ProblemSet9-AnirudhPenmatcha

2023-11-02

Question 1 (Problem Set D)

1.

The measurements don't appear to be from a symmetric distribution because the values are growing towards the right. This means that the data is heavily concentrated to the right and hence it will be left skewed.

```
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, 28.8, 33.9, 40.7, 51.3, 56.2, 61.7, 69.2)
```

For a distribution to be symmetric, their mean and median have to be close to each other

```
print(paste(mean(normal), median(normal)))
```

```
## [1] "13.5333333333333 10.95"
```

We can see that the mean and median are not close for the normal data. Hence it is not symmetric.

```
print(paste(mean(diabetic), median(diabetic)))
```

```
## [1] "35.275 31.35"
```

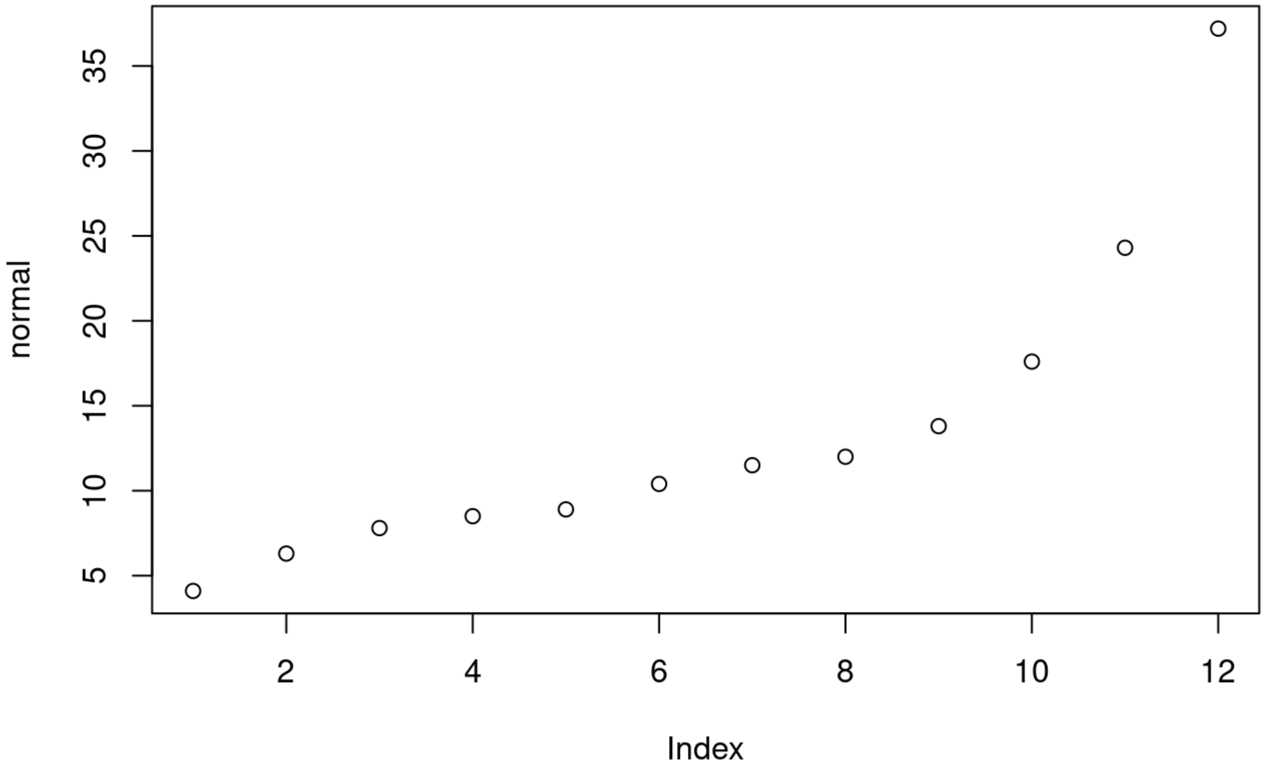
Again, mean and median are not close. Hence, they are not symmetric

2.

