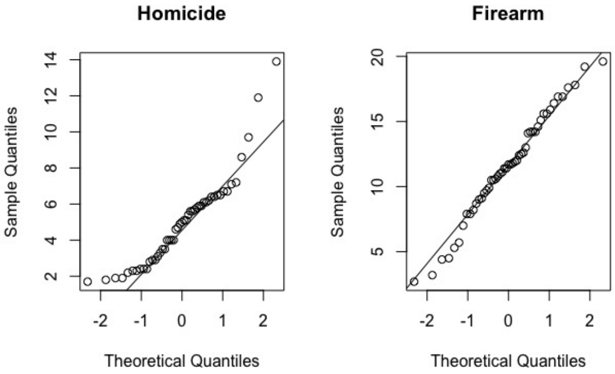

Fig. 9. Histogram of (c)

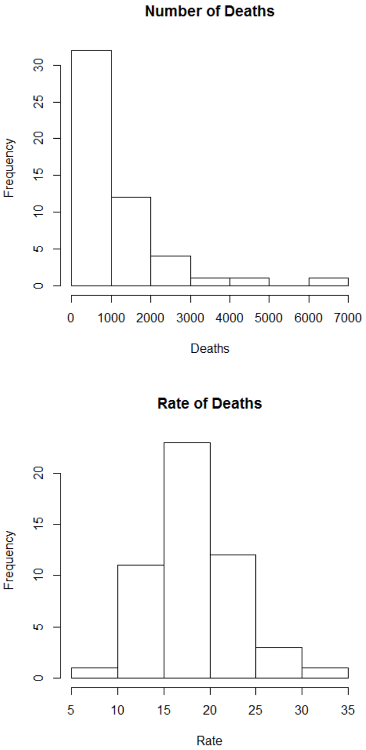

Fig. 10. Histogram of (d)

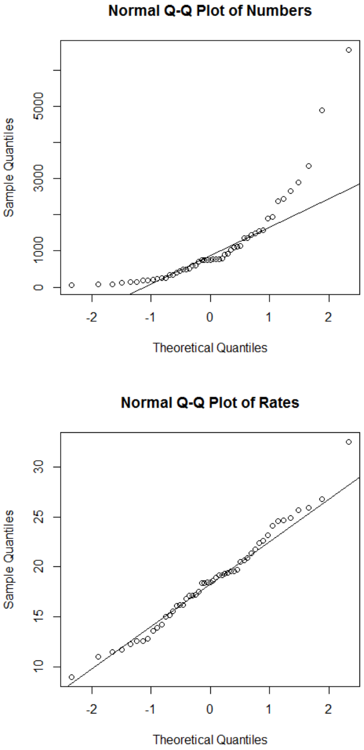

Fig. 11. Q-Q Plot of (d)

III. Methodology

A. Wilcoxon-Rank Sum

To see if there is a difference in total mortality between the states which have higher income and lower income, the Wilcoxon rank sum test would be performed. Before performing the Wilcoxon test, the permutation test on deviances would be performed to see if there is equal variance between higher income group and the lower group. It is a nonparametric method to see the differences in the deviance among groups by permuting all possible outcomes with minimal assumptions. For this test, we only assume that each sample is identical and comes from the same population. After the permutation test on deviances is performed, the Wilcoxon rank sum test will be conducted to see if there is a difference in mean of the mortality between the higher income group and the lower group. Wilcoxon test is similar to the permutation test but ranks are assigned instead of exact values. The Wilcoxon test is preferred than permutation test because the Wilcoxon test filters out outliers by assigning ranks. For this test, we assume that each sample data come from the same population, and each pair is chosen randomly and independently.

B. Kruskal-Wallis Test

The test we will perform is a Kruskal-Wallis test. The null hypothesis for the test is that the mean ranks of the different treatments are the same. A low (< 0.05) p-value would mean that at least one group is different from the others. This test does not require normally-distributed data. It does however require that the groups have equal variances. In order to test this assumption we will first use a permutation test on deviances. This test will check each pairwise combination of regions for unequal variance. Note that if even one pair of regions have different variances then the Kruskal-Wallis test cannot be used. If the Kruskal-Wallis test does indicate there exists a difference between means, then pairwise permutation tests are performed with a Bonferroni adjustment. This pairwise test will determine which specific groups have different means from one another. The assumptions for this test are satisfied by the previous two tests.

