Problem Set - 9 -DilipNikhil

Dilip Nikhil Francies

2023-11-02

Question 2

Given : The heights of men and women are approximately normally distributed. We have to estimate the average difference in heights between men and women.

n =7

mu0_men = 68.5

sd0_men = 3.0

mu0_women =65.5

sd0_women = 2.5

Question 2a

Answer:

We use standard normal distribution to calculate p-values and confidence intervals when we have a fairly large number of samples. In this case, our sample size in 7, which makes it difficult for the central limit theorem to kick in.The t-distribution on the other can handle uncertainty through degrees of freedom.Higher the number of samples, closer it gets to the original standard normal distribution which is not the case here with only 7 samples.

Secondly, we do not know the standard deviation of the population, but the sample.If we did, we could use pnorm() and qnorm().Yes, it is impossible to get the true standard deviation of the population, but one as mentioned above, one cannot rely on the inference of a sample that has only 7 observations.Hence, using a t-distribution can help account for this variability, especially when calculating the tail probabilities which do not die away as quickly as the normal distribution.

For the above mentioned reasons, one can use t-distribution to calculate p-values and confidence intervals when the gives sample size is fairly small.

Question 2b

Given:

```
degrees_of_freedom <- 11.6
sd_men <- 3
sd_women <- 2.5
n<- 7
```

the difference in their mean:

delta_hat : 68.5 - 65.5 = 3

```
delta_Ht <- 68.5 - 65.5
```

The estimated standard error is given by:

```
se <- sqrt((sd_men^2/n) + sd_women^2/n)
se
```

```
## [1] 1.475998
```

Because we already know the value of degrees of freedom, we can calculate the 95% confidence interval by:

```
q <- qt(.975,df=11.62)
lower <- delta_Ht - q * se
upper <- delta_Ht + q * se
print(paste("The 95% confidence interval of the mean of difference of average
            height of men and women going to the uni is", lower, " and" ,upper))
```

```
## [1] "The 95% confidence interval of the mean of difference of average \n            height of men and
```

Question 2c: From the previous question, we have calculated a 95% confidence interval that the mean difference between average height of men and women would be between -0.228 and 6.23. Just because one of the values is 0 does not mean that there is no difference.It means that i am 95% confident that average mean of difference of heights of men and women will be between this range. Will there be times where the difference is 0, yes. will there be times where the difference is 6 inches, yes! Will there be times where a a women will be taller than men, yes. But just because one of the values in the range is 0, does not mean there is no difference at all.

Question 3

Given:

The samples were large and right skewed.

Wish to test that the treatment and control group will on average result in the same number of weeks worked. which means:

Null hypotheses: delta = 0 Alternate hypotheses : delta != 0

```
n_treat <- 592
n_control <- 154

mu_treat <- 16.8
sd_treat <- 15.9

mu_control <- 24.3
sd_control <- 17.3
```

```

var_treat <- 15.9^2
var_control <- 17.3^2

print(paste("The variance of control group is",var_control,
            "and that of treatment group is",var_treat))
```

Question 3a:

```
## [1] "The variance of control group is 299.29 and that of treatment group is 252.81"
```

One can clearly see that the difference in variance of the groups is large enough to say that, the population variances are not equal either. To a certain extent, the sample size that we have, 592 and 154 if not large, it is big. Assuming we do not have any outliers, and the data is symmetric, Welch's test would be robust compared to students two sample t-test. Hence, Unless i am 100% sure that the population variances are equal, a safe bet would be to perform a Welch's test rather than students t-test.

Though the data is right skewed, transformation such as log will help us reach that normal distribution, hence the variance between the two samples should be a factor while choosing the type of test one would opt for.

Question 3b: Lets first calculate degrees of freedom :

```
deg_freedom <- (var_treat/n_treat + var_control/n_control)^2 /
((var_treat/n_treat)^2/(n_treat-1)+
   (var_control/n_control)^2/(n_control-1))

deg_freedom
```

```
## [1] 224.8164
```

Welch's t statistic is given by:

```
se <- sqrt(var_treat/n_treat + var_control/n_control)
delta_ht <- mu_control - mu_treat
t.Welch <- delta_ht/se
t.Welch
```

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

Hence, for two tailed test, the p value is given by:

```
2*(1-pt(t.Welch,df=deg_freedom))
```

```
## [1] 2.091618e-06
```

Because the p value we obtained is tiny, we can reject the null hypotheses and conclude that there is enough evidence that the difference between number of weeks worked between control and treatment group is more than 0.

```

lower <- delta_ht - qt(0.975,df=deg_freedom) * se
upper <- delta_ht + qt(0.975,df=deg_freedom) * se

print(paste("Hence, the 95% confidence interval of the average difference between number of weeks worke
```

Question 3c:

```
## [1] "Hence, the 95% confidence interval of the average difference between number of weeks worked by
```

Question 4:

Given:

Scores are approximately normal:

```
n_A <- 100
n_B <- 100

mu_A <- 61
sd_A <- 10

mu_B<- 59
sd_B <- 13
```

Question 4a: Lets formulate our hypotheses as:

Null hypotheses H0 : mu=0

Alternate hypotheses H1 : mu!= 0

```
var_A <- sd_A^2
var_B <- sd_B^2
print(paste("Variance of group A is" ,var_A))
```

```
## [1] "Variance of group A is 100"
```

```
print(paste("Variance of group B is", var_B))
```

```
## [1] "Variance of group B is 169"
```

Because there is difference in variance and only around 100 observations in the sample, lets be safe and perform Welch's t-test.

```
delta_hat <- mu_A - mu_B
se <- sqrt(var_A/n_A + var_B/n_B)
t.Welch <- delta_hat/se
t.Welch
```

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

```
degrees_freedom <- ((var_A/n_A + var_B/n_B)^2) /((var_A/n_A)^2/(n_A-1) + (var_B/n_B)^2/(n_B-1))
degrees_freedom
```

```
## [1] 185.7768
```

```
p_value <- 2 * (1-pt(t.Welch,df = degrees_freedom))
p_value
```

```
## [1] 0.2242302
```

