problem set 9

1)

a)

The data should be analysed using a one sample t-test because the values of two samples depend on each other as the

measurement of glycemic levels are done on the same patient with a gap of several days .

b)

mu = mean of the difference between the measurements (glyecmic level with coffee and without coffee)

H0 : mu=0 H1 : mu not equal to zero

```
xbar<- 11.5
sd<- 21
n<-10
mu=0

t.stat <- 11.5/(21/sqrt(10))

#P-VALUE FOR THE TWO TAILED TEST

2*(1-pt(t.stat,df=n-1))
```

```
## [1] 0.1173671
```

c)

95% CI

```
#upper limit
xbar+qt(0.975,df=n-1)*21/sqrt(10)
```

```
## [1] 26.5225
```

```
#Lowe limit
xbar-qt(0.975,df=n-1)*21/sqrt(10)
```

```
## [1] -3.522495
```

95% CI (-3.522495,26.5225)

d)

we cant be exactly sure that dates have the same glycemic index with or without coffee because the 95%

confidence interval for average(mean) differnce between glyceminc levels for after consuming dates with and without

coffee lies between -3.6 and 27 .

2)

a)

we should use a t-distribution because the sample size is very small and population standard deviation is unknown

b)

```
xbar.men<-68.5
xbar.women<-65.5
sd.men<-3
sd.women<-2.5
xbar.men-xbar.women
```

```
## [1] 3
```

Welchs t statistic

```
w.tstat2 <-(xbar.men-xbar.women)/sqrt((sd.men*sd.women)/7 + (sd.men*sd.women)/7)
w.tstat2
```

```
## [1] 2.04939
```

95% CI

```
degrees = 11.6 #(degress of freedom)
#upper limit
xbar.men-xbar.women+(sqrt(sd.men*sd.men/7 + sd.women*sd.women/7)*qt(.975,df=11.62))
```

```
## [1] 6.227627
```

```
#Lower limit
xbar.men-xbar.women-(sqrt(sd.men*sd.men/7 + sd.women*sd.women/7)*qt(.975,df=11.62))
```

```
## [1] -0.2276273
```

95% CI (-0.2276273,6.227627)

c)

lets do a hypothesis test

H0 :mean difference in heights of men and women = 0 H1 : mean difference in heights of mean and women not equal to 0

```
#Two tailed test

2*(1-pt(w.tstat2,df=11.62))
```

```
## [1] 0.06369575
```

The P value is very small so we can reject the null hypothesis and conclude that there is a difference in heights

between men and women ????

###based on the confidence interval we can say that average height of men - avg height of women can vary from

3)

1)

a)

the experimental unit is middle aged men

b)

The experimental units are drawn from two populations

they are -

Type A - behavior is characterized by urgency,aggression,ambition

Type B - behavior is non competitive, more relaxed ,less hurried

20 units were drawn from each population

this is a two sample problem

c)

one measurement is taken on each experimental unit - cholestrol levels

d)

as this is a two sample problem the parameter of interest is delta

where delta is —

Let Xi cholestrol level in type A men, and Yj be the cholestrol level in type B men Let the population mean cholestrol level in type A men be μ1 and the population mean cholestrol level in type B men μ2. Let $\Delta = \mu 1 - \mu 2$.

e)

H0 : delta <=0 H1 : delta > 0

2)

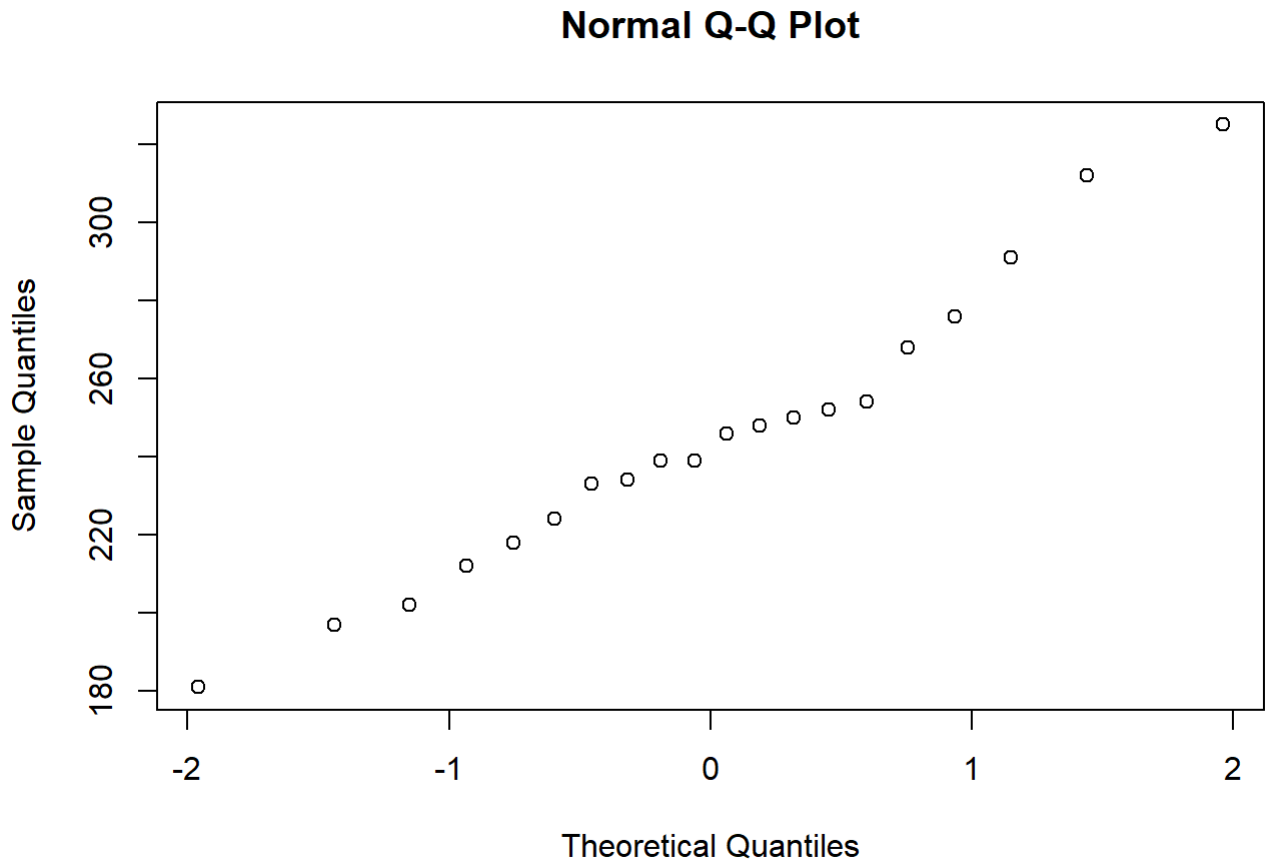
```
typeA <- c(233, 291, 312, 250, 246, 197, 268, 224, 239, 239,
254 , 276, 234, 181, 248, 252, 202, 218, 212, 325)

typeB <- c(344 , 185 , 263 , 246 , 224 , 212 , 188 ,  250, 148, 169 ,
226 , 175 , 242 , 252 , 153 ,  183 , 137 , 202, 194, 213)
```