C. Pearson's

To determine the degree to which a states homicide rate is related to its firearm rate (death via firearm), the Spearman rank correlation test is performed. This non-parametric method allows us to delve further into a linear relationship by measuring the extent to which Homicide increases with Firearm by comparing the ranks of Homicide and Firearm. First, the Spearmans rank correlation statistic (rs), which measures how one variable is monotonically dependent on the other variable, was calculated for the observed values. Given there are no ties, a rs value of 1 (or -1) indicates that one group is a perfect monotone increasing (or decreasing) function of the other. Subsequently, all possible outcomes were permuted and compared to the observed Spearman statistic to see whether, based upon this sample, there is any or no evidence to suggest that linear correlation is present in the population. For this test, we assume that the data are not normal, paired, linearly related, and that the data are in internal or ratio level.

D. Bootstrap

Bootstrap methods are applied with 1000 iterations. The mean values of these 1000 iterations are plotted in a histogram and the new distributions of the bootstrap means are normally distributed in the following figure. By determining the mean of the bootstrap distribution a point estimate of the population mean is found.

Then, a 95% confidence interval is found for the point estimate by calculating the sample standard deviation of the bootstrap distribution. Then 100 point estimates and corresponding 95% confidence intervals are calculated and then compared against 100 point estimates and their corresponding 95% confidence intervals determined by non-bootstrap methods.

Finally, by taking the difference between the 95% confidence intervals of the bootstrap and non-bootstrap methods we can observe if there exists a difference between the confidence intervals calculated through the bootstrap and non-bootstrap methods.

IV. Results

A. Wilcoxon-Rank Sum

Test	p-value	α	Conclusion
Permutation Test on Deviances	0.1475	0.05	Fail to Reject H_0 (equal variances)
Wilcoxon-Rank Sum Test	0.2264	0.05	Fail to Reject H_0 (equal rank sums)

TABLE VI

RESUTS OF (A)

B. Kruskal-Wallis Test

Test	p-value	α	Conclusion
Permutation Test on Deviances (Minimum p)	0.2946	0.05	Fail to Reject H_0 (equal variances)
K-W Test	0.0005	0.05	Reject H_0 (Different means)

TABLE VII

RESUTS OF (B)

	Northeast	Midwest	South	West
Northeast	-	3.12	3.84	0.00
Midwest	-	-	1.44	0.42
South	-	-	-	0.06
West	-	-	-	-

TABLE VIII

P-VALUES OF PAIRWISE PERMUTATION TESTS, BONFERRONI ADJUSTED.

C. Pearson's

Test	p-value	α	Conclusion
Spearman Correlation Permutation Test	≈ 0	0.05	Reject H_0 (Exists a correlation)

TABLE IX

RESUTS OF (C)

D. Bootstrap

	Non-Bootstrap	Bootstrap
Number	(785.56,1886.68)	(802.90,1860.49)
Rate	(16.19,20.27)	(16.14,20.35)

TABLE X

CONFIDENCE INTERVAL FOR ONE ITERATION.

For comparison, the actual mean of the number of deaths is 1117.26 and for rate of deaths is 18.53.

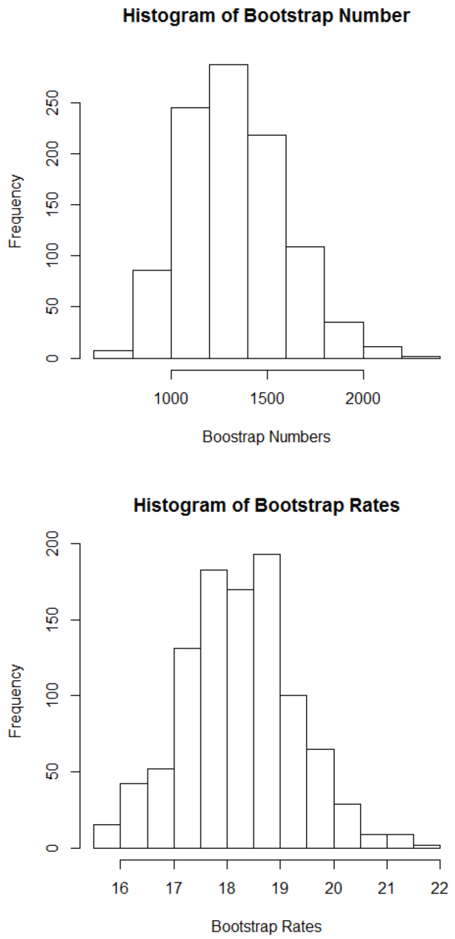

Fig. 12. Histogram of Bootstrap Samples (d)

