Homework 4

Caitlin Jagla

03/25/2022

1. Estrogen

A study of estrogen levels in two different groups of women finds the following results (in pg/mL):

```
est <- tibble(group = c(rep("A", 19), rep("B", 8)),
              estrogen = c(18.7,  20.6,  20.7,  19.7,  19.9,
                           19.4,  20.2,  21.6,  18.8,  14.1,
                           21.6,  16.2,  21.7, 20.8,  19.3,
                           21.3,  19.9,  20.8,  23.2, 15.2,
                           36.2,  27.5,    4.7,  24.5,  29.4,
                           25.9,  62.8))
```

We want to know if there is a difference in the mean estrogen levels between the two groups. Use an $\alpha = 0.05$ for all of the hypothesis tests.

A. Use a Welch's t-test (unpooled variance) to test the difference between the groups. Interpret what you found.

```
t.test(estrogen ~ group, data = est, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  estrogen by group
## t = -1.3925, df = 7.0864, p-value = 0.2059
## alternative hypothesis: true difference in means between group A and group B is not equal to 0
## 95 percent confidence interval:
##  -22.505347   5.797452
## sample estimates:
## mean in group A mean in group B
##        19.92105        28.27500
```

> The Welch's t-test with unpooled variance supports the null hypothesis (that there is no true difference in mean estrogen levels between the two groups), as `p > 0.05`.

B. Conduct a pooled test instead. Interpret what you have found.

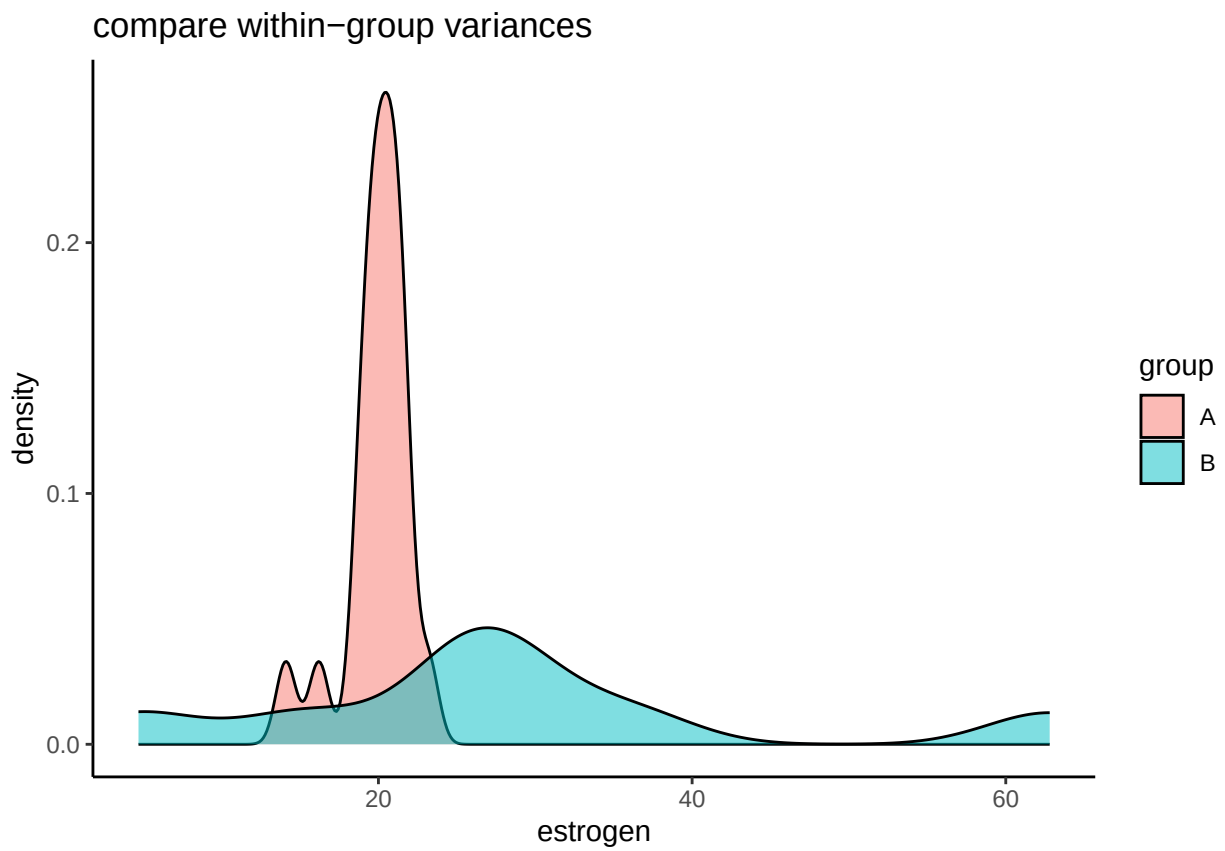
```r
t.test(estrogen ~ group, data = est, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  estrogen by group
## t = -2.1738, df = 25, p-value = 0.0394
## alternative hypothesis: true difference in means between group A and group B is not equal to 0
## 95 percent confidence interval:
##  -16.2688606  -0.4390341
## sample estimates:
## mean in group A mean in group B
##        19.92105        28.27500
```

