Lab Assignment 5 Updated

Samantha Chen

5/6/2021

Question 1: Dandelions and the habitat edge

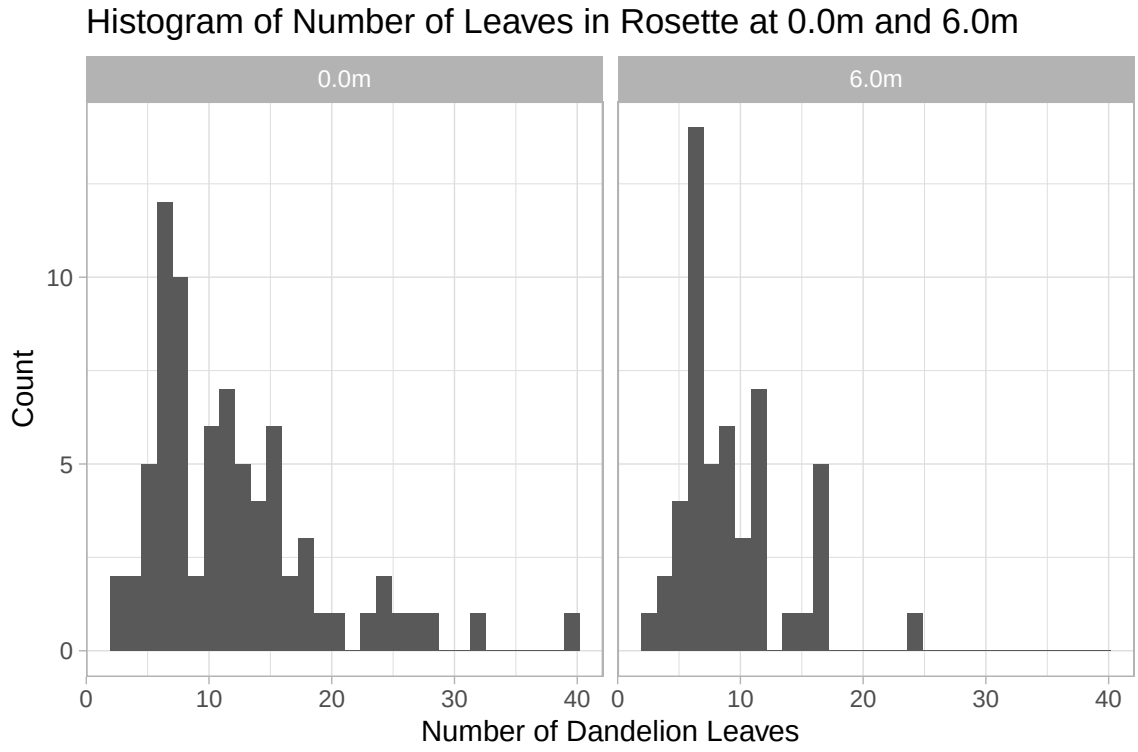
Is there a difference in the number of dandelion leaves per rosette between dandelion 0m from the habitat edge and 6m from the habitat edge?

Look at histograms and boxplots of num_leaves_in_rosette at 0.0m and 6.0m. Based on these boxplots, are there any outliers? What is the shape of each distribution?

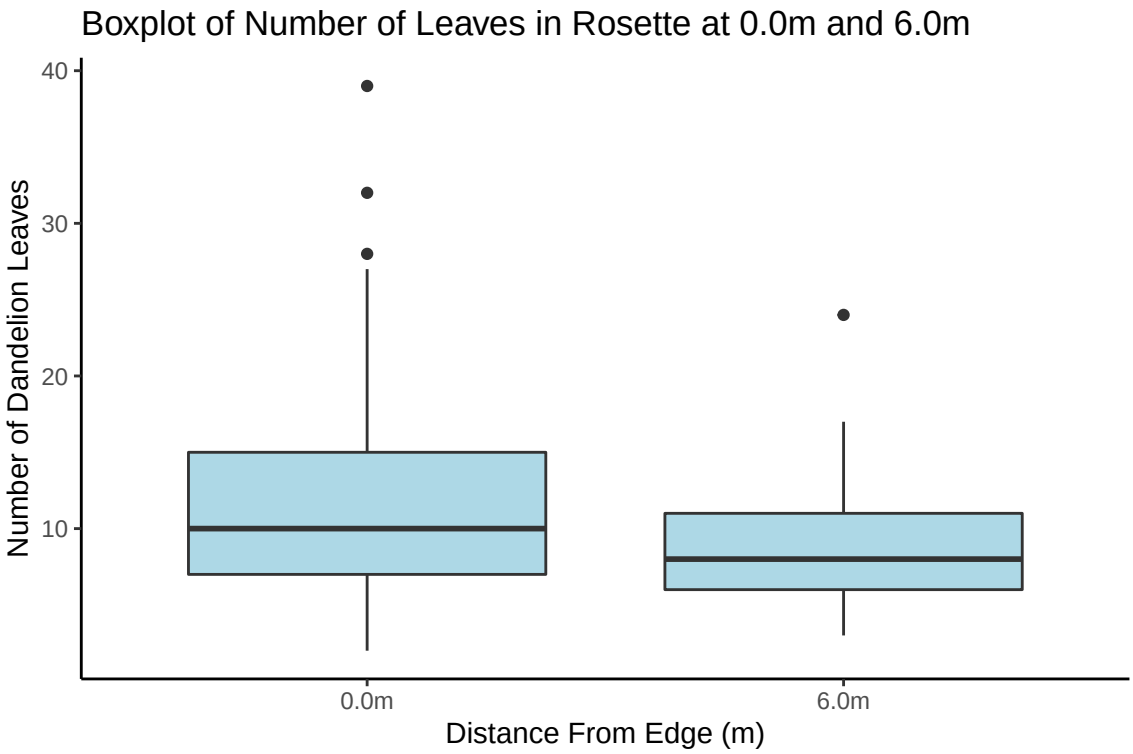
```
# modifying data
plantdata <- read.csv("~/EEMB 146 Lab Files/Lab 5 Data/plant_data.csv")
sub_plant_data <- plantdata[plantdata$dist_from_edge_m != "3.0m",]
levels(sub_plant_data$dist_from_edge_m)
```

```
## NULL
```

```
# visualizing data --histogram
ggplot(sub_plant_data, aes(x = num_leaves_in_rosette)) +  #histogram
  geom_histogram() +
  facet_wrap(~ dist_from_edge_m) +  #separates histogram based on a certain grouping
  theme_light() +
  stat_bin(bins = 30) +
  labs(x = "Number of Dandelion Leaves", y = "Count") +
  ggtitle("Histogram of Number of Leaves in Rosette at 0.0m and 6.0m")
```




```
#boxplot
ggplot(sub_plant_data, aes(x = dist_from_edge_m, y = num_leaves_in_rosette)) +
  geom_boxplot(fill = "lightblue") +
  theme_classic() +
  labs(x = "Distance From Edge (m)", y = "Number of Dandelion Leaves") +
  ggtitle("Boxplot of Number of Leaves in Rosette at 0.0m and 6.0m")
```



