Exploratory Data Analysis Assignment2

FNU Anirudh

September 16, 2015

Solution 1:-

1. E [Xi]= ?? and Var[Xi]= ??2

* \_ d ????? (X- ??)-------> N (0, ??^2)

Given ??2 = 1/??2 We can write \_ d ???n ( X- 1/??) -->??? N (0, 1/??2)

1. We know that the asymptotic distribution of the sample median is

* \_ d ???n ( X- X0.5) --???> N (0, 1/4 {f(X0.5) }2) f (X0.5 )= f (log2/ ??) = ?? e (-?? log 2/??) = ??/2 We have , 1/ 4 { {f(X0.5) }2 = 1 /4 (??2/4d)--->??? N (0,5) -

1. Var (T2)= Var (X/log2) \_ = (1/ln2)^2 Var (X) = 2.081 ??^2
2. ARE(T1,T2)= ??^2/n/2.081 ??^2/n = 0.481= 48.1%
3. Comparing the two statistics T1 and T2, T1 is the better because when we compare the variance of these two methods T1 has a lesser variance, so this would be the preferred statistic method of the given two

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

Depth Lower Upper Mid Spread pseudo-s M 500.5 0.6932 0.6932 0.6932 0.0000 0.0000 F 250.5 0.2929 1.3877 0.8403 1.0948 0.8116 E 125.5 0.1291 2.0037 1.0664 1.8746 0.8148 D 63.0 0.0617 2.6406 1.3511 2.5790 0.8405 C 32.0 0.0289 3.3680 1.6985 3.3391 0.8963 B 16.5 0.0149 4.3889 2.2019 4.3741 1.0154 A 8.5 0.0083 4.6239 2.3161 4.6157 0.9546 Z 4.5 0.0050 5.2925 2.6488 5.2874 0.9938 Y 2.5 0.0009 5.6520 2.8264 5.6511 0.9792 X 1.5 0.0006 5.8699 2.9352 5.8692 0.9475 W 1.0 0.0004 5.9646 2.9825 5.9642 0.9044

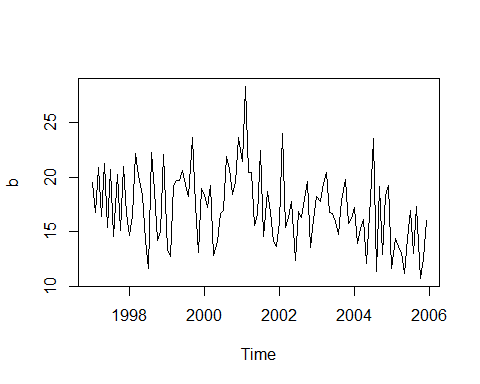
Data is skewed to the right.

Solution 2:-

library(aplpack)

## Loading required package: tcltk

# Letter Value  
lval <- function(x) {  
 #tag <- c("M ","F ","E ","D ","C ","B ","A ","Z ","Y ","X ","W ","V","U","T",  
 # "S","R","Q","P","O","N")  
 # gau <- abs(qnorm(c(.25,.125,1/16,1/32,1/64,1/128,1/256,1/512,1/1024,1/2048,  
 # 1/4096, 1/8192, 1/16384, 1/32768, 1/65536)))  
 tag <- c("M",LETTERS[6:1],LETTERS[26:14])  
   
 gau <- abs(qnorm(1/2^(2:20)))  
   
 # col 1 = depth; 2 = lower; 3 = upper; 4 = mid; 5 = spread; 6 = pseudo-s  
   
 y <- sort(x[!is.na(x)])  
 n <- length(y)  
 m <- ceiling(log(n)/log(2)) + 1  
 depth <- rep(0,m)  
 depth[1] <- (1 + n)/2  
   
 for (j in 2:m) {depth[j] <- (1 + floor(depth[j-1]))/2 }  
   
 ndepth <- n+1 - depth  
 out <- matrix(0, m, 6)  
 dimnames(out) <- list(tag[1:m],  
 c("Depth", "Lower","Upper", "Mid", "Spread","pseudo-s"))  
 out[1,2:3] <- median(y)  
 out[,1] <- depth  
   
   
 for (k in 2:m) {  
 out[k,2] <- ifelse(depth[k] - round(depth[k]) == 0,  
 y[depth[k]], (y[depth[k]-.5]+y[depth[k]+.5])/2 )  
 out[k,3] <- ifelse(ndepth[k] - round(ndepth[k]) == 0,  
 y[ndepth[k]], (y[ndepth[k]-.5]+y[ndepth[k]+.5])/2 )  
 }  
   
 out[1:m,4] <- (out[1:m,2] + out[1:m,3])/2  
 out[2:m,5] <- out[2:m,3] - out[2:m,2]  
 out[2:m,6] <- out[2:m,5]/(2\*gau[1:(m-1)])  
 round(out,4)  
}  
  