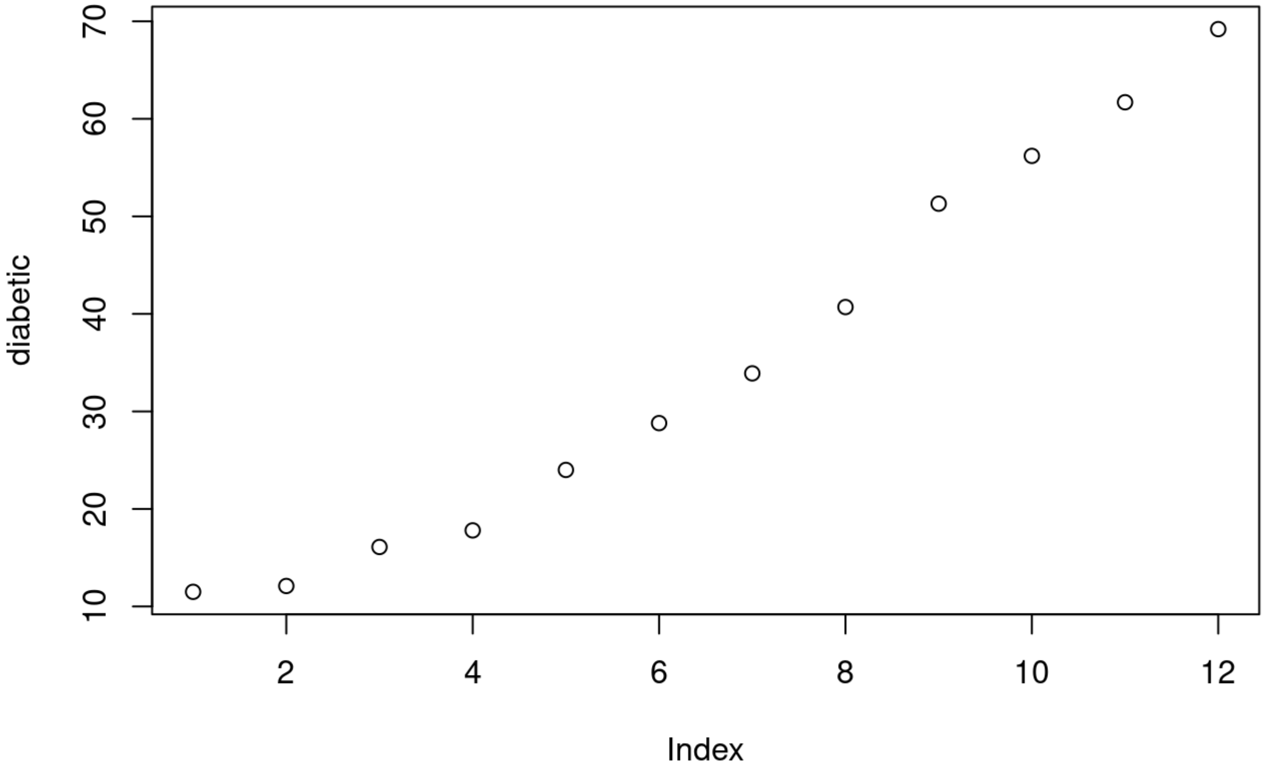
plotting

plots of original data

```
plot(normal)
```



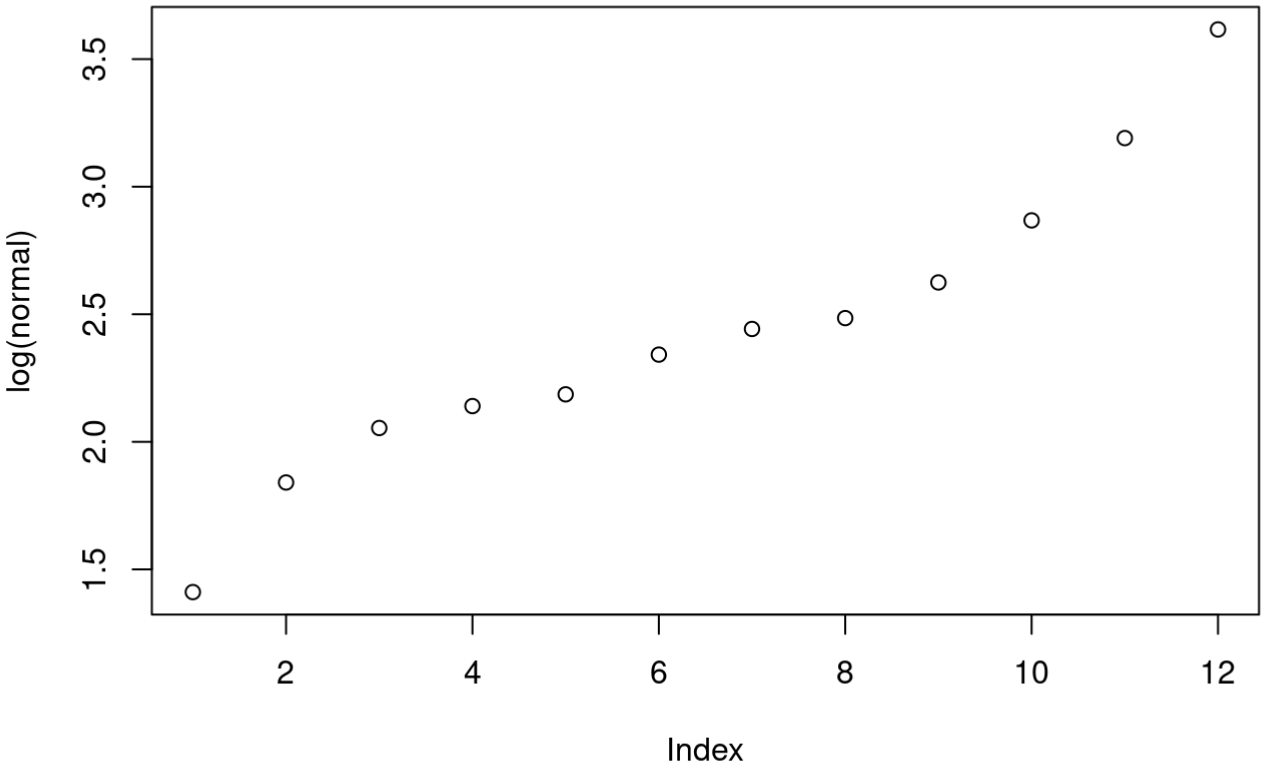
```
plot(diabetic)
```



#

plots of data transformed with log

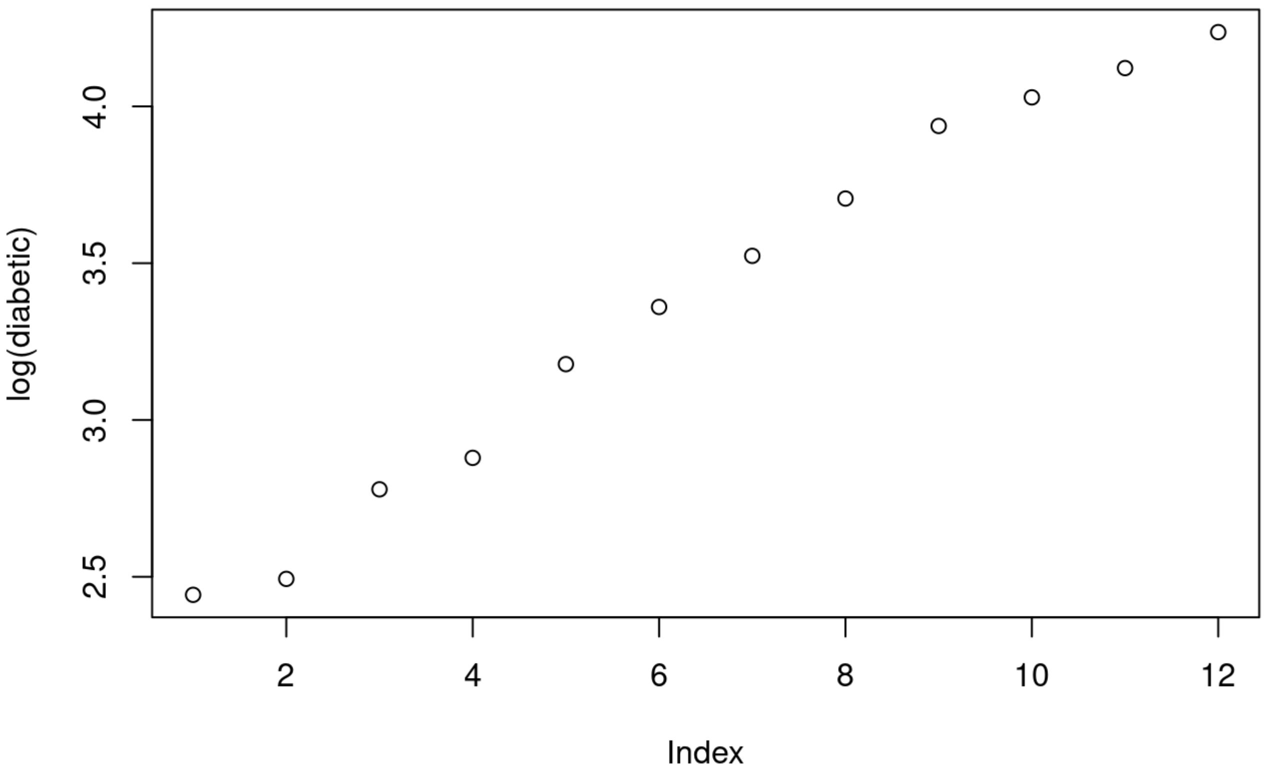
```
plot(log(normal))
```



```
log(normal)
```

```
