STAT S 670 - Exploratory Data Analysis - Homework #2

Ganesh Nagarajan

September 15, 2015

1. A) Let X1,X2,…. Xn are i.i.d random variables with mean μ and variance σ2. Hence the sample mean can be defined as follows,

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

When n →∞, the difference between $\overline{X_n}$ and the real mean μ decreases and this distribution will be near 0. [1] ie Var($\overline{X_n}$) = σ2/n → 0 as n → ∞, hence by the L.L.N, this distribution is normal and by the central limit theorem as n → ∞, the $\overline{X_n}$ collapses around the real mean μ and the asymptotic distribution can be given by N(0,σ^2)

B) From the lecture Notes, $\sqrt{n}(\tilde{X} - X_{0.5}) \xrightarrow{d} N(0, \frac{1}{4\{f\{\mu\}\}^2})$ ---(1)

Substituting this to $f\left(\frac{1}{\lambda}\right) = \lambda * e^{-\lambda * \frac{\ln(2)}{\lambda}} = \frac{\lambda}{2}$ ---(2)

Substituting it back to the (1),
The distribution is normal N(0,σ^2)

C) From Wikipedia, the ratio of mean and median can be given as

$\frac{mean}{median} = \frac{1}{\ln 2}$ ie $median = \ln 2 * mean.$

$Variance(\overline{X_n}) = \frac{\left(\frac{1}{\lambda}\right)^2}{n}$,

Hence by property, $Variance(T_2) = \frac{\left(\frac{1}{\lambda}\right)^2}{(\ln 2)^2 * n} = \frac{\sigma^2}{(\ln 2)^2\, n}$

D) Hence ARE(T1,T2) = $\frac{Variance(T_1)}{Variance(T_2)} = \frac{(\ln 2)^2 * Variance(\overline{X_n})}{Variance(\overline{X_n})} = (\ln 2)^2 = 0.480453$,
≈ 48.04% More efficient.

E) Statistic which is based on Median is more robust than the statistic based on mean. Moreover in this case, the ARE supports using T2 statistic for estimates.

F) Letter value plots for the generated random numbers are as follows,

```
source("lvalprogs.r")
x<- rexp(1000,1)
lval(x)
```

```
  Depth  Lower  Upper    Mid Spread pseudo-s
M 500.5 0.7177 0.7177 0.7177 0.0000   0.0000
F 250.5 0.2733 1.3920 0.8326 1.1187   0.8293
E 125.5 0.1299 1.9394 1.0346 1.8095   0.7865
D  63.0 0.0561 2.7505 1.4033 2.6944   0.8782
C  32.0 0.0309 3.4653 1.7481 3.4344   0.9219
B  16.5 0.0187 4.0796 2.0491 4.0609   0.9427
A   8.5 0.0095 5.2219 2.6157 5.2124   1.0780
Z   4.5 0.0038 6.4907 3.2472 6.4869   1.2193
Y   2.5 0.0031 6.7232 3.3631 6.7201   1.1644
X   1.5 0.0015 7.0963 3.5489 7.0948   1.1453
W   1.0 0.0001 7.3037 3.6519 7.3036   1.1076
```

As seen in the letter value plot, the mid value increases, hence the distribution is skewed right.

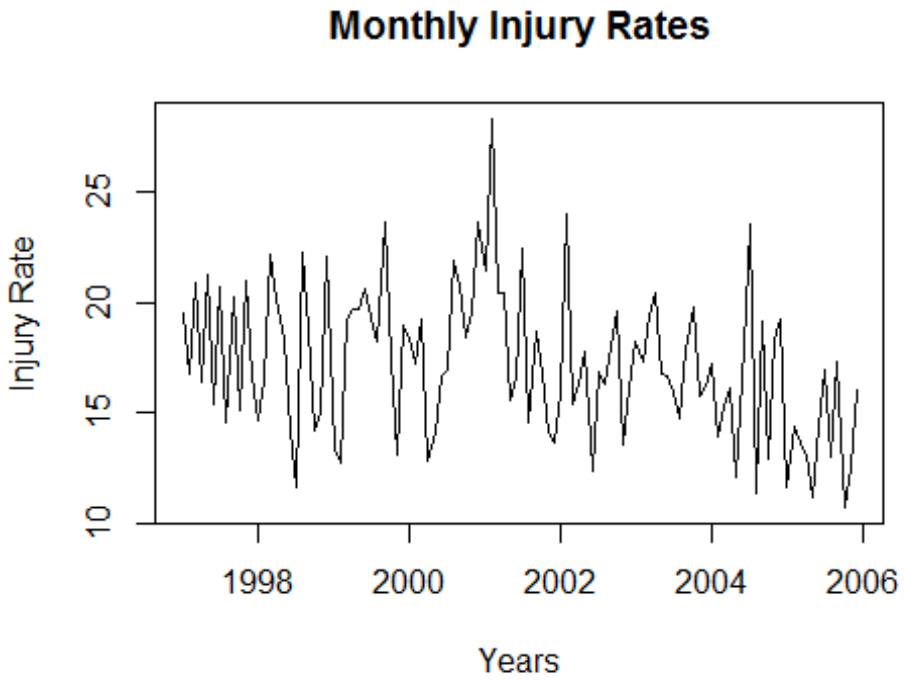
2. Following is the R code for analyzing the Colorado Department of Transport Data.

