S520 Problem Set 10 Solutions

Arturo Valdivia

Due on 04/04/2023

Q1

a. We do not know whether distance traveled is normal, but it's not necessary for our test, as we are mainly interested in the mean distance traveled. Given the large sample sizes, the sample mean would be approximately normal even if the original population has a distribution that is not close to normal.

b.
```
xbar = 23.4 #orange
ybar = 21.9 #blue
s1 = 5.7 #orange
s2 = 7.2 #blue
n1 = 235 #orange
n2 = 197 #blue
SE = sqrt(s1^2/n1 + s2^2/n2)
Deltahat = xbar - ybar

nu.hat = (s1^2/n1+s2^2/n2)^2/((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
alpha = 1 - 0.98
q = qt(1 - alpha/2, nu.hat)

Deltahat - q*SE
```

```
## [1] 0.01970644
```

```
Deltahat + q*SE
```

```
## [1] 2.980294
```

We are 98% confident that the difference in average distance traveled (orange - blue) is between 0.02 and 2.98 feet.

c. We should use Welch's 2-sample t-test as we do not know whether the population variances are equal (clearly the sample variances are not)

d. No precise guidelines were given in terms of what the researcher wanted to find. So, let's find whether different colors produce different results. The hypotheses would be $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$. The test is:

```
t.w  = (Deltahat -  0)/SE
t.w
```

```
## [1] 2.367561
```

```
2*(1 - pt(abs(t.w), nu.hat))
```

```
## [1] 0.0184189
```

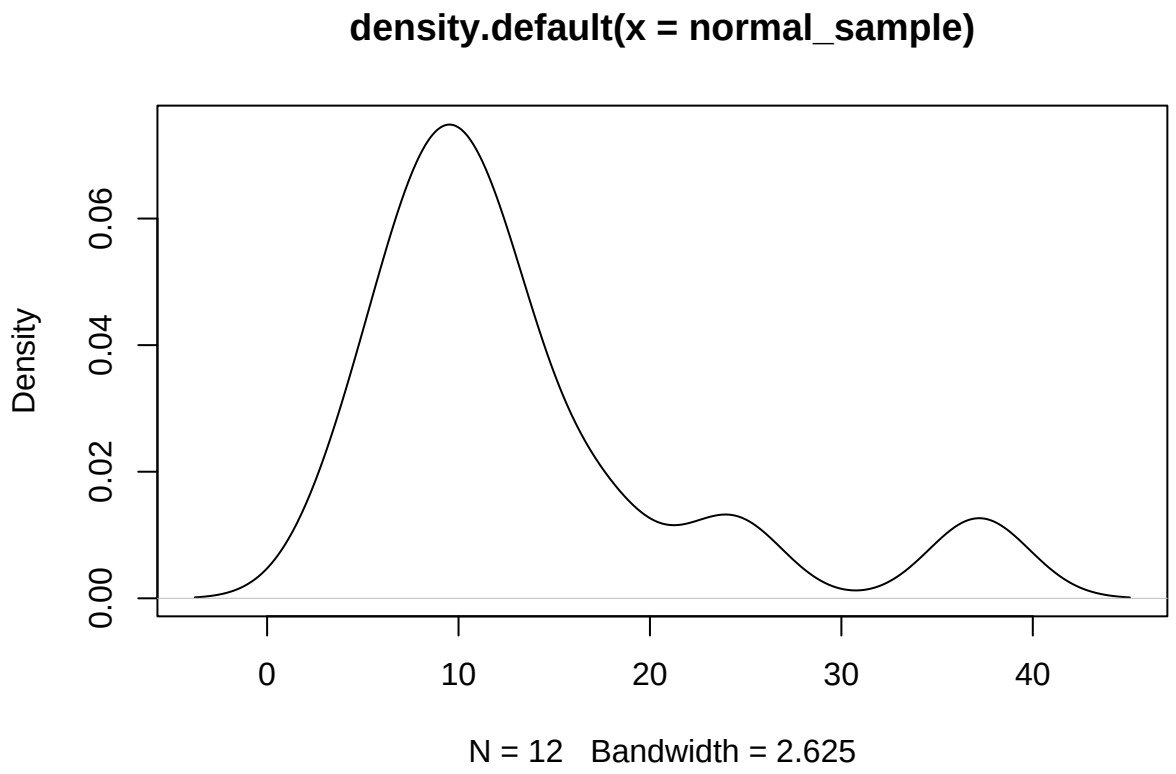
We reject the null hypothesis. There is enough evidence to conclude that color background may enhance online readability.