This is not a small p-value, Hence we can safely conclude that the difference in score between group A and group be is not that huge.

```

q <- qt(0.95,df = degrees_freedom)
lower <- delta_hat - q * se
upper <- delta_hat + q * se
print(paste("Hence, the 90% confidence interval of the average score difference between  group A and gr
```

Question 4b

```
## [1] "Hence, the 90% confidence interval of the average score difference between  group A and group B
```

Question 4C: sample size: We have a sample size of 100 in each group, which is not good at the same time not too bad. We are assuming that the population would be normal based on the fact that the sample distribution is normal.hence, for such a small sample we cant prove that the entire population would have the same effect.

p-value:P-value we obtained was 0.0012 which is small enough to say that the data is not consistent with alternate hypotheses.Hence, the average score difference between group A and group B is equal to zero.

conf interval: if there is any difference between the mean in scores, then it would fall in the range of -0.711 to 4.711.

Question 5:

Given :

```
treatment_group <- 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_group <- 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)
```

5a: We formulate the hypotheses as:

delta = mu_treatment - mu_control

H0 : delta = 0

H(1): delta != 0

```
var(treatment_group)
```

```
## [1] 5.562045
```

```
var(control_group)
```

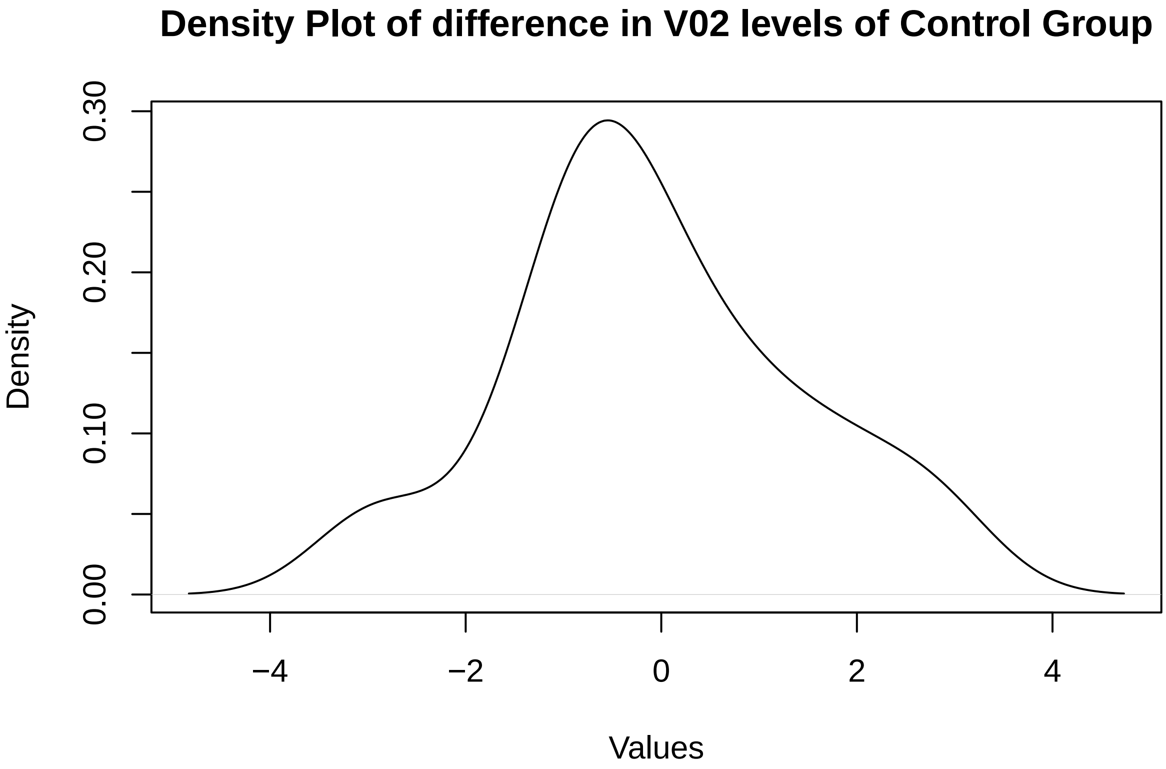
```
## [1] 2.375152
```

We could use a two sample Welch's t-test as the samples are independent but we are not sure if they have normally distributed populations, and we do not know if the population variances of these two samples are the same, hence its a safer bet to use Welch's test.