  
  
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, 17.79, 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)  
  
# Plot Time series  
b= ts(A, frequency = 12, start = c(1997, 1))  
plot(b)



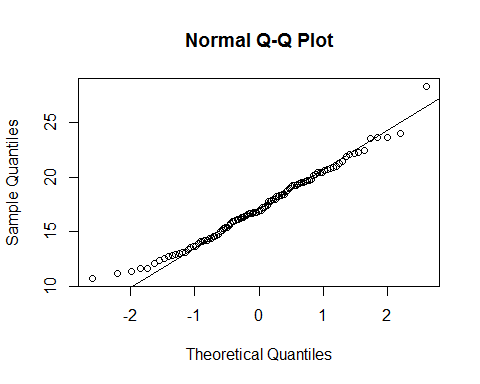
# Stem and Leaf  
stem.leaf(b)

## 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 | 123477899  
## 43 18 | 22334477  
## 35 19 | 0112223555777  
## 22 20 | 1244466899  
## 12 21 | 249  
## 9 22 | 0124  
## 5 23 | 566  
## 2 24 | 0  
## HI: 28.34

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

# QQ Plot  
qqnorm(b)  
qqline(b)



From stem and leaf we can see that data is skewed to the left. d) NotNormally distributed e) Yes, there is outlier.

Solution 3:-

1. Single Batch n=120

0.4 + 0.007\* n= 0.4+ 0.007 \* 120= 1.24

1. Two batches n=60

a= 0.4+0.007\*n= 0.4+0.42=0.82

Similarly,b= 0.82

Total outside values= a+b= 0.82+0.82= 1.64

1. n=40

a= 0.4+ 0.007\* 40= 0.68

n=30

b= 0.4+ 0.007\* 30= 0.61

n=20

c= 0.4+0.007\* 20= 0.54

n=10

c= 0.4+0.007\* 10= 0.47

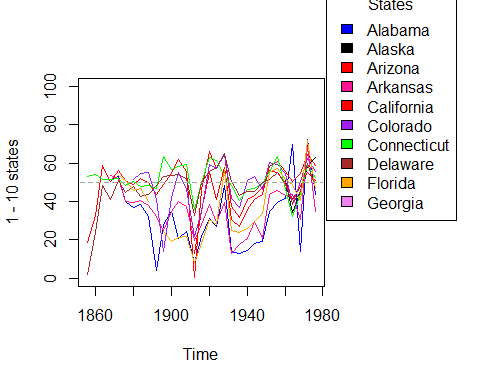
n=5

c= 0.4+ 0.007\*5= 0.435

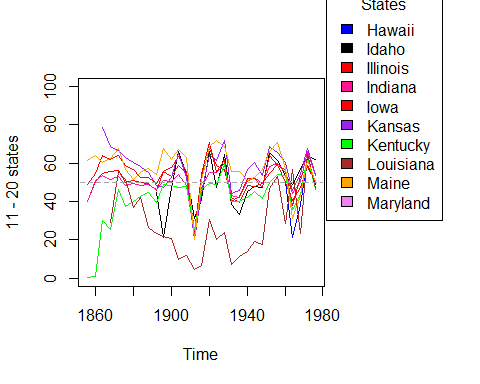
Total outside value= 0.68 + 0.61 + 0.54 + 0.47 + 4\* 0.435 = 4.04

Solution 4

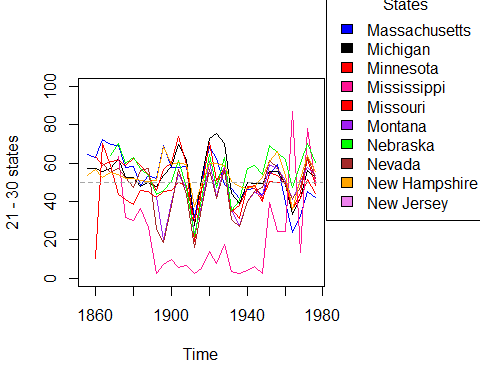
library(cluster)  
vote=votes.repub  
vote\_calc= function(vote,lower,upper)  
{  
 colours=c("blue","black","red","deeppink","red","purple","green","brown",  
 "orange","violet")  
 par(xpd=NA,oma=c(0,0,0,6))  
 for(i in lower:upper)  
 {  
 vote\_t=vote[i,]  
 vote\_t=as.data.frame(t(vote\_t))  
 vote\_ts=ts(vote\_t,frequency=0.25,start=1856)  
 if(i==lower)  
 {  
 plot(vote\_ts,col=colours[1],ylim=range(0,100),xlim=range(1856,1976),  
 ylab=paste(lower,"-",upper,"states"))  
 }  
 else  
 lines(vote\_ts,col=colours[i%%10])  
 }  
 legend(1982,150,legend=row.names(vote[lower:upper,]),  
 fill=colours,title="States")  
 segments(1852,50,1980,50,col="grey65",lty=2)  
}  
vote\_calc(vote,1,10)