Q2. Problem Set D.

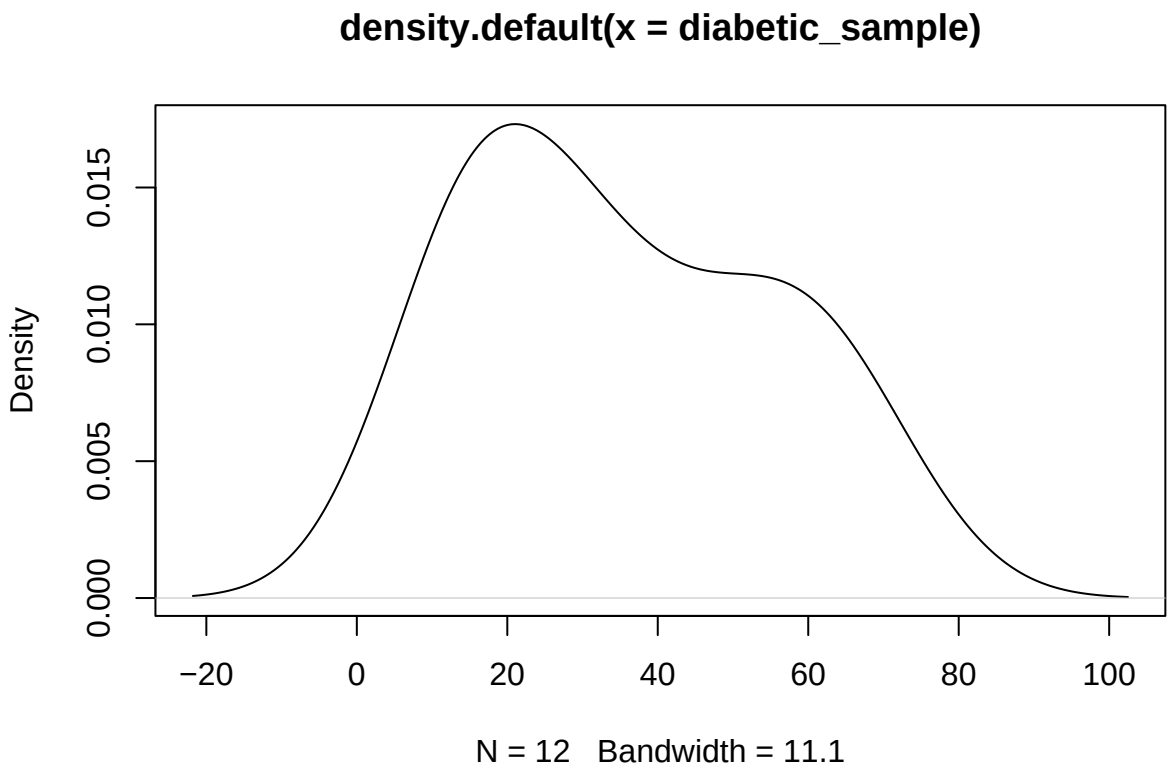
```
normal_sample <- c(4.1,6.3,7.8,8.5,8.9,10.4,11.5,12.0,13.8, 17.6,24.3, 37.2)
diabetic_sample <- c(11.5,33.9,12.1,40.7,16.1,51.3,17.8,56.2,24.0,61.7,28.8,69.2)
```

2.1.

```
plot(density(normal_sample))
```



```
plot(density(diabetic_sample))
```