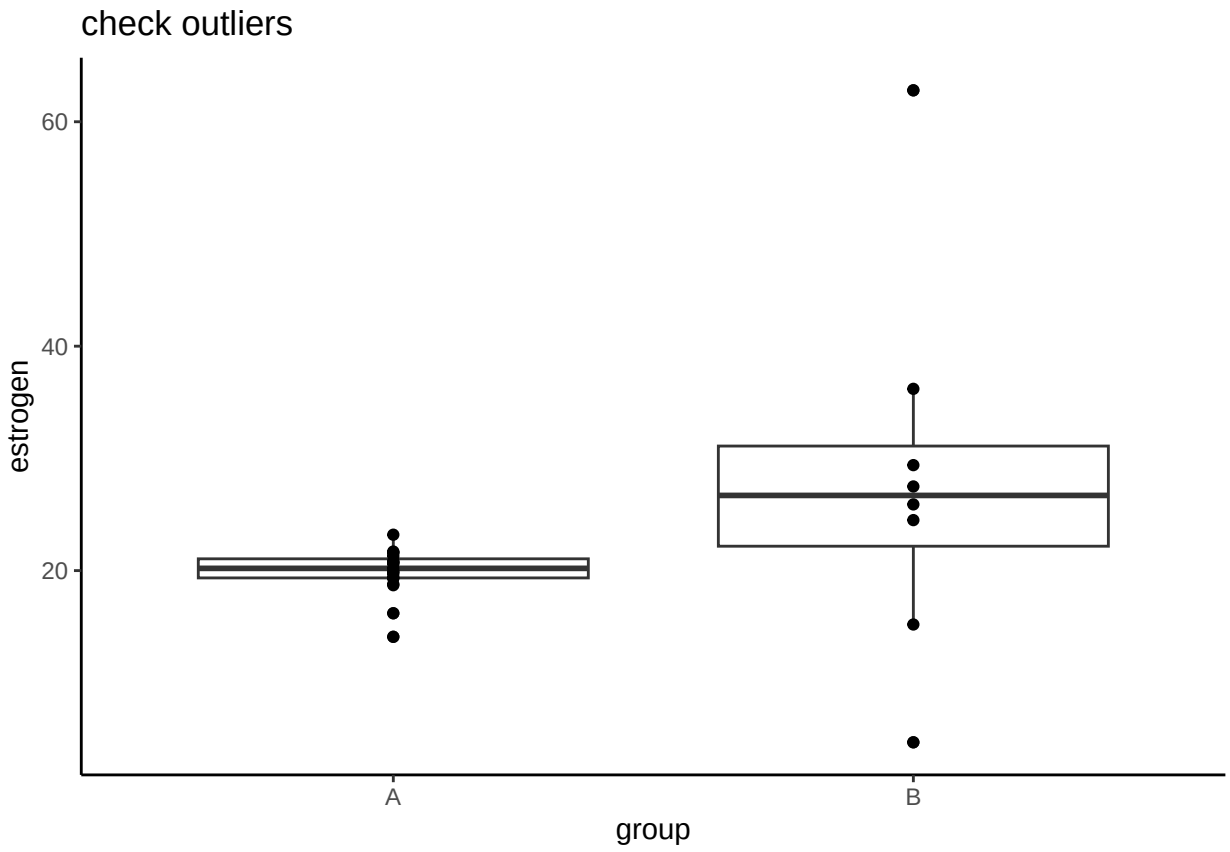
> The pooled t-test supports the alternative hypothesis, indicating that there is a true difference in mean estrogen levels between the two groups, since `p < 0.05`.

C. Create a plot to explore the differences between the groups.

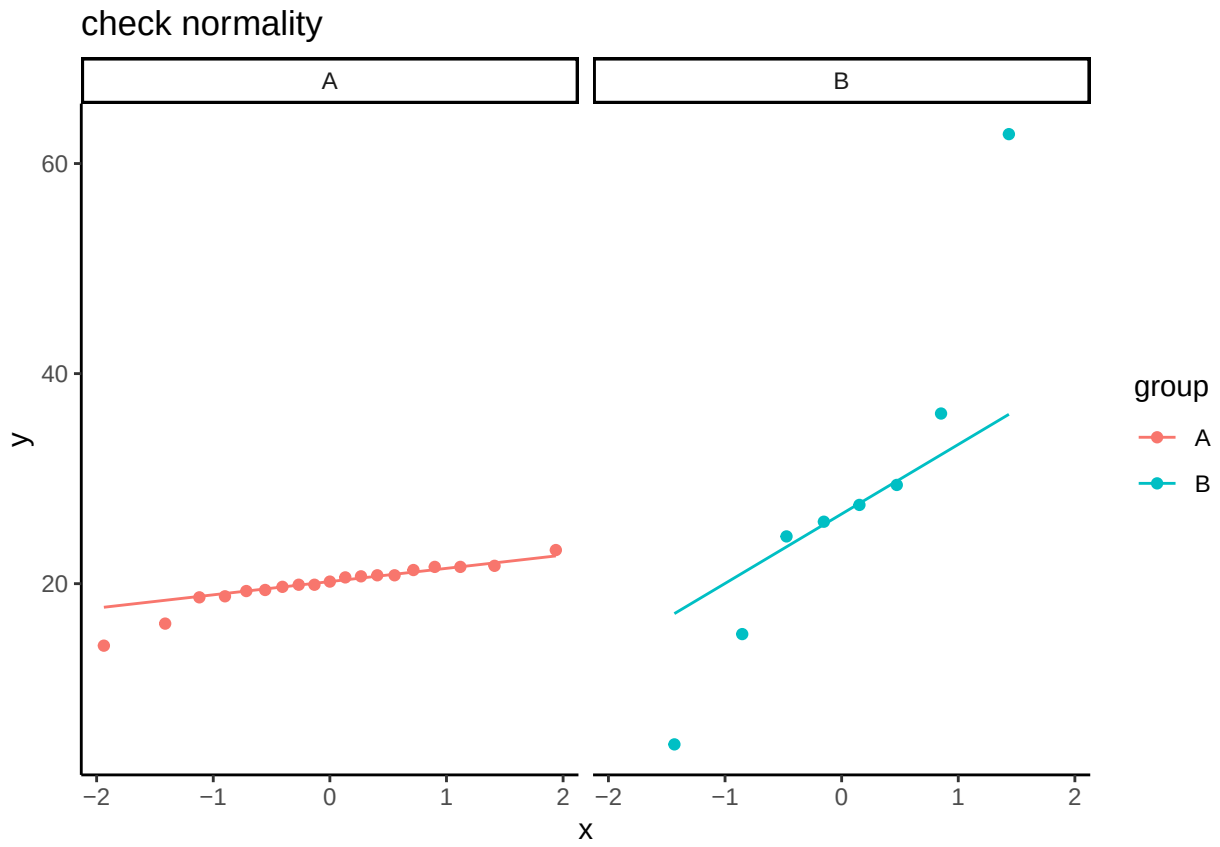
```r
#compare within-group variance
est |>
ggplot(aes(estrogen, fill = group)) +
  geom_density(alpha = 0.5) +
  labs(title = "compare within-group variances") +
  theme_classic()
```




```
#check outliers
est |>
ggplot(aes(x = group, y = estrogen)) +
  geom_boxplot() +
  geom_point() +
  labs(title = "check outliers") +
  theme_classic()
```




```
#check normality
est |>
  ggplot(aes(sample = estrogen, color = group)) +
  geom_qq() + geom_qq_line() +
  facet_wrap(~group) +
  labs(title = "check normality") +
  theme_classic()
```



D. Which type of test should be used for these data? Why?

The Welch's (unpooled) t-test should be used for these data. The two groups do not have equivalent variances - group B has a much higher variance than group A. This can be seen in the "compare within-group variance" plot above.

