Problem Set 10

Keith Hickman

November 6, 2017

1. Problem 1

Trosset 11.4 C 1-3 **Item 1 - (problem B 1-3)**

```
chol_a <- c(233, 291, 312, 250, 246, 197, 268, 224, 239, 239, 254, 276, 234, 181, 248, 252, 202, 218, 2
summary(chol_a)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   181.0   222.5   242.5   245.0   257.5   325.0
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   137.0   181.0   207.0   210.3   243.0   344.0
```

(a) What is the experimental unit?

The experimental unit here is the patient.

(b) From how many populations were the experimental units drawn? Identify the population(s). How many units were drawn from each population? Is this a 1- or a 2-sample problem?

The patients belong to one of two populations: Type A and Type B personalities. 20 men were selected who are Type A personality, and 20 men were selected who are Type B personalities. Let X_i note subject i with Type A personality, and Y_j note subject j with Type B personality. $EX = \mu_1$ and $EY = \mu_2$

(c) How many measurements were taken on each experimental unit? Identify them.

The measure we're concerned with (cholesterol) was taken once on each unit after the person was selected for the study.

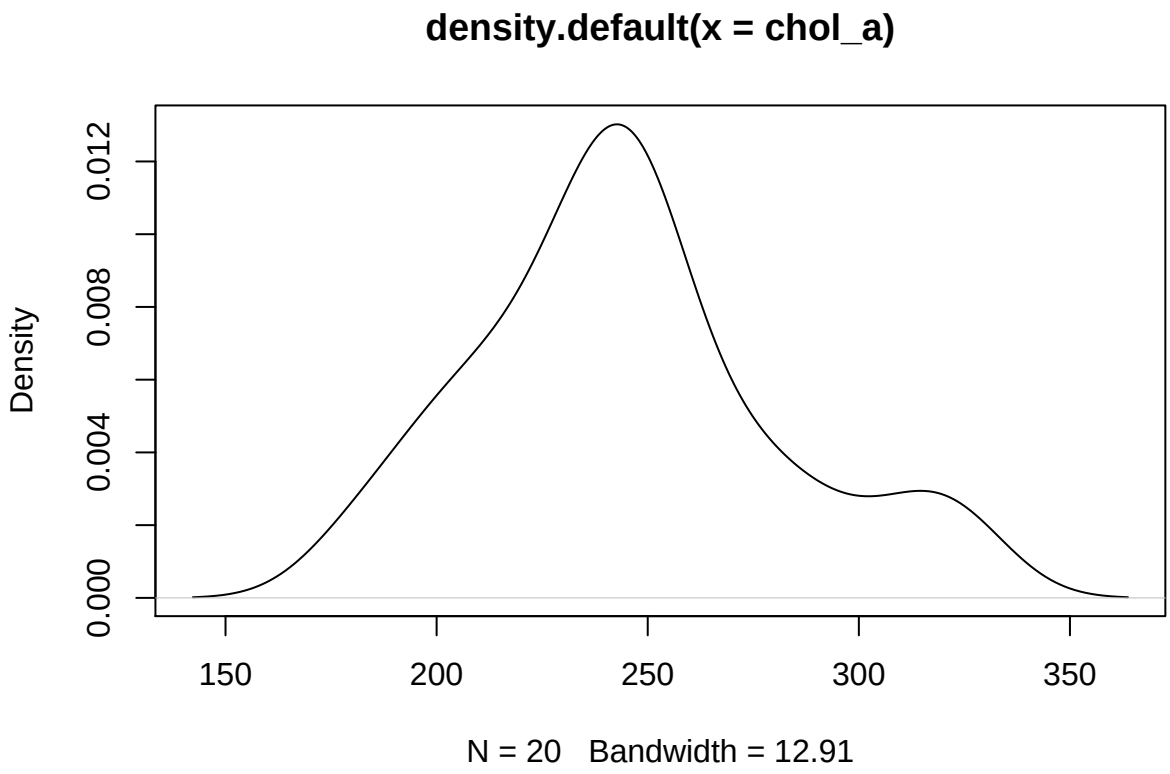
(d) Define the parameter(s) of interest for this problem. For 1- sample problems, this should be μ; for 2-sample problems, this should be δ.

The parameter of interest is mean δ. We are interested in whether $\mu_1 - \mu_2 > 0$.

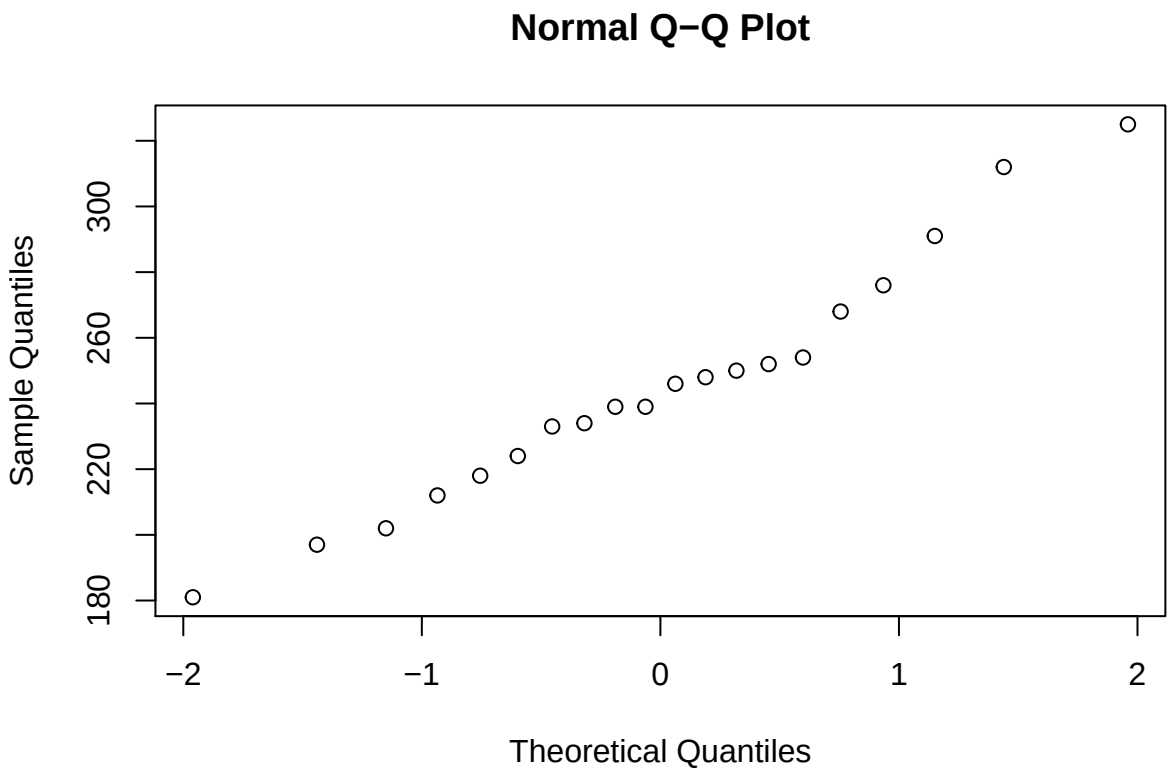
(e) State appropriate null and alternative hypotheses. Let's couch the problem by considering whether personality type has a measureable effect on cholesterol levels; namely that Type A (P_1) will have higher cholesterol than Type B (P_2). Stated as the alternative hypothesis $H_1 : \mu_1 - \mu_2 > 0$, which is the same as saying $H_1 : \Delta > 0$. This leaves us with $H_0 : \Delta \leq 0$.

Trosset 11.4 C Item 2 2. Is it reasonable to assume the two populations drawn from normal distributions - why or why not?
With a relatively small sample size for both X and Y, it might be a bit of a risk to assume a completely normal distribution. We can, however depict the distribution of our samples with density and QQnorm plots. The kernel density estimates indicate approximate normality, but the qq plots do show the presence of outliers (possibly only one), in variable Y or Type B personality group. We could either eliminate that from our variable, choose median as our parameter, transform the variable, or use a non-parametric model. The presence of outliers in a small sample makes the assumption of normality less plausible.