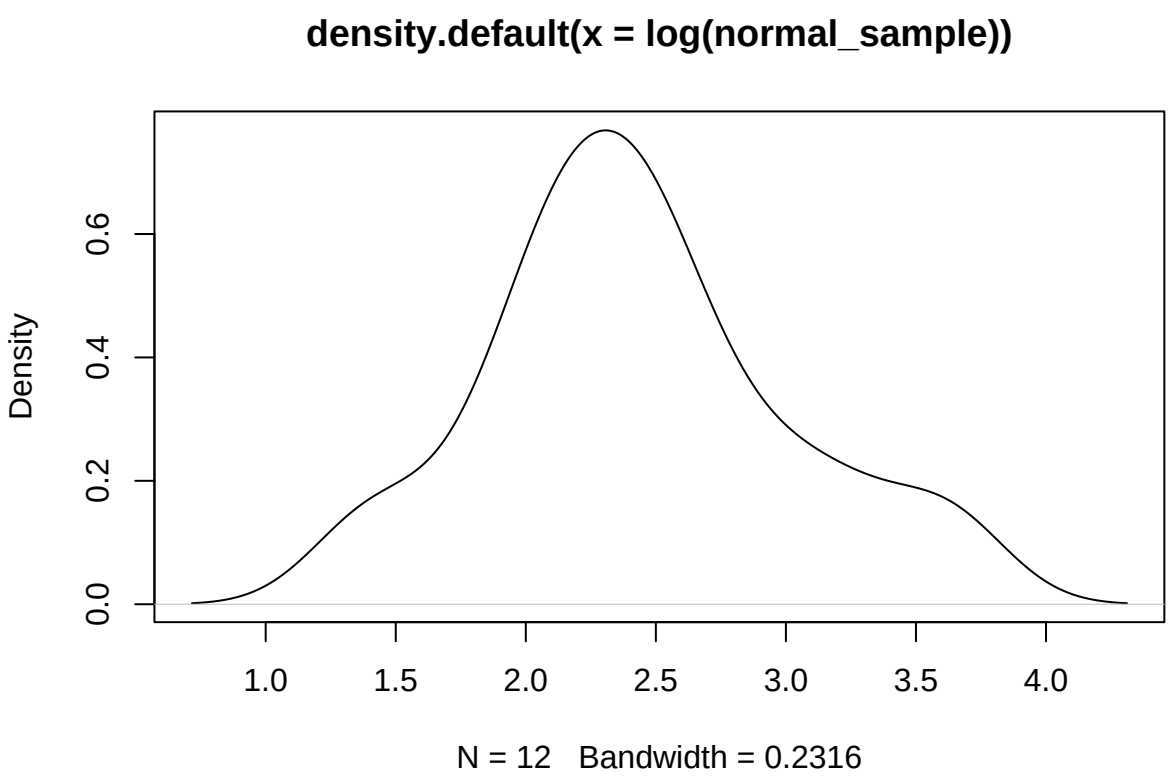


They do not seem to be sampled from a symmetric distribution because of a tail on the right side. They are right skewed.

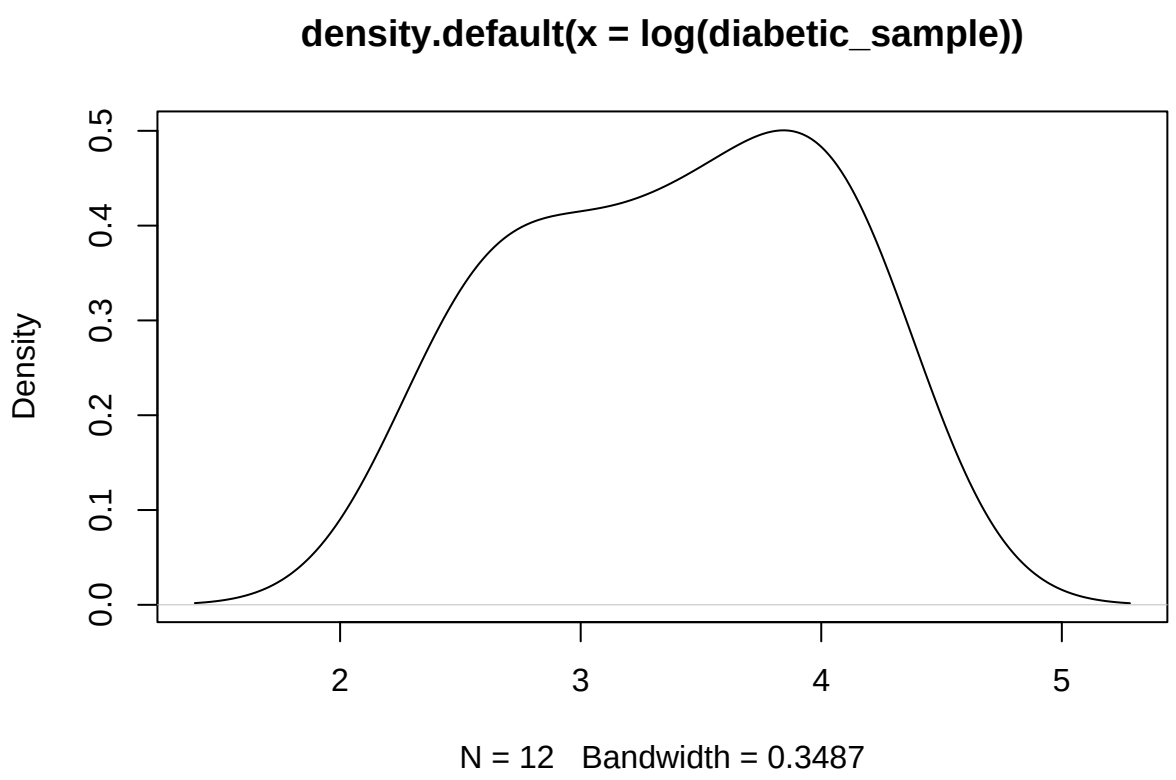
2.2.

Log transform:

```
plot(density(log(normal_sample)))
```

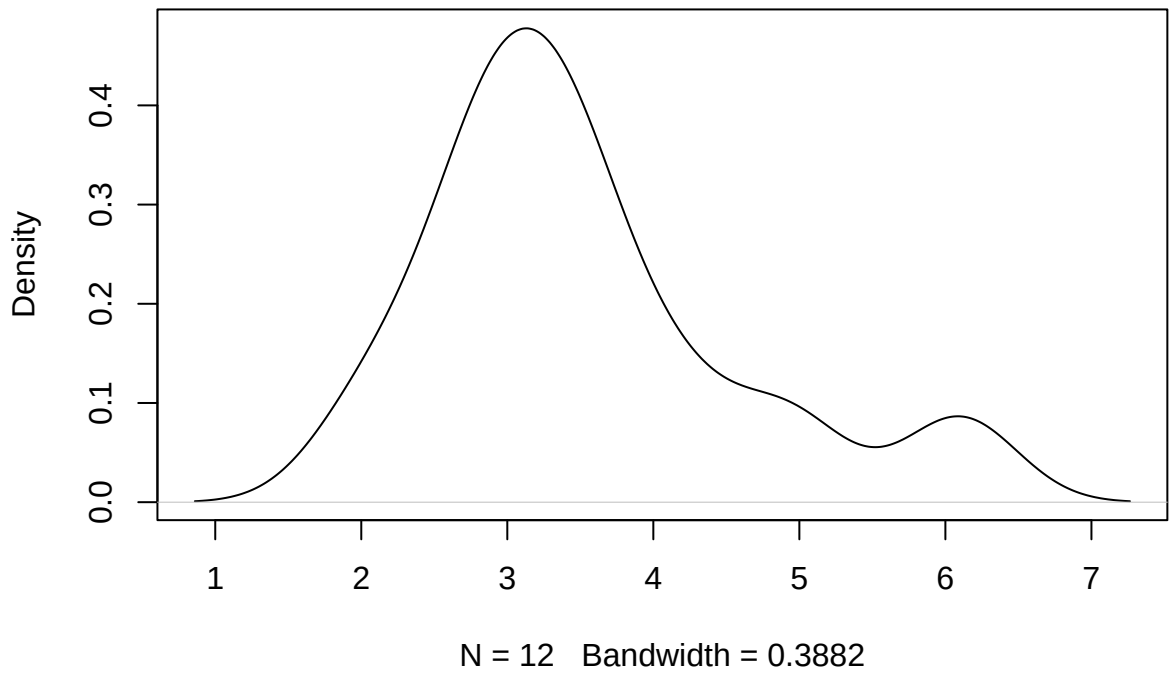



```
plot(density(log(diabetic_sample)))
```