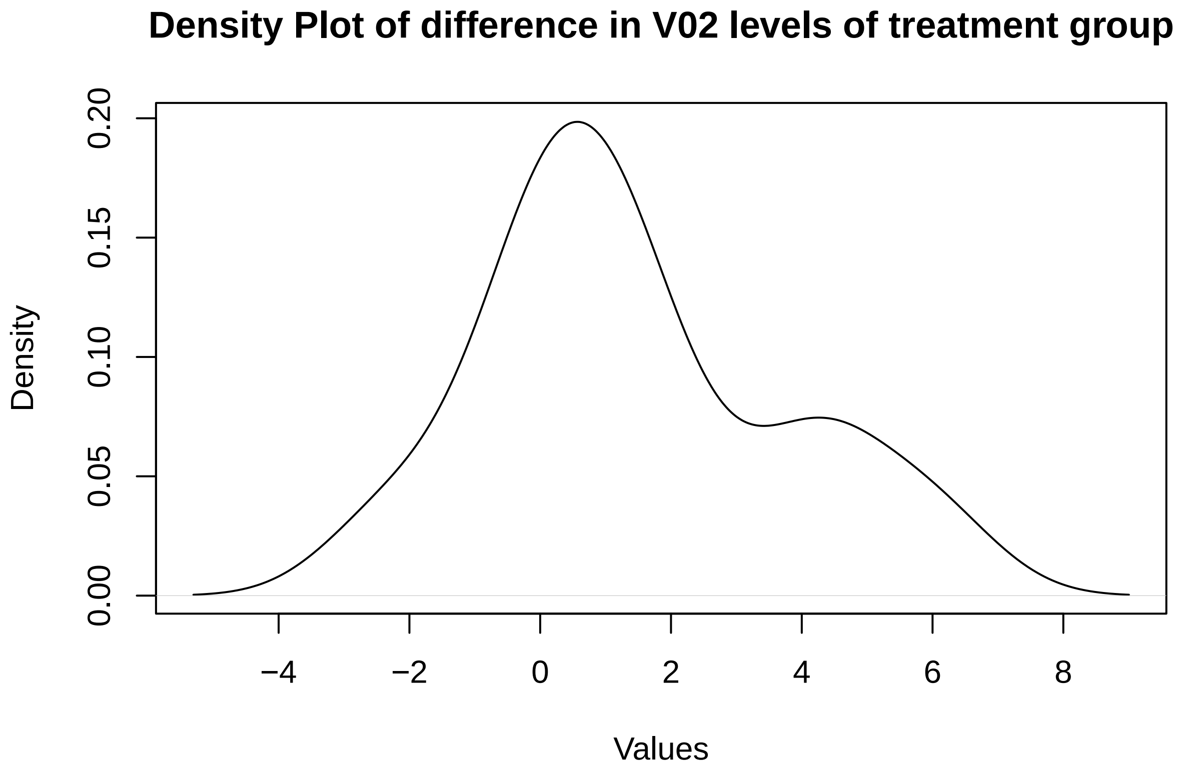
Yes, the sample size of 12 is way too less for the test to be carried out, but looking at the plots below, we can see that the distribution is fairly symmetric[does require some transformation] and there are no outliers.

hence, if not all assumptions, some of them are met, and the test is fairly robust to minor violations of its assumptions.

```
plot(density(control_group), main = "Density Plot of difference in V02 levels of Control Group", xlab =
```



```
plot(density(treatment_group), main = "Density Plot of difference in V02 levels of treatment group", xla
```



Question 5b Lets find the p-value:

```
n_treatment <- 12
n_control <- 12
mean_treatment <- mean(treatment_group)
mean_control <- mean(control_group)
sd_treatment <- sd(treatment_group)
sd_treatment <- sd(control_group)
var_treatment <- var(treatment_group)
var_control <- var(control_group)

se <- sqrt(var_treatment/12 + var_control/12)

delat_hat <- mean_treatment - mean_control

print(paste("standard error of the differences is given by",se))
```

```
## [1] "standard error of the differences is given by 0.813285362470075"
```

```
se
```

```
## [1] 0.8132854
```

```
t.Welch <- delat_hat/se
t.Welch
```

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

```
delta_hat <- mean_treatment - mean_control
delta_hat
```

```
## [1] 1.491667
```

```
degrees_free <- (var_treatment/12 + var_control/12)^2 /(((var_treatment/12)^2/(12-1)) + ((var_control/1

degrees_free
```

```
## [1] 18.9457
```

```
p <- (1 - pt(abs(t.Welch), df = degrees_free))
p
```

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

Because the p-value is less[considering 0.05], we can conclude based the mean difference in V02 levels between control and treatment group is not zero. But are we 100% sure? we would need more information and perhaps more data to draw final conclusions.

```

q <- qt(0.975,df = degrees_free)
lower <- delta_hat - q *se
upper <- delta_hat + q * se
print(paste("The 95% confidence interval of the mean of difference in V02 level of treatment and contro
```

Question 5c

```
## [1] "The 95% confidence interval of the mean of difference in V02 level of treatment and control grou
```

Question 5d: Doctor, so the idea behind the confidence interval is that, " 95% of the time, one can be sure that difference between the after-before VO2 levels of treatment and control group will fall in the range of -0.21 and 3.1942. There will be instances where the difference will be negative, there will be instances where the difference will be 0, but 95% of the time, it will be in the range of -0.2 to 3.2.

Question 6:

Given: The differences between the measurements ("without coffee" minus "with coffee") were approximately normal.

```
n = 10

mu_no_coffee <- 53
sd_no_coffee <- 19

mu_coffee <- 41.5
sd_coffee <- 17

mu_dif <- 11.5
sd_dif <- 21
```

Question 6a: The experimental unit is an individual with type 2 diabetes. There are two measurements being recorded in the experiment. One after consuming dates without coffee and one after consuming dates and coffee.This is a problem with one independent sample, provided the experiment was well run.

Question 6b: Null Hypotheses : mu = 0

Alternate Hypotheses : mu!= 0

That is, in null hypotheses, there would not be any increase in glycemic index when dates and coffee was taken together, and the alternate hypotheses would be, either there would be increase or decrease in the glycemic index.

Assuming that the population is normal, lets calculate one sample t-statistic

```
t.stat<- (mu_dif-0)/(sd_dif/sqrt(n))
t.stat
```

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

p value is given by:

right tailed test:

```
right <- 1-pt(t.stat, df = n - 1)
right
```

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

left tailed test:

```
left <- pt(t.stat, df = n - 1)
right
```

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

p value :

```
2*min(right,left)
```

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

Hence, we reject the alternate hypotheses that the difference between the two measurements is not equal to zero, based on the data that was provided.

Lets calculate 95% confidence interval:

```
q <- qt(.975, df = n-1)
mu_dif - q * sd_dif / sqrt(n)
```

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

```
mu_dif + q * sd_dif / sqrt(n)
```

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

Question 6c: The P-value (significance probability) was calculated to be 0.12, so the null hypothesis was not rejected. From this and the other information given, is it correct to conclude that we are sure that on average, dates have the same glycemic index with or without coffee?