##  [1] 1.410987 1.840550 2.054124 2.140066 2.186051 2.341806 2.442347 2.484907
##  [9] 2.624669 2.867899 3.190476 3.616309
```

```
plot(log(diabetic))
```

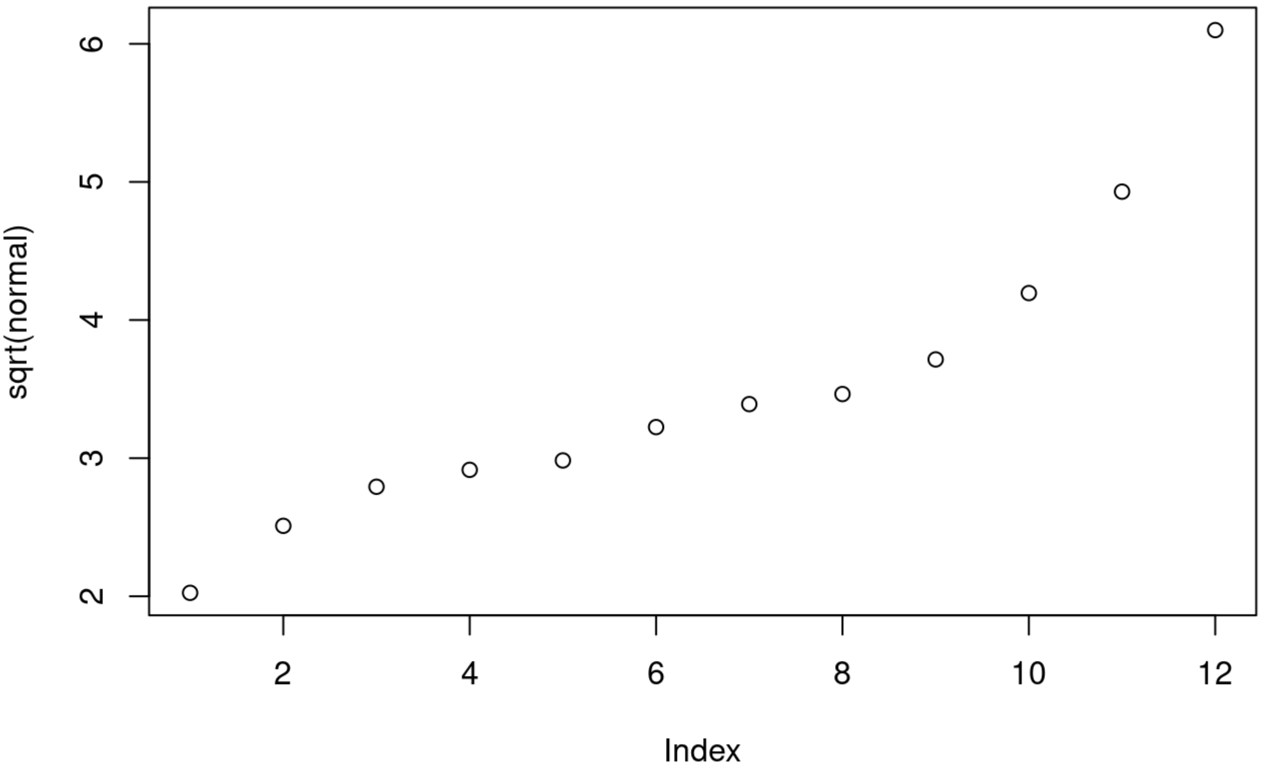


```
log(diabetic)
```

```
##  [1] 2.442347 2.493205 2.778819 2.879198 3.178054 3.360375 3.523415 3.706228
##  [9] 3.937691 4.028917 4.122284 4.237001
```