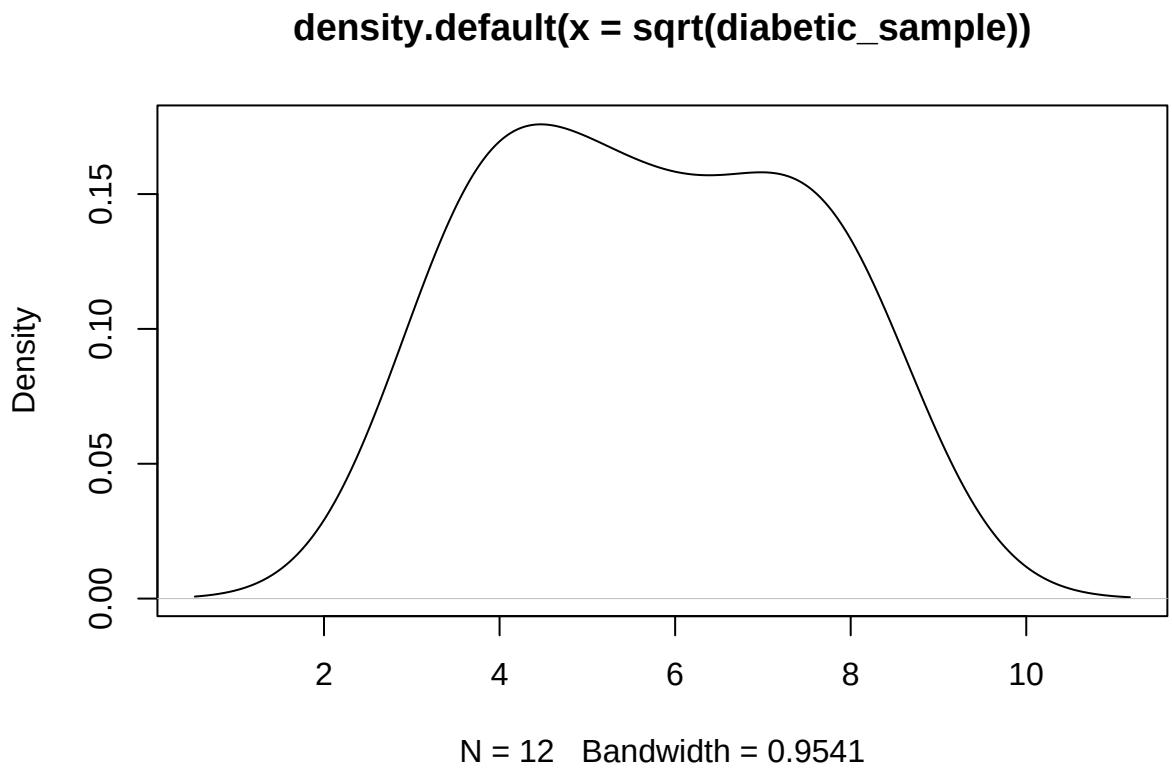


Square root transform:

```
plot(density(sqrt(normal_sample)))
```



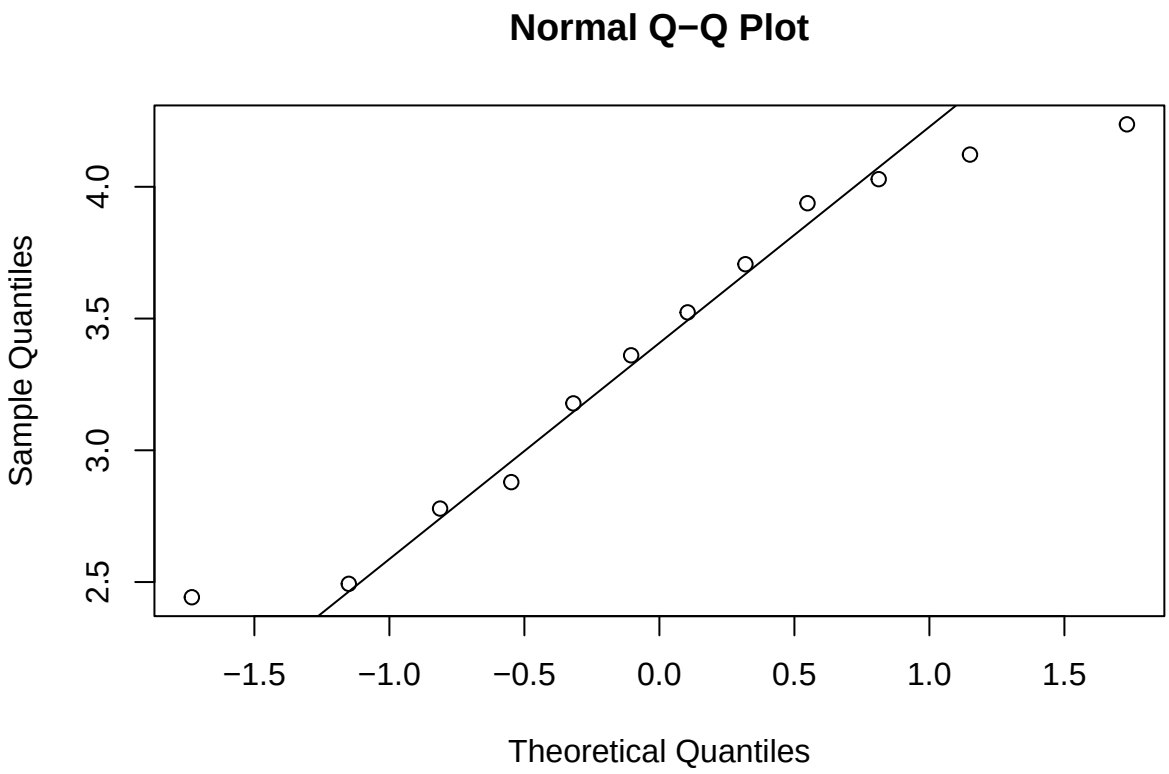

```
plot(density(sqrt(diabetic_sample)))
```