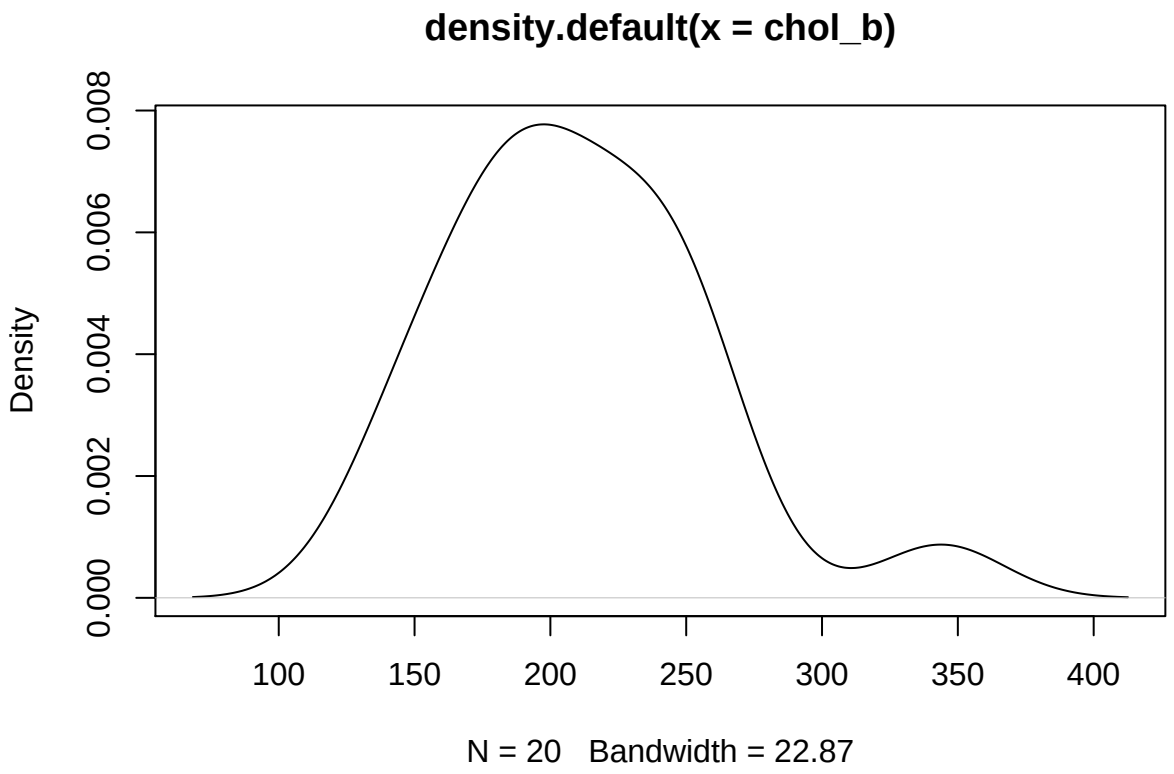
```
plot(density(chol_a))
```



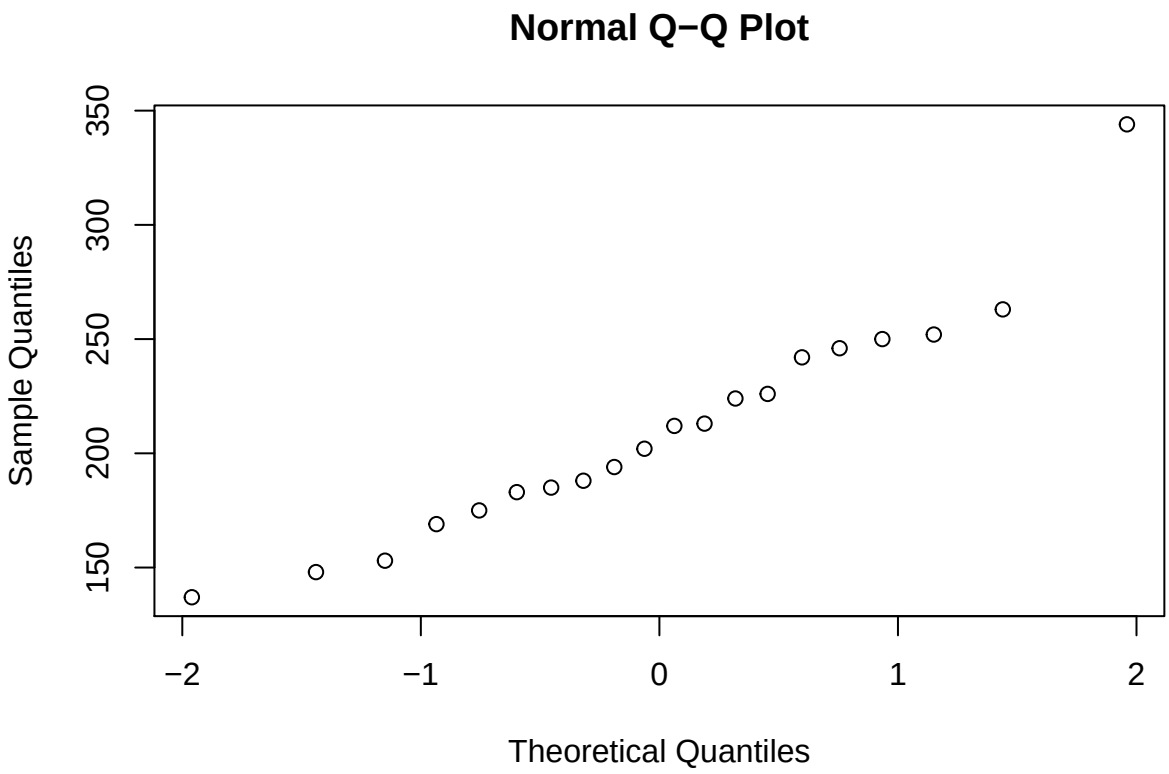

```
qqnorm(chol_a)
```




```
plot(density(chol_b))
```




```
qqnorm(chol_b)
```



Trosset 11.4 C Item 3 a. Test the null hypothesis using Welch's approximate t-test. What is the significance probability? Should we reject the null at .05? In order to conduct the test, we need to find a point estimate and standard error, find the t-statistic, the degrees of freedom, and then a P-value.

Point Estimate:

```
delta.hat <- mean(chol_a) - mean(chol_b)
delta.hat
```

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

Standard Error of Delta Hat:

```
se <- sqrt(var(chol_a)/length(chol_a) + var(chol_b)/length(chol_b))
se
```

```
## [1] 13.56303
```

Degrees of Freedom:

```
nu <- (var(chol_a)/20 + var(chol_b)/20)^2/((var(chol_a)/20)^2/19 + (var(chol_b)/20)^2/19)
nu
```

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

T statistic

```
t.Welch <- delta.hat/se
t.Welch
```

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

```r
p.value <- 2 * (1 - pt(abs(t.Welch), 35.413))
p.value
```

```
## [1] 0.01481051
```

```r
t.test(chol_a, chol_b)
```

```
##
##  Welch Two Sample t-test
##
## data:  chol_a and chol_b
## t = 2.5621, df = 35.413, p-value = 0.01481
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   7.227071 62.272929
## sample estimates:
## mean of x mean of y
##    245.05    210.30
```

b. Construct a 2-sided confidence interval for Δ with a confidence coefficient of .90.

```r
q <- qt(.95, nu)
lower.90 <- delta.hat - q*se
upper.90 <- delta.hat + q*se
lower.90
```

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

```r
upper.90
```

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

Our 90% confidence interval for Δ is 11.84 to 57.65.

With a p-value of .01481 and significance level of .05, we do not have enough evidence to reject the null hypothesis out-of-hand.

Problem 2

Trosset 11.4 D 1-4 data:

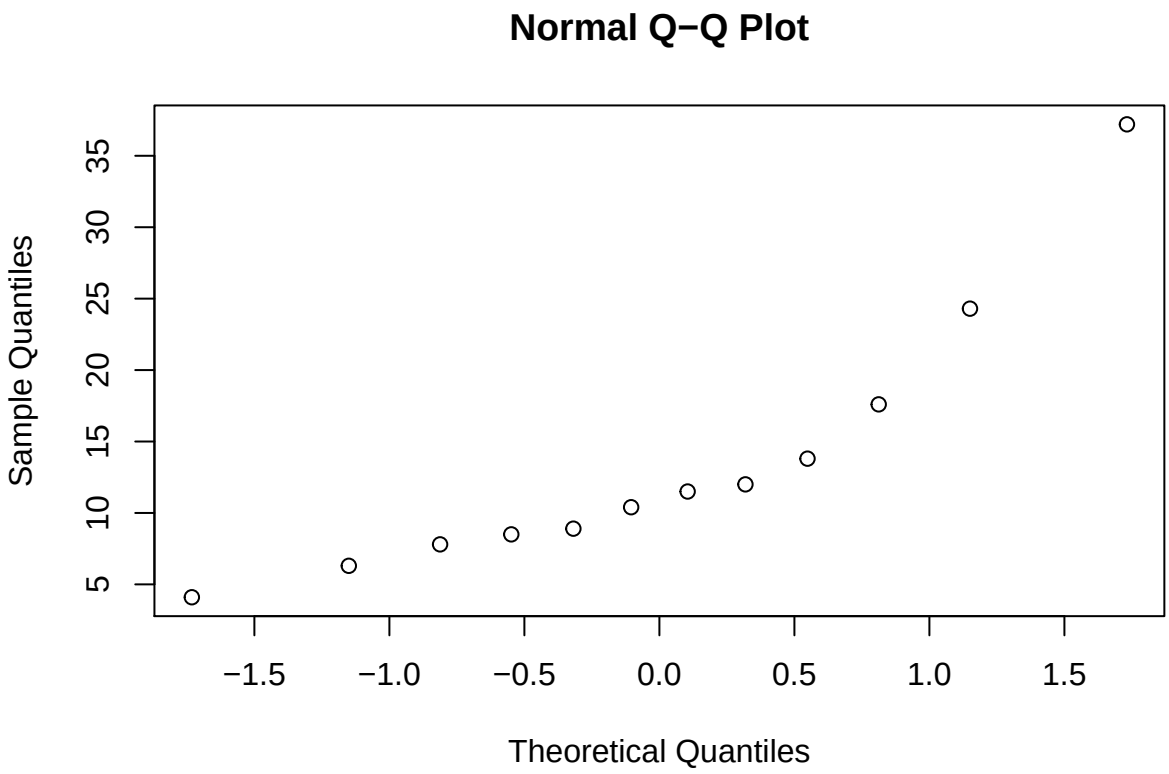
```r
normal <- (c(4.1, 6.3, 7.8, 8.5, 8.9, 10.4, 11.5, 12.0, 13.8, 17.6, 24.3, 37.2))
diabetic <- (c(11.5, 12.1, 16.1, 17.8, 24.0, 28.8, 33.9, 40.7, 51.3, 56.2, 61.7, 69.20))
summary(normal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.100   8.325  10.950  13.530  14.750  37.200
```