plots of data transformed with square root

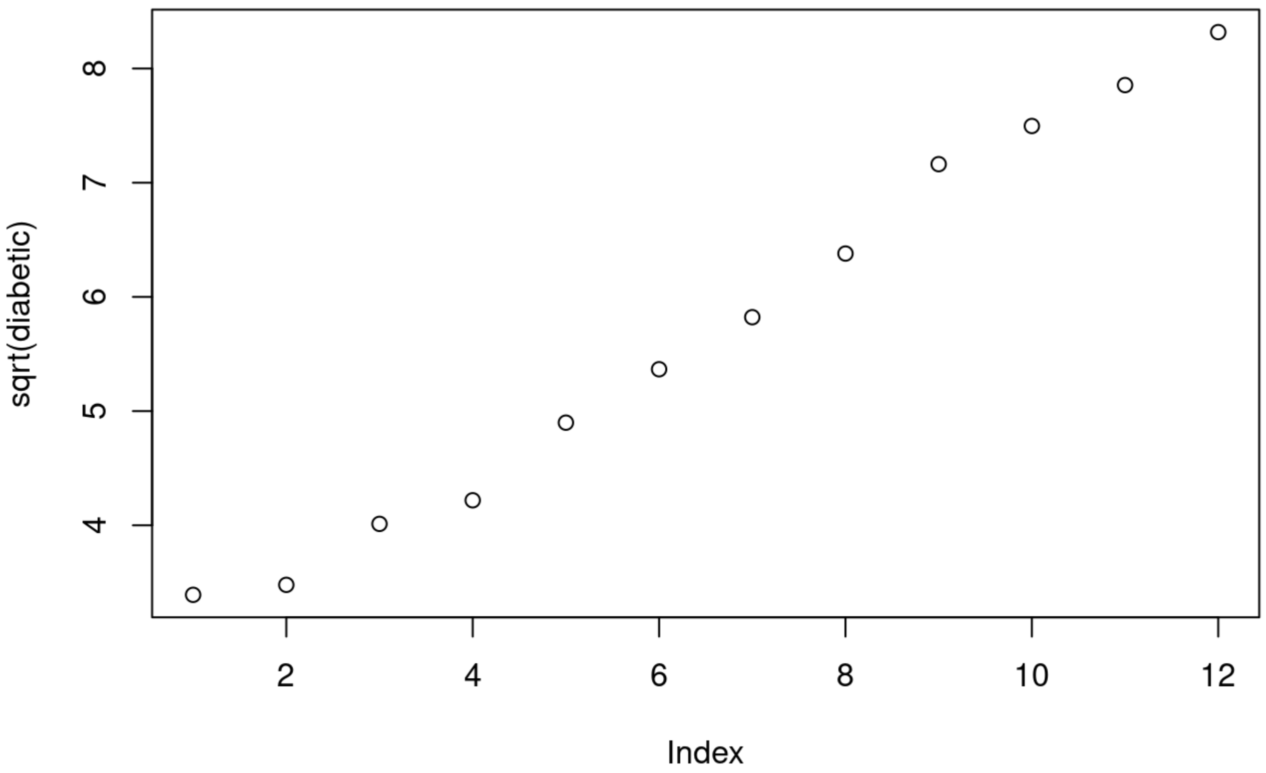
```
plot(sqrt(normal))
```



```
sqrt(normal)
```

```
##  [1] 2.024846 2.509980 2.792848 2.915476 2.983287 3.224903 3.391165 3.464102
##  [9] 3.714835 4.195235 4.929503 6.099180
```

```
plot(sqrt(diabetic))
```