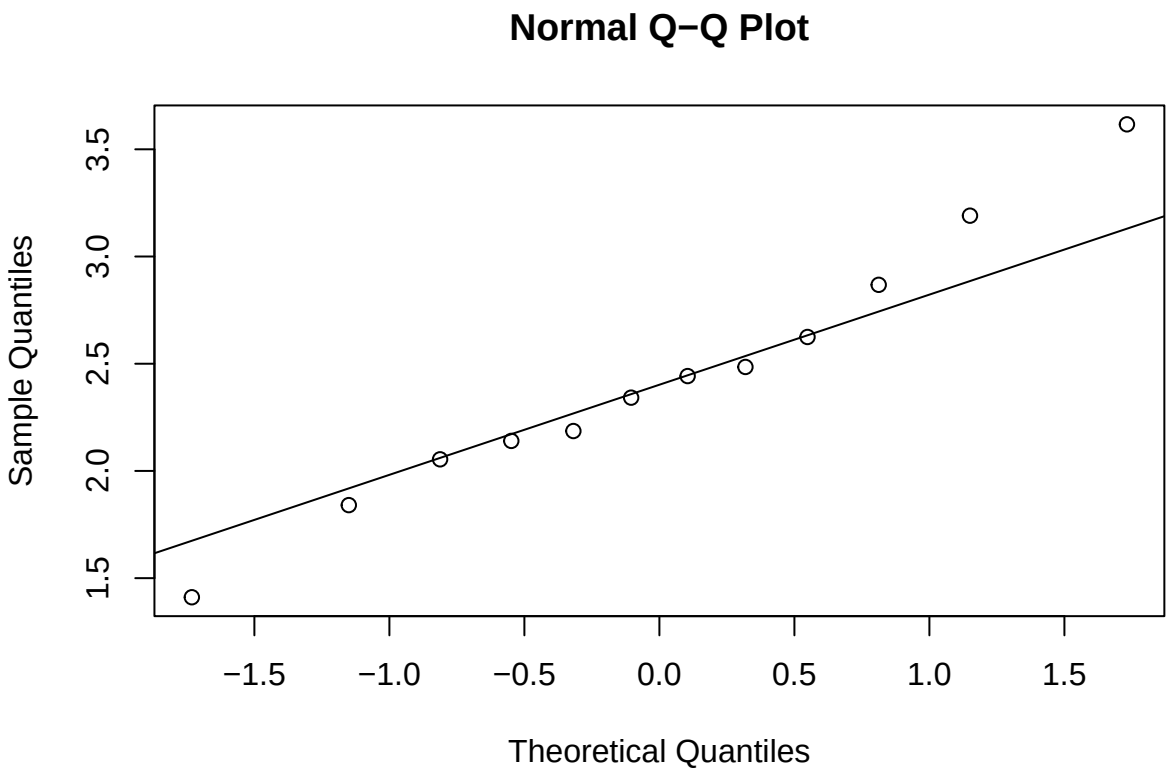
The log transform seems slightly better and preferrable as it makes them symmetric.