Answer : No it is not correct to conclude that we are 100% sure that on average dates have the same glycemic index with or without coffee. Yes, the p-value we obtained was less, but we only a handful of 10 subjects, which is nowhere enough to draw clear cut conclusions. We have performed one sample t-test under the assumption that the distribution of population was normal, but is that the case?

we need more data, and better quality data? and run the experiment in a controlled manner so that cause and effect relation can be found out in the right manner.

Question 1:

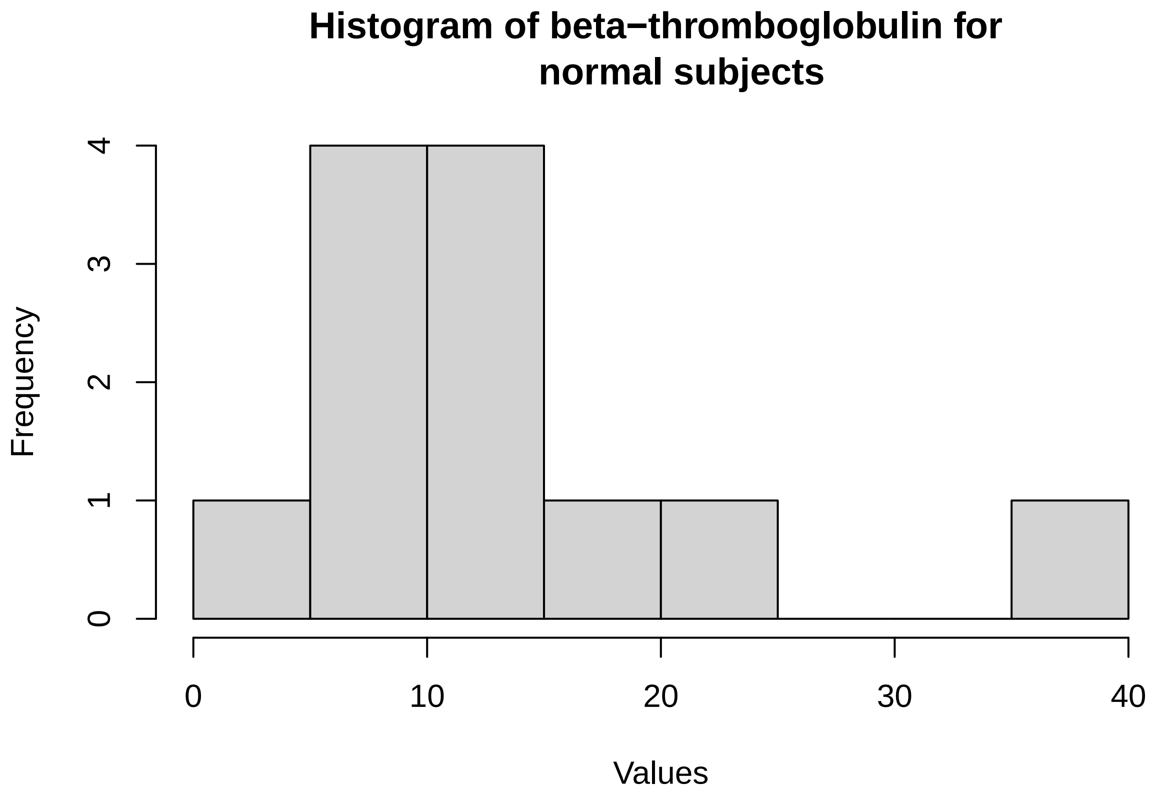
```
normal <- c(4.1 ,6.3, 7.8, 8.5 ,8.9 ,10.4,11.5 ,12.0 ,13.8 ,17.6 ,24.3 ,37.2)
diabetic <- c(11.5,12.1 ,16.1, 17.8 ,24.0, 28.8,33.9 ,40.7, 51.3, 56.2 ,61.7, 69.2)
```

Question 1.1 Do these measurements appear to be samples from symmetric distributions? Why or why not?

lets plot a histogram and density plots for the values provided:

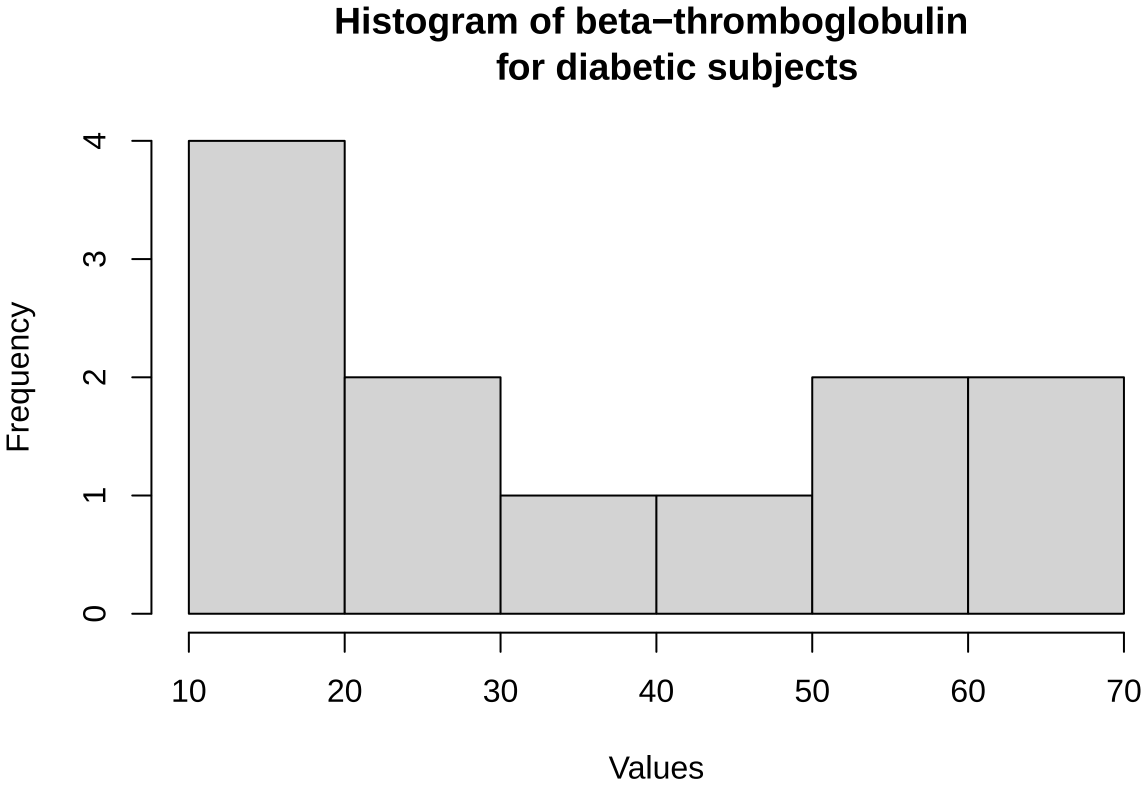
i. beta-thromboglobulin Values for Normal subjects :

```
hist(normal, main = "Histogram of beta-thromboglobulin for
     normal subjects ", xlab = "Values", col = "lightgrey", border = "black")
```