Draw qqplots for both data

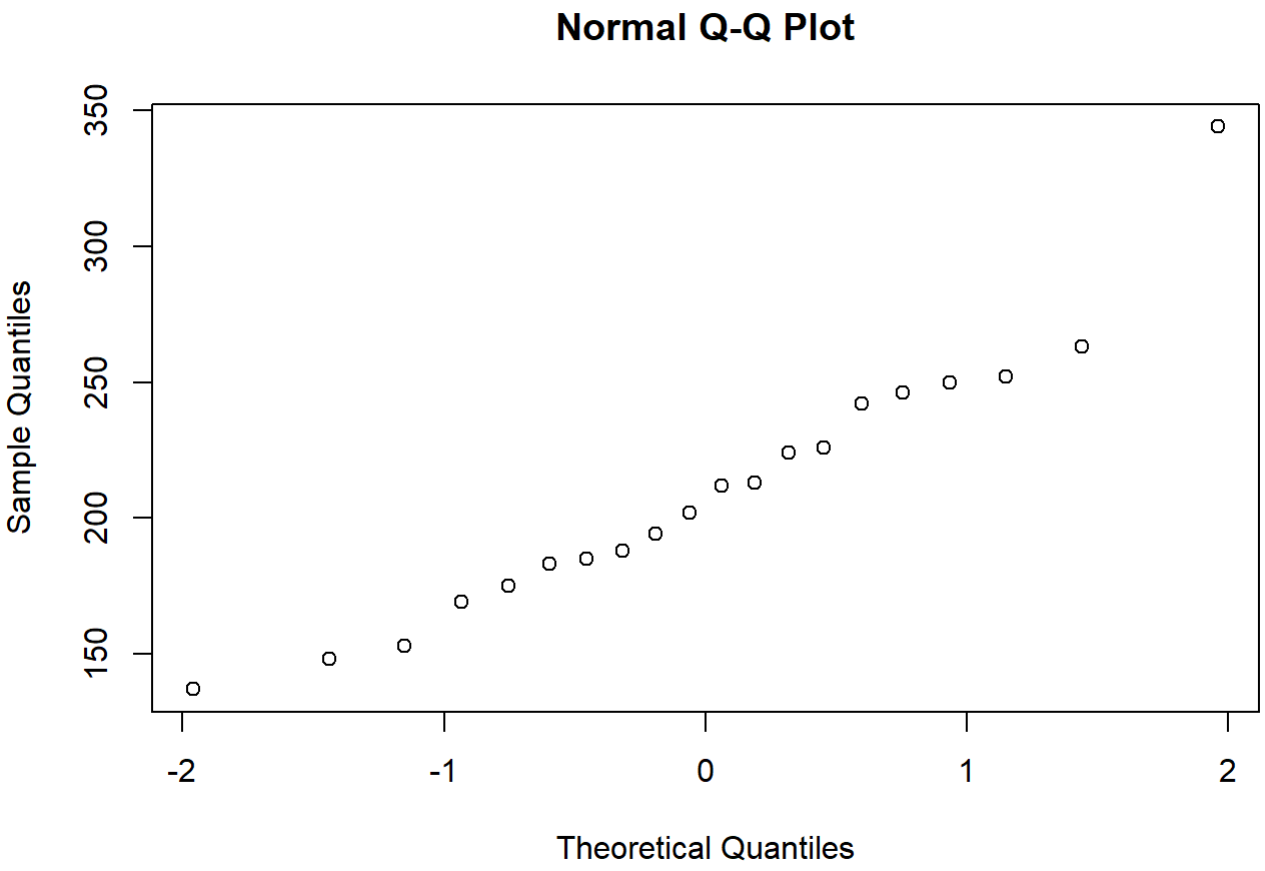
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
#TYPE A MEN
qqnorm(typeA)
```




```
#TYPE B MEN
qqnorm(typeB)
```



From the above graphs we can infer that the observed values of both samples are kind of drawn from approximately normal distributions

3)

a)

welchs t-test

Welch's t-statistic

```
xbarA <- mean(typeA)
xbarB <- mean(typeB)
sdA <- sd(typeA)
sdB <- sd(typeB)
w.tstat3 <-(xbarA-xbarB)/sqrt((sdA*sdA)/length(typeA) + (sdB*sdB)/length(typeB))
# welchs t-statistic
w.tstat3
```