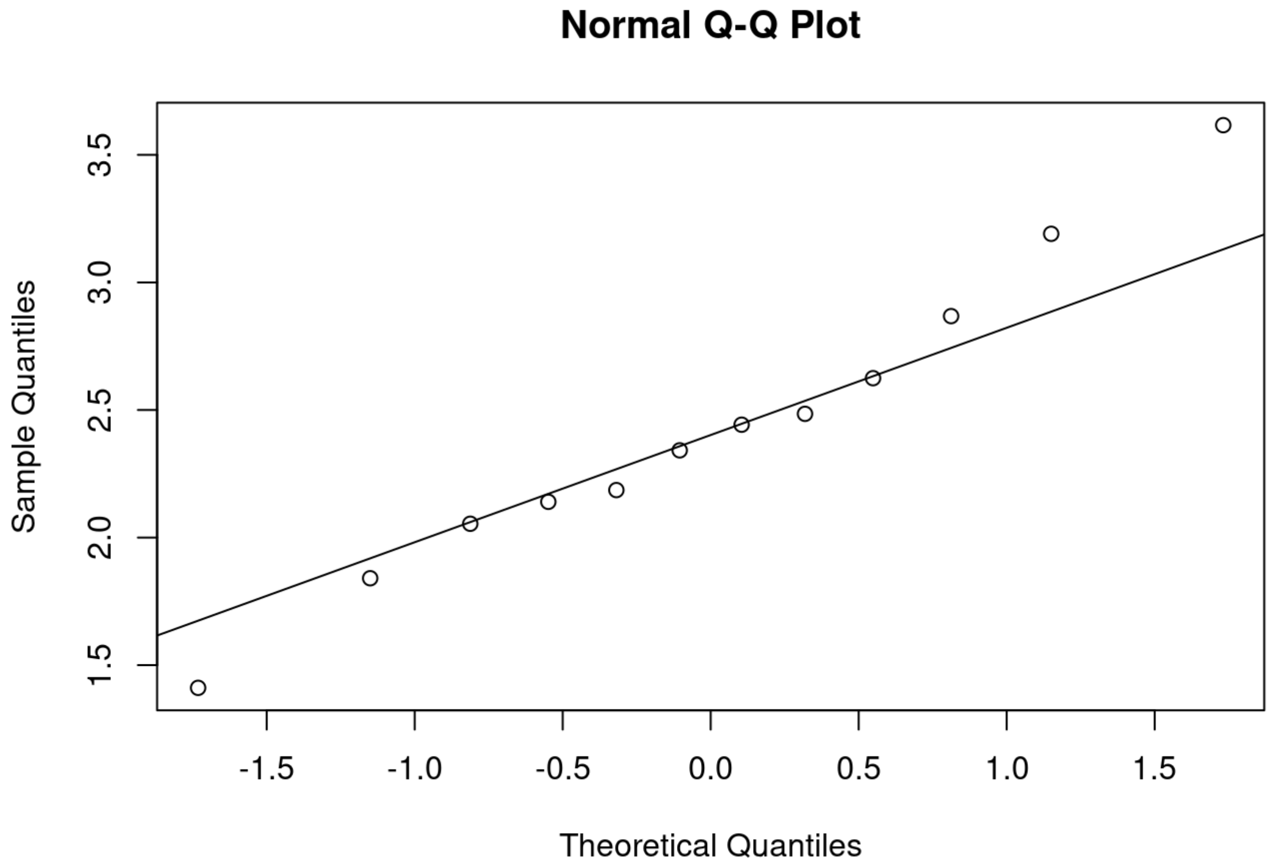
```
sqrt(diabetic)
```

```
##  [1] 3.391165 3.478505 4.012481 4.219005 4.898979 5.366563 5.822371 6.379655
##  [9] 7.162402 7.496666 7.854935 8.318654
```