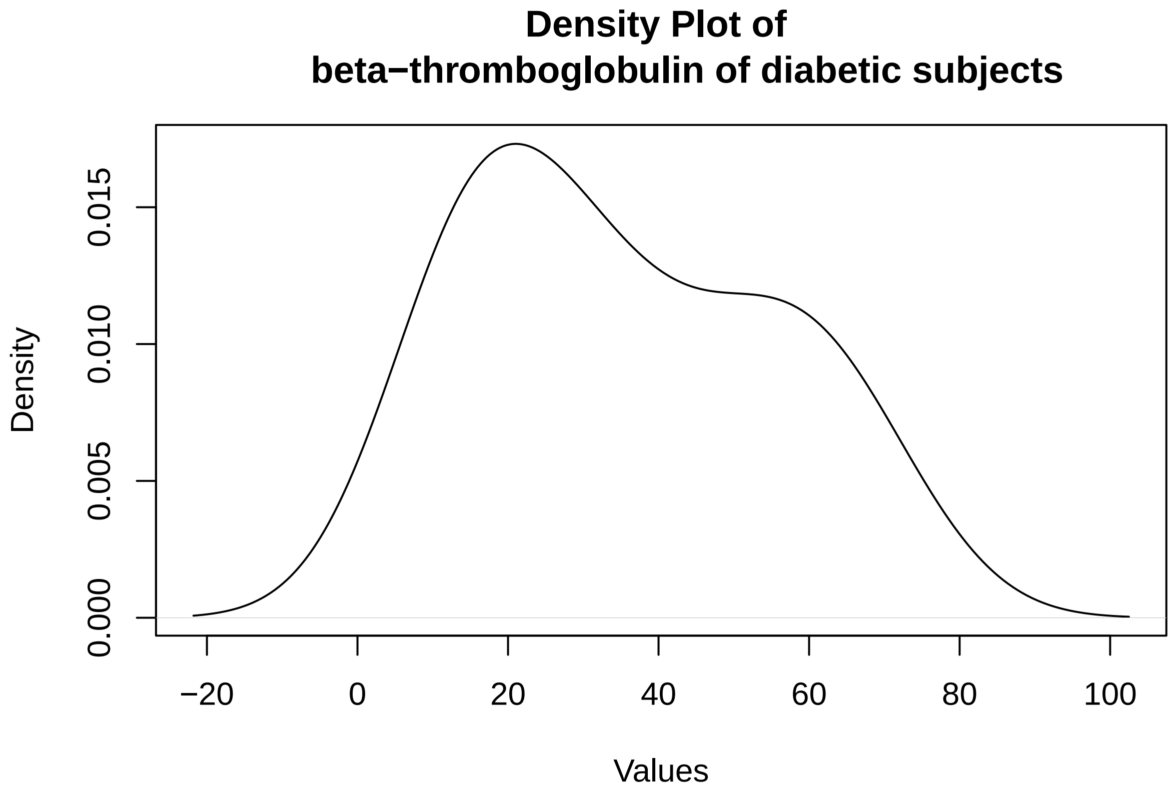
ii.beta-thromboglobulin Values for diabetic subjects :

```
hist(diabetic, main = "Histogram of beta-thromboglobulin
     for diabetic subjects ", xlab = "Values", col = "lightgrey", border = "black")
```