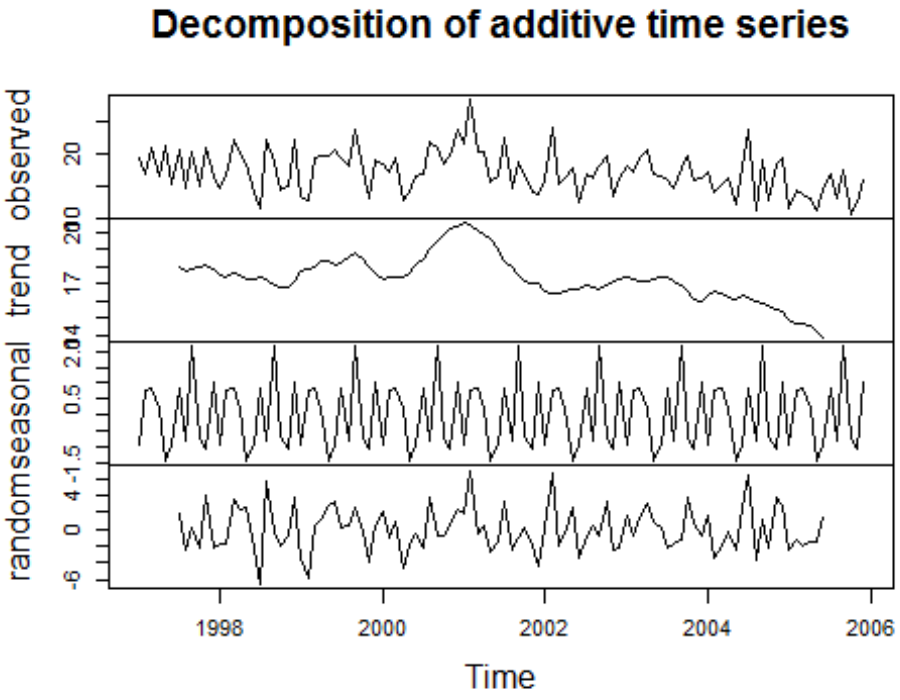
```
library(aplpack)

## Warning: package 'aplpack' was built under R version 3.2.2

## Loading required package: tcltk

library(outliers)

## Warning: package 'outliers' was built under R version 3.2.2

A<-c(19.50,16.72,20.92,16.42,21.22,15.40,20.68,14.55,20.23,
     15.11,20.95,16.68,14.67,16.50,22.15,20.14,18.33,14.20,
     11.61,22.24,18.75,14.22,15.03,22.07,13.34,12.73,19.23,
     19.74,19.74,20.60,19.29,18.22,23.65,17.44,13.07,19.00,
     18.44,17.25,19.19,12.77,14.10,16.69,16.92,21.92,20.84,
     18.43,19.54,23.61,21.40,28.34,20.43,20.43,15.58,16.58,
     22.44,14.59,18.70,16.79,14.12,13.67,15.94,24.04,15.42,
     16.26,17.74,12.37,16.87,16.28,17.97,19.56,13.56,16.13,
     18.20,17.29,19.38,20.47,16.75,16.69,15.93,14.73,17.83,
     19.78,15.78,16.17,17.18,13.90,15.33,16.10,12.03,17.92,
     23.56,11.35,19.10,12.91,18.32,19.24,11.57,14.33,13.60,
     13.12,11.19,14.33,16.91,13.03,17.32,10.70,12.56,16.04)
B<- ts(A,frequency = 12,start=c(1997,1))
plot(B,main="Monthly Injury Rates",xlab="Years",ylab="Injury Rate")

tsDecompose <- decompose(B)
plot(tsDecompose)

hist(B,main = "Histogram of Injury Rates",xlab="Injury Rates",freq = F)
lines(density(B,kernel = "epanechnikov"),col=2,lty=2)
legend("topright",legend = c("Distribution Density"),
  col = c(2),lty = 2,cex = 0.5)

qqnorm(B,main="QQ Plot for DoT Dataset")
qqline(B,lty=2,col=2)
```

```
outlier(B)
```

```
## [1] 28.34
```

a) Time Series Plot for the Data Set is given above Since the dataframe is decomposed, t he decomposed trends plot is as follows,





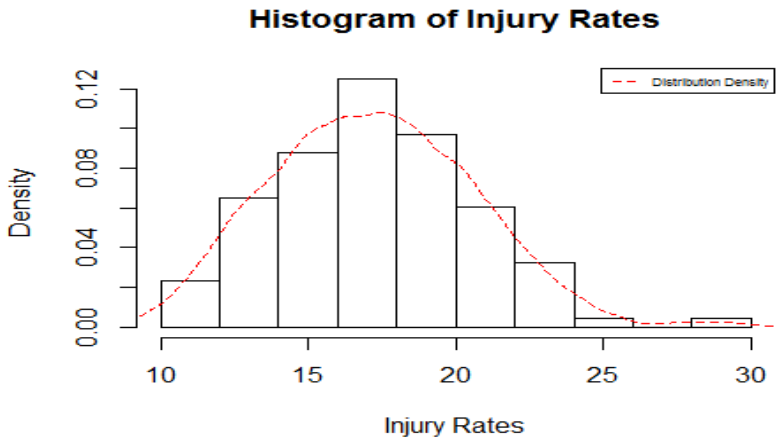
b) From the stem-leaf plot, since the stems below the median is greater than stems above the median, this suggests that mean is to the right of the median. Also the distribution seems to be skewed to right.

```
stem.leaf(B,rule.line = "Dixon")
```