E. Explain why you got different answers to A and to B. Consider power of each test and hypothesis testing errors that could have occurred.

The answers were different because the variance of the two groups was so different that pooling them made a dramatic difference in the power of the t-test. In this case, using a pooled t-test when you should have used an unpooled t-test would result in a Type I error (false positive). This is evident when comparing the results of the two t-tests. The pooled t-test gives $p = 0.0393974$, which is lower than $\alpha = 0.05$, meaning the null hypothesis would be rejected. In contrast, the unpaired t-test gives $p = 0.2059145$, which is higher than α=0.05, meaning the null hypothesis would not be rejected.

2. Beta-thromboglobulin and diabetes data cleaning

A. Load the btg.csv dataset into R using the `read_csv()` function from the tidyverse. These data are the excretion of β-thromboglobulin (β-TG) in the urine of diabetic and non-diabetic mice. Print the dataset by calling its name.

```
btg <- read_csv("btg.csv")
btg
```

```
## # A tibble: 24 x 2
##    status   btg
##    <chr>  <dbl>
##  1 normal   4.1
##  2 normal   6.3
##  3 normal   7.8
##  4 normal   8.5
##  5 normal   8.9
##  6 normal  10.4
##  7 normal  11.5
##  8 normal  12
##  9 normal  13.8
## 10 normal  17.6
## # i 14 more rows
```

B. Create a new column called `logbtg` that is the natural log of btg. Save the new column to the dataset and print the dataset to demonstrate the change.

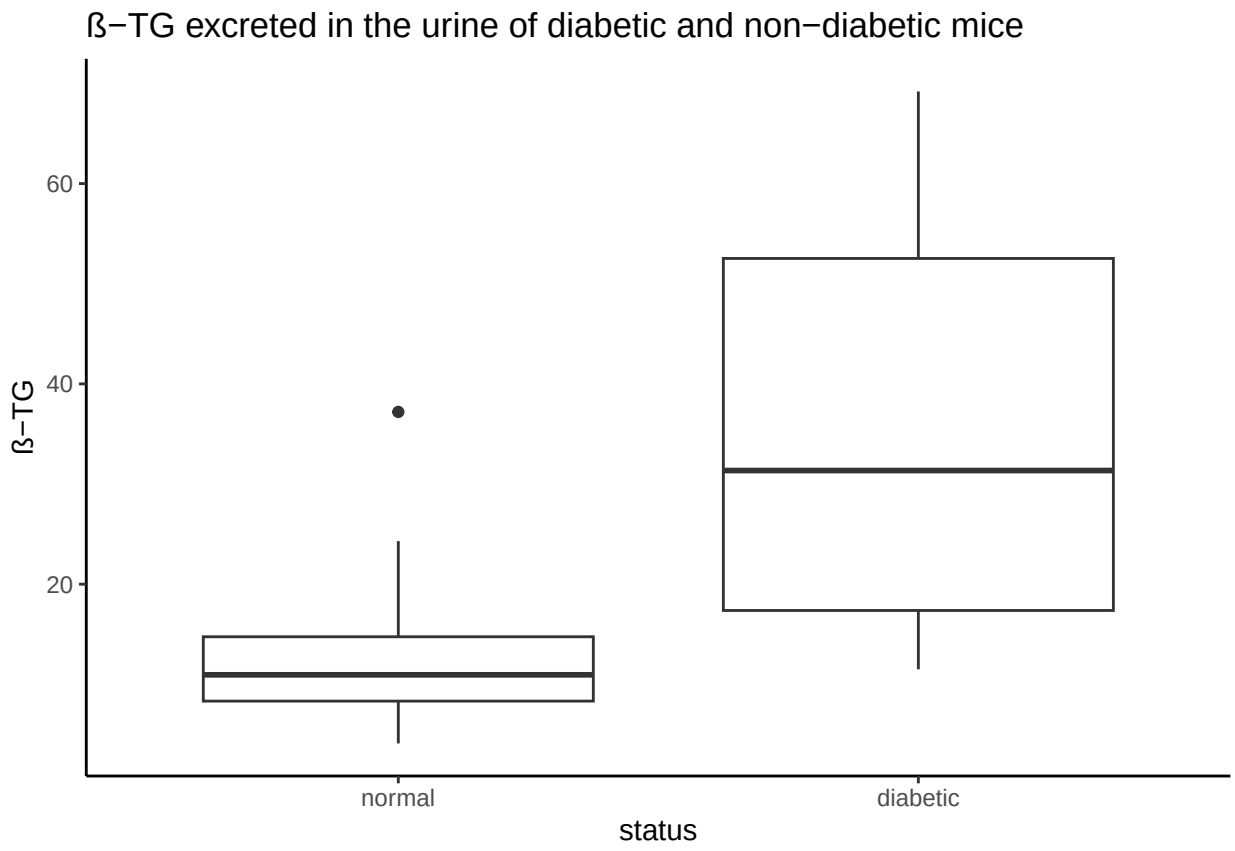
```
btg <- btg |> mutate(logbtg = log(btg))
btg
```

```
## # A tibble: 24 x 3
##    status   btg logbtg
##    <chr>  <dbl>  <dbl>
##  1 normal   4.1   1.41
##  2 normal   6.3   1.84
##  3 normal   7.8   2.05
##  4 normal   8.5   2.14
##  5 normal   8.9   2.19
##  6 normal  10.4   2.34
##  7 normal  11.5   2.44
##  8 normal  12     2.48
##  9 normal  13.8   2.62
## 10 normal  17.6   2.87
## # i 14 more rows
```