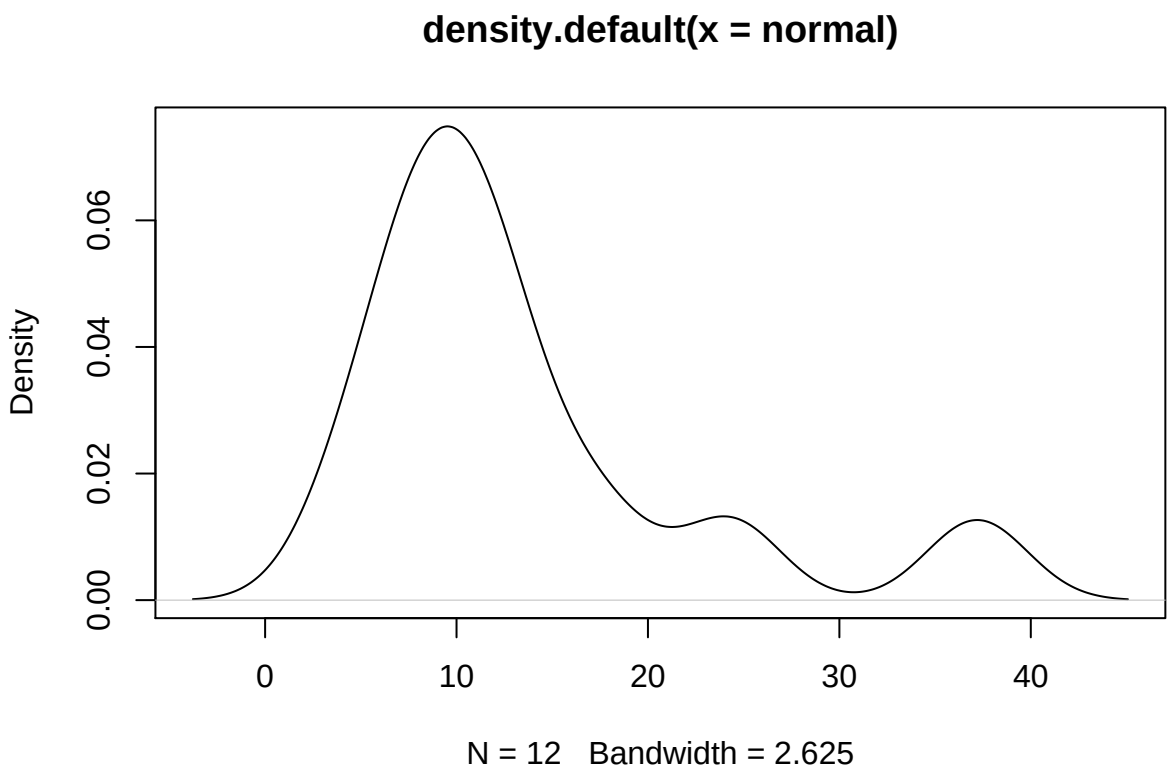
```r
summary(diabetic)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.50   17.38   31.35   35.28   52.52   69.20
```

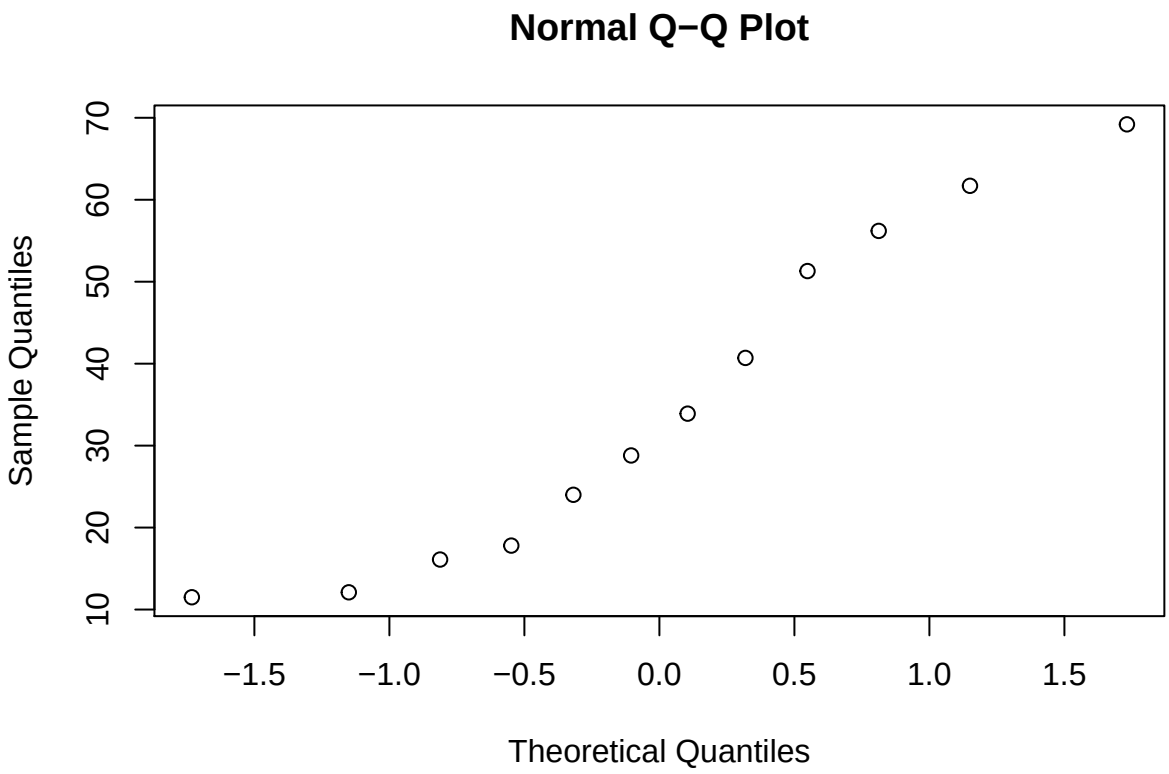
```r
qqnorm(normal)
```



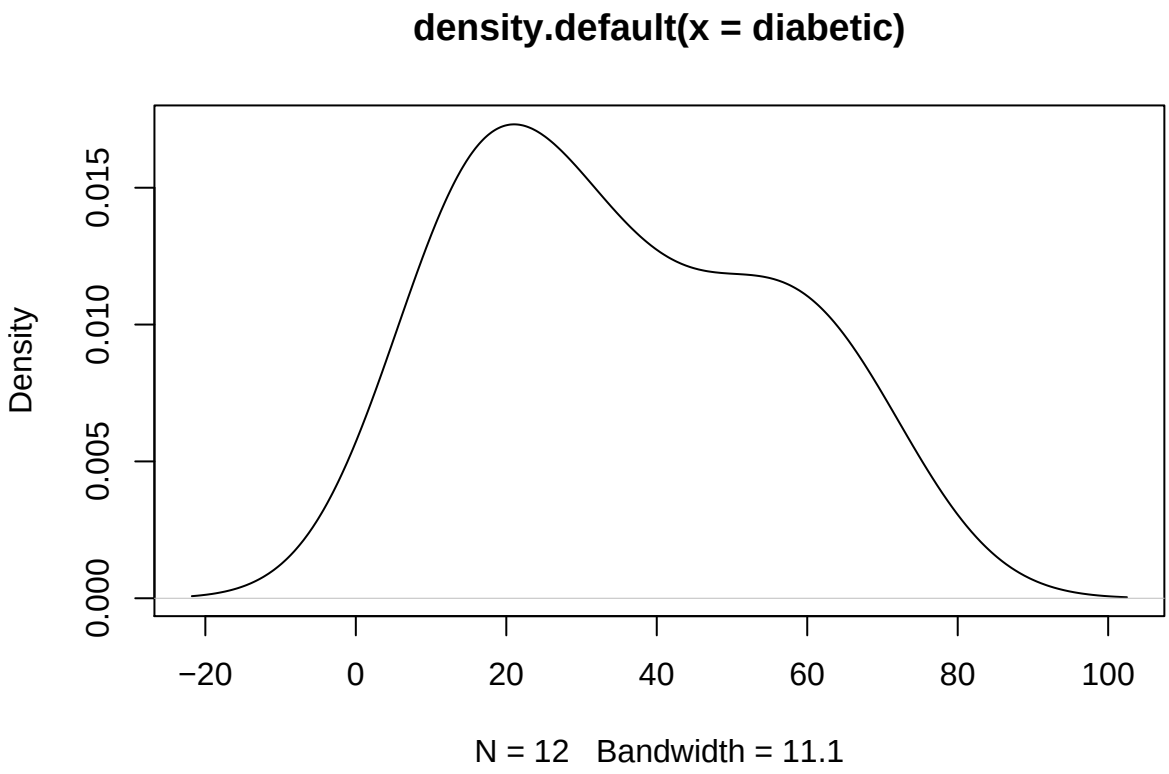

```
plot(density(normal))
```




```
qqnorm(diabetic)
```



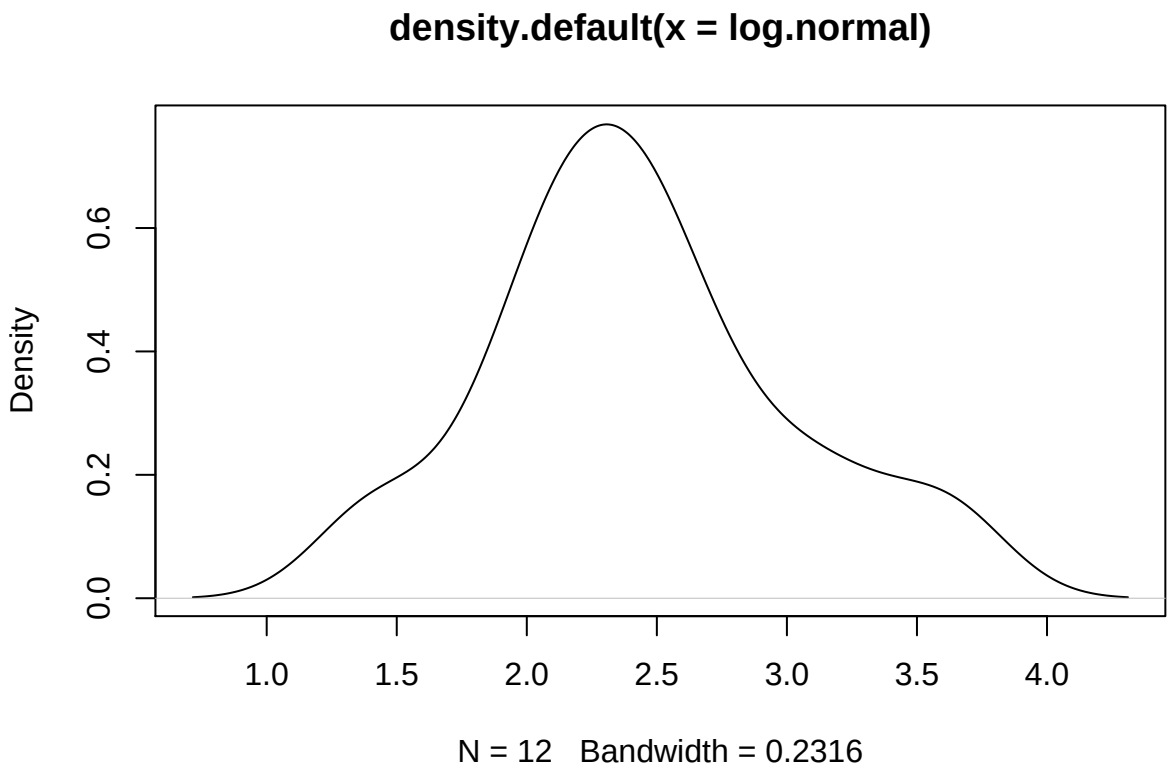

```
plot(density(diabetic))
```



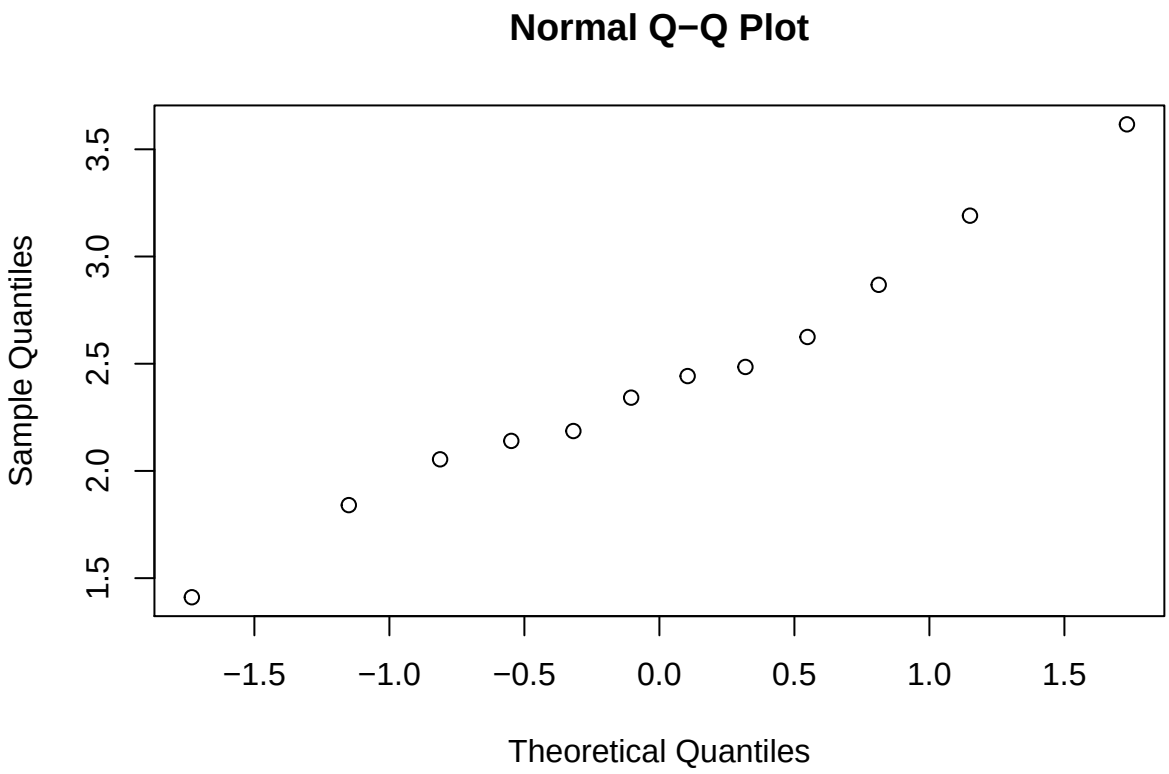
(A) These distributions are almost certainly not symmetric about a mean. Both the `normal` and `diabetic` variables have values which begin to approximate right-skewed distributions. Both distributions do have this in common, however. Another note is that the values are on different scales, so some normalization would be necessary.

(B) Natural log of each and square root of each.

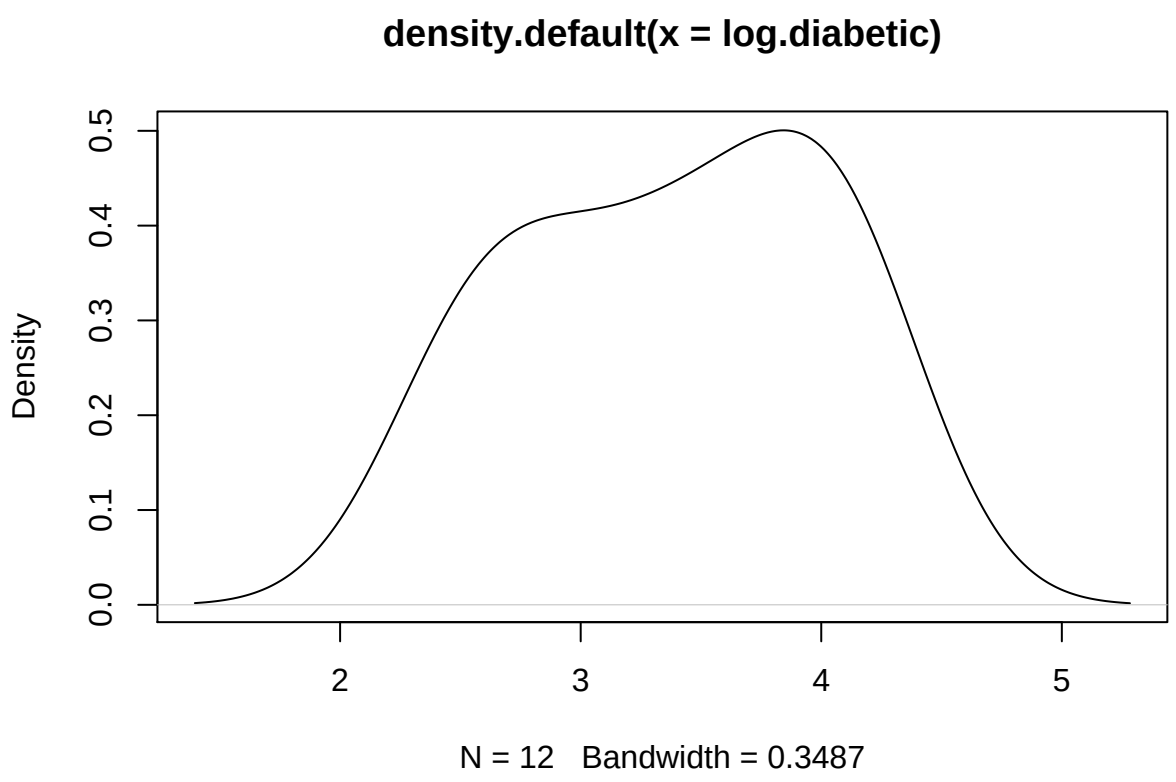
```
log.normal <- log(normal)
log.diabetic <- log(diabetic)
plot(density(log.normal))
```