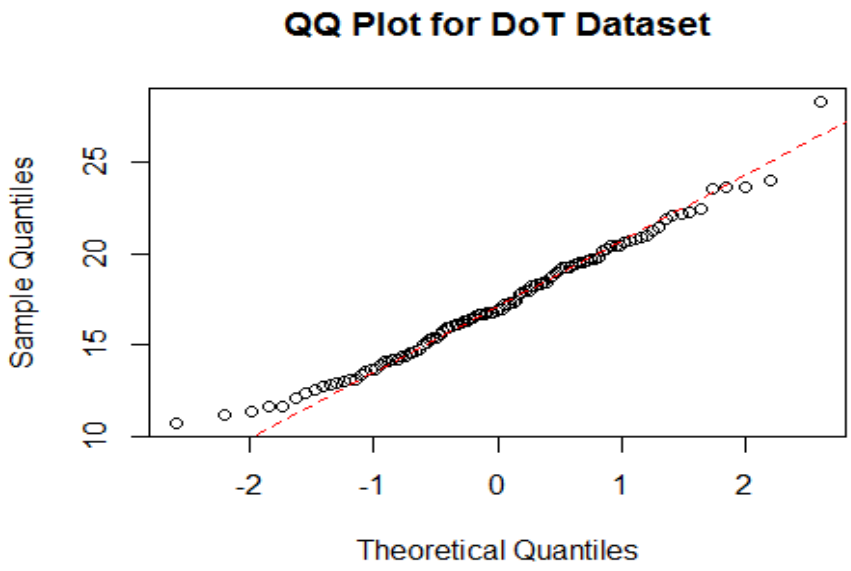
```
## 1 | 2: represents 1.2
##  leaf unit: 0.1
##             n: 108
##    1    10 | 7
##    5    11 | 1356
##   11    12 | 035779
##   19    13 | 00135669
##   29    14 | 1122335567
##   38    15 | 013445799
##   56    16 | 011122455666777899
##   (9)   17 | 122347899
##   43    18 | 22334477
##   35    19 | 0112223555777
##   22    20 | 1244466899
##   12    21 | 249
##    9    22 | 0124
##    5    23 | 566
##    2    24 | 0
## HI: 28.34
```

c) Letterbox plot of the dataset is as follows. It can be seen from distance between the median increases in F_U than F_L, hence clearly shows that the distribution is skewed right. This can also be verified using the histogram as shown below.

```
source("./lvalprogs.r")
lval(B)

##   Depth  Lower  Upper     Mid Spread pseudo-s
## M  54.5 16.890 16.890 16.8900  0.000   0.0000
## F  27.5 14.630 19.520 17.0750  4.890   3.6250
## E  14.0 13.120 20.920 17.0200  7.800   3.3903
## D   7.5 12.465 22.195 17.3300  9.730   3.1712
## C   4.0 11.570 23.610 17.5900 12.040   3.2318
## B   2.5 11.270 23.845 17.5575 12.575   2.9192
## A   1.5 10.945 26.190 18.5675 15.245   3.1530
## Z   1.0 10.700 28.340 19.5200 17.640   3.3157
```



d) QQ Plot From the QQ plot it can be seen that there is a light left tail and light right tail, the distribution seems to be almost normal, however there is an outlier at the top right corner.



e) Yes, There are Outliers, This can be verified using the histogram aswell as the QQ Plot. Quantitatively, outliers package can be used to identify the outliers,

```
outlier(B)
```

```
## [1] 28.34
```

3) Outliers based on Gaussian Theory

According to Tukey, the outliers of any single batch of observation can be estimated by the function, outliers = 0.4+0.07n.

Below is the R program finding the outliers,

```
calculateOutliers<-function (sampleSize){
  i = length(sampleSize);
  outlier=0;
  for ( i in seq(1:i)){
    outlier=outlier+0.4+(0.00698*sampleSize[i])
  }
  print(outlier)
}
sampleSize<-c(120)
calculateOutliers(sampleSize)
```

```
## [1] 1.2376
```

```
sampleSize<-c(60,60)
calculateOutliers(sampleSize)
```

```
## [1] 1.6376
```

```
sampleSize<-c(40,30,20,10,5,5,5,5)
calculateOutliers(sampleSize)

## [1] 3.6027
```

Hence the outliers are,

1) 1.2376 for n=120

2) 1.6376 for n=60,60

3) 3.6027 for n=40,30,20,10,5,5,5,5

Another interpretation can be made from the theory of normal distribution.

BY 68-95-97-9.7 rule, considering a standard deviation of ±3σ covers 99.7 of the population, there is a possibility that all 1-99.7 is outlier probability.

1-99.7 (Outlier Probability)	Data set length	Probability being outlier * no of observations	Outlier probability for each point in the observation
0.0025%	100	0.25%	0.25%
0.0025%	60	0.15%	0.30%
0.0025%	60	0.15%	
0.0025%	30	0.075%	0.2625%
0.0025%	40	0.1%	
0.0025%	20	0.05%	
0.0025%	10	0.025%	
0.0025%	5	0.0125$	