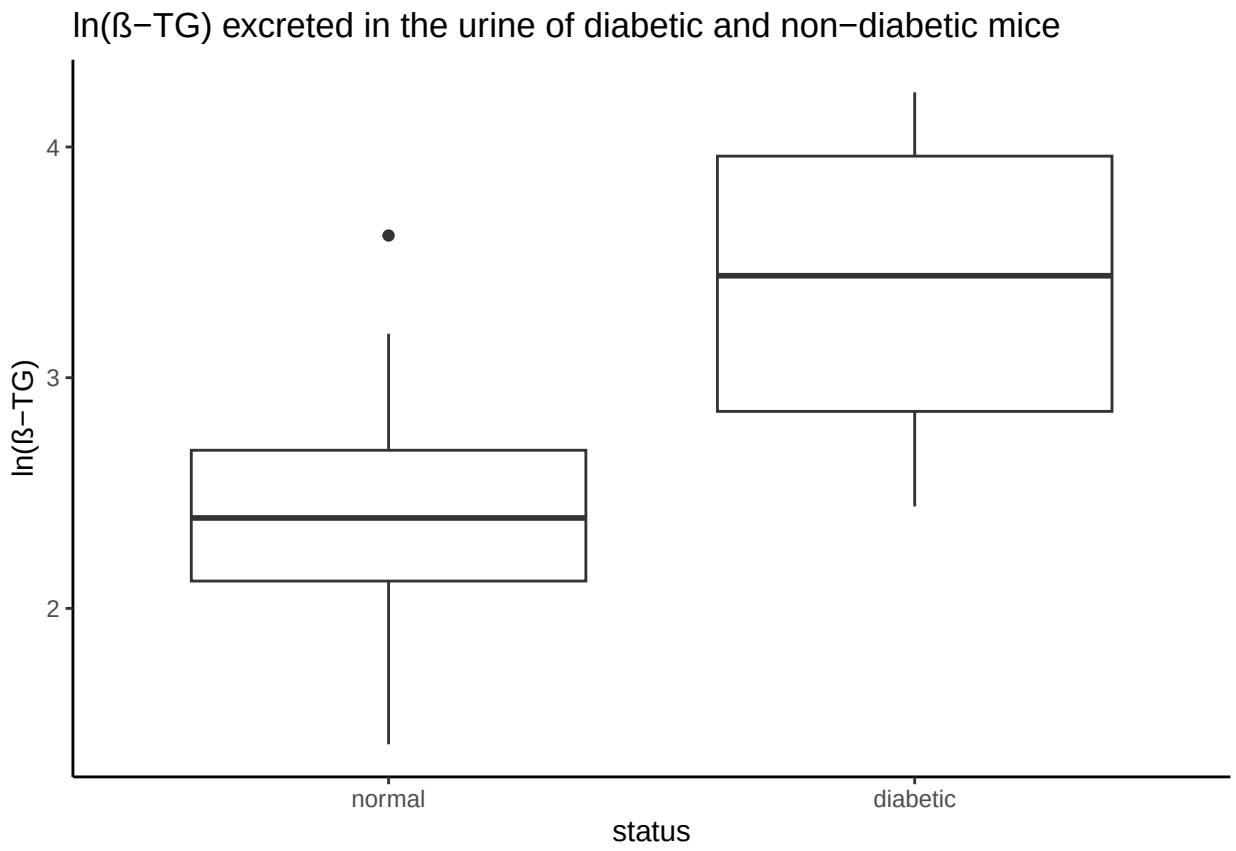
C.

Create 2 plots: - show btg values for diabetic and non-diabetic mice - show logbtg values for diabetic and non-diabetic mice

```
btg |>
ggplot(aes(x = reorder(status, btg), y = btg)) +
  geom_boxplot() +
  labs(title = "\u03B2-TG excreted in the urine of diabetic and non-diabetic mice",
       x = "status",
       y = "\u03B2-TG") +
  theme_classic()
```




```
btg |>
ggplot(aes(x = reorder(status, logbtg), y = logbtg)) +
  geom_boxplot() +
  labs(title = "ln(\u03B2-TG) excreted in the urine of diabetic and non-diabetic mice",
       x = "status",
       y = "ln(\u03B2-TG)") +
  theme_classic()
```



D. Calculate the mean and sd for diabetic and non-diabetic mice on the original btg scale and on the log scale.

```
btg |> group_by(status) |> summarize(btg_mean = mean(btg), btg_sd = sd(btg),
                                     logbtg_mean = mean(logbtg), logbtg_sd = sd(logbtg))
```

```
## # A tibble: 2 x 5
##   status   btg_mean btg_sd logbtg_mean logbtg_sd
##   <chr>       <dbl>  <dbl>       <dbl>     <dbl>
## 1 diabetic     35.3  20.3         3.39     0.637
## 2 normal       13.5   9.19        2.43     0.595
```

E. What is the effect of the log transformation on these data?

The log transformation decreases the nominal difference between groups and variance within each group.

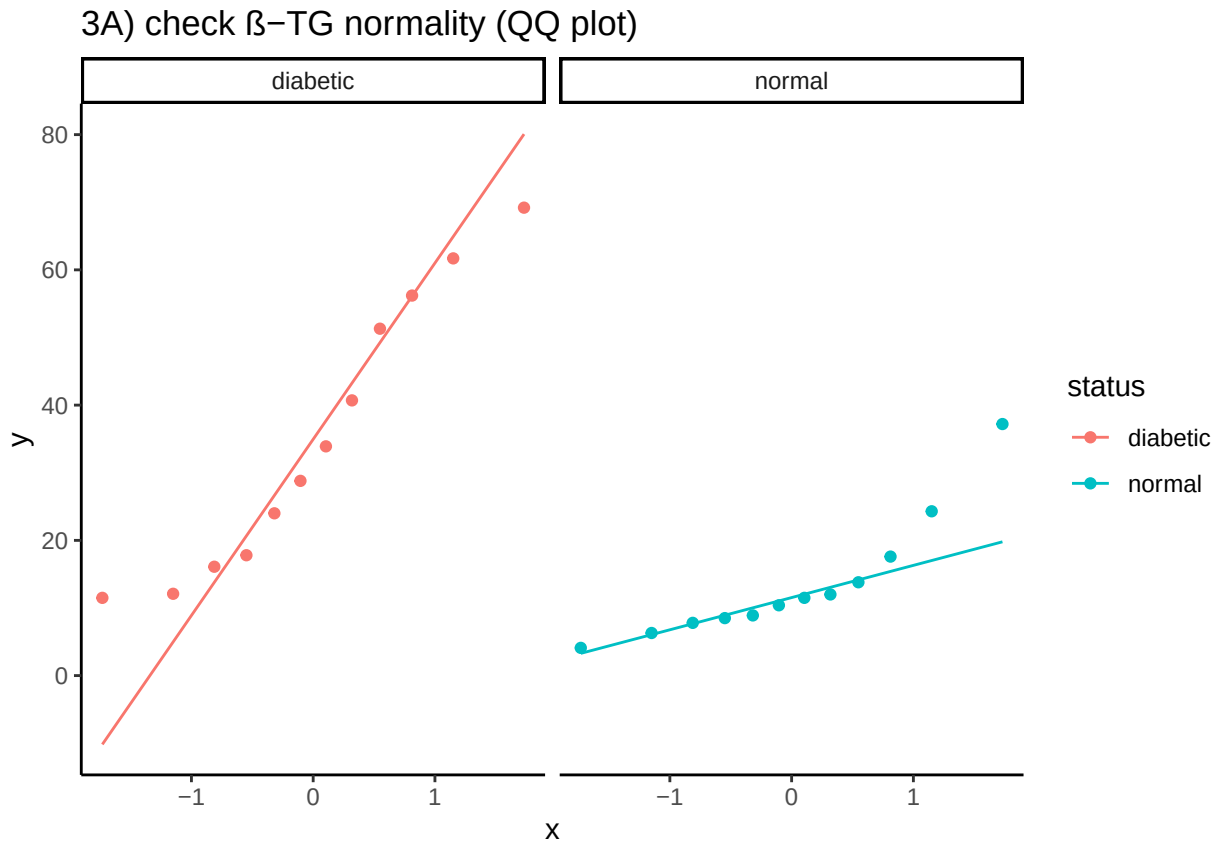
3. Beta-thromboglobulin and diabetes t-test

