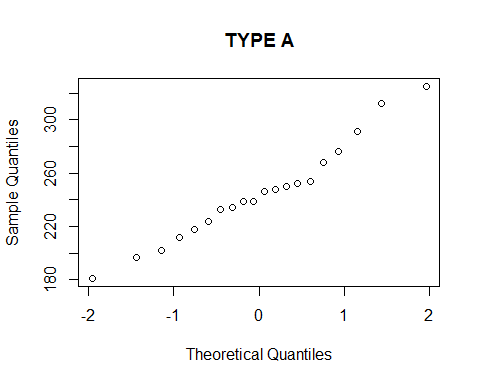
Assignment 9 (S-520)

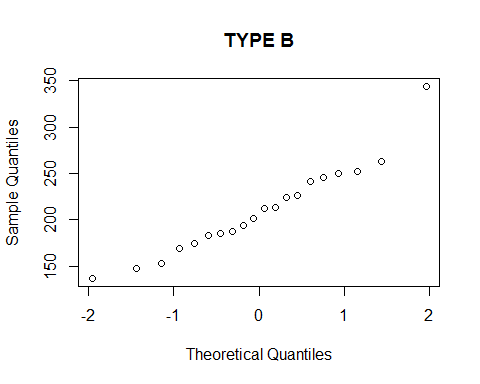
FNU Anirudh

November 5, 2015

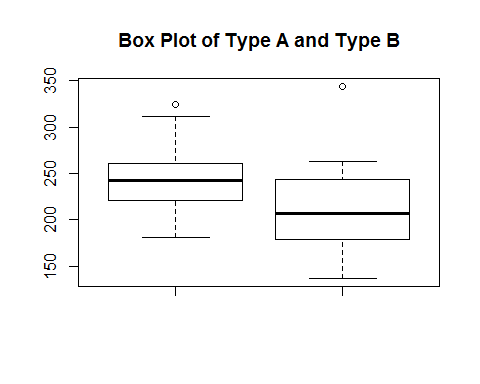
typea<- c(233,291,312,250,246,197,268,224,239,239,254,276,234,181,248,252,  
 202,218,212,325)  
typeb<- c(344,185,263,246,224,212,188,250,148,169,226,175,242,252,153,183,  
 137,202,194,213)  
qqnorm(typea,main = "TYPE A")



qqnorm(typeb,main = "TYPE B")



boxplot(typea,typeb,main="Box Plot of Type A and Type B")



# QQplot for both Type A and Type B suggests some values may be inconsistent  
# with normal distribution specially largest in each set as seen in boxplot.  
a=IQR(typea)/sqrt(var(typea))  
a

## [1] 0.9552842

b=IQR(typeb)/sqrt(var(typeb))  
b

## [1] 1.282584

# Ratio for Type B suggest sample more close to normal distribution but  
# also has large outlier hence I would not assume data was drawn from  
# normal distribution although there is slight chance of being picked up  
# from normal distribution.  
delta<- mean(typea)- mean(typeb)  
n1=length(typea)  
n2=length(typeb)  
va=var(typea)/n1  
vb=var(typeb)/n2  
se=sqrt(va+vb)  
nu<- (va+vb)^2/(va^2/(n1-1)+vb^2/(n2-1))  
# If we let alpha= 0.05 then  
1- pt(2.5621,nu)

## [1] 0.00740548

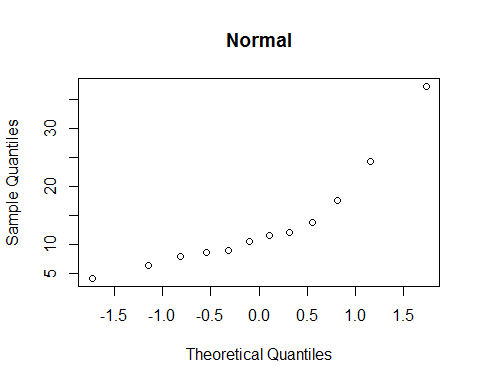
# 0.007405 < 0.05 = alpha -> reject H0  
# b) We want 90% confidence interval for delta,  
qt=qt(0.95,nu)  
lower=delta-qt\*se  
upper=delta+qt\*se  
lower