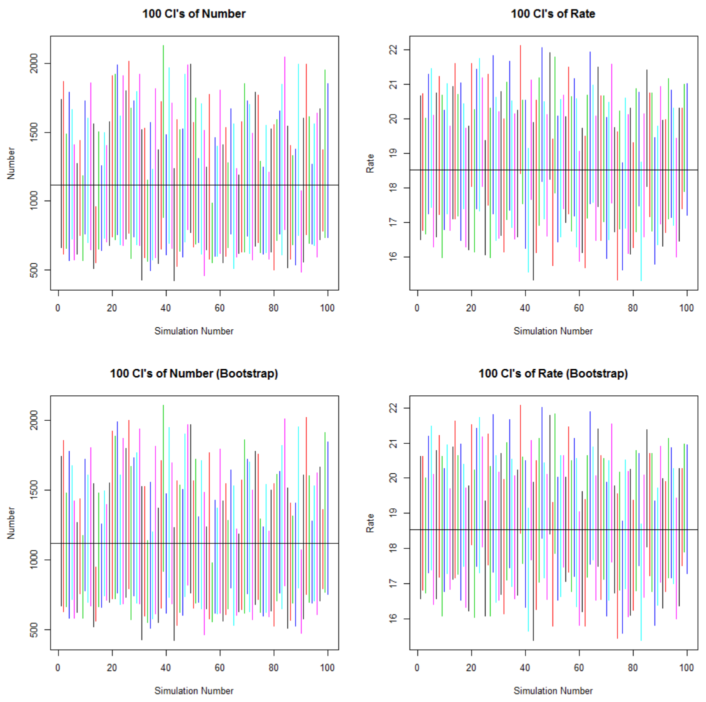

Fig. 13. 100 Confidence Intervals

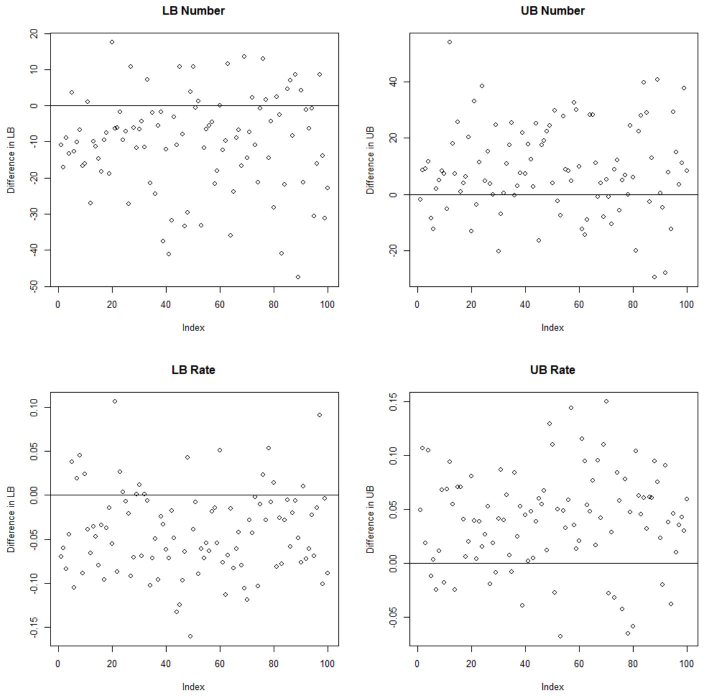

Fig. 14. Difference in CI bounds between methods.

V. DISCUSSION

The following are the hypotheses for the tests performed in (a). First, for the permutation test on deviances, the null hypothesis indicates that the variances of mortalities across the low and high income levels are equal while the alternative hypothesis rejects this. The test concludes with a p-value of 0.1475, therefore we fail to reject the null hypothesis at a significance level of 0.05. Since we have now proved that the variances between the two groups are equal, we can proceed with the Wilcoxon-Rank Sum test. For the Wilcoxon-Rank Sum test, the null hypothesis indicates that the mortality rates for the two income groups are the same while the alternative hypothesis indicates that they are different. The test concludes with a p-value of 0.2264, therefore we fail to reject the null hypothesis at a significance level of 0.05. This concludes that there is no significant difference in general mortality levels between the states.