A. Follow all steps of hypothesis testing (1. Declare hypotheses, 2. choose alpha, 3. check assumptions using exploratory plots, 4. calculate test statistic, 5. compare p to alpha and conclude) to conduct the most appropriate test of two means for btg explained by status

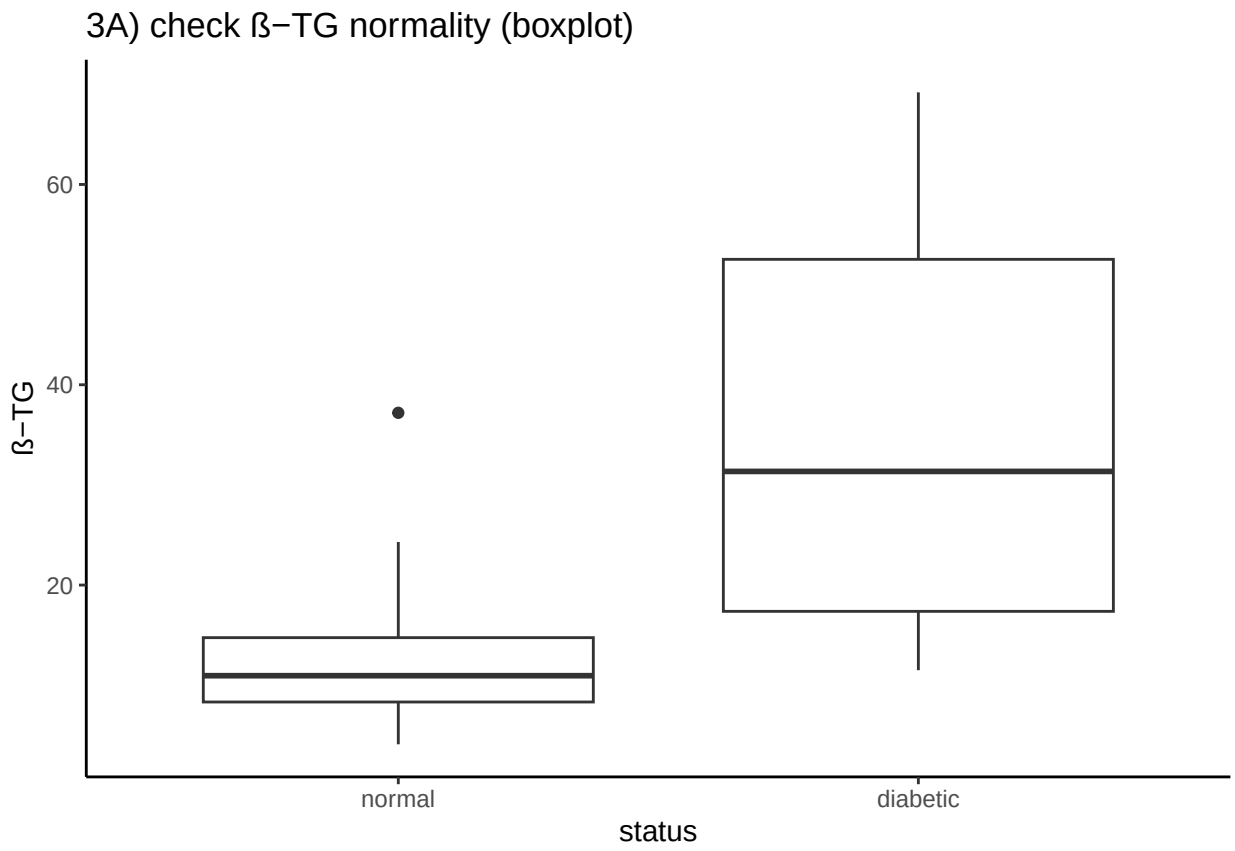
> note that if normality is not met, please indicate that, but then proceed with a t-test rather than the non-parametric alternative that we will learn next week

- Declare hypotheses:
  - $H_0 : \mu_{normal} = \mu_{diabetes}$
  - $H_1 : \mu_{normal} \neq \mu_{diabetes}$
- Choose alpha:
  - $\alpha = 0.05$
- Check assumptions using exploratory plots:
  - normality: both groups are fairly normally distributed, though there is one outlier in the non-diabetic group
  - independent samples: these data are collected from separate groups of mice, so the samples are independent
  - variance: the variance in the two groups is different enough that it seems safer to *not* rely on them being homogeneous