The histogram of both 0.0m and 6.0m show asymmetry suggesting the data is not normal. Histogram of 0.0m has a tail indicating a right skew while histogram of 6.0m doesn't have a tail but still looks like it is right skewed as well.

Based on the boxplots 0.0m data show has 3 outliers while 6.0m has 1. It makes sense that 0.0m has more outliers because it is right skewed.In addition, 0.0m show a larger spread than 6.0m.

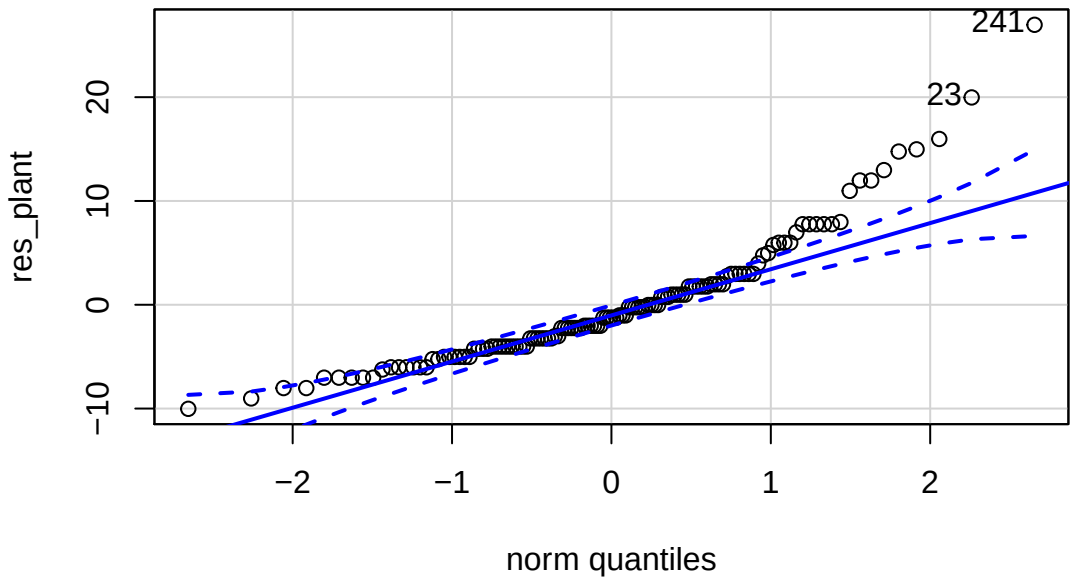
Check the assumptions of normality of num_leaves_in_rosette at 0.0m and 6.0m from the habitat edge using the residuals (see appendix for how to do this). Are the residuals normal or not normal? Show a QQ-plot and a Shapiro-Wilk statistic.

```
# checking normality of data through residuals
fit_plant <- lm(num_leaves_in_rosette~dist_from_edge_m, data = sub_plant_data)
res_plant = fit_plant$residuals

shapiro.test(res_plant)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_plant
## W = 0.89044, p-value = 3.647e-08
```

```
# p-value = 3.647e-08
qqPlot(res_plant)
```




```
## 241  23
## 121   3
```

A Shapiro-Wilk test was done on the residuals of dist_from edge_m vs. num_leaves_in_rosette. The null hypothesis (H_0) is that the residual data is normal and the alternative hypothesis (H_A) is that the residