**Use R to Analyze the Data related to Cause of Death in the United States in 2012**

**Statistical Computing- Project 1**

**Objective:** The purpose of this project is to examine the many different causes of death in the United States during the year 2012. Note that the District of Columbia is treated as a separate state. Therefore, we have fifty-one states total. There were thirteen variables used in this data set, namely, the number of heart attacks (HEART), strokes (STROKE), respiratory infections (RESPIR), “freak” accidents (ACCID), automobile accidents (VEHICLE), deaths by cancer (CANCER), diabetes (DIABETES), Alzheimer’s (ALZHEIM), influlenza (FLU), suicides (SUICIDE), homicides (HOMICIDE), AIDS (AIDS), and the total number of deaths (TOTAL).

> data<-read.table("D:/STAT/DeathData.csv",header=TRUE,sep=",")

> x<-data$HEART > y<-data$CANCER

> z<-data$STROKE > u<-data$RESPIR

> v<-data$ACCID > w<-data$VEHICLE

> a<-data$DIABETES > b<-data$ALZHEIM

> c<-data$FLU > d<-data$SUICIDE

> e<-data$HOMICIDE > f<-data$AIDS

> g<-data$TOTAL

We started analyzing this data by creating several side-by-side boxplots comparing the number of deaths (total in the U.S.) for similar reasons. This first boxplot compares the number of deaths by heart attacks by the number of deaths by strokes and respiratory related infections. We note that heart attacks are the cause of more deaths than strokes and respiratory infections by a significant amount.

> boxplot(x,z,u,col=c(4,2,3),main="Sudden Deaths for Health Reasons")

> text(1,75, "Heart Attacks")

> text(2,100, "Strokes")

> text(3,100,"Respiratory")



This next boxplot compares the number of deaths by cancer, diabetes, Alzheimer’s disease, Influenza, and AIDS. As we can see, cancer is the leading cause of death between these five diseases.

> boxplot(y,a,b,c,f,col=c(2,3,4,5,6),main="Deaths by Diseases")> text(1,125,"Cancer")

> text(2,60,"Diabetes")

> text(3,65,"Alzheimer's")

> text(4,45,"Influenza")

> text(5,50,"AIDS")



This next boxplot compares the number of deaths by “freak” accidents and the number of deaths by accidents relating to a vehicle. As we can see, there are much more deaths caused by “freak” accidents and vehicle accidents. However, the definition of a “freak” accident is quite subjective. Once might assume that a “freak” accident could be any accident that wasn’t caused by a vehicle of some sort.

> boxplot(v,w,col=c(2,3),main="Deaths by Accidents")

> text(1,15,"Freak Accidents")

> text(2,35,"Vehicle Accidents")



This last boxplot compares the number of deaths by suicides and the number of deaths by homicides. As we can see, there are many more deaths by homicides than suicides.

> boxplot(d,e,col=c(3,4),main="Deaths by Killing")

> text(1,21,"Suicides")

> text(2,28,"Homicides")



These next five graphs each represent ten states and their respective number of deaths. Note that the fifth graph has eleven states (the first one contains the District of Columbia).

> count<-c(1018.1,465.5,752.1,1000.1,647.8,615.2,836.7,860.2,985.3,971.5)

> names(count)<-c("AL","AK","AZ","AR","CA","CO","CT","DE","DC","FL")

> barplot(count,col=c("1","2","3","4","5","6","7","8","pink","orange"),main="Total Deaths in Alabama-Florida")



> count<-c(745.4,715.1,719.7,807.6,869.1,910.4,870.7,932.1,934.8,944.6)

> names(count)<-c("GA","HI","ID","IL","IN","IA","KS","KY","LA","ME")

> barplot(count,col=c("1","2","3","4","5","6","7","8","pink","orange"),main="Total Deaths in Georgia-Maine")



> count<-c(777.8,849.5,842.2,726,960.1,937.5,873.3,838.9,767.9,778.1)

> names(count)<-c("MD","MA","MI","MN","MS","MO","MT","NE","NV","NH")

> barplot(count,col=c("1","2","3","4","5","6","7","8","pink","orange"),main="Total Deaths in Maryland-New Hampshire")



> count<-c(820.5,751.2,794.1,847.5,882.9,927.5,978.6,843.3,1028.8,904)

> names(count)<-c("NJ","NM","NY","NC","ND","OH","OK","OR","PA","RI")

> barplot(count,col=c("1","2","3","4","5","6","7","8","pink","orange"),main="Total Deaths in New Jersey-Rhode Island")



> count<-c(887.9,886.4,946.1,679.7,558,803.8,758.1,721.7,1145.4,827.7,780.8)

> names(count)<-c("SC","SD","TN","TX","UT","VT","VA","WA","WV","WI","WY")

>barplot(count,col=c("1","2","3","4","5","6","7","8","pink","orange","white"),main="Total Deaths in South Carolina-Wyoming")



After examining these bar graphs, we see that the state with the largest number of totals deaths was West Virginia. Now, let’s examine the average number of total deaths.