The following are the hypotheses for the tests performed in (b). First, for the permutation test on deviances, the null hypothesis indicates that the variances of mortalities for pairwise groups chosen from the four geographic regions are equal while the alternative hypothesis rejects this. Pairwise tests are performed for every single permutation and the lowest p-value is chosen. The test concludes with a minimum p-value of 0.2946, therefore we fail to reject the null hypothesis at a significance level of 0.05. Since we have now proved that the variances between the four groups are equal, we can proceed with the Kruskal-Wallis test. For the Kruskal Wallis test, the null hypothesis indicates that the mean mortality rates are the same between all four groups while the alternative hypothesis indicates that there it at least one different mean. We reject the null hypothesis at a significance level of 0.05 (p=0.0005). This concludes that there is a significant difference between the mean mortality rates between the states regarding heart disease. Furthermore, permutation tests performed between the groups and liberally adjusted by the Bonferroni method indicates that there exists a significant difference in means only between the Northeast and West. There is almost a significant difference between the West and South with a p-value of 0.06, which under a more conservative adjustment may be considered statistically significant. The tests conclude that the mean rate of heart disease in the Western states is significantly different than that of the Northeastern states. Further investigation should be done to determine what factors may give arise to this difference.

The following are the hypothesis for the tests performed in (c). First, the correlation coefficient value of 0.6145 confirms what was indicated in the graph; there appears to be a positive correlation between Firearm and Homicide in the sample. Thus, large values of Homicide are associated with large values of death by firearm. Second, for the permutation test for the Spearman coefficient, the null hypothesis indicates that there is no evidence to suggest that there is a relationship between homicide rates and firearm deaths in the US. The alternative hypothesis indicates that there is evidence to suggest that there is a monotomically correlated relationship between homicide rates and firearm deaths in the US. Since the p-value is 0, we can say that we have very strong evidence to believe the alternative hypothesis, and conclude that there is a strong, positive, monotonic correlation between homicide and firearm deaths.

The following is a discussion on the findings in (d). Both methods, bootstrap and non-bootstrap, generated a confidence interval from a sample of the 50 states and both contained the actual mean of both number and rate of deaths due to influenza in the United States. By repeating the bootstrap methods repeatedly it is determined that the bootstrap methods yield a desirable tighter confidence intervals compared to non-bootstrap methods. This is determined by iterating both methods 100 times and determinign their corresponding confidence intervals. By taking the difference in confidence intervals generated by the two methods for these 100 iterations, it is observed that the lower bounds are higher for the bootstrap method and upper bounds are lower for the bootstrap method, on average. By using the bootstrap interval, we check for the states whose values are greater than the upper bound of the 95% confidence interval as these are statistically significant deviations from the mean. The states which satisfy these criteria and are then determined to be more prone to influenza and also impact a large number of people. These states are: New York, Ohio, and Pennsylvania. Notably, the most populated cities within each state are New York, Columbus, and Philadelphia, respectively. If further investigations are to be performed in the United States on mortality due to influenza, these are prime locations.

N.B. For reference: The states which have significant values less than the lower tail are: Alaska, Arizona, Colorado, DC (District of Columbia), Idaho, Minnesota, Oregon, South Carolina, Utah, Vermont, and Washington. These would be states of least concern.

VI. CONCLUSION

Overall the non parametric tests display the variation in mortality patterns across the states. Our findings include: Although there is no difference in the total mortality between the states in terms of the amount of income, the significance difference is found among different region of states. There is a difference between heart disease death rate and region, especially the Western region is significantly different from the Northeastern region. A significant, positive, monotonic relationship between homicide rate and injury by firearms between states. Furthermore, based on the confidence intervals determined by the bootstrap methods, two of the three states which are more prone to influenza are located in the Northeast region. Generally, the Northeastern states exhibit statistically significant variance from the other states. Based on this information, it is advisable to further investigate mortalities in this region to determine what factors may cause it to be different from the other regions.