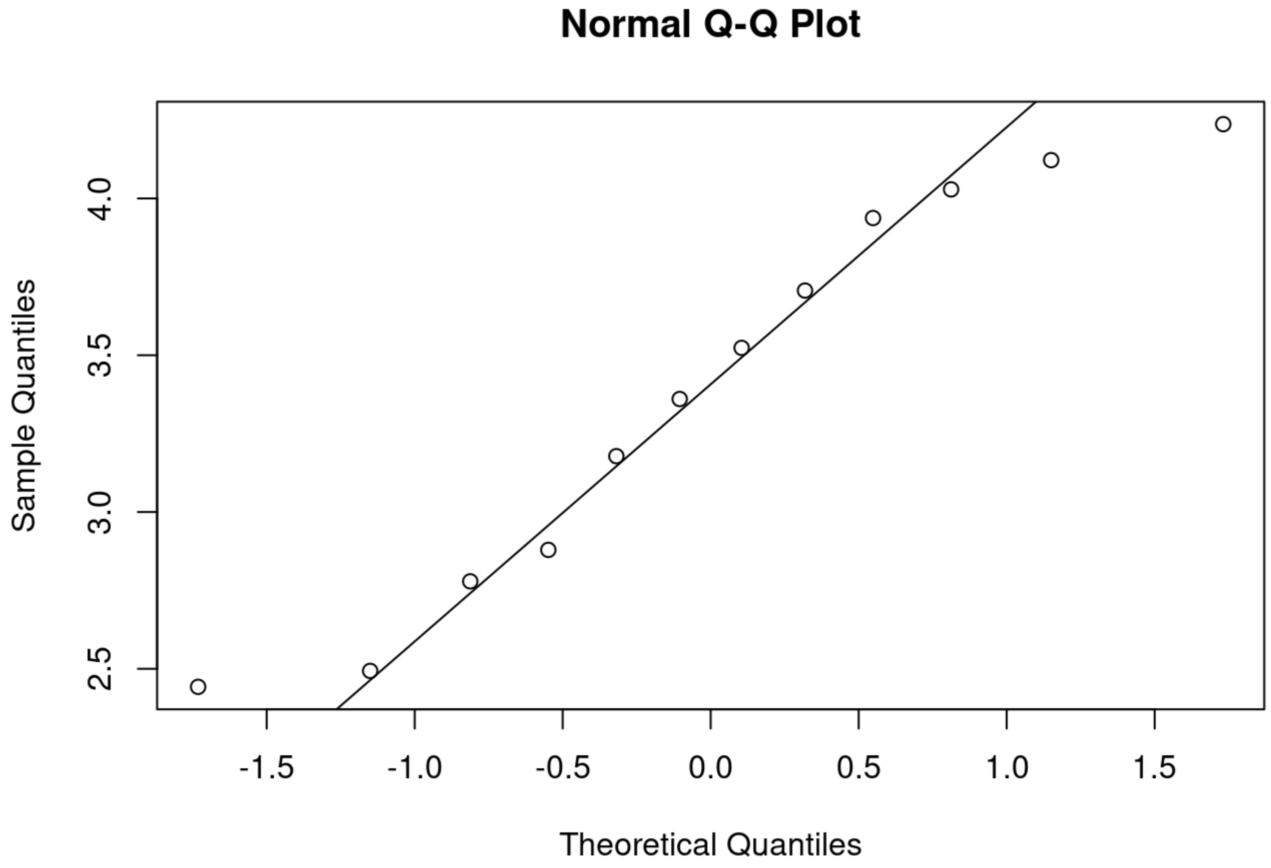
The plots of the data transformed with log look closer to a sample obtained from a symmetric distribution. The data transformed with square roots has one issue which is the sqrt(normal) looks less symmetric than log(normal). Hence I would prefer log transformation here.

3.

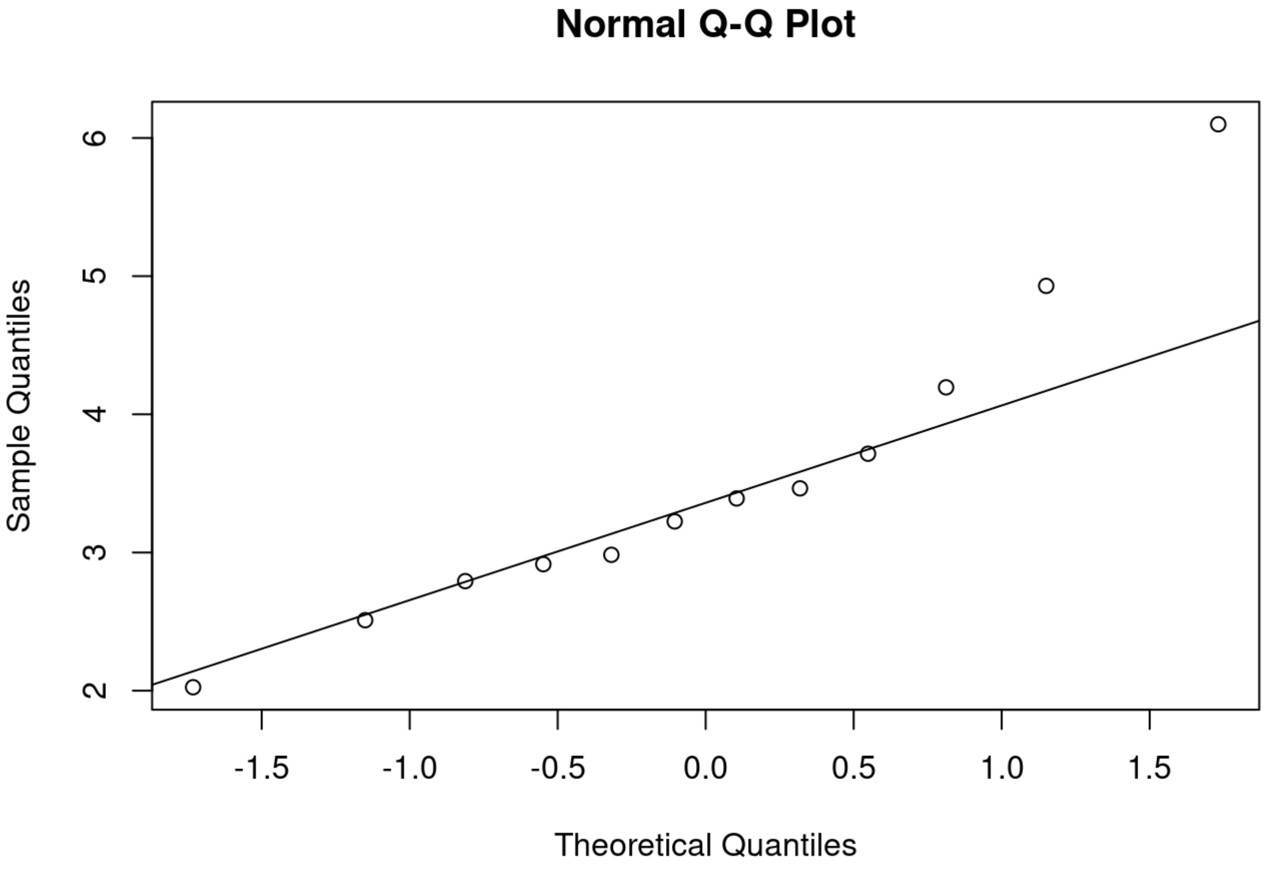
```
qqnorm(log(normal))
qqline(log(normal))
```