4) R code for the voters dataset analysis is as follows,

```
library(cluster)
library(reshape2)
library(ggplot2)
plotAggregate <- function(statesList,df1,xlabc,ylabc,mainc,colc){
  dataFrame1<-df1[which(df1$StateName %in% statesList),2:3]
  dataFrame1<-aggregate(dataFrame1,by=list("yearTrend"=dataFrame1$variable),m
ean)
  plot(dataFrame1$yearTrend,dataFrame1$value,type="b",xlab=xlabc,ylab = ylabc
,main=mainc,col=colc)
  abline(h=50,lty=2)
}
plotqq <- function(statesList,df1,xlabc,ylabc,mainc,colc){
  dataFrame1<-df1[which(df1$StateName %in% statesList),2:3]
  #dataFrame1<-aggregate(dataFrame1,by=list("yearTrend"=dataFrame1$variable),
mean)
  qqnorm(dataFrame1$value,xlab=xlabc,ylab = ylabc,main=mainc,col=colc)
  qqline(dataFrame1$value)
```

```
  #boxplot(dataFrame1$value,xlab=xlabc,ylab = ylabc,main=mainc,col=colc)
}
data("votes.repub")
df<-as.data.frame(votes.repub[,26:30])
colnames(df)<-c(1956,1960,1964,1968,1972)
df["StateName"]<-rownames(df)
df1<-melt(df,id.vars = "StateName")
df1$StateName<-as.factor(df1$StateName)
df1$variable<-as.integer(as.character(df1$variable))
df1$value[is.na(df1$value)]<-0
NorthEast=c("Connecticut","Delaware","Maine","Massachusetts","New Hampshire",
"New Jersey","New York","Pennsylvania","Rhode Island","Vermont")
MaEc<-c("Kentucky","Maryland","North Carolina","South Carolina","Tennessee","
Virginia","West Virginia")
South<-c("Alabama","Arkansas","Florida","Georgia","Louisiana","Mississippi","
Oklahoma","Texas")
midwest<-c("Illinois","Indiana","Iowa","Kansas","Michigan","Minnesota","Misso
uri","Nebraska","Ohio","Wisconsin")
rockies<-c("Colorado","Idaho","Montana","North Dakota","South Dakota","Utah",
"Wyoming")
west<-c("Alaska","Arizona","California","Hawaii","Nevada","New Mexico","Orego
n","Washington")
```

```
plot(df1$variable[df1$StateName=="Alabama"],df1$value[df1$StateName=="Alabama"],main=
"Republican Votes Percentages",xlab = "Years",ylab="Percentages",xlim = c(1953,1974))
for (i in seq(1:length(StateNames))){
lines(df1$variable[df1$StateName==StateNames[i]],df1$value[df1$StateName==StateNames[
i]],col=i,type="b")
}
abline(h=50,lty=2)
#par(xpd=TRUE)
legend("bottomleft",legend=StateNames,lty=1,cex=0.5,col=1:10)
```

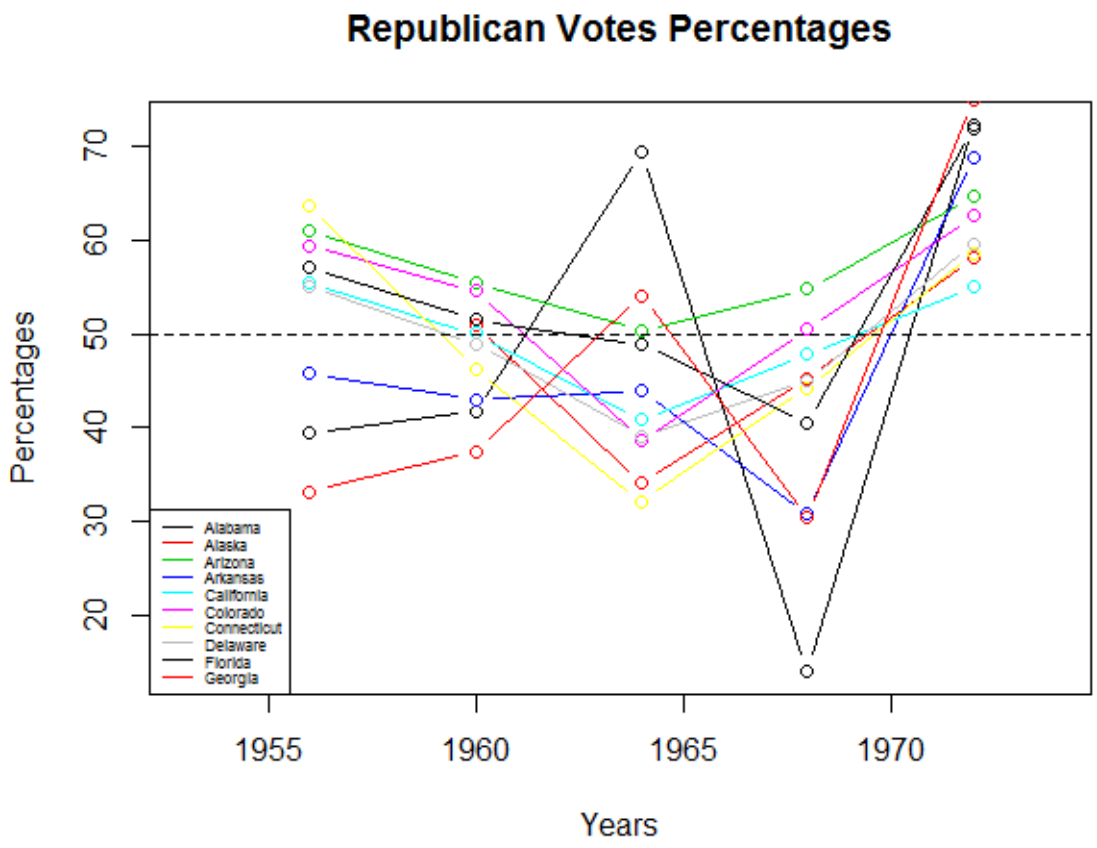
```
par(mfrow=c(2,3))
plotAggregate(NorthEast,df1,"Years","Votes Percentage","North East Region",2)
plotAggregate(MaEc,df1,"Years","Votes Percentage","Mid Atlantic/East Central
Region",3)
plotAggregate(South,df1,"Years","Votes Percentage","Southern Region",4)
plotAggregate(midwest,df1,"Years","Votes Percentage","Mid West Region",2)
plotAggregate(rockies,df1,"Years","Votes Percentage","Rockies Region",3)
plotAggregate(west,df1,"Years","Votes Percentage","West Region",4)
```

```
par(mfrow=c(2,3))
plotqq(NorthEast,df1,"Years","Votes Percentage","North East Region",2)
plotqq(MaEc,df1,"Years","Votes Percentage","Mid Atlantic/East Central Region"
,3)
plotqq(South,df1,"Years","Votes Percentage","Southern Region",4)
plotqq(midwest,df1,"Years","Votes Percentage","Mid West Region",2)
plotqq(rockies,df1,"Years","Votes Percentage","Rockies Region",3)
plotqq(west,df1,"Years","Votes Percentage","West Region",4)
```