CODE

```
##################################################################################
#(A)
new.data <- read.delim2("C:/Users/Jimin Lee/Desktop/STAT 3480/final/new.data.txt")
attach(new.data)
#########
high2 = high[!is.na(high)]

par(mfrow = c(1,2))
hist(high);hist(low)

par(mfrow = c(1,1))
boxplot(new.data)

summary(high);summary(low)

sd(high, na.rm = TRUE);sd(low)

mean(high2);mean(low)

par(mfrow = c(1,2))
qqnorm(high);qqline(high)
qqnorm(low);qqline(low)

#########

## test equal variance
library("jmuOutlier")
rmd.test(high2, low, "two.sided")

## wilcoxon test
wilcox.test(high, low, alternative = "two.sided")

mediantst <- function(x, y, nreps=100)
{
  d.obs <- abs(median(x) - median(y))
  nx <- length(x)
  ny <- length(y)
  tail.prob <- 0

  for(i in 1:nreps)
  {
    xy <- sample(c(x,y)) # permute combined list
    x <- xy[1:nx] # first nx are assigned to x
    y <- xy[seq(nx+1,nx+ny)] # next ny to y
    d.sim <- abs(median(x) - median(y))
    if(d.sim >= d.obs) # increment tail prob
      tail.prob <- tail.prob + 1
  }
  tail.prob <- tail.prob / nreps
  return(tail.pro
}

mediantst(data.2[[1]][!is.na(data.2[1])],data.2[[2]][!is.na(data.2[2])])
mediantst(data.2[[1]][!is.na(data.2[1])],data.2[[3]][!is.na(data.2[3])])
mediantst(data.2[[1]][!is.na(data.2[1])],data.2[[4]][!is.na(data.2[4])])
mediantst(data.2[[2]][!is.na(data.2[2])],data.2[[3]][!is.na(data.2[3])])
mediantst(data.2[[2]][!is.na(data.2[2])],data.2[[4]][!is.na(data.2[4])])
mediantst(data.2[[3]][!is.na(data.2[3])],data.2[[4]][!is.na(data.2[4])])

##################################################################################
# (B)
data = read.csv('heart_disease.csv', header = T)
attach(data)
####################
### Arrange data ###
####################

# Split into vectors by region
northeast <- subset(data, data$Region=="Northeast")
northeast.rate <- northeast[,4]

midwest <- subset(data, data$Region=="Midwest")
midwest.rate <- midwest[,4]

south <- subset(data, data$Region=="South")
south.rate <- south[,4]

west <- subset(data, data$Region=="West")
west.rate <- west[,4]


# Create group vector: 1 Northeast, 2 Midwest, 3 South, 4 West
division_vec <- c(rep(1, length(northeast.rate)),
  rep(2, length(midwest.rate)), rep(3, length(south.rate)),
                 rep(4, length(west.rate)))

rate_vec <- c(northeast.rate, midwest.rate, south.rate, west.rate)

#####################################
### Summary Statistics and plots  ###
#####################################
summary(Rate[1:51])
summary(northeast.rate)
summary(midwest.rate)
summary(south.rate)
summary(west.rate)

par(mfrow=c(1,1))
boxplot(northeast.rate,midwest.rate, south.rate, west.rate,
        names = c("Northeast", "Midwest", "South", "West"))

par(mfrow=c(2,2))
hist(northeast.rate, main = "Northeast", xlab = "Rate")
hist(midwest.rate, main = "Midwest", xlab = "Rate")
hist(south.rate, main = "South", xlab = "Rate")
hist(west.rate, main = "west", xlab = "Rate")

#####################################
### Permutation Test on Deviances ###
#####################################
data.2 = read.csv('region_state_rates.csv', header = T)
```

```
attach(data.2)

test.deviance = function(trt1,trt2)
{
  dev1 = trt1 - median(trt1);
  dev2 <- trt2 - median(trt2)
  all = c(dev1, dev2)
  index = seq(along=all)
  indexIntrt1 = combn(index, 3)