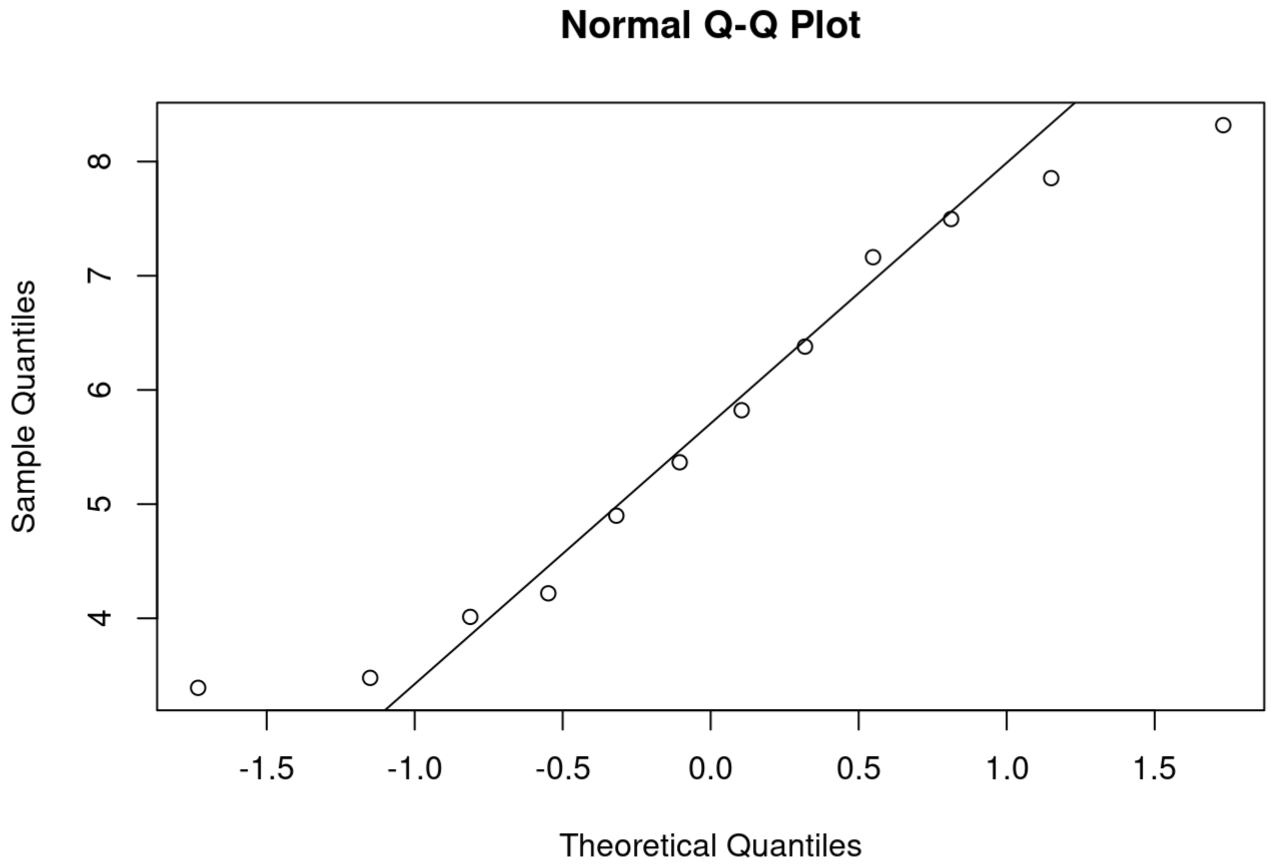
```
qqnorm(log(diabetic))
qqline(log(diabetic))
```



```
qqnorm(sqrt(normal))
qqline(sqrt(normal))
```



```
qqnorm(sqrt(diabetic))
qqline(sqrt(diabetic))
```



The transformed measurements of log appear to be from a normal distribution more than the transformed measurements of the square root.

4.

We can check this by performing a hypothesis test

```
t.test(log(diabetic), log(normal), alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  log(diabetic) and log(normal)
## t = 3.8041, df = 21.9, p-value = 0.0004888
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5250797       Inf
## sample estimates:
## mean of x mean of y
##  3.390628  2.433349
```

The p-value of 0.0004 support the alternative hypothesis and hence we can say that the patients have an increased urinary excretion

Question 2

a.

The problem we are working with here has provided a very small sample size which is 7. Hence, using the t-distribution becomes imperative

b.

```
delta.hat <- 68.5 - 65.5
var.male <- 3^2
var.female <- 2.5^2
SE <- sqrt(var.male/7 + var.female/7)
```

We are given the degrees of freedom as 11.6 and the value of R code qt(.975,df=11.62) as 2.187

```
# confidence intervals of the t-distributions for the given samples of male and female heights

print(paste("The upper interval = ", delta.hat + 2.187 * SE, " The lower interval = ", delta.hat - 2.187 * SE))
```

```
## [1] "The upper interval =  6.22800861323864  The lower interval =  -0.228008613238641"
```

c.