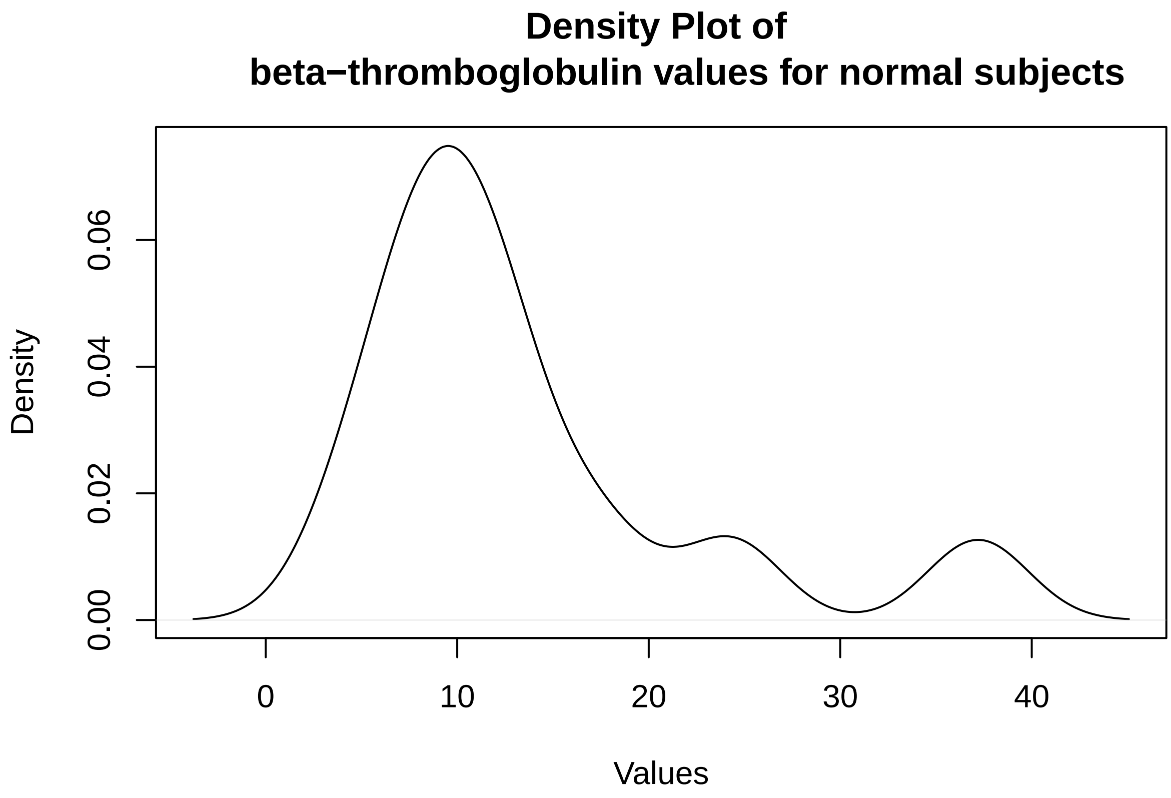
iii.

```
plot(density(diabetic), main = "Density Plot of
     beta-thromboglobulin of diabetic subjects", xlab = "Values")
```



iv.

```
plot(density(normal), main = "Density Plot of
     beta-thromboglobulin values for normal subjects", xlab = "Values")
```

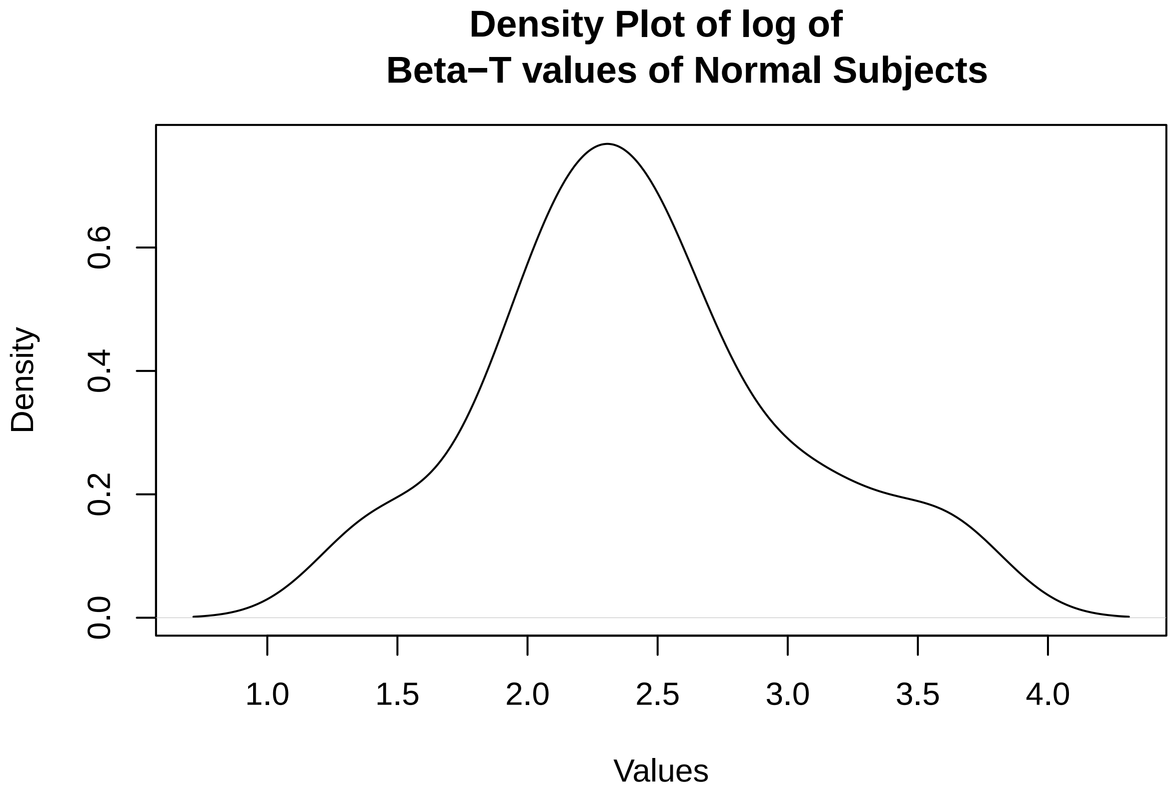


From the above plots, it is clear that the data is not from symmetric distribution. We have outliers, and the data looks positively skewed as well. Perhaps, the population is skewed as well, or we may not have enough observations [only 12 in this case] to draw conclusions about the population distribution.