  RMD = NULL

  for(i in 1:dim(indexIntrt1)[2])
  {
    RMD[i] = mean(abs(all[indexIntrt1[, i]]))/mean(abs(all[-indexIntrt1[, i]]))
  }

  p_value = sum(RMD >= RMD[1])/length(RMD)
  p_value
}

regions = list(0,0,0,0)

for (i in 1:4)
{
  regions[[i]] = data.2[[i]][!is.na(data.2[i])]
}

p.values.list = NULL

for (i in 1:4)
{
  for(j in 1:4)
  {
    p.values.list = c(p.values.list, test.deviance(regions[[i]],regions[[j]]))
  }
}

min(p.values.list)<=0.05
min(p.values.list)

###########################
### Kruskal-Wallis test ###
###########################
kruskal.test(rate_vec, division_vec) # kruskal.test(rate_vec˜divison_vec) gives the same result
# Kruskal-Wallis rank sum test
#
# data:  rate_vec.2 and division_vec.2
# Kruskal-Wallis chi-squared = 17.589, df = 3, p-value = 0.0005346

##################################################################################
# (C)

deathdata<-read.csv("homicide.csv", header=T)
attach(deathdata)

par(mfrow = c(1,2))
hist(Homicide)
hist(Firearm)

par(mfrow = c(1,2))
qqnorm(Homicide, main="Homicide")
qqline(Homicide)
qqnorm(Firearm, main = "Firearm")
qqline(Firearm)

summary(Homicide)
summary(Firearm)

sd(Homicide, na.rm=T)
sd(Firearm, na.rm=T)

par(mfrow = c(1,2))
qqnorm(high)
qqline(high)

qqnorm(low)
qqline(low)

par(mfrow = c(1,1))
plot(Firearm,Homicide, main="Homicide vs Firearm")

##### Spearman test

(Firearm = rank(Firearm))
(Homicide = rank(Homicide))

## Spearman correlation
(rs.obs = cor(Firearm, Homicide))

## permutation test for the Spearman correlation
perm.approx.r <- function(x,y,R)
{
  ## approximate permutation distribution of sample correlation coefficient r
  n <- length(x)
  results <- rep(NA,R)
  for (i in 1:R) results[i] <- cor(x,y[sample(1:n,n)])
  results
}
permr <- perm.approx.r(Firearm, Homicide, 1000)
mean(permr >= rs.obs)
mean(abs(permr) >= abs(rs.obs))

##################################################################################
# (D)
set.seed(6)
data.inf = read.csv("inf.csv");data.inf

# Check to see if the histograms are normal.
inf = data.inf[2:52,];inf
par(mfrow = c(2,1))
hist(inf$Number, main = "Number of Deaths", xlab = "Deaths")
hist(inf$Rate, main = "Rate of Deaths", xlab = "Rate")
qqnorm(inf$Number, main = "Normal Q-Q Plot of Numbers")
```

```
qqline(inf$Number)
qqnorm(inf$Rate, main = "Normal Q-Q Plot of Rates")
qqline(inf$Rate)
summary(inf)

# Take a random sample.
sam = sample(1:51, 25)
inf.sample.num = inf$Number[sam]
inf.sample.rate = inf$Rate[sam]

# Find CI for Number of Deaths
x.bar = mean(inf.sample.num); s = sd(inf.sample.num)
x.bar - 2*s/sqrt(25); x.bar + 2*s/sqrt(25)
# Find CI for Rate of Deaths
x.bar = mean(inf.sample.rate); s = sd(inf.sample.rate)
x.bar - 2*s/sqrt(25); x.bar + 2*s/sqrt(25)

# Perform a boostrap on both Number and Rates.
bootmeans.num = rep(NA, 1000)
for (i in 1:1000) {
  bootsample = sample(inf.sample.num, 25, replace=T)
  bootmeans.num[i] = mean(bootsample)
}
hist(bootmeans.num, main = "Histogram of Bootstrap Number",
     xlab = "Boostrap Numbers")

bootmeans.rate = rep(NA, 1000)
for (i in 1:1000) {
  bootsample = sample(inf.sample.rate, 25, replace=T)
  bootmeans.rate[i] = mean(bootsample)
}
hist(bootmeans.rate, main = "Histogram of Bootstrap Rates",
     xlab = "Bootstrap Rates")

# Calculate means and sd for both bootstrap samples.
num.mean = mean(bootmeans.num);num.sd = sd(bootmeans.num)