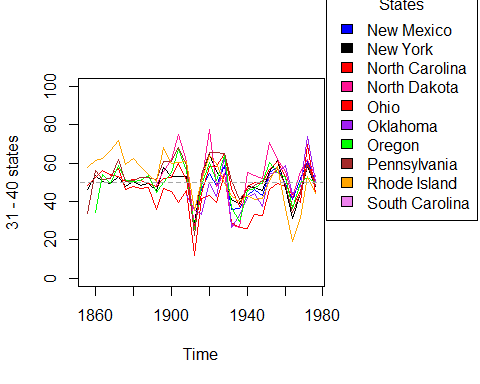
vote\_calc(vote,11,20)



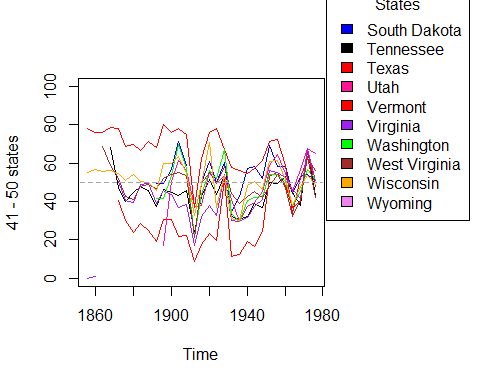
vote\_calc(vote,21,30)



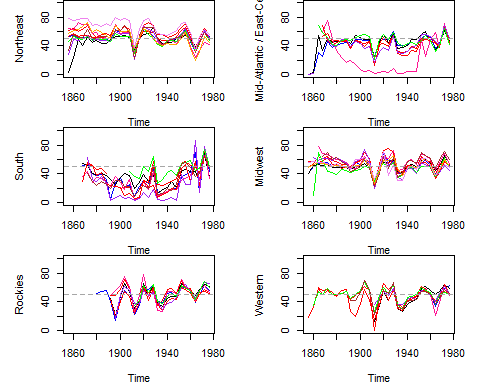
vote\_calc(vote,31,40)



vote\_calc(vote,41,50)



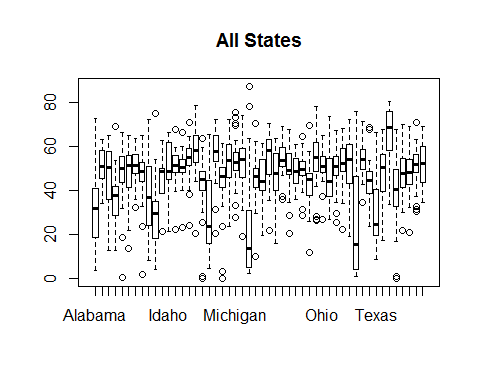
library(cluster)  
vote = votes.repub  
par(mfrow = c(3, 2))  
vote\_calc = function(vote, reg, name)  
{  
   
 colours=c("blue","black","red","deeppink","red","purple","green","brown",  
 "orange","violet")  
 par(mar=c(4,5,0,2))  
 j = 1  
 for(i in reg)  
 {  
 vote\_t = vote[i,]  
 vote\_t = as.data.frame(t(vote\_t))  
 vote\_ts = ts(vote\_t, frequency = 0.25, start = 1856)  
 if(i==reg[1])  
 {  
 plot(vote\_ts, col = colours[1], ylim = range(0,100),  
 ylab = name)  
 }  
 else  
 j = j + 1  
 lines(vote\_ts, col = colours[j])  
 }  
 segments(1852,50,1980,50, col = "grey65", lty = 2)  
}  
  