Question 1, 2a: Lets apply log to our data to see if it gets better:

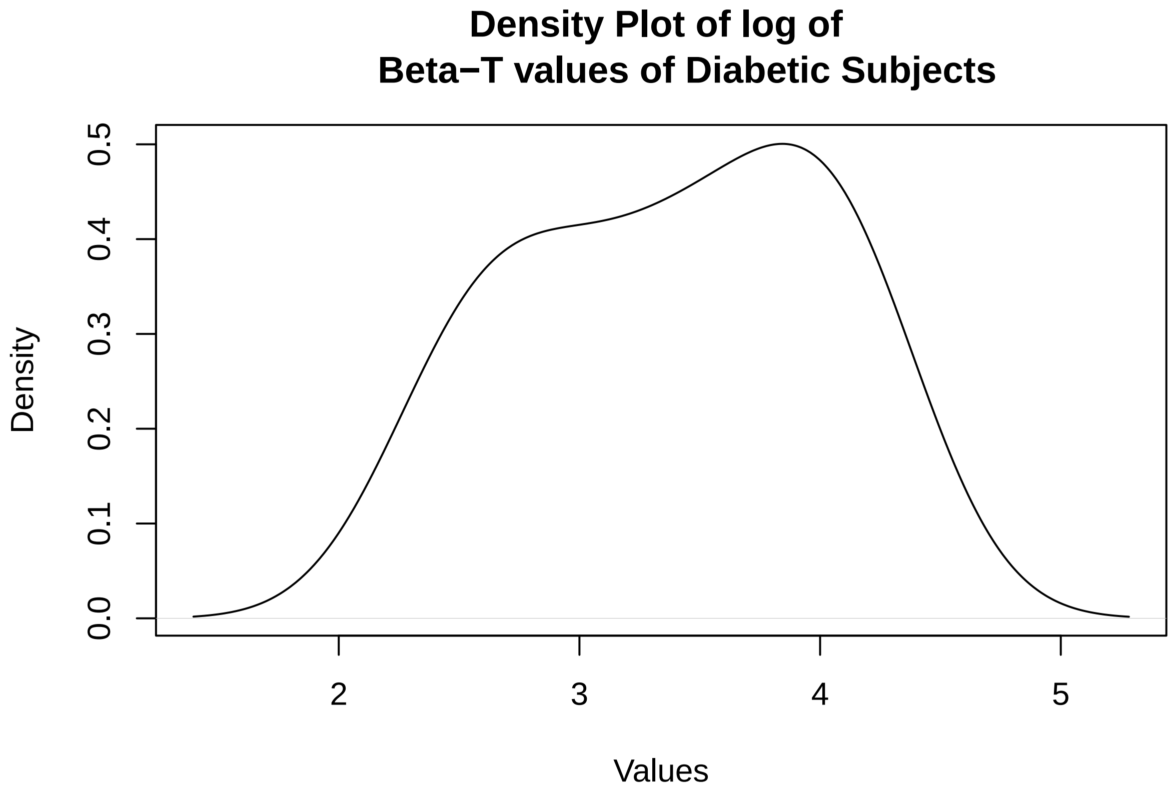
i. Normal subjects:

```
plot(density(log(normal)), main = "Density Plot of log of
     Beta-T values of Normal Subjects", xlab = "Values")
```



ii. diabetic:

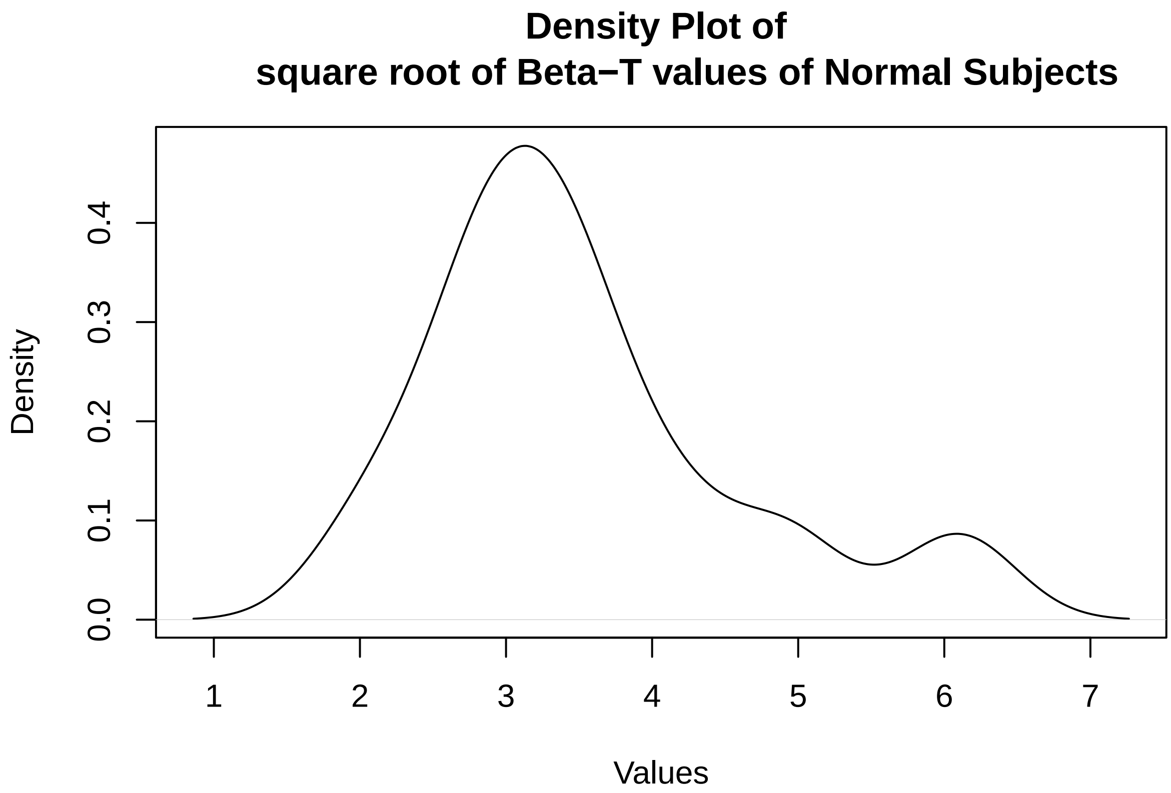
```
plot(density(log(diabetic)), main = "Density Plot of log of
     Beta-T values of Diabetic Subjects", xlab = "Values")
```



Question 1, 2b: Lets take square root of our data to see if it gets better:

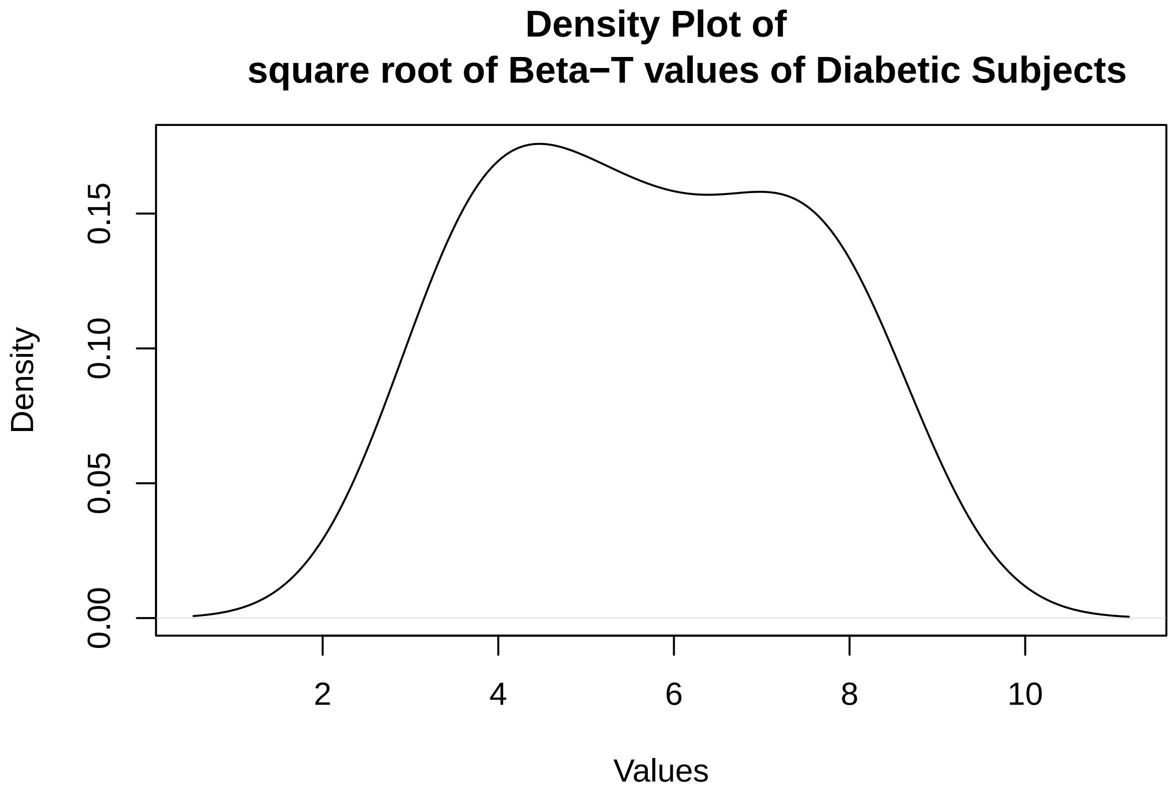
i. Normal subjects:

```
plot(density(sqrt(normal)), main = "Density Plot of
     square root of Beta-T values of Normal Subjects", xlab = "Values")
```



ii. Diabetic:

```
plot(density(sqrt(diabetic)), main = "Density Plot of
     square root of Beta-T values of Diabetic Subjects", xlab = "Values")
```



its clear that applying log gets us close to the symmetric distribution than square root transformation. The square root transformations still looks positively skewed for the normal subjects, and to a certain extent, does not make any difference in the diabetic subjects.

I prefer log transformations compared to square root.

Question 1, 3: yes, the log transformations looks very close to the symmetric distribution with minor violations. Logs compress the data, and they compress the larger values more compared to smaller values, hence bringing down the positively skewed data within the range of other observations.Not only they are better than square root transformations, they are more interpretable as well, though subjective.

question 1, 4 Lets perform hypotheses test to see if the researchers are right:

Welch’s two sample Test:

delta = mean(diabetic) - mean(normal)

null hypotheses H(0): delta =0

Alternative Hypotheses H(1) != 0

```
normal_log <- log(normal)       # apply log
diabetic_log <- log(diabetic)
mean_normal <- mean(normal_log) # find mean
mean_diabetic <- mean(diabetic_log)
Delta.hat <- mean_diabetic - mean_normal # calculate difference
```

calculate the test statistic:

```
var1 <- var(normal_log)
var2 <- var(diabetic_log)
n1 <- 12
n2 <- 12
se <- sqrt(var1/n1 + var2/n2)
print(paste("standard error for log of two sample is given by", se))
```

```
## [1] "standard error for log of two sample is given by 0.251645798825659"
```

```
t.Welch = Delta.hat / se
print(paste("T static for Welchs test is",t.Welch))
```

```
## [1] "T static for Welchs test is 3.80407203949253"
```

degrees of freedom:

```
nu <- (var1/n1 + var2/n2)^2 /((var1/n1)^2/(n1-1) + (var2/n2)^2/(n2-1))
print(nu)
```

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

if the doctors are right, then for right tailed test, the p-value has to be less than 0.05.Let's check.

```
P.value <- (1 - pt(abs(t.Welch), df = nu))
P.value
```

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

lets find the two-tailed p-value

```
P.value <- 2 * (1 - pt(abs(t.Welch), df = nu))
P.value
```

```
## [1] 0.0009776127
```

Hence, researchers are actually right. the Beta-thromboglobulin for diabetics patients is more than that of normal patients. But again? we only have 12 subjects in each group. We would need more data to be 100% sure.

Lets calculate the 95% confidence interval for it.

```
q<- qt(0.975, df=nu)


lower <- Delta.hat - q * se
upper <- Delta.hat + q * se

#convert back to original scale

lower <- exp(lower)
upper <- exp(upper)

print(paste("The 95% confidence interval of the mean of difference in Beta-thromboglobulin level of dial
```

```
## [1] "The 95% confidence interval of the mean of difference in Beta-thromboglobulin level of diabetic
```