> T<-data$TOTAL

> T

[1] 1,018.10 465.5 752.1 1,000.10 647.8 615.2 836.7 860.2

[9] 985.3 971.5 745.4 715.1 719.7 807.6 869.1 910.4

[17] 870.7 932.1 934.8 944.6 777.8 849.5 842.2 726

[25] 960.1 937.5 873.3 838.9 767.9 778.1 820.5 751.2

[33] 794.1 847.5 882.9 927.5 978.6 843.3 1,028.80 904

[41] 887.9 886.4 946.1 679.7 558 803.8 758.1 721.7

[49] 1,145.40 827.7 780.8

First, we will determine if the data set T, or the total number of deaths in all fifty-one states is normal.

> shapiro.test(T)

Shapiro-Wilk normality test

data: T

W = 0.9006, p-value = 0.2926

Since our p-value is greater than 0.05, we fail to reject the null hypothesis of normality. Therefore, our data (T) is normal. Now, let’s test the hypothesis at a 95% confidence level that the average number of total deaths in the United States in 2012 was 900 (per state).

> t.test(T,alternative="greater",mu=900)

One Sample t-test

data: T

t = -3.9431, df = 43, p-value = 0.9999

alternative hypothesis: true mean is greater than 900

95 percent confidence interval:

792.1919 Inf

sample estimates:

mean of x

824.4159

We see that our p-value is greater than our desired level of alpha (0.05). Therefore, we fail to reject the null hypothesis, and conclude that the mean number of total deaths is less than or equal to 900 in each state. Furthermore, we examine this histogram of total deaths:

> hist(T, col="pink",xlab="number of deaths",main="Histogram of Total Number of Deaths")



Now, we will test the hypothesis that the average number of deaths by heart attacks is larger than the average number of deaths from cancer. We will use a 95% confidence level. But first, we will test whether or not the data for the number of deaths by heart attacks (x) and that data for the number of deaths by cancer (y) are normal.

> shapiro.test(x)

Shapiro-Wilk normality test

data: x

W = 0.9816, p-value = 0.6089

> y<-data$RESPIR

> shapiro.test(y)

Shapiro-Wilk normality test

data: y

W = 0.9869, p-value = 0.8405

Note that both of their respective p-values are significantly greater than 0.05. Therefore, we fail to reject the null hypothesis of normality.

> t.test(x,y,alternative="greater")

Welch Two Sample t-test

data: x and y

t = 3.6055, df = 85.988, p-value = 0.0002608

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

15.02809 Inf

sample estimates:

mean of x mean of y

220.8745 192.9843

We can see that our p-value is less than our desired . Therefore, we reject the null hypothesis. In other words, we conclude that on average, there are more deaths by heart attacks than cancer.

Next, we will test whether or not the average number of deaths from car accidents is larger than the number of deaths by suicide. Again, we will use a 95% confidence level. However, we will first test whether or not the data for the number of car accidents (w) and the data for the number of suicides (d) are normal.

> shapiro.test(d)

Shapiro-Wilk normality test

data: d

W = 0.9693, p-value = 0.2073

> shapiro.test(w)

Shapiro-Wilk normality test

data: w

W = 0.9577, p-value = 0.06647

Since the p-values for both w and d are greater than 0.05, we conclude that both of these data sets are normal.

> t.test(w,d,alternative="greater")

Welch Two Sample t-test

data: w and d

t = 4.7075, df = 83.201, p-value = 4.948e-06

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

2.916273 Inf

mean of x mean of y

16.95098 12.44118

Since our p-value is much less than , we will reject the null hypothesis. Therefore, we conclude that there are more deaths by car accidents on average than by suicides.

Therefore, after analyzing this data set, we conclude that one of the leading causes of death in the United States in the year 2012 was heart attacks. Furthermore, the state with the largest total number of deaths was West Virginia.