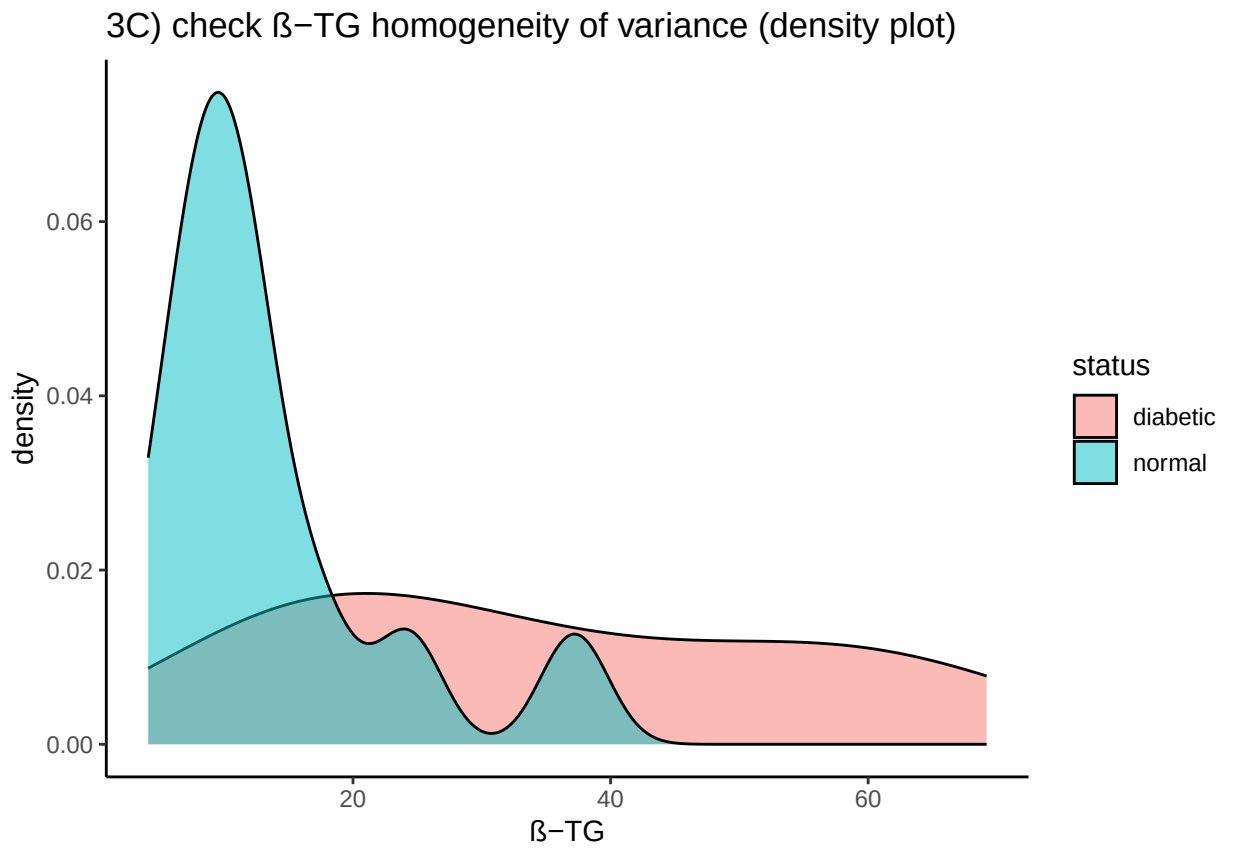
```
# check normality with qq plot
btg |>
  ggplot(aes(sample = btg, color = status)) +
  geom_qq() + geom_qq_line() +
  facet_wrap(~status) +
  labs(title = "3A) check \u03B2-TG normality (QQ plot)") +
  theme_classic()
```




```
# check normality with boxplot
btg |>
  ggplot(aes(x = reorder(status, btg), y = btg)) +
  geom_boxplot() +
  labs(title = "3A) check \u03B2-TG normality (boxplot)",
       x = "status",
       y = "\u03B2-TG") +
  theme_classic()
```




```
# check variance with density plot
btg |>
  ggplot(aes(btg, fill = status)) +
  geom_density(alpha = 0.5) +
  labs(title = "3C) check \u03B2-TG homogeneity of variance (density plot)",
       x = "\u03B2-TG") +
  theme_classic()
```




```
# check with descriptive statistics
btg.stats <- btg |> group_by(status) |> summarize(btg_mean = mean(btg),
                                                  btg_sd = sd(btg),
                                                  btg_var = var(btg),
                                                  n = n())
btg.stats
```

```
## # A tibble: 2 x 5
##   status   btg_mean btg_sd btg_var     n
##   <chr>       <dbl>  <dbl>   <dbl> <int>
## 1 diabetic     35.3  20.3    411.      12
## 2 normal       13.5   9.19    84.5     12
```

- Calculate test statistic:

```
t.test(btg ~ status, data = btg, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  btg by status
## t = 3.3838, df = 15.343, p-value = 0.003982
## alternative hypothesis: true difference in means between group diabetic and group normal is not equal
## 95 percent confidence interval:
```

```
##   8.07309 35.41024
## sample estimates:
## mean in group diabetic   mean in group normal
##               35.27500               13.53333
```