```
## [1] 2.562113
```

Right tailed test

```
#degrees of freedom



df3<-((var(typeA)/length(typeA) + var(typeB)/length(typeB))**2/((var(typeA)/length(typeA))**2/(l
ength(typeA)-1) + ((var(typeB)/length(typeB))**2/(length(typeB)-1))))

df3
```

```
## [1] 35.41308
```

Significance probability P-value

```
1-pt(w.tstat3,df=df3)
```

```
## [1] 0.007405252
```

The P value is very less than the significance level so we can reject the null hypothesis

b)

CI of 90%

```
delta3.hat=xbarA-xbarB
delta3.hat
```

```
## [1] 34.75
```

```
#upper limit
xbarA-xbarB+(qt(0.95,df=df3)*sqrt(var(typeA)/length(typeA) + var(typeB)/length(typeB)))
```

```
## [1] 57.65845
```

```
#Lower limit
xbarA-xbarB-qt(0.95,df=df3)*sqrt(var(typeA)/length(typeA) + var(typeB)/length(typeB))
```

```
## [1] 11.84155
```

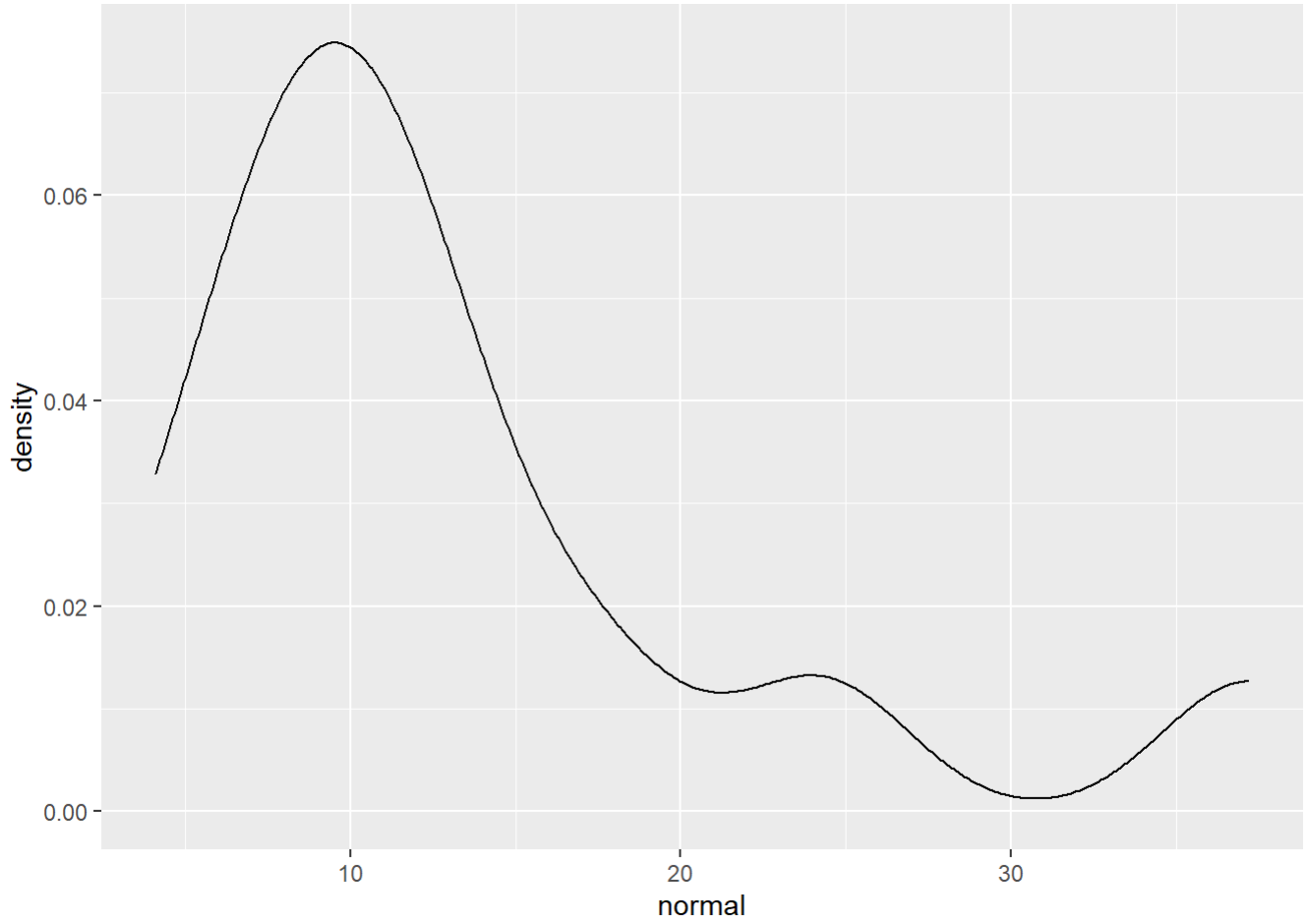
90% CI (11.84155,57.65845)

4)

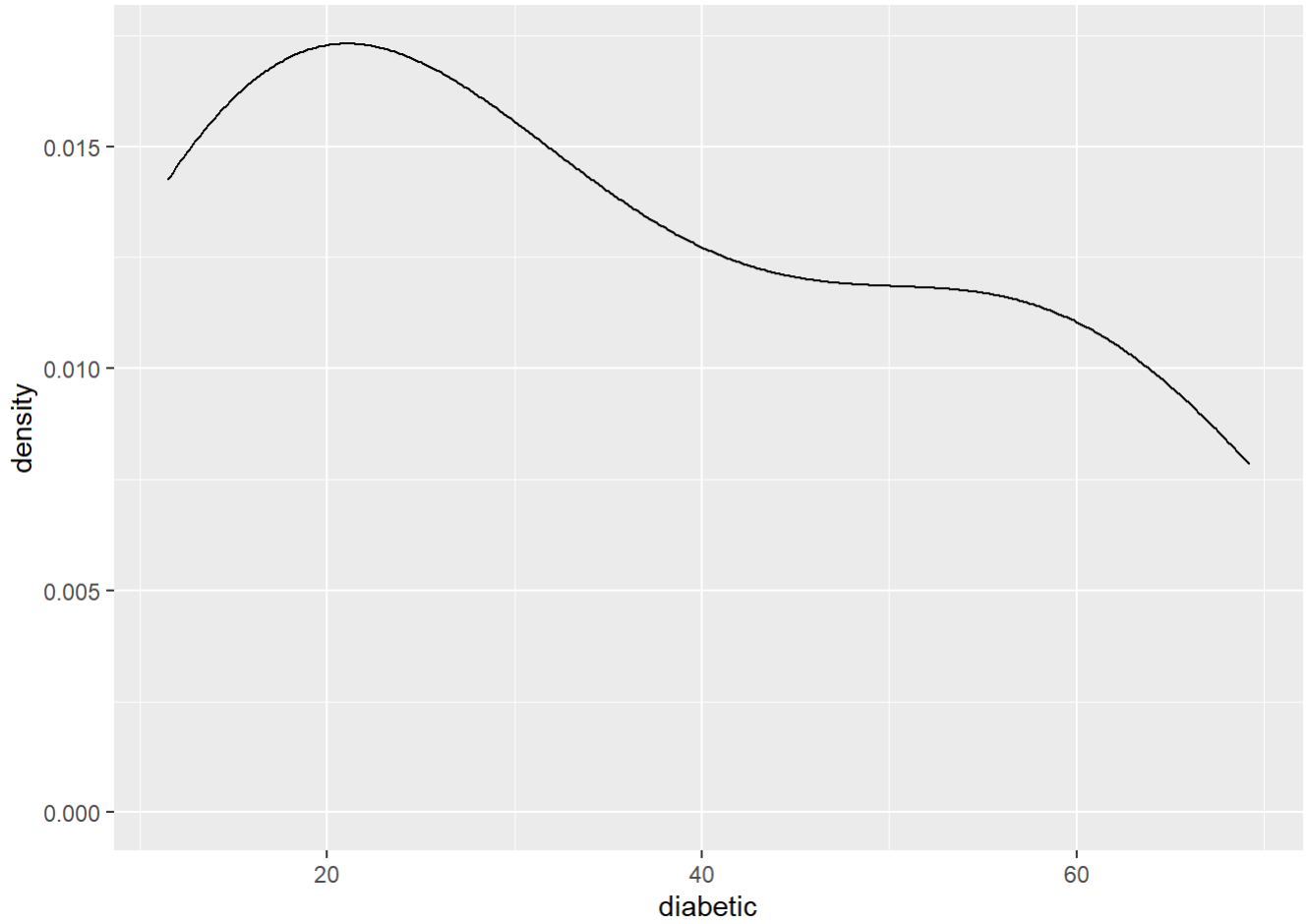
```
normal <- c(4.1, 6.3, 7.8, 8.5, 8.9, 10.4,
11.5 ,12.0 ,13.8 ,17.6 ,24.3, 37.2)