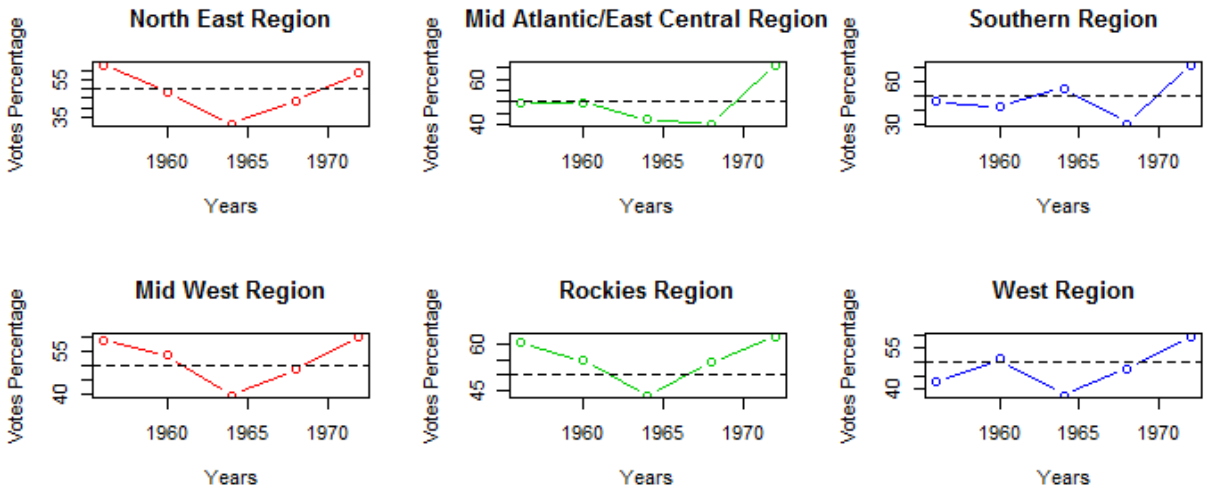
```
par(mfrow=c(1,1))
qqnorm(df1$value,main="QQ plot for the entire dataset")
qqline(df1$value)

qplot(variable,value,data=df1,facets = Region~.,color=StateName,geom="line")+
theme(legend.position="bottom")+guides(colour=guide_legend(nrow=5))+ggtitle("
Votes by Country by Region")

df1$Region[which(df1$StateName %in% NorthEast)]<-"North East"
df1$Region[which(df1$StateName %in% South)]<-"South"
df1$Region[which(df1$StateName %in% west)]<-"West"
df1$Region[which(df1$StateName %in% rockies)]<-"Rockies"
df1$Region[which(df1$StateName %in% midwest)]<-"Mid West"
df1$Region[which(df1$StateName %in% MaEc)]<-"Mid Atlantic / East Central"
boxplot(df1$value,main="Box plot of the Republican votes from 1956 to 1972")
boxplot(df1$value~df1$Region, main="Box plot of Republican votes by Region")
qplot(variable,value,data=df1,facets = Region~.,color=StateName,geom="line")
```

The above code was just a consolidated code for illustration. The code uses Melt function to melt the data frame from wide to long format, creates a column called region and checks if the states comes in the region using the which and in command. Average votes is plotted using the base plotting system and the individual votes of the states by region are plotted through the ggplot2 plotting system.

a) Following is the x-y plot for the republican votes percentage. From the plot, there seems to be an interesting pattern with Alabama and Georgia. While the republican vote percentages dropped from other states for year 1964, it has actually increased in Alabama and Georgia.