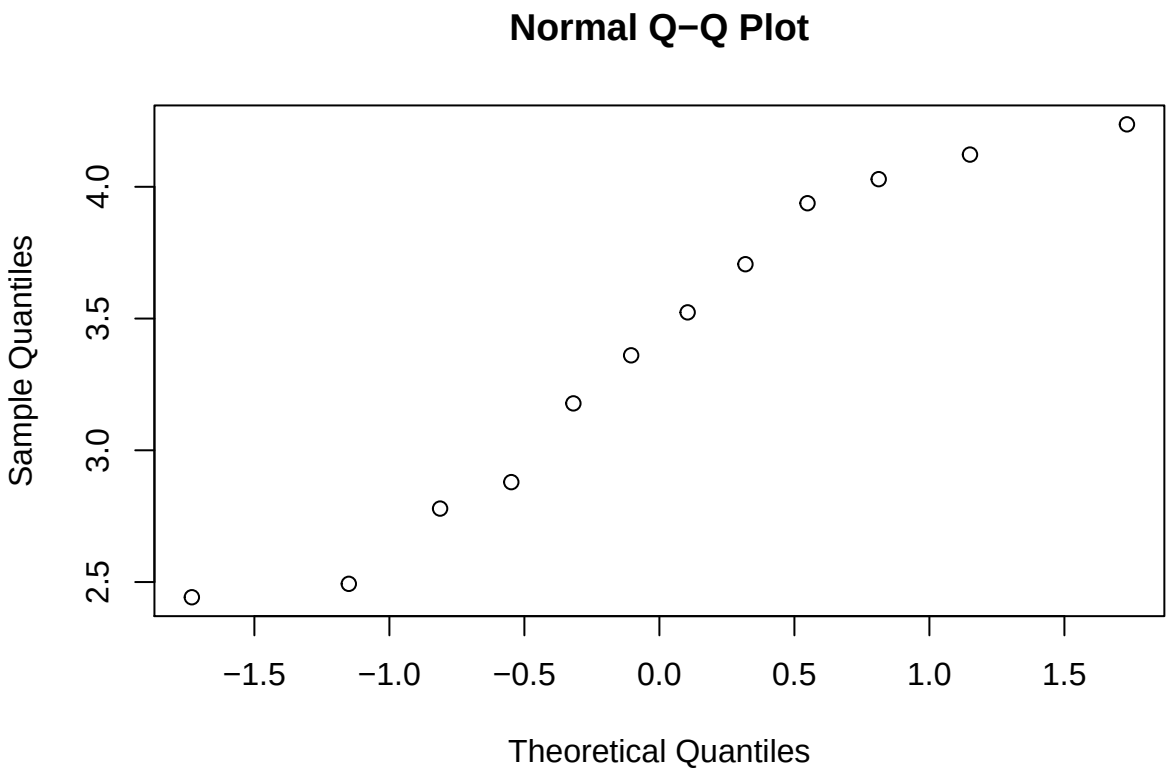

```
qqnorm(log.normal)
```



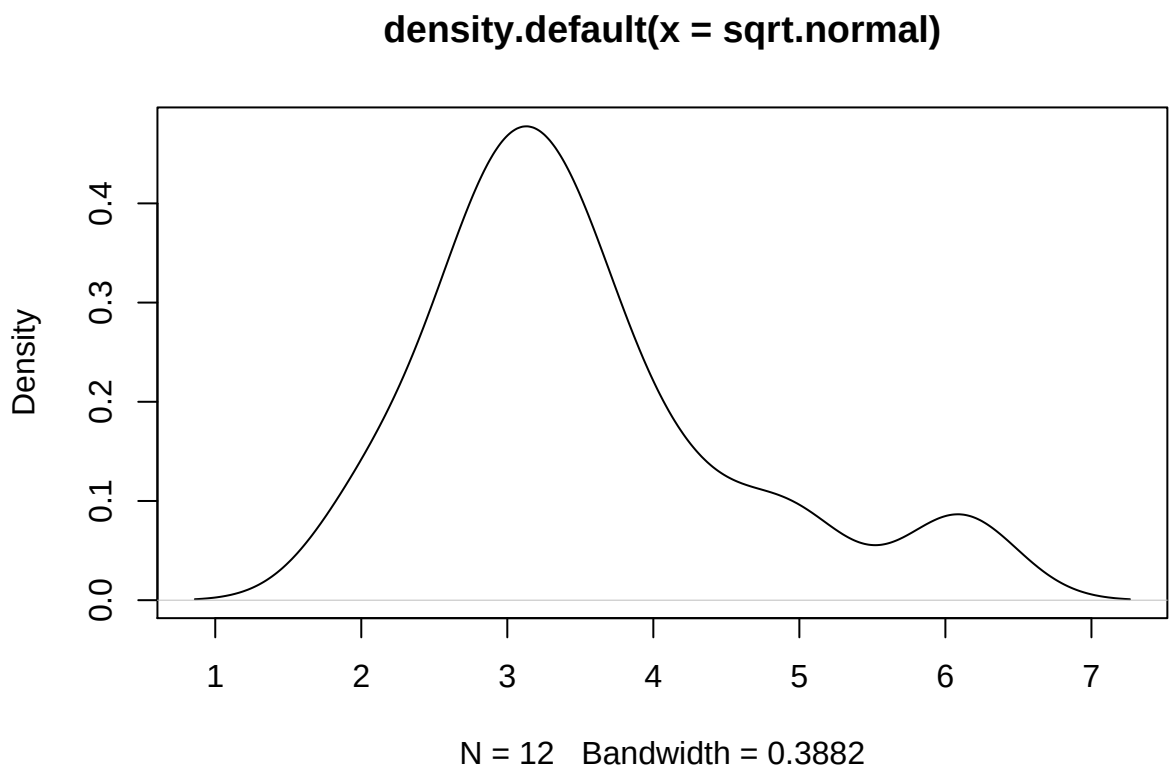

```
plot(density(log.diabetic))
```



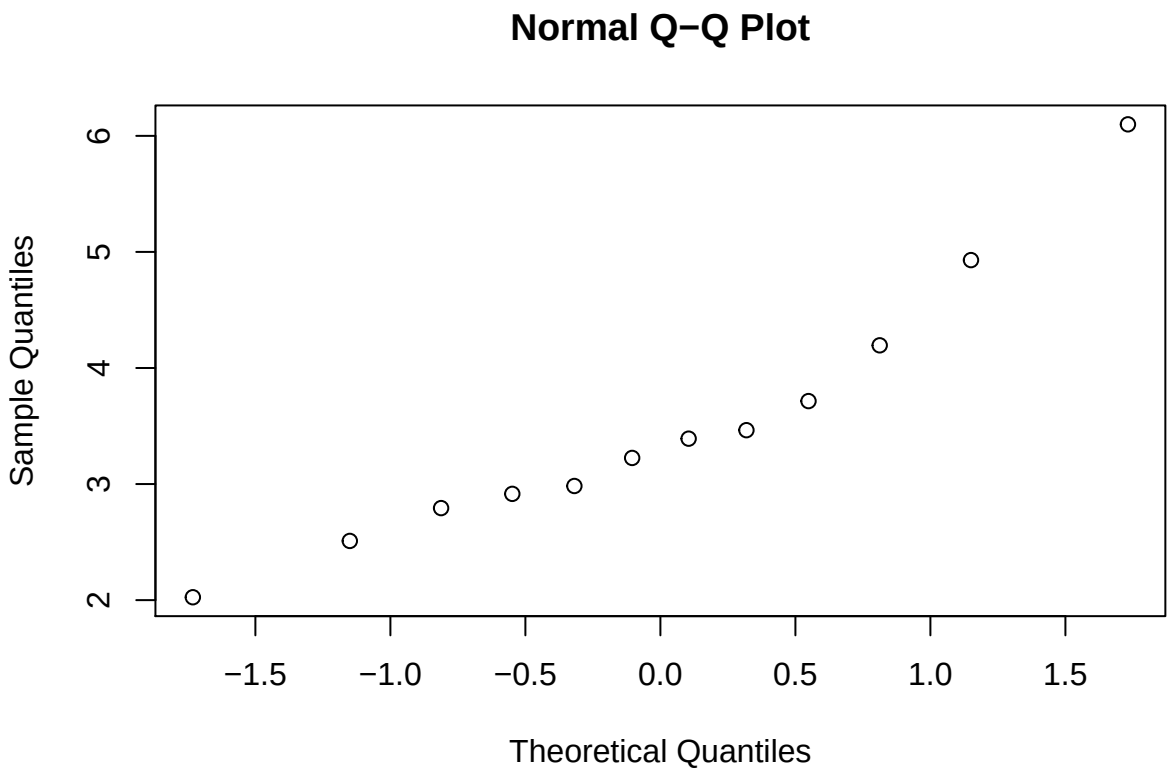

```
qqnorm(log.diabetic)
```



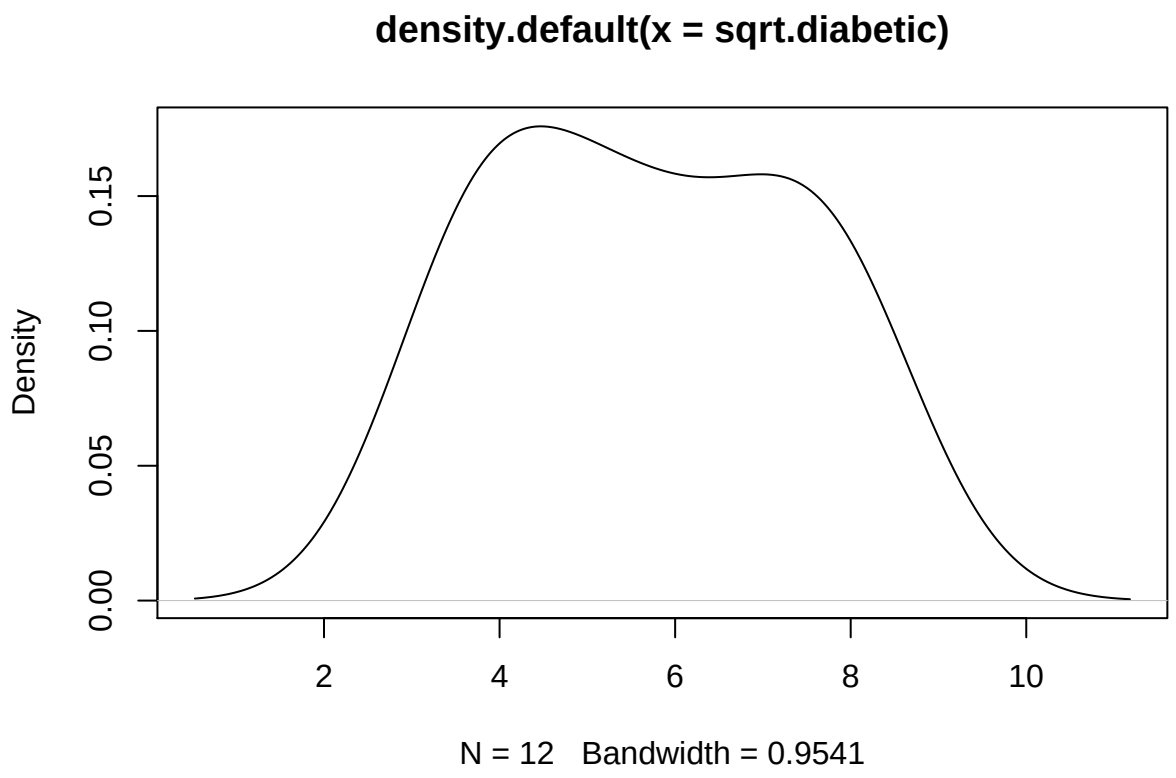

```
sqrt.normal <- sqrt(normal)
sqrt.diabetic <- sqrt(diabetic)
plot(density(sqrt.normal))
```



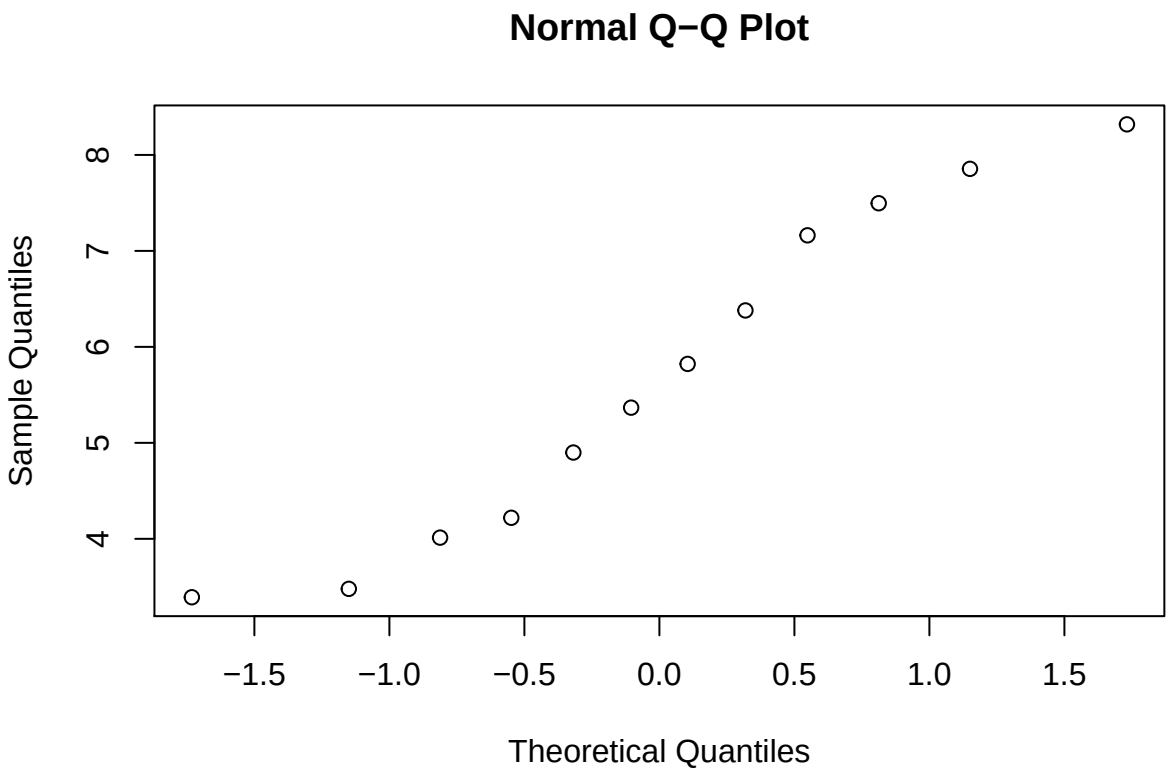

```
qqnorm(sqrt.normal)
```




```
plot(density(sqrt.diabetic))
```




```
qqnorm(sqrt.diabetic)
```



In both of these cases, the distributions are closer approxmations of normal distributions. The `diabetic` variable never gets close enough to symmetrical, though the other variable (`normal`) does. I would choose the log transformation here, as it gets the `normal` variable much closer to a symmetric distribution. Even though the transformed variables appear closer to symmetric, the underlying variables are clearly not. (I realized a bit too late that I probably should have used a less-confusing variable name.)

(3) The transformed measurements do appear to be approaching normal variables, as they are symmetric about a mean. Additionally, both variables do not have gross outliers and are close to a straight line on the qq plot.

(4) This is a two-population problem where we need to find the difference between the two variables, X and Y, represented by the `diabetic` and ‘`normal` variables in R above. Our goal is to find out whether the patients (experimental unit) with diabetes have increased thromboglobulin levels (what’s being measured), which will provide our alternative hypothesis. The null hypothesis, $H_0 : \Delta_0 = 0$, and the Alternative is that $H_1 : \Delta_1 \neq 0$, or that patients with diabetes do not have increased thermoglobulin levels.

Since we don’t know the population variances, we can estimate with a Welch’s t-test. First, we find Δ:

Delta hat:

```
p2delta.hat <- mean(log.diabetic) - mean(log.normal)
p2delta.hat
```

```
## [1] 0.9572787
```

The standard error of delta hat:

```
p2.se <- sqrt(var(log.diabetic)/length(log.diabetic) + var(log.normal)/length(log.normal))
p2.se
```

```
## [1] 0.2516458
```

Degrees of Freedom:

```
p2.nu <- (var(log.diabetic)/12 + var(log.normal)/12)^2/((var(log.diabetic)/12)^2/11 + (var(log.normal)/
p2.nu
```

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

The t-statistic and p-value:

```
p2t.Welch <- p2delta.hat/p2.se
p2t.Welch
```

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

```
p2p.value <- 2 * (1 - pt(abs(p2t.Welch), 21.8992))
p2p.value
```

```
## [1] 0.0009776511
```

Now we can transform our variables back to the raw value scale via the exponent function and perform the t-test:

```
t.test(log.normal, log.diabetic)
```

```
##
##  Welch Two Sample t-test
##
## data:  log.normal and log.diabetic
## t = -3.8041, df = 21.9, p-value = 0.0009776
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4792986 -0.4352589
## sample estimates:
## mean of x mean of y
##  2.433349  3.390628
```

Based on a small p-value of .003, I would consider this enough compelling evidence to invesgigate further. We can't reject the null hypothesis outright because of the small n of 12 in each sample.

Problem 3:

Trosset 11.4 E part 3

Let's read in the data:

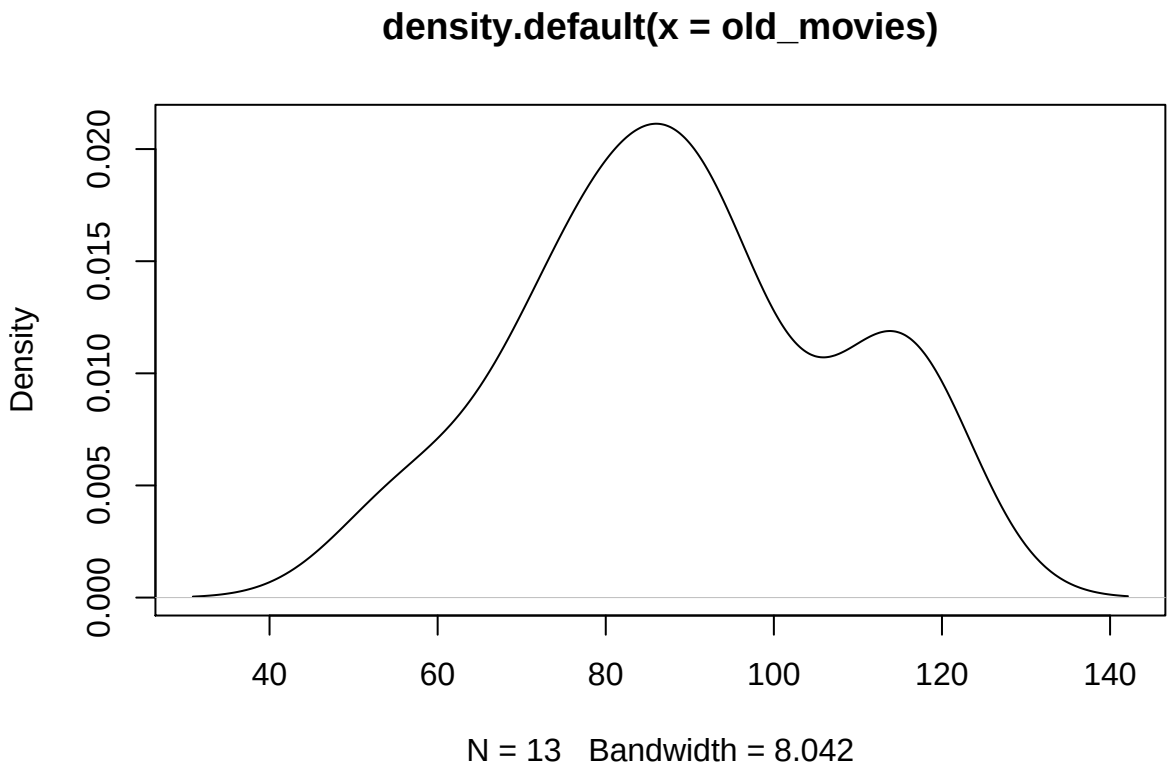
```
old_movies <- c(74, 114, 114, 87, 92, 55, 67, 118, 79, 79, 92, 99, 87)
new_movies <- c(70, 98, 90, 95, 88, 108, 110, 96, 91, 88, 120, 96, 90, 90)
summary(old_movies)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      55      79      87      89      99     118
```

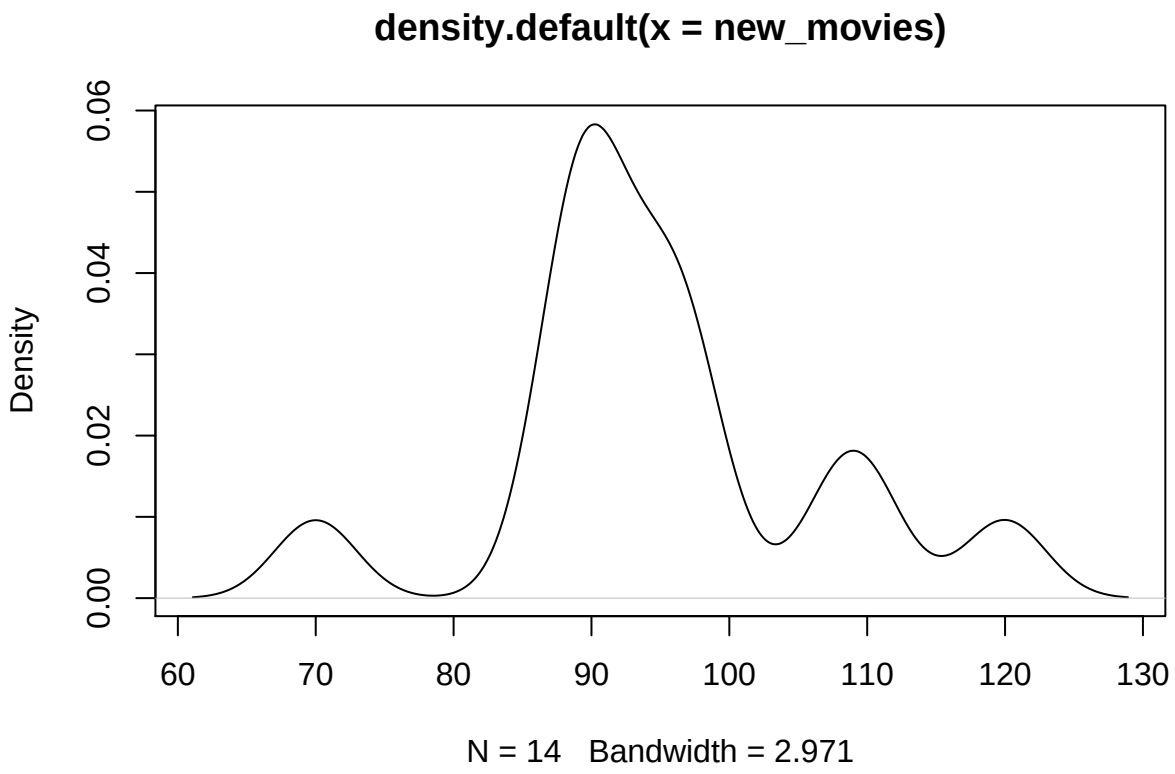
```
summary(new_movies)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    70.0    90.0    93.0    95.0    97.5   120.0
```