rate.mean = mean(bootmeans.rate); rate.sd = sd(bootmeans.rate)

# CI for Bootstrap Number of Deaths
num.mean-2*num.sd;num.mean+2*num.sd
# CI for Bootstrap Rate of Deaths
rate.mean-2*rate.sd;rate.mean+2*rate.sd

# Actual mean for number.
mean(inf$Number)
# Actual mean for rates.
mean(inf$Rate)

# CI using the "Standard" Method
# Number: (785.56, 1886.68)
# Rate  : ( 16.19,   20.27)

# CI using the Bootstrap Method
# Number: (802.90, 1860.49)
# Rate  : ( 16.14,   20.35)

# Actual means.
# Number: 1117.26
#   Rate:   18.53

# Repeat both tests to plot multiple CI’s.

# Begin with the "standard" method.

ci.lb.num = c(NULL, 100)
ci.ub.num = c(NULL, 100)

ci.lb.rate = c(NULL, 100)
ci.ub.rate = c(NULL, 100)

ci.lb.num.b = c(NULL, 100)
ci.ub.num.b = c(NULL, 100)

ci.lb.rate.b = c(NULL, 100)
ci.ub.rate.b = c(NULL, 100)

for (j in 1:100){
  sam = sample(1:51, 25)
  inf.sample.num = inf$Number[sam]
  inf.sample.rate = inf$Rate[sam]

  x.bar = mean(inf.sample.num); s = sd(inf.sample.num)
  lb.num = x.bar - 2*s/sqrt(25)
  ub.num = x.bar + 2*s/sqrt(25)

  ci.lb.num[j] = lb.num
  ci.ub.num[j] = ub.num

  x.bar = mean(inf.sample.rate); s = sd(inf.sample.rate)
  lb.rate = x.bar - 2*s/sqrt(25)
  ub.rate = x.bar + 2*s/sqrt(25)

  ci.lb.rate[j] = lb.rate
  ci.ub.rate[j] = ub.rate

  bootmeans.num = rep(NA, 1000)
  for (i in 1:1000) {
    bootsample = sample(inf.sample.num, 25, replace=T)
    bootmeans.num[i] = mean(bootsample)
}

  bootmeans.rate = rep(NA, 1000)
  for (i in 1:1000) {
    bootsample = sample(inf.sample.rate, 25, replace=T)
    bootmeans.rate[i] = mean(bootsample)
}

  num.mean = mean(bootmeans.num);num.sd = sd(bootmeans.num)
  rate.mean = mean(bootmeans.rate); rate.sd = sd(bootmeans.rate)

  ci.lb.num.b[j] = num.mean-2*num.sd
  ci.ub.num.b[j] = num.mean+2*num.sd
  ci.lb.rate.b[j] = rate.mean-2*rate.sd
  ci.ub.rate.b[j] = rate.mean+2*rate.sd
}

ymat.num = rbind(ci.lb.num,ci.ub.num)
xmat.num = rbind(1:100,1:100)
ymat.rate = rbind(ci.lb.rate,ci.ub.rate)
xmat.rate = rbind(1:100,1:100)
ymat.num.b = rbind(ci.lb.num.b,ci.ub.num.b)
xmat.num.b = rbind(1:100,1:100)
ymat.rate.b = rbind(ci.lb.rate.b,ci.ub.rate.b)
xmat.rate.b = rbind(1:100,1:100)

par(mfrow = c(2,2))
matplot(xmat.num, ymat.num, type = "l", lty = 1,
        xlab = "Simulation Number", ylab = "Number")
abline(h = 1117.26)
title(main = "100 CI’s of Number")

matplot(xmat.rate, ymat.rate, type = "l", lty = 1,
        xlab = "Simulation Number", ylab = "Rate")
abline(h = 18.53)
title(main = "100 CI’s of Rate")

matplot(xmat.num.b, ymat.num.b, type = "l", lty = 1,
        xlab = "Simulation Number", ylab = "Number")
abline(h = 1117.26)
title(main = "100 CI’s of Number (Bootstrap)")

matplot(xmat.rate.b, ymat.rate.b, type = "l", lty = 1,
        xlab = "Simulation Number", ylab = "Rate")
abline(h = 18.53)
title(main = "100 CI’s of Rate (Bootstrap)")

# See the difference in CI between the two tests.

par(mfrow = c(2,2))
plot(ci.lb.num-ci.lb.num.b, ylab = "Difference in LB");abline(h=0)
title("LB Number")
plot(ci.ub.num-ci.ub.num.b, ylab = "Difference in UB");abline(h=0)
title("UB Number")
plot(ci.lb.rate-ci.lb.rate.b, ylab = "Difference in LB");abline(h=0)
title("LB Rate")
plot(ci.ub.rate-ci.ub.rate.b, ylab = "Difference in UB");abline(h=0)
title("UB Rate")
```