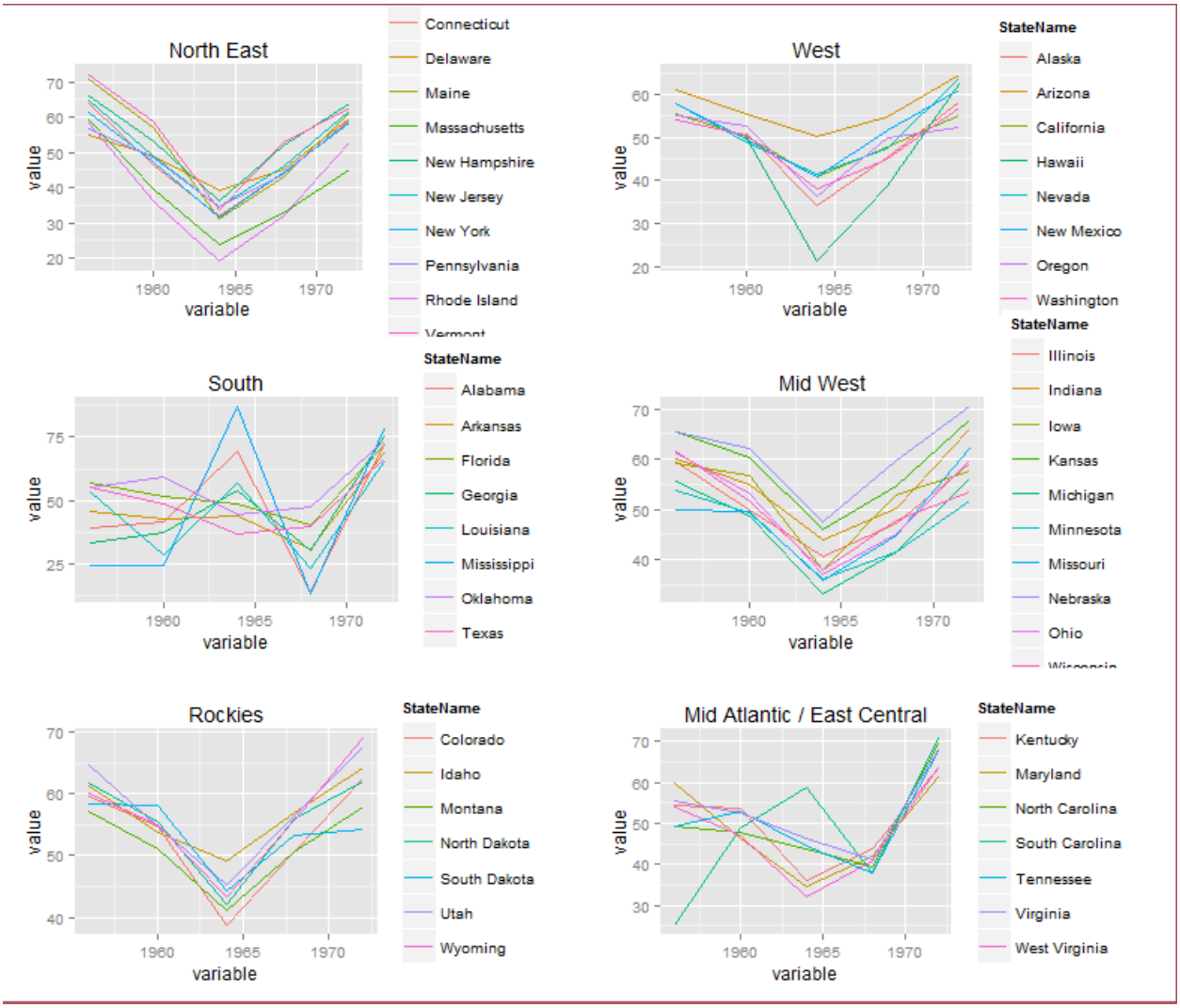
b) Following is the plot for aggregated vote percentages in all the regions for years 1962 through 1972. It can be seen that the party has majority votes after 1974 elections and during 1963 election, the party had minimal votes percentage in all the areas except the southern part of the country.



The votes breakup is shown by the below plot plotted using ggplot2 package. Multiplot function is credited to [2] Location for the code.

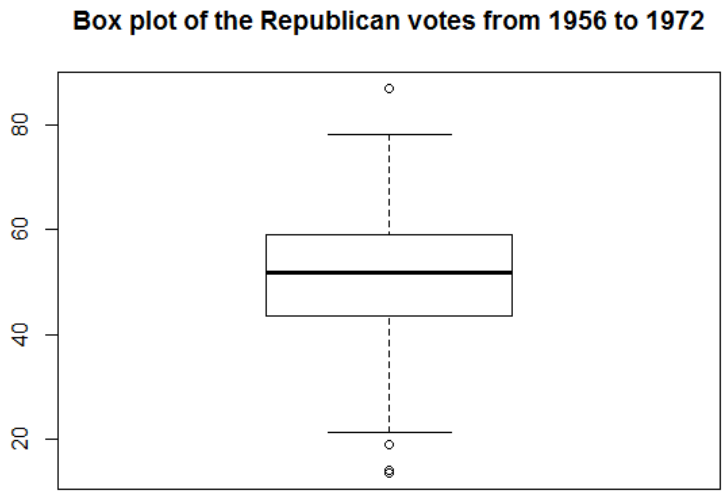
R Code :

```
a<-ggplot(subset(df1,Region == "North
East"),aes(x=variable,y=value,color=StateName))+geom_line()+ggtitle("North East")
b<-ggplot(subset(df1,Region ==
"South"),aes(x=variable,y=value,color=StateName))+geom_line()+ggtitle("South")
c<-ggplot(subset(df1,Region ==
"Rockies"),aes(x=variable,y=value,color=StateName))+geom_line()+ggtitle("Rockies")
d<-ggplot(subset(df1,Region ==
"West"),aes(x=variable,y=value,color=StateName))+geom_line()+ggtitle("West")
e<-ggplot(subset(df1,Region == "Mid
West"),aes(x=variable,y=value,color=StateName))+geom_line()+ggtitle("Mid West")
f<-ggplot(subset(df1,Region == "Mid Atlantic / East
Central"),aes(x=variable,y=value,color=StateName))+geom_line()+ggtitle("Mid Atlantic / East Central")
multiplot(a,b,c,d,e,f, cols=2)
```