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)
```

1)

```
ggplot(data.frame(normal),aes(x=normal))+geom_density()
```




```
ggplot(data.frame(diabetic),aes(x=diabetic))+geom_density()
```



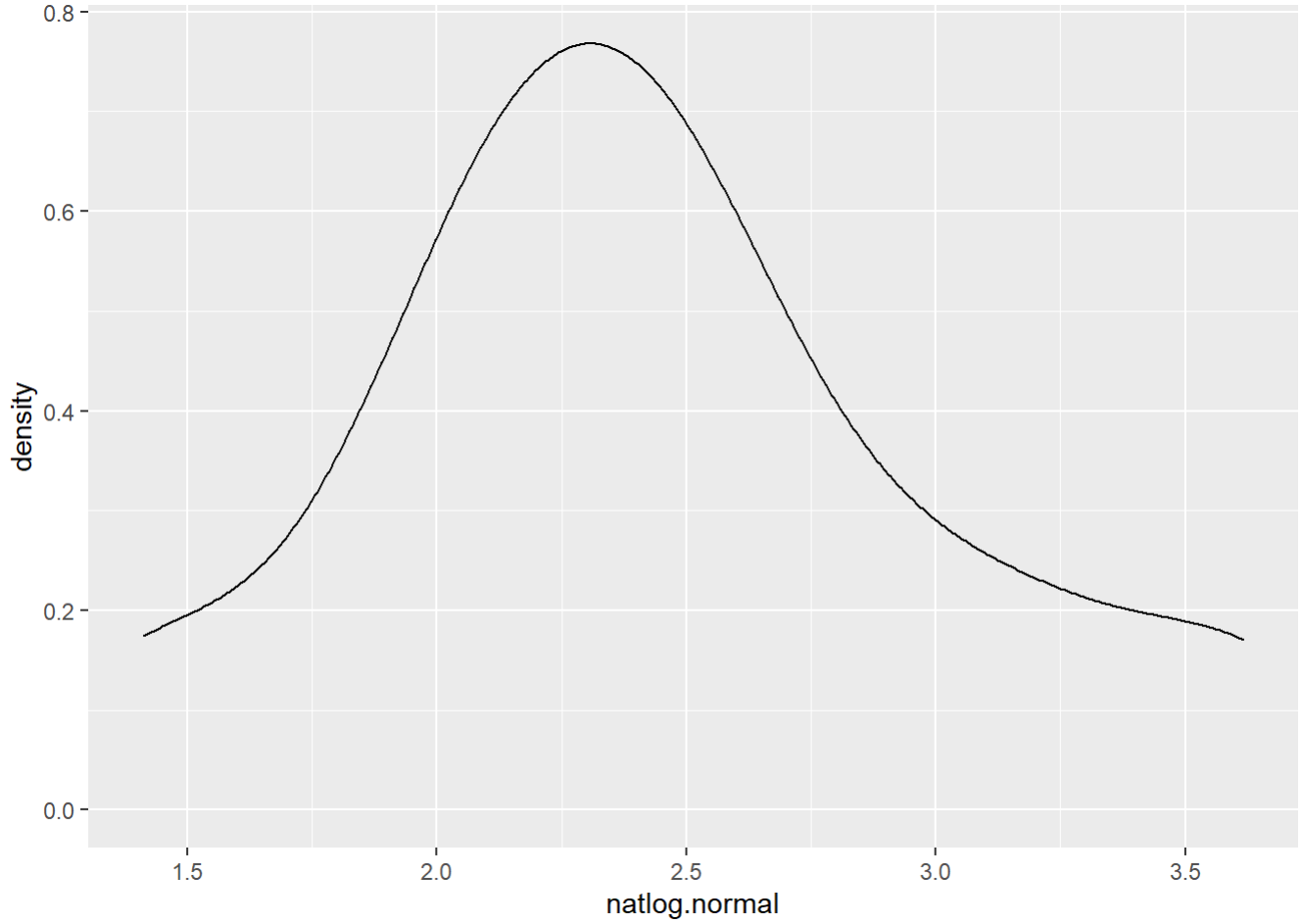
From the above density plots we can say that the samples are not from symmetric distribution and appear to be skewed to the right

2)

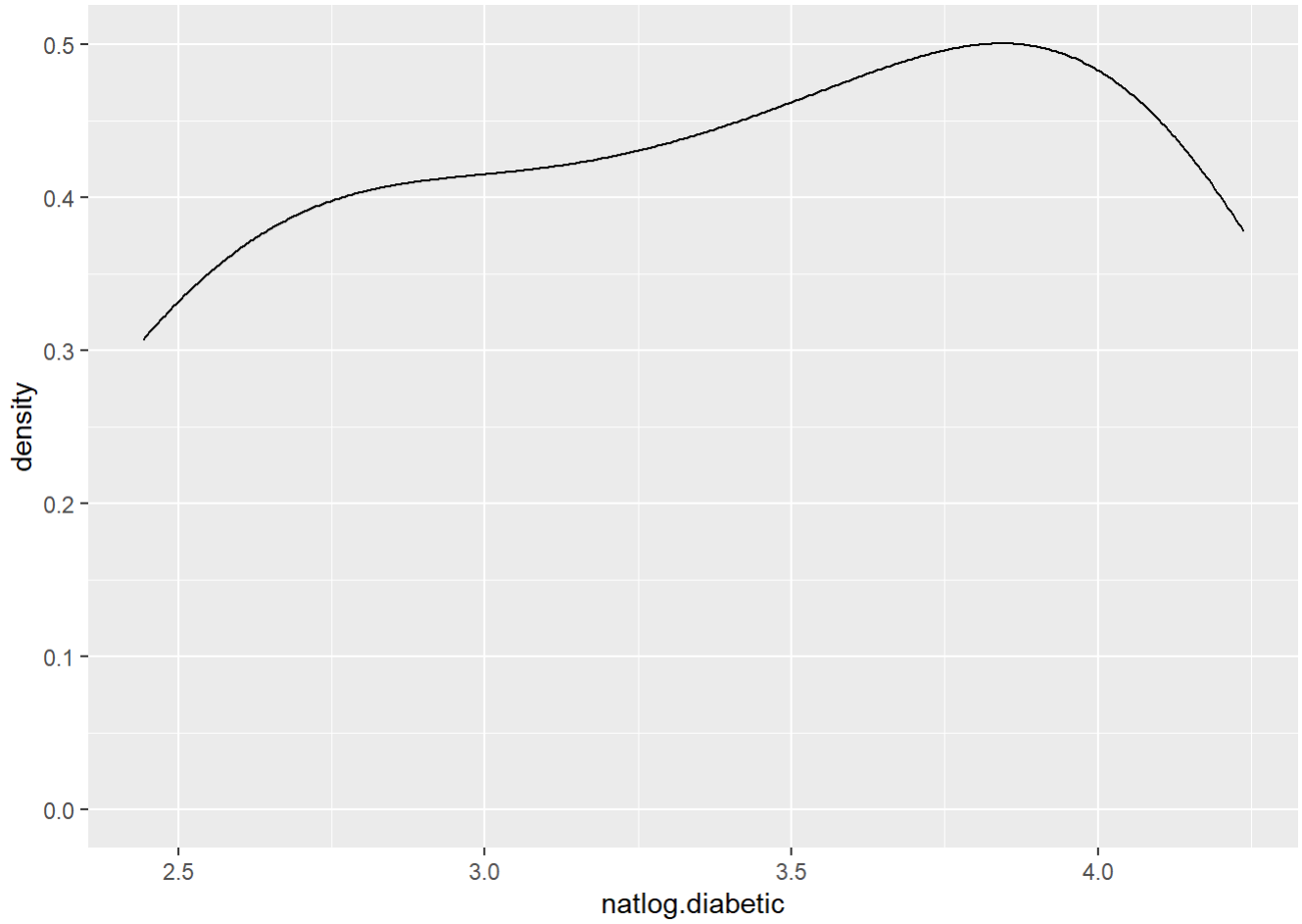
```
natlog.normal<-log(normal)

natlog.diabetic<-log(diabetic)

sqroot.normal<-sqrt(normal)