We needn’t conclude that there is no difference because there is a possibility for a combination where their differences in height can be 0 but also more and less than that. Generally speaking we shouldn’t be hasty in jumping to conclusions in such cases especially when dealing with a small dataset.

Question 3

a.

The rule for student’s test is that their variances must be equal. However, we are given the SD’s as 15.9 and 17.3. Meaning, variances will equal 252.81 and 299.29. Therefore, we’ll need to use welch’s test.

b.

```
delta.hat <- 24.3 - 16.8
SE <- sqrt( (15.9^2)/592 + (17.3^2)/154 )
T.Welch <- delta.hat/SE
T.Welch
```

```
## [1] 4.871275
```

p-value

```
2 * (1 - pt(abs(T.Welch), df = 225))
```

```
## [1] 2.090604e-06
```

c.

95% confidence interval

```
print(paste("The upper interval = ", delta.hat + qt(.975, df = 225) * SE, " The lower
interval = ", delta.hat - qt(.975, df = 225) * SE))
```

```
## [1] "The upper interval =  10.5339544731288  The lower interval =  4.4660455268712
3"
```

Therefore, we can say that we are 95% confident that the average difference in weeks worked between treament and control groups is between 10.5 and 4.4.

Question 4

a.

Because we are taking two samples out of one population and given that the standard deviations in the results of the two groups are not equal, it is valid to use welch’s test here

```
delta.hat <- 61 - 59
SE <- sqrt(((10^2)/100) + ((13^2)/100))
T.Welch <- delta.hat/SE
T.Welch
```

```
## [1] 1.219422
```

```
df <- (((10^2)/100+(12^2)/100)^2)/ ((((10^2)/100)^2/99)+(((13^2)/100)^2/99))
```

p-value

```
2 * (1 - pt(abs(T.Welch), df = df))
```

```
## [1] 0.2245626
```

b.

#lower interval

```
delta.hat - qt(.95, df = df) * SE
```

```
## [1] -0.7142111
```

#upper interval

```
delta.hat + qt(.95, df = df) * SE
```

```
## [1] 4.714211
```

c.

I can conclude that the p-value being high enough we can accept null and reject alternative. The confidence interval is includes negative and positive values which means that there is a possibility for an increased performance when looking at their averages. However, even a decreased performance can be seen. Hence, we need a bigger sample size to work with in order to make a clearer conclusion about their average performance having increased or decreased or even remaining the same.

Question 5

a.

I would use t test here because the sample is quite small (12). The assumption made in t test is that the data follows a normal distribution. In test scores, generally, we can observe a normal distribution. So it does come close to satisfying the assumption of this test.

b.

Let u be the expected change. So H0: u0 <or= 0; H1: u1 >= 0

```
treatment <- c(-2.3, -0.7, -0.2, 0.1, 0.5, 0.8, 0.9, 1.6, 2.0, 3.9, 4.5, 6.0)
control <- c(-2.9, -1.5, -0.9, -0.8, -0.7, -0.5, -0.2, 0.2, 0.6, 1.2, 1.9, 2.8)
T.stat <- t.test(treatment, control, alternative = "greater")

change <- mean(treatment) - mean(control)
SE <- sqrt((var(treatment)/12) + (var(control)/12))
df <- ((var(treatment)/12+var(control)/12)^2) / (((var(treatment)/12)^2/11)+((var(con
trol)/12)^2/11))
T.stat <- change / SE
T.stat
```

```
## [1] 1.834125
```

p-value

```
(1 - pt(abs(T.stat), df = df))
```

```
## [1] 0.04119593
```

c.

95% confidence interval

```
print(paste("The upper interval = ", change + qt(.975, df = 11) * SE , " The lower in
terval = ", change - qt(.975, df = 11) * SE ))
```

```
## [1] "The upper interval =  3.28169568038305  The lower interval =  -0.298362347049
717"
```

d.

The above information says that it’s possible the mean changes in VO2 between the treatment and control groups is from -0.29 to 3.2. Although this is based on 24 patients we can say the differences will be within this range. However, it’s also worth noting that the changes are not consistently positive as we can see that it will be even between 0 and -0.29 which means that there is a negative effect possible. One reason for this is that the number of experimental units is low. Perhaps having a larger group to perform this experiment on can give us an interval that is more representative of what kind of an impact whether positive or negative the aerobic exercise is having on people.

Question 6

a.

The experimental unit here is a type 2 diabetes patient. The measurements taken here are of patients’ glycemic index when they had only dates and when they had dates with coffee. This problem is only with one independent sample

b.

The null hypothesis will be H0: u0 = 0 and H1: u0 != 0

```
t.stat <- (11.5 - 0) / (21/sqrt(10))
t.stat
```

```
## [1] 1.731723
```

right-tailed

```
1-pt(t.stat, df = 9)
```

```
## [1] 0.05868355
```

left-tailed

```
pt(t.stat, df = 9)
```

```
## [1] 0.9413165
```

The minimum here is right-tailed

```
2*(1-pt(t.stat, df = 9))
```

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

The probability is decently high for us to accept null and reject alternative hypothesis

C.

The p-value is not high enough first of all to say conclusively that dates' glycemic index changes with or without coffee. Second of all the sample size is not large enough. We need more data.