2.3.

```
qqnorm(log(diabetic_sample))
qqline(log(diabetic_sample))
```




```
qqnorm(log(normal_sample))
qqline(log(normal_sample))
```



They seem like they are sampled from a normal distribution from the qq plots. Most points of both these samples lie on the 45 degree line.

2.4.

mu1 be the mean of the diabetic sample and mu2 be the mean of the normal sample. delta_hat = mu1 - mu2 H0: delta_hat <= 0
H1: delta_hat > 0

Theory based approach:

```
diabetic_sample = log(diabetic_sample)
normal_sample = log(normal_sample)
Delta.hat = mean(diabetic_sample) - mean(normal_sample)
se = sqrt(var(diabetic_sample)/12 + var(normal_sample)/12)
df = nu = (var(diabetic_sample)/12+var(normal_sample)/12)^2/
((var(diabetic_sample)/12)^2/11+(var(normal_sample)/12)^2/11)
t.welch = (Delta.hat - 0) / se
1 - pt(t.welch,df=df)
```

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

The p value is less than 0.1%, so we have evidence to reject the null and say that diabetic patients have increased urinaly excretion than normal patients.

Simulation based approach:

First reformat the data

```
exc = c(diabetic_sample,normal_sample)
group = c(rep("diabetic",12),rep("normal",12))
data = data.frame(group,exc)
```

Now create bootstrap:

```
library(infer)
null_dist = data %>%
  specify(exc ~ group) %>%
  hypothesise(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("diabetic","normal"))
```

We get again the estimate difference based on the original samples

```
delta_hat <- data %>%
  specify(exc ~ group) %>%
  calculate(stat = "diff in means", order = c("diabetic", "normal"))
delta_hat
```

```
## Response: exc (numeric)
## Explanatory: group (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.957
```

```
null_dist %>%
get_p_value(obs_stat = delta_hat , direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.003
```

The p value is less, so we have evidence to reject the null and say that diabetic patients have increased urinary excretion than normal patients.

Q3

3.1.

Here is the code:

```
set.seed(100)
df_pass <- bechdel |>
  na.omit() |>
  filter(binary == "PASS") |>
  slice_sample(n = 60) |>
  mutate(profit = domgross - budget)
df_fail <- bechdel |>
  na.omit() |>
  filter(binary == "FAIL") |>
  slice_sample(n = 72)|>
  mutate(profit = domgross - budget)

df_final <- rbind(df_pass, df_fail)
```

3.2.

- Experimental unit is a film.
- 2 populations, films that pass (1) or fail (2) the Bechdel test.
- This is a 2-sample problem, $n_1 = 60$ and $n_2 = 72$.
- Two measurements were taken per experimental unit in order to obtain the profit: domgross and budget. So, profit for films that pass the Bechdel test can be represented by $X_i = D_i - B_i$ for $i = 1, \ldots, 60$ and those that fail the test would be given by $Y_j = D_j - B_j$ for $j = 1, \ldots, 72$ where D represents domgross and B budget.
- The parameter of interest is $\Delta = \mu_1 - \mu_2$ the difference of average profit for those films that pass minus average profit for those films that fail the Bechdel test, and the hypotheses are $H_0 : \Delta \geq 0$ versus $H_1 : \Delta < 0$ (we want to find evidence whether films that fail the test are more profitable).

3.3.

```
xbar = mean(df_pass$profit)
n1 = length(df_pass$profit)
s1 = sd(df_pass$profit)
ybar = mean(df_fail$profit)
n2 = length(df_fail$profit)
s2 = sd(df_fail$profit)
SE = sqrt(s1^2/n1 + s2^2/n2)
Deltahat = xbar - ybar

nu.hat = (s1^2/n1+s2^2/n2)^2/((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))

t.w  = (Deltahat -  0)/SE
t.w
```

```
## [1] 1.553952
```

```
pt(t.w, nu.hat)
```

```
## [1] 0.9386524
```

3.5.

```
alpha = 1 - 0.97
q = qt(1 - alpha/2, nu.hat)

Deltahat - q*SE
```

```
## [1] -5380192
```

```
Deltahat + q*SE
```

```
## [1] 31470090
```