sqroot.diabetic<-sqrt(diabetic)
```

Natural Logarithm

```
#Normal
ggplot(data.frame(natlog.normal),aes(x=natlog.normal))+geom_density()
```

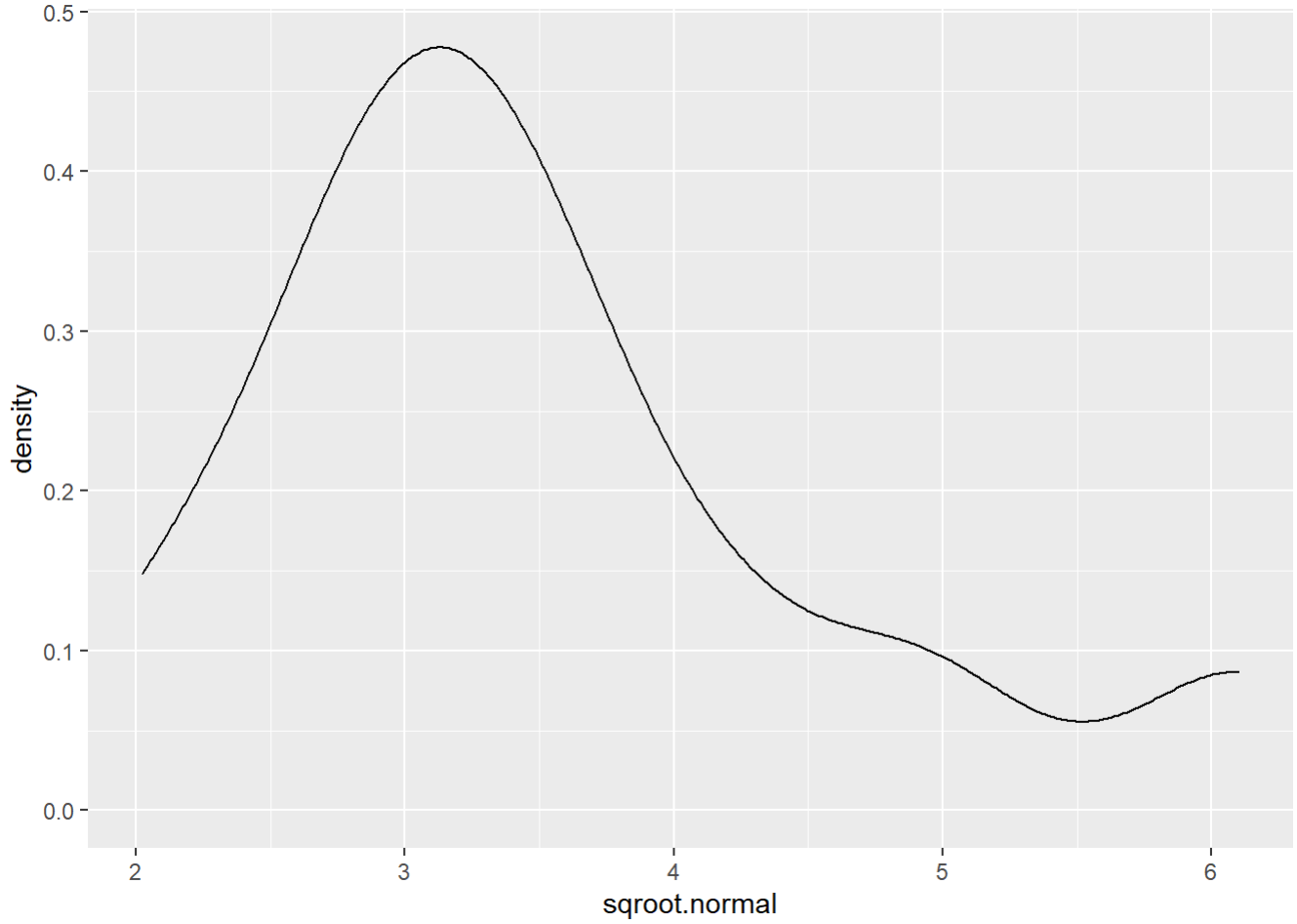



```
#diabetic
ggplot(data.frame(natlog.diabetic),aes(x=natlog.diabetic))+geom_density()
```

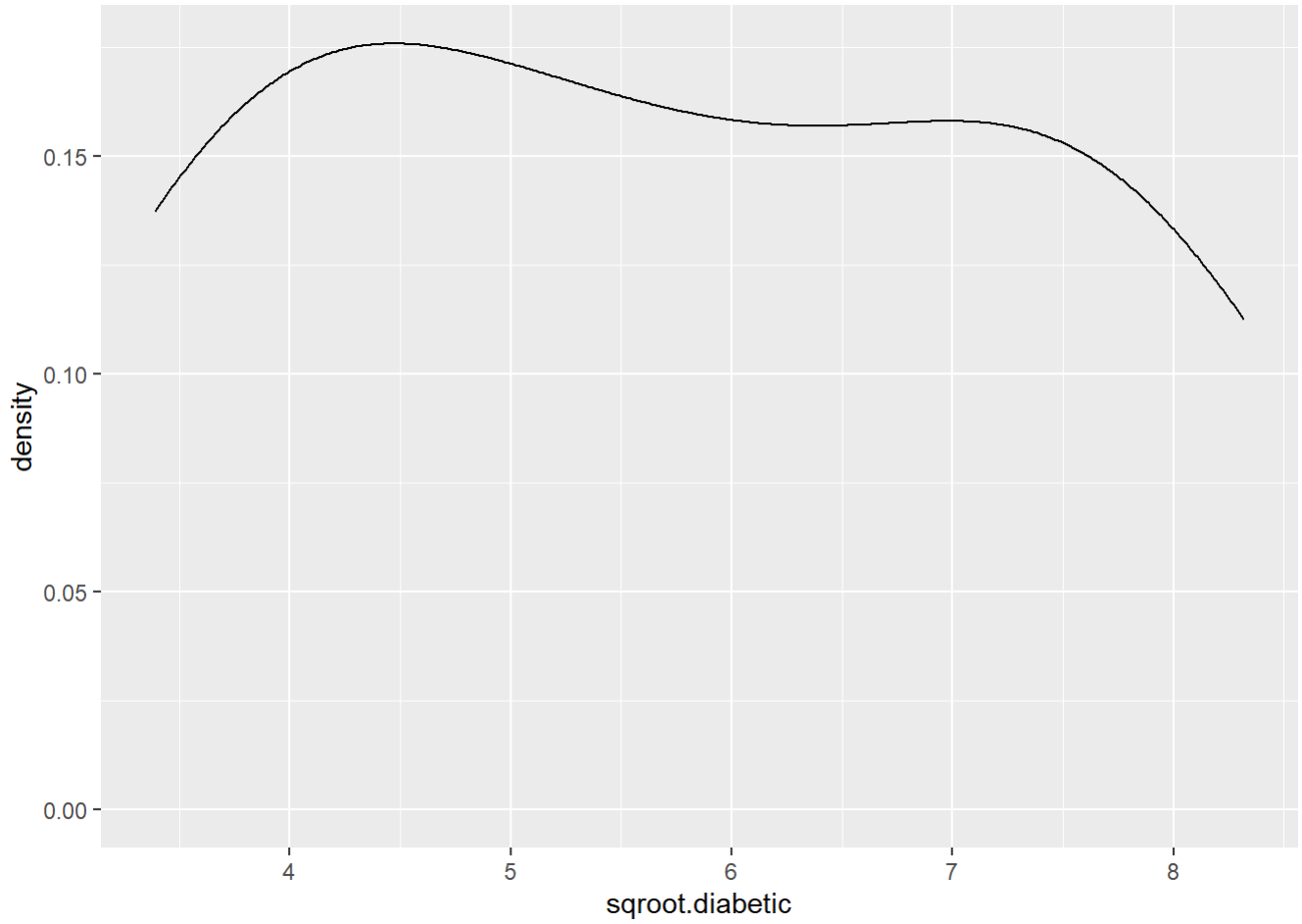


square root