## [1] 11.84155

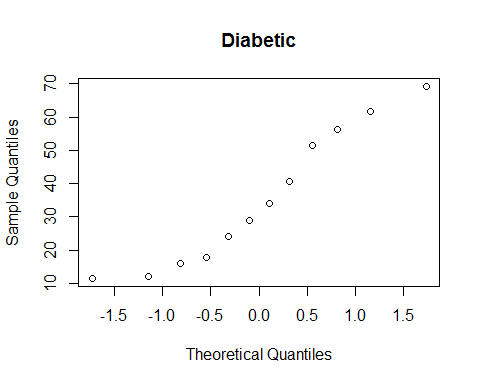
upper

## [1] 57.65845

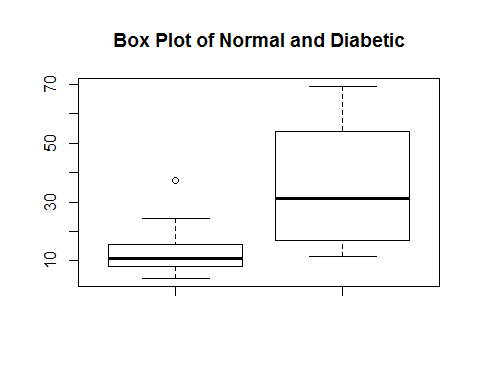
normal<- c(4.1,6.3,7.8,8.5,8.9,10.4,11.5,12.0,13.8,17.6,24.3,37.2)  
diabetic<- c(11.5,12.1,16.1,17.8,24.0,28.8,33.9,40.7,51.3,56.2,61.7,69.2)  
qqnorm(normal,main="Normal")



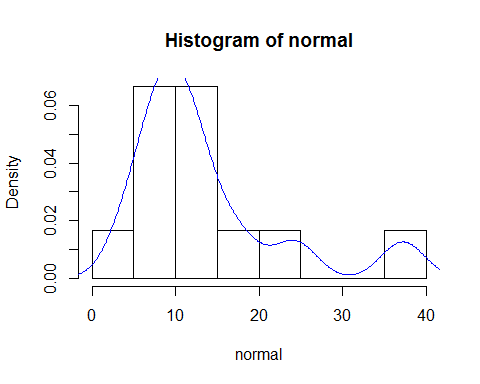
qqnorm(diabetic,main = "Diabetic")



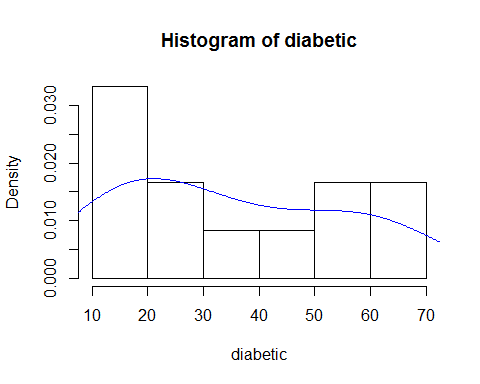
boxplot(normal,diabetic,main="Box Plot of Normal and Diabetic")



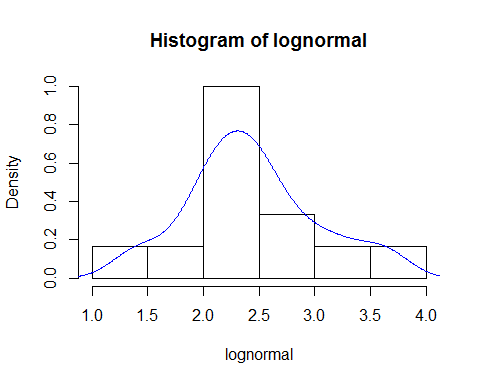
hist(normal,prob=TRUE)  
lines(density(normal),col="blue")



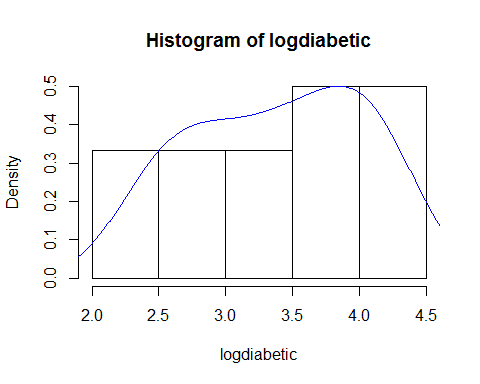
hist(diabetic,prob=TRUE)  
lines(density(diabetic),col="blue")