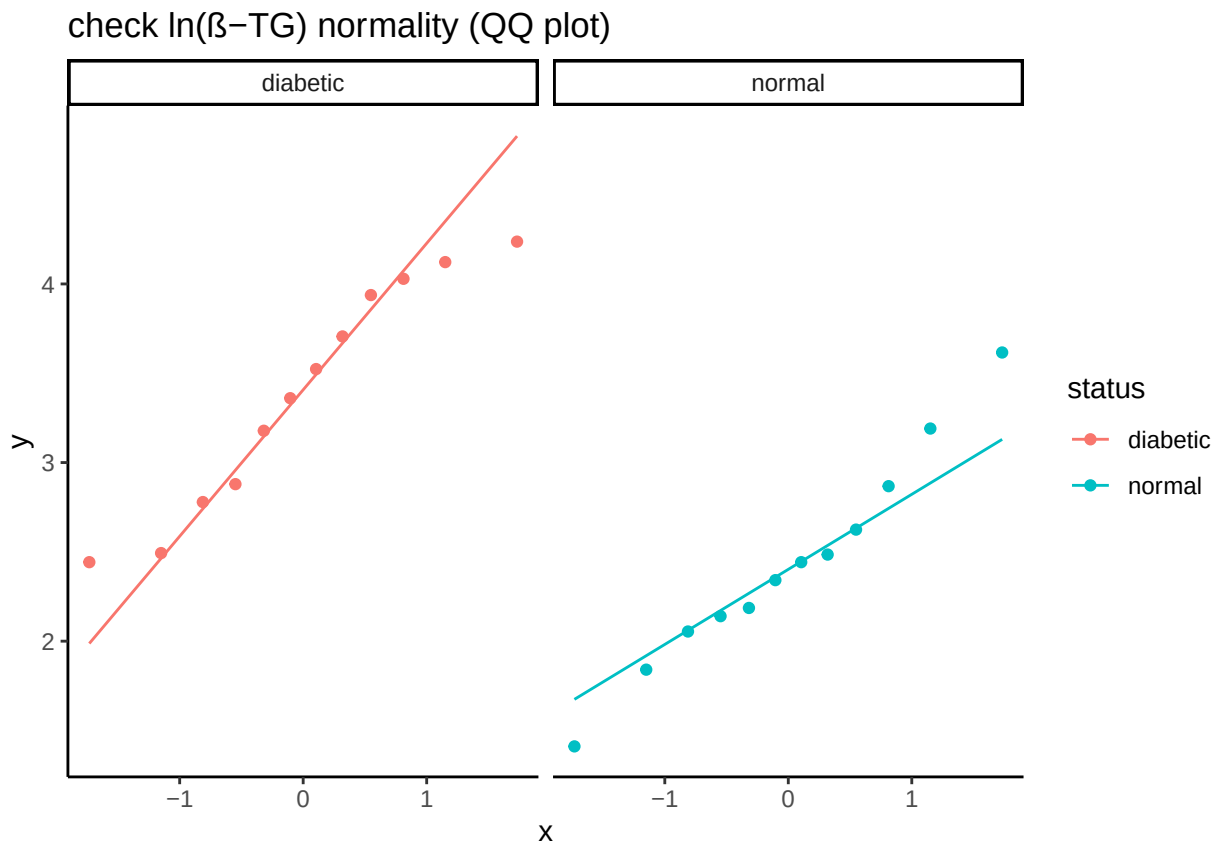
- Compare p to α and conclude:
  - $p = 0.0039817$, which is less than $\alpha = 0.05$
  - Therefore we reject the null hypothesis and conclude that there is a difference in β-TG excreted in the urine of diabetic vs. non-diabetic mice

B. Follow all steps of hypothesis testing (1. Declare hypotheses, 2. choose alpha, 3. check assumptions using exploratory plots, 4. calculate test statistic, 5. compare p to alpha and conclude) to conduct the most appropriate test of two means for logbtg explained by status

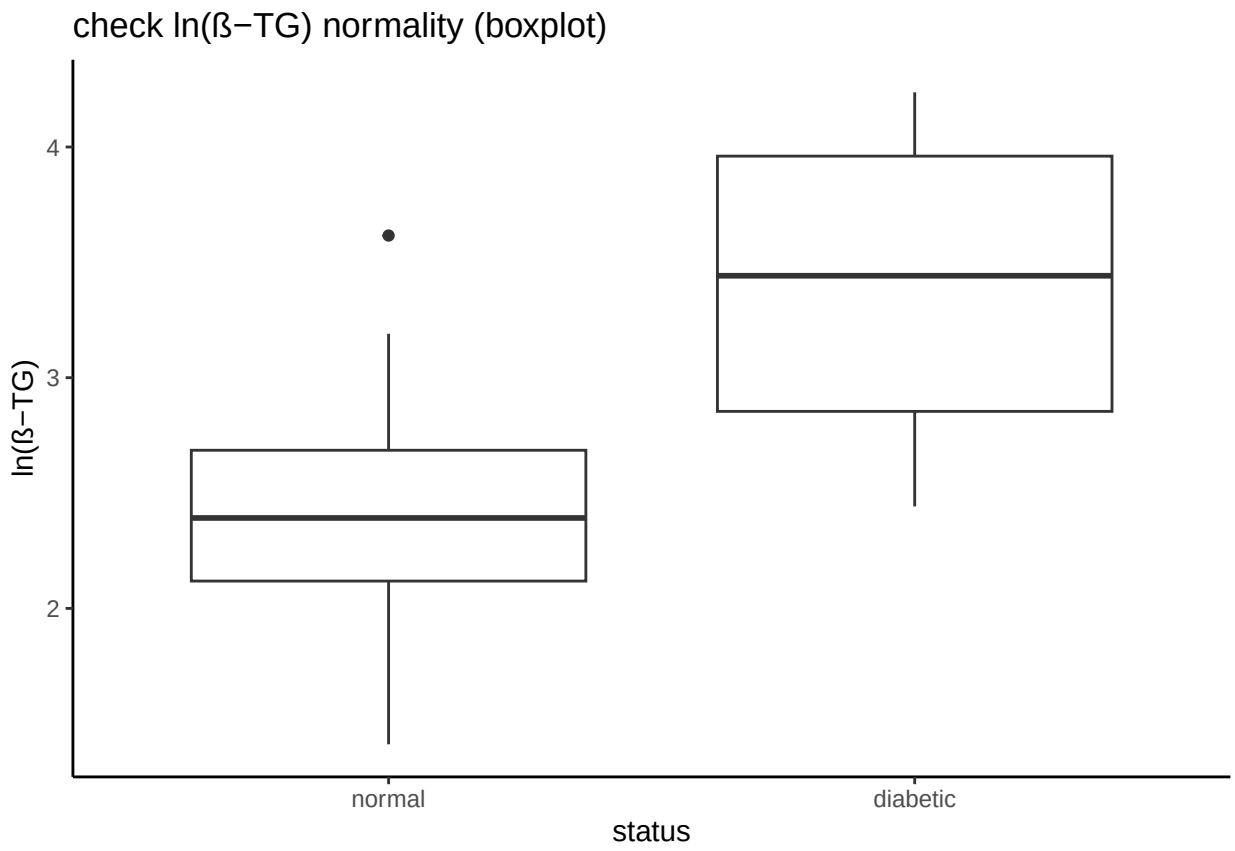
> note that if normality is not met, please indicate that, but then proceed with a t-test rather than the non-parametric alternative that we will learn next week

- Declare hypotheses:
  - $H_0 : \mu_{normal} = \mu_{diabetes}$
  - $H_1 : \mu_{normal} \neq \mu_{diabetes}$
- Choose alpha:
  - $\alpha = 0.05$
- Check assumptions using exploratory plots:
  - normality: both groups are fairly normally distributed, though there is one outlier in the non-diabetic group
  - independent samples: these data are collected from separate groups of mice, so the samples are independent
  - variance: the two groups have equal variance

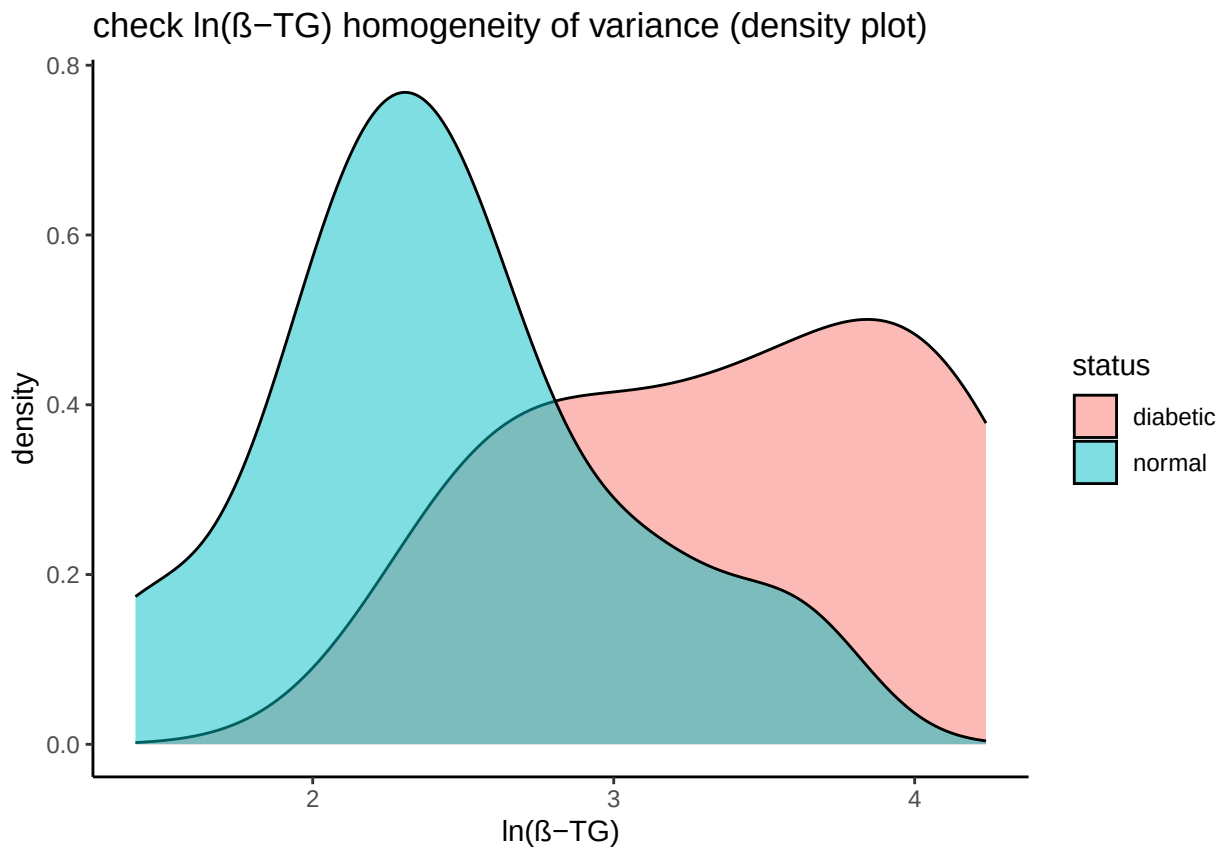
```
# check normality with qq plot
btg |>
  ggplot(aes(sample = logbtg, color = status)) +
  geom_qq() + geom_qq_line() +
  facet_wrap(~status) +
  labs(title = "check ln(\u03B2-TG) normality (QQ plot)") +
  theme_classic()
```




```
# check normality with boxplot
btg |>
  ggplot(aes(x = reorder(status, logbtg), y = logbtg)) +
  geom_boxplot() +
  labs(title = "check ln(\u03B2-TG) normality (boxplot)",
       x = "status",
       y = "ln(\u03B2-TG)") +
  theme_classic()
```




```
# check variance with density plot
btg |>
  ggplot(aes(logbtg, fill = status)) +
  geom_density(alpha = 0.5) +
  labs(title = "check ln(\u03B2-TG) homogeneity of variance (density plot)",
       x = "ln(\u03B2-TG)") +
  theme_classic()
```




```
# check with descriptive statistics
logbtg.stats <- btg |> group_by(status) |> summarize(logbtg_mean = mean(logbtg),
                                                     logbtg_sd = sd(logbtg),
                                                     logbtg_var = var(logbtg),
                                                     n = n())
logbtg.stats
```

```
## # A tibble: 2 x 5
##   status   logbtg_mean logbtg_sd logbtg_var     n
##   <chr>          <dbl>     <dbl>      <dbl> <int>
## 1 diabetic        3.39     0.637      0.406    12
## 2 normal          2.43     0.595      0.354    12
```

- Calculate test statistic:

```
t.test(logbtg ~ status, data = btg, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  logbtg by status
## t = 3.8041, df = 22, p-value = 0.0009714
## alternative hypothesis: true difference in means between group diabetic and group normal is not equal
## 95 percent confidence interval:
```

```
##  0.4353973 1.4791602
## sample estimates:
## mean in group diabetic   mean in group normal
##               3.390628               2.433349
```

- Compare p to α and conclude:
  - $p = 9.7143973 \times 10^{-4}$, which is less than $\alpha = 0.05$
  - Therefore we reject the null hypothesis and conclude that there is a difference in β-TG excreted in the urine of diabetic vs. non-diabetic mice

C. Compare and contrast the results from A and B.

The results from both t-tests lead to the rejection of the null hypothesis ($p < \alpha$) with the conclusion that there is a true difference in β-TG excreted in the urine of diabetic vs. non-diabetic mice. Taking the natural log of measured β-TG values allows us to use a pooled t-test because the groups have equal variance after the transformation, which greatly increases the power of the test.