```
#normal
ggplot(data.frame(sqroot.normal),aes(x=sqroot.normal))+geom_density()
```




```
#diabetic
ggplot(data.frame(sqroot.diabetic),aes(x=sqroot.diabetic))+geom_density()
```

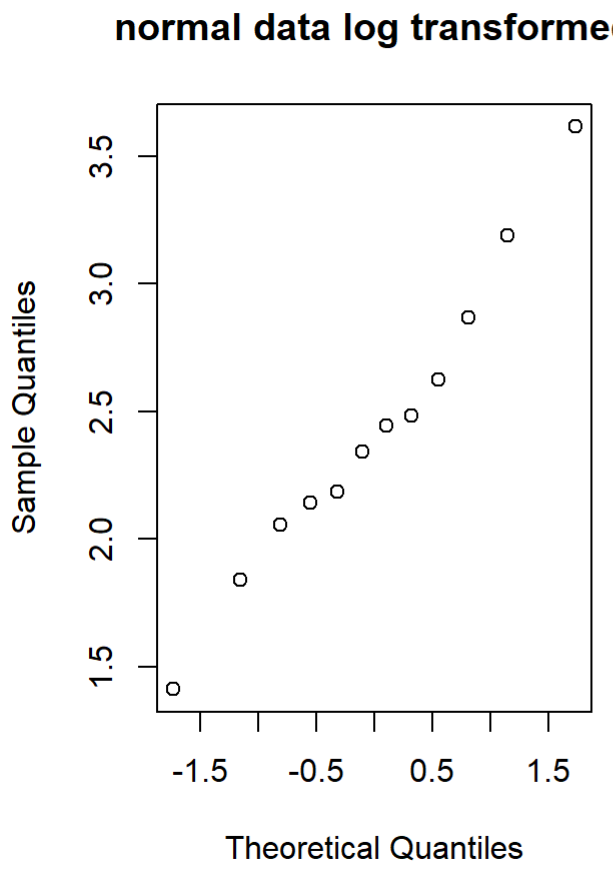


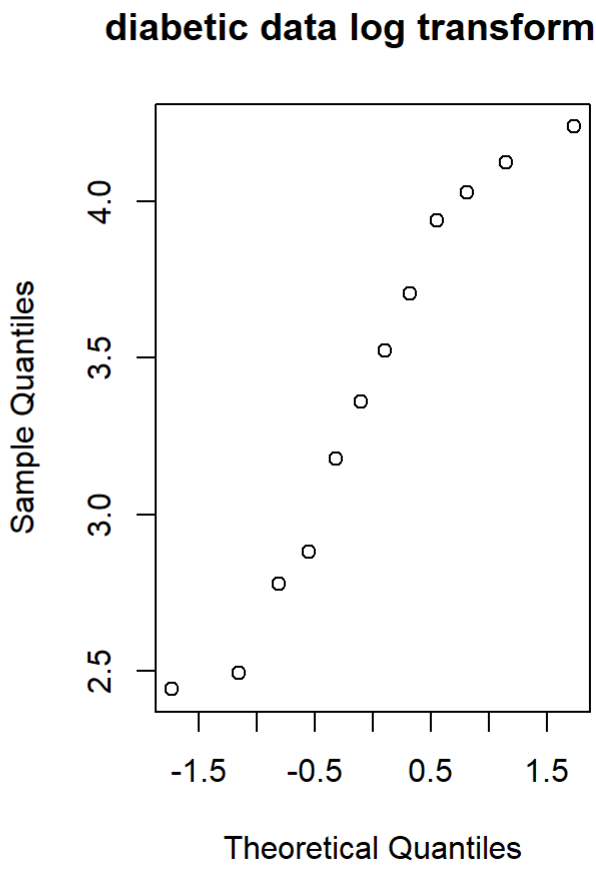
from the above two transformations the log transform approximately more symmetric than sqrt transformation so i would prefer log transformation

3)

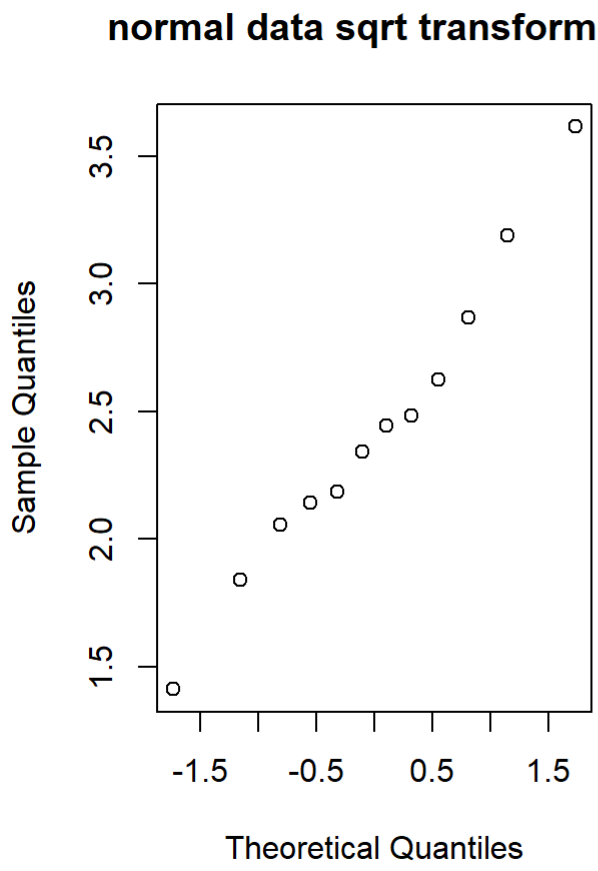
natural logarithm

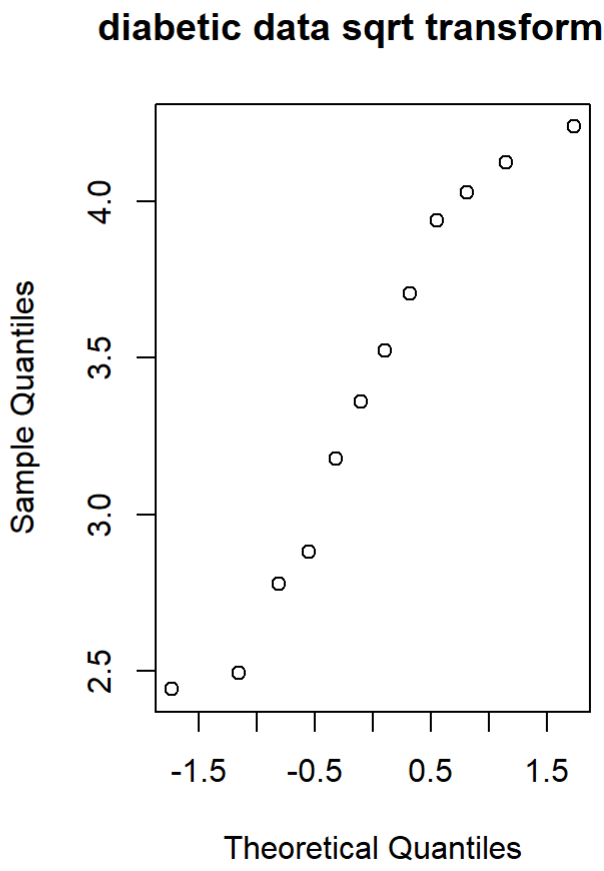
```
par(mfrow=c(1,2))
qqnorm(natlog.normal,main = 'normal data log transformed')