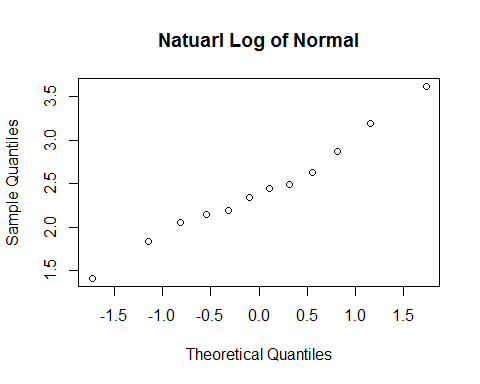
# 1 )  
#After seeing qqplot,boxplot and histogram we can say that samples are   
# not drawn from normal distribution.  
# 2)  
# (a) Natural Logarithm  
lognormal<- log(normal)  
logdiabetic<- log(diabetic)  
hist(lognormal,prob=TRUE)  
lines(density(lognormal),col="blue")



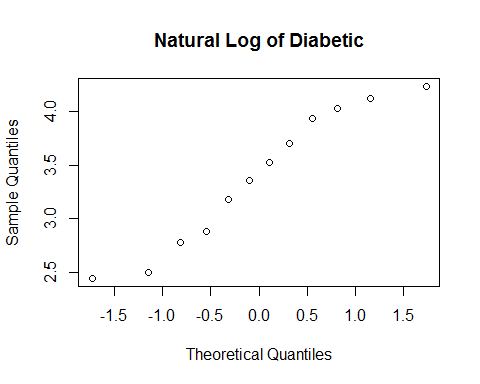
hist(logdiabetic,prob=TRUE)  
lines(density(logdiabetic),col="blue")



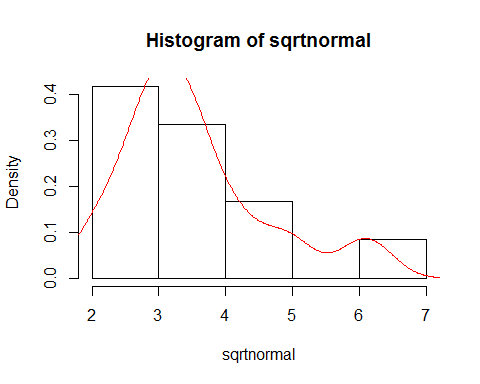
qqnorm(lognormal,main="Natuarl Log of Normal")



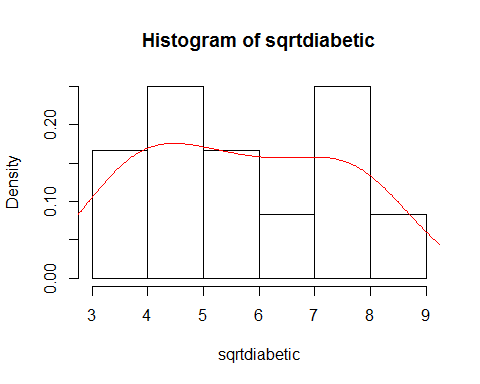
qqnorm(logdiabetic,main ="Natural Log of Diabetic")



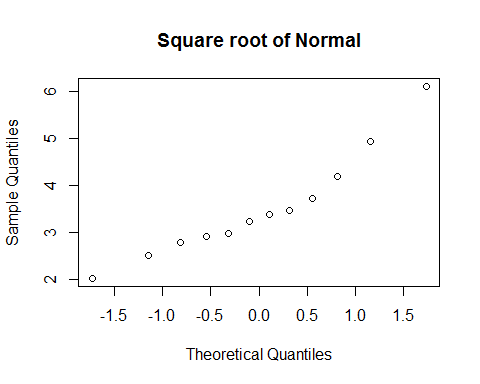
# (b) Square Root  
sqrtnormal<-sqrt(normal)  
sqrtdiabetic<-sqrt(diabetic)  
hist(sqrtnormal,prob=TRUE)  
lines(density(sqrtnormal),col="red")



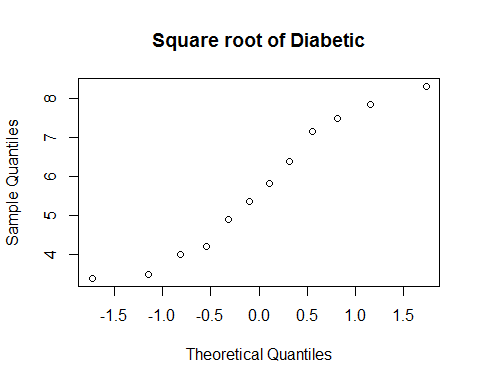
hist(sqrtdiabetic,prob=TRUE)  
lines(density(sqrtdiabetic),col="red")



qqnorm(sqrtnormal,main="Square root of Normal")



qqnorm(sqrtdiabetic,main ="Square root of Diabetic")