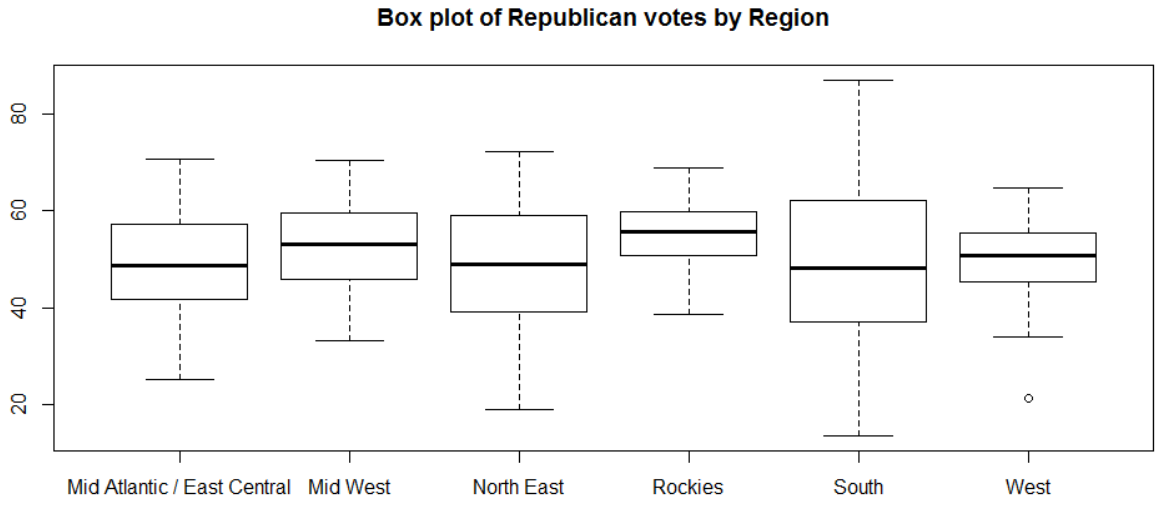


3) Box plots and QQ plots

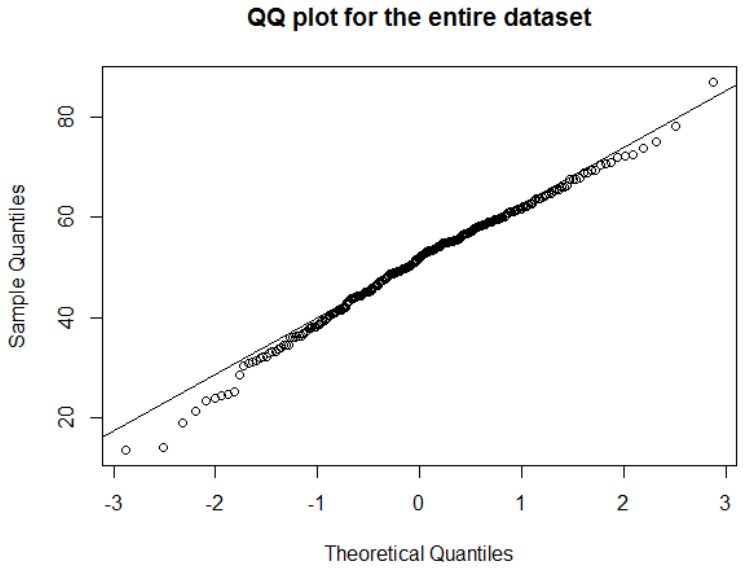
Box plot for the entire dataset,



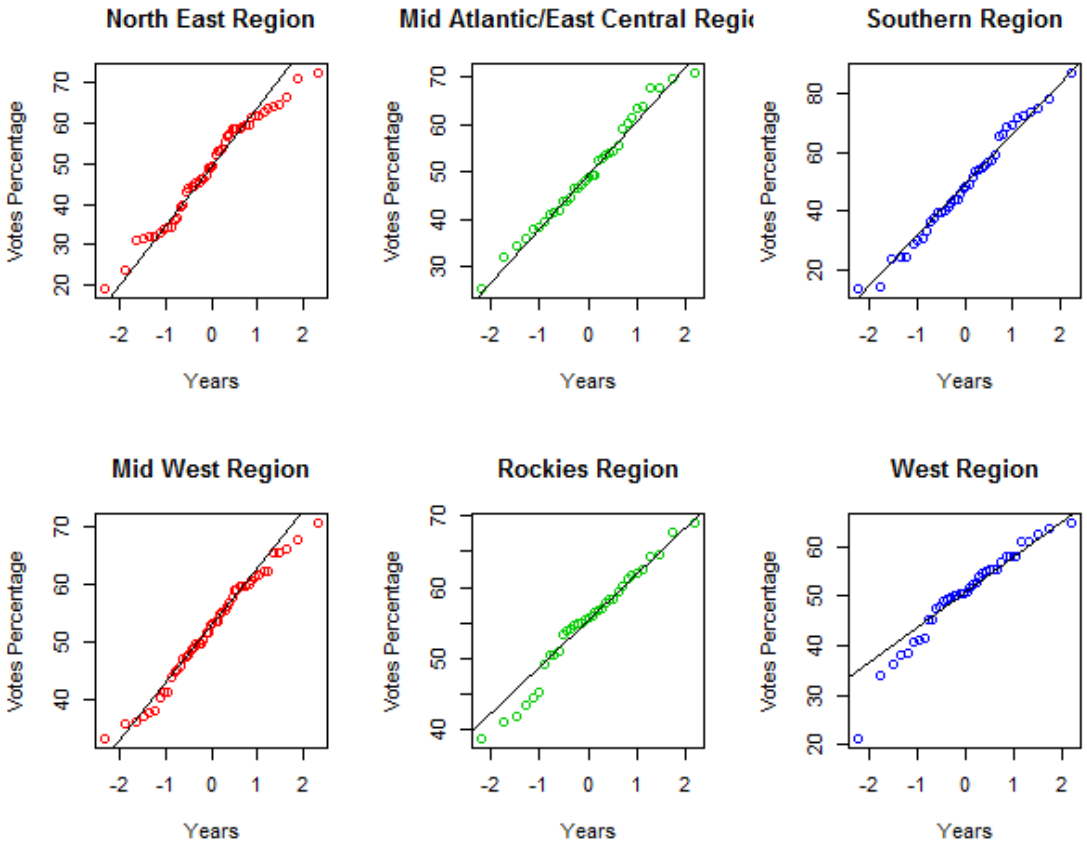
Box plot by the regions,



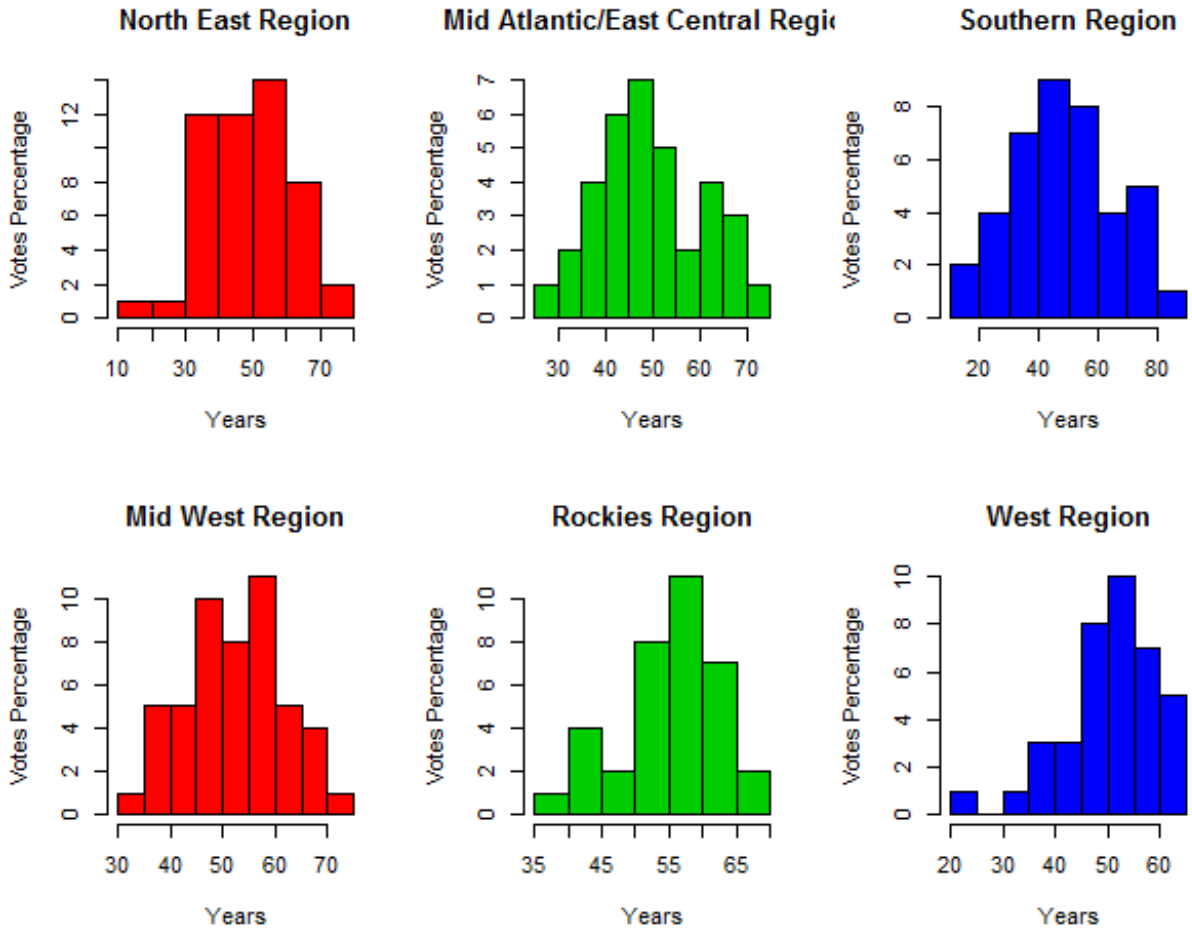
QQ plot of the entire dataset,



QQ plot for dataset by regions,



Below histogram explains the distribution visually,



References and Acknowledgements

[1]http://faculty.ksu.edu.sa/73125/Publications/%5B3%5D%20Convergence%20of%20Random%20Variables.pdf
[2] http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_%28ggplot2%29/ for the multiplot function
[3] http://eml.berkeley.edu/~powell/e240b_sp10/asynotes.pdf
[4] http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture3.pdf