qqnorm(natlog.diabetic,main = 'diabetic data log transformed')
```






```
par(mfrow=c(1,2))
qqnorm(natlog.normal,main = 'normal data sqrt transformed')

qqnorm(natlog.diabetic,main = 'diabetic data sqrt transformed')
```





from

above plots both the transformations appear approximately normal

4)

delta = mean of log of diabetic transformed data - mean of log of normal transformed data

Null hypothesis H0: delta.hat<=0

alternate hypothesis H1: delta.hat >0

```
xbar.normal <-mean(natlog.normal)
xbar.diabetic<-mean(natlog.diabetic)
sd.normal<-sd(natlog.normal)
sd.diabetic<-sd(natlog.diabetic)
normal.len<-length(natlog.normal)
diabetic.len<-length(natlog.diabetic)
```

wlechs test

```
wlech.4tstat<-(xbar.diabetic-xbar.normal)/sqrt((sd.normal*sd.normal)/normal.len + (sd.diabetic*s
d.diabetic)/diabetic.len)
wlech.4tstat
```

```
## [1] 3.804072
```

Degree of freedom

```
df4 <- ((var(natlog.normal)/length(natlog.normal)+var(natlog.diabetic)/length(natlog.diabetic))*
*2/((var(natlog.normal)/length(natlog.normal))**2/(length(natlog.normal)-1) + ((var(natlog.diabe
tic)/length(natlog.diabetic))**2/(length(natlog.diabetic)-1))))

df4
```

```
## [1] 21.89982
```

P-VALUE

```
1-pt(wlech.4tstat,df=df4)
```

```
## [1] 0.0004888064
```

the p-value is very small so we can reject the null hypothesis

95% CI

```
#upper limit in transformed space
xbar.diabetic- xbar.normal+qt(0.975,df=df4)*sqrt(var(natlog.normal)/length(natlog.normal) + var
(natlog.diabetic)/length(natlog.diabetic))
```

```
## [1] 1.479299
```

```
#Lower limit in original space
xbar.diabetic- xbar.normal-qt(0.975,df=df4)*sqrt(var(natlog.normal)/length(natlog.normal) + var
(natlog.diabetic)/length(natlog.diabetic))
```

```
## [1] 0.4352589
```

```
#upper limit in original space
exp(1.479299)
```

```
## [1] 4.389867
```

```
#Lower limit in original space

exp(0.4352589)
```

```
## [1] 1.545363
```

our 95% CI has only positive values and the p-value is also very small so we have

pretty strong evidence supporting the researchers claim