# I would prefer log transformation over square root transformation since  
# log transformation is more symmetric to normal distribution.  
# 3)  
# As seen from histograms, density plots and qqplots, log transformed  
# measurements appear closer to normal distribution.  
# 4)  
# Welch's t-test  
Delta = mean(logdiabetic) - mean(lognormal)  
se = sqrt(var(logdiabetic)/12 + var(lognormal)/12)  
Tw = Delta/se  
nu = (var(logdiabetic)/12+var(lognormal)/12)^2/((var(logdiabetic)/12)^2/11+(var(lognormal)/12)^2/11)  
Pvalue = 2\*(1-pt(abs(Tw),df=nu))  
Pvalue

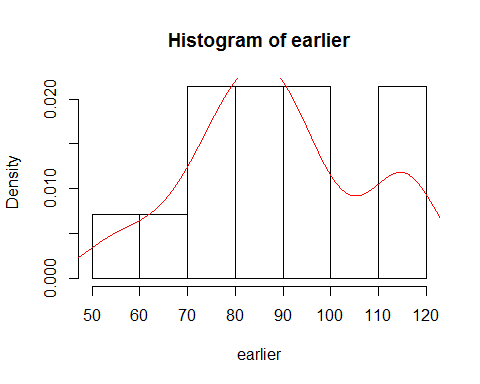
## [1] 0.0009776127

# Welch 95% confidence interval  
q = qt(0.975, df=nu)  
lower = Delta - q\*se  
upper = Delta + q\*se  
CI<-c(lower,upper)  
CI

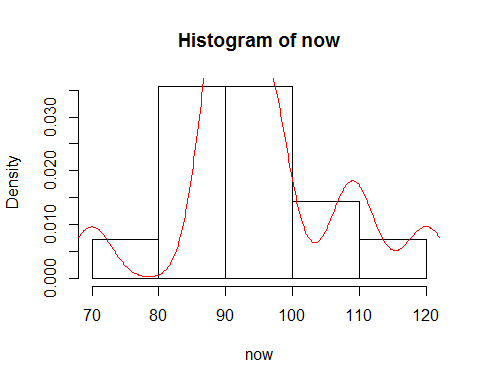
## [1] 0.4352589 1.4792986

# Since P-value is quite low we can reject H0 in favor of Ha.

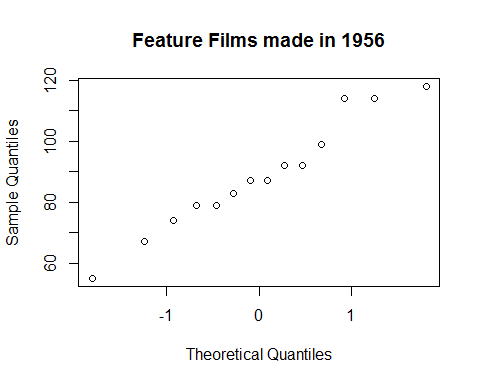
earlier<-c(74,114,114,87,92,55,67,118,79,83,79,92,99,87)  
now<-c(70,98,90,95,88,108,110,96,91,88,120,96,90,90)  
hist(earlier,prob=TRUE)  
lines(density(earlier),col="red")



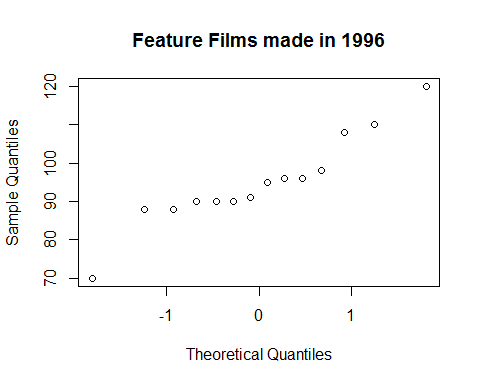
hist(now,prob=TRUE)  
lines(density(now),col="red")



qqnorm(earlier,main="Feature Films made in 1956")



qqnorm(now,main="Feature Films made in 1996")



t.test(now,earlier)

##   
## Welch Two Sample t-test  
##   
## data: now and earlier  
## t = 1.105, df = 22.395, p-value = 0.2809  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -5.623821 18.480963  
## sample estimates:  
## mean of x mean of y   
## 95.00000 88.57143

# P-value =0.2809 which shows that mean length of movies in 1956 is more  
# than mean length of movies in 1996 with 28% probability assuming null  
# hypothesis is true.  
# 95% Confidence Interval is (-5.623,18.480) shows that difference in movie  
# is not necessarily above 0.  
# 0.01 and 0.05 are ideal significance values in hypothesis testing which  
# is read as probability of null hypothesis being true.  
# After conducting welch's t-test which is based on normality of samples  
# datasets are normally distributed as seen from plots to conduct experiment.