northeast = c("Connecticut","Delaware","Maine", "Massachusetts","New Hampshire","New Jersey",  
 "New York","Pennsylvania","Rhode Island","Vermont")  
east\_central = 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","Missouri","Nebraska",  
 "Ohio","Wisconsin")  
rockies = c("Colorado","Idaho","Montana","North Dakota","South Dakota","Utah","Wyoming")  
west = c("Alaska","Arizona","California","Hawaii","Nevada","New Mexico","Oregon","Washington")  
  
vote\_calc(vote, northeast, "Northeast")  
vote\_calc(vote, east\_central, "Mid-Atlantic / East-Central")  
vote\_calc(vote, south, "South")  
vote\_calc(vote, midwest, "Midwest")  
vote\_calc(vote, rockies, "Rockies")  
vote\_calc(vote, west, "Western")



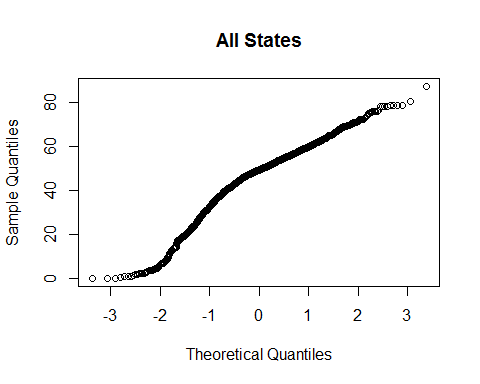
Republicans recieved less votes from South.

3).

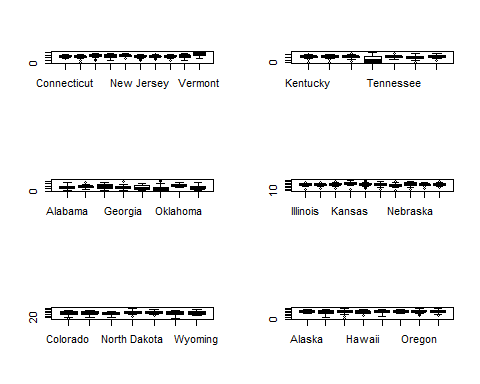
library(cluster)  
vote = votes.repub  
par(mfrow = c(3, 2))  
vote\_calc = function(vote, reg, name)  
{  
   
 colours=c("blue","black","red","deeppink","red","purple","green","brown",  
 "orange","violet")  
 par(mar=c(4,5,0,2))  
 j = 1  
 for(i in reg)  
 {  
 vote\_t = vote[i,]  
 vote\_t = as.data.frame(t(vote\_t))  
 vote\_ts = ts(vote\_t, frequency = 0.25, start = 1856)  
 if(i==reg[1])  
 {  
 plot(vote\_ts, col = colours[1], ylim = range(0,100),  
 ylab = name)  
 }  
 else  
 j = j + 1  
 lines(vote\_ts, col = colours[j])  
 }  
 segments(1852,50,1980,50, col = "grey65", lty = 2)  
}  
  
northeast = c("Connecticut","Delaware","Maine", "Massachusetts","New Hampshire","New Jersey",  
 "New York","Pennsylvania","Rhode Island","Vermont")  
east\_central = 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","Missouri","Nebraska",  
 "Ohio","Wisconsin")  
rockies = c("Colorado","Idaho","Montana","North Dakota","South Dakota","Utah","Wyoming")  
west = c("Alaska","Arizona","California","Hawaii","Nevada","New Mexico","Oregon","Washington")  
  
# Box Plot for All States  
par(mfrow=c(1,1))  
boxplot(t(votes.repub), main ="All States")



qqnorm(t(votes.repub), main ="All States")



par(mfrow=c(3,2))  
# Box Plot for Northwest  
boxplot(t(votes.repub[northeast,]))  
  
# Box Plot for East central  
boxplot(t(votes.repub[east\_central,]))  
  
#Box Plot for South  
boxplot(t(votes.repub[south,]))  
  
#Box Plot for Midwest  
boxplot(t(votes.repub[midwest,]))  
  
# Box Plot for Rockies  
boxplot(t(votes.repub[rockies,]))  
  
# Box Plot for West  
boxplot(t(votes.repub[west,]))



par(mfrow=c(3,2))  
qqnorm(t(votes.repub[northeast,]), main ="North East")  
qqnorm(t(votes.repub[east\_central,]), main = "East central")  
qqnorm(t(votes.repub[south,]), main="South")  
qqnorm(t(votes.repub[midwest,]), main="Midwest")  
qqnorm(t(votes.repub[rockies,]), main="Rockies")  
qqnorm(t(votes.repub[west,]), main="West")