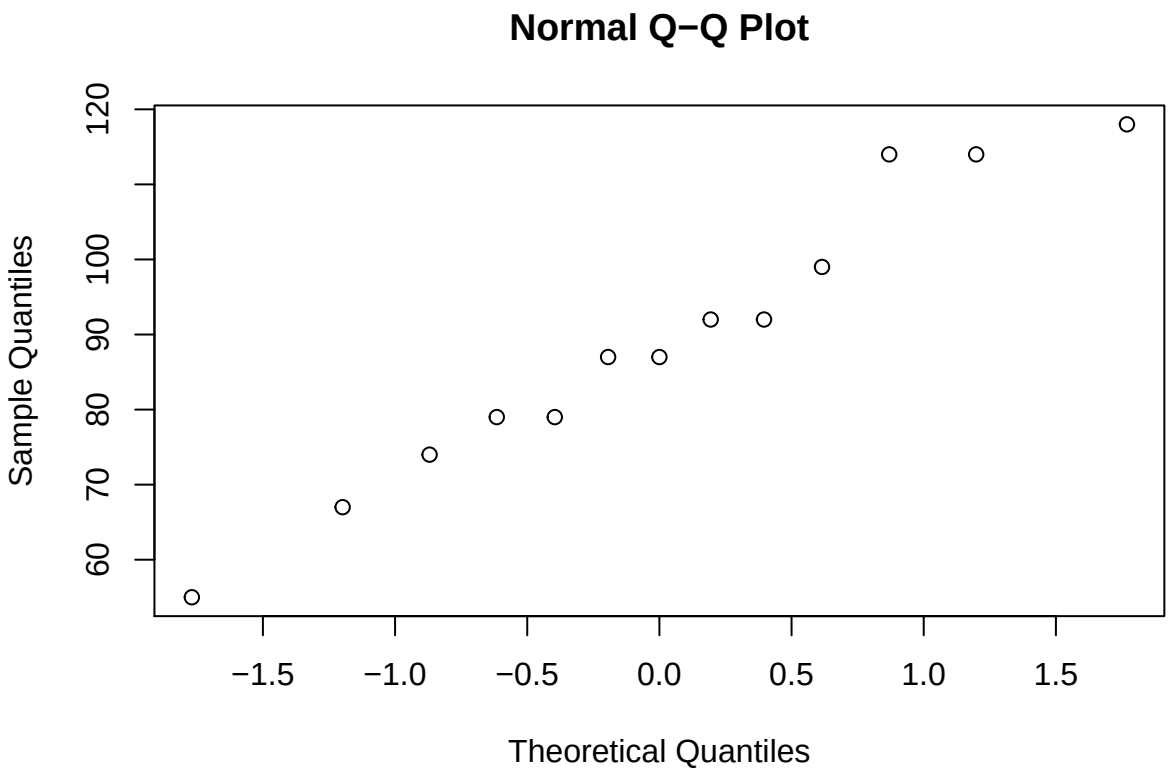
```
plot(density(old_movies))
```



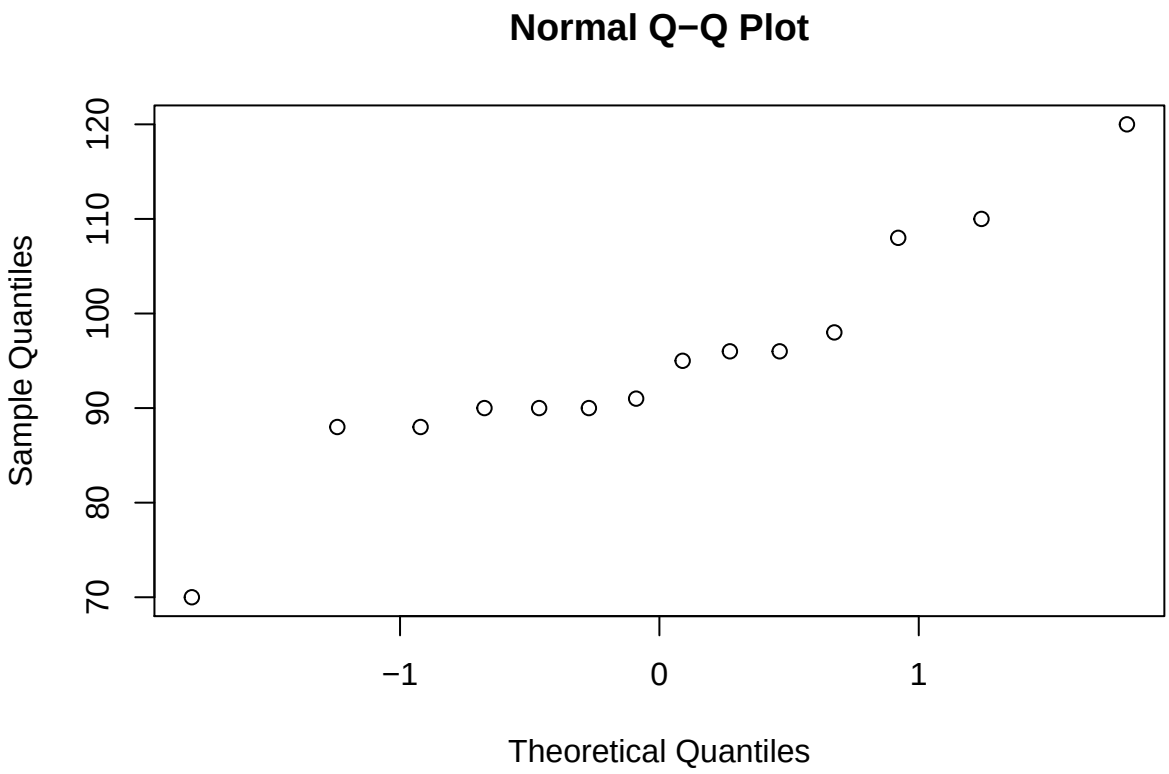

```
plot(density(new_movies))
```




```
qqnorm(old_movies)
```




```
qqnorm(new_movies)
```



The two variables we'd like to compare, `old_movies` and `new_movies` are both somewhat symmetrical and normal, though both have outliers and it appears that `old_movies` might be bi-modal. The variables will require transformation.

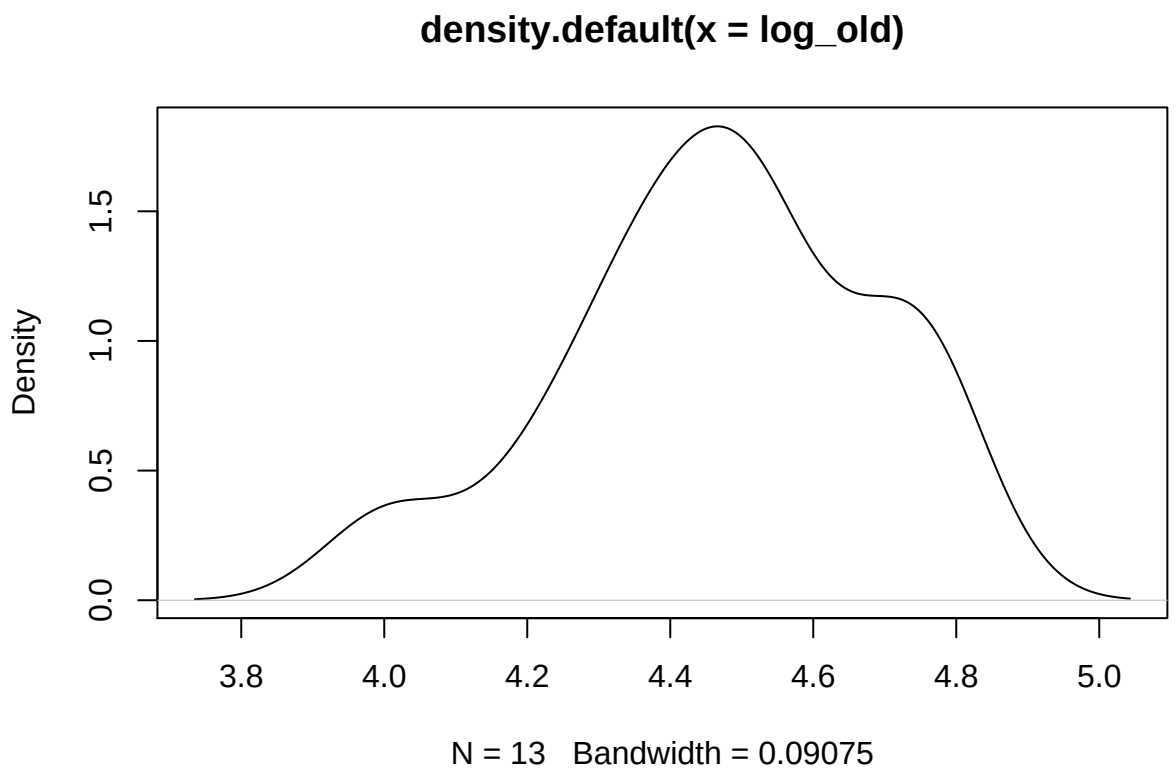
```
log_old <- log(old_movies)
log_new <- log(new_movies)

summary(log_old)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.007   4.369   4.466   4.467   4.595   4.771
```

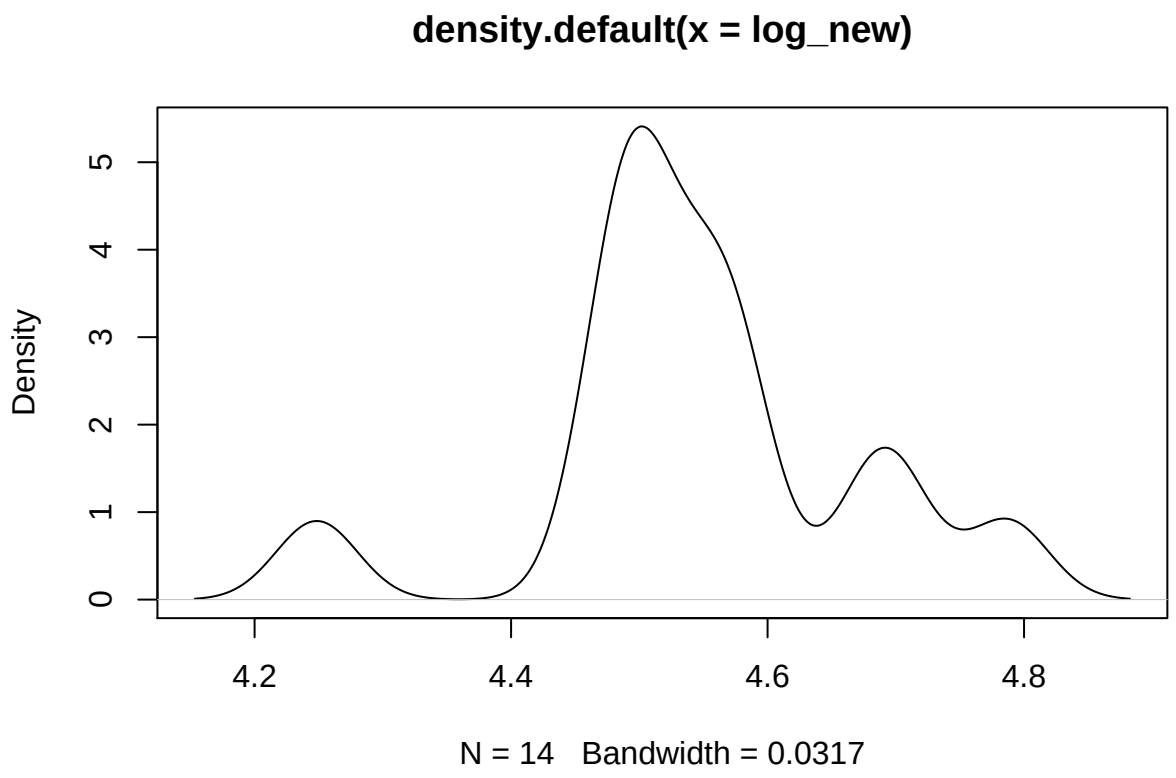
```
summary(log_new)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.248   4.500   4.532   4.547   4.580   4.787
```

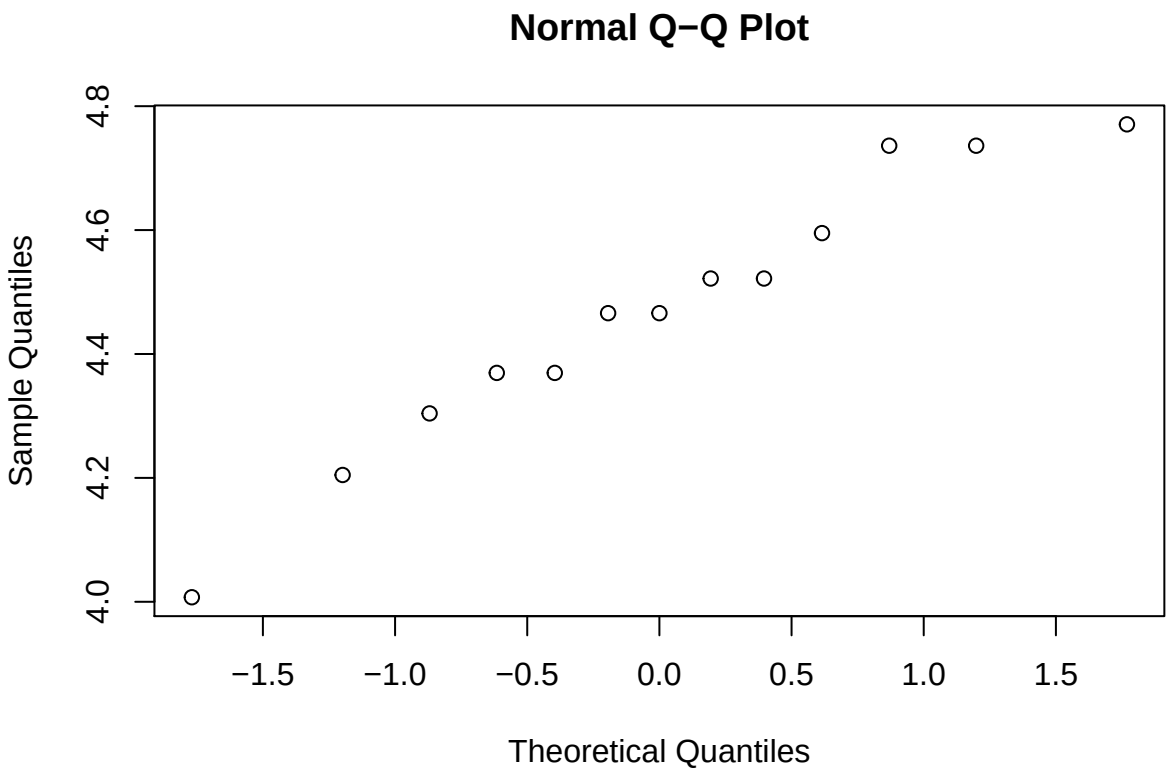
```
#sqrt_old <- sqrt(old_movies)
#sqrt_new <- sqrt(new_movies)

plot(density(log_old))
```



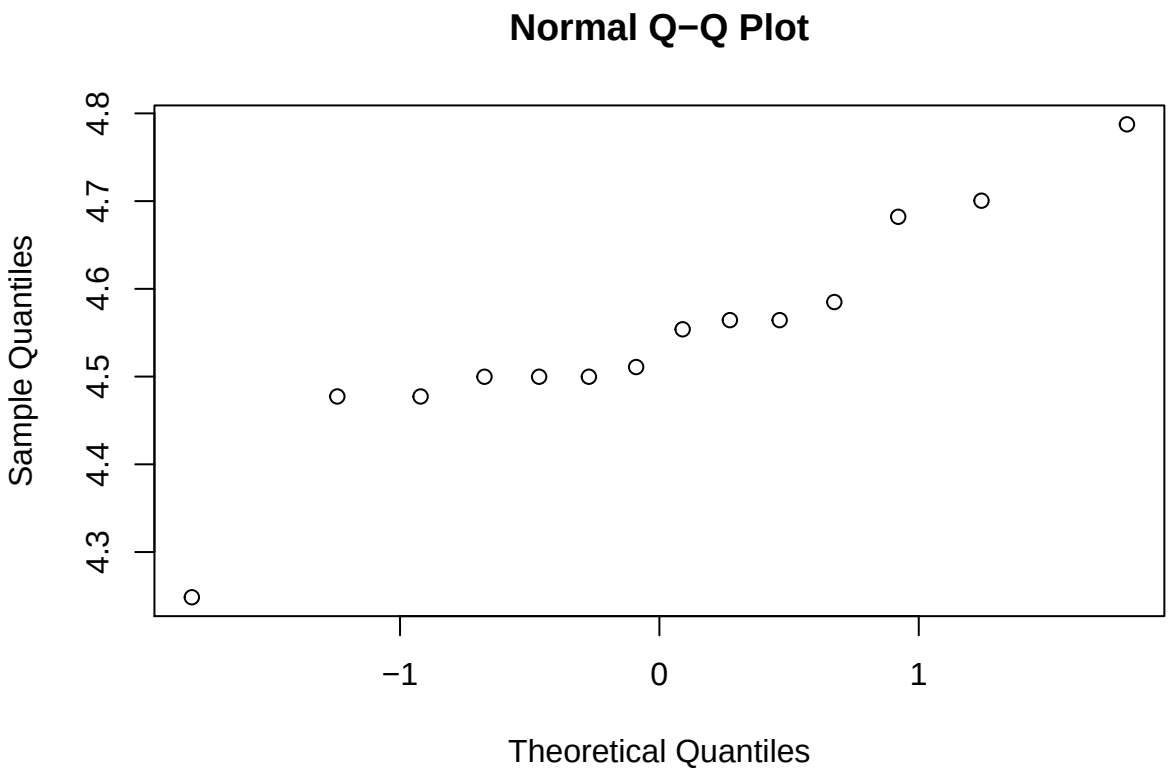

```
plot(density(log_new))
```




```
qqnorm(log_old)
```




```
qqnorm(log_new)
```




```
#plot(density(sqrt_old))
#plot(density(sqrt_new))
```

With the log transforms, the variables start to approach normality and symmetry. We can carry on with our analysis under the assumption of approximate normality. We're interested in the change in movie run times between these two populations of movies. We don't know the entire population variance, so we can use the Welch's t-test, avoiding the Student's t-test yet again. The variable X represents (log) runtimes of 1996 movies (log_new) and the variable Y represents runtimes of 1956 movies (log_old), again transformed. Our experimental unit is an individual movie, and the measurement taken is runtime. The samples were taken from one population, and both are identically distributed. Finally, the parameter of interest is the difference between the runtimes or Δ.

We're trying to show whether movies made or released in 1996 actually run longer than movies made in 1956, which is the Null hypothesis: $H_0 : \Delta = 0$ and alternative hypothesis $H_1 \Delta \neq 0$.

First, find Δ:

```
p3delta_hat <- mean(log_new) - mean(log_old)
p3delta_hat
```

```
## [1] 0.0796933
```

The standard error of the log variables:

```
p3.se <- sqrt(var(log_new)/length(log_new) + var(log_old)/length(log_old))
p3.se
```

```
## [1] 0.06998198
```

The degrees of freedom:

```
p3.nu <- (var(log_old)/14 + var(log_new)/14)^2/((var(log_old)/14)^2/13 + (var(log_new)/14)^2/13)
p3.nu
```

```
## [1] 20.71681
```

Finally, we're ready for the t-test and p-value and to construct a confidence interval:

```
p3t.Welch <- p3delta_hat/p3.se
p3t.Welch
```

```
## [1] 1.138769
```

```
p3p.value <- 2 * (1 - pt(abs(p3t.Welch), p3.nu))
p3p.value
```

```
## [1] 0.2677972
```

The p-value looks pretty large at .26. Even if we assume a significance level of .10, this p-value is much larger than I would be comfortable rejecting the null hypothesis.

```
t.test(log_old, log_new)
```

```
##
##  Welch Two Sample t-test
##
## data:  log_old and log_new
## t = -1.1388, df = 18.828, p-value = 0.2691
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.22625782  0.06687122
## sample estimates:
## mean of x mean of y
##  4.466814  4.546507
```

```
## and just for fun,
t.test(old_movies, new_movies)
```

```
##
##  Welch Two Sample t-test
##
## data:  old_movies and new_movies
## t = -0.97874, df = 19.981, p-value = 0.3394
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.788367   6.788367
## sample estimates:
## mean of x mean of y
##        89        95
```

Do we even need to construct a confidence interval when our p-value is so large? Probably not, but I'm going to do it anyway.

```
q = qt(0.975, df = p3.nu)
lower = p3delta_hat - q * p3.se
upper = p3delta_hat + q * p3.se
lower
```

```
## [1] -0.06596344
```

```
upper
```

```
## [1] 0.22535
```

To construct the confidence interval back to the original scale, we need to calculate the exponent of the confidence interval:

```
exp(c(-.22535, .06596344))
```

```
## [1] 0.7982368 1.0681877
```

This is an interesting result to find a p-value that is clearly larger than any significance level we might adopt. Examining the summary values at the beginning of this problem indicated that the movies in the 1996 sample were longer than the 1956 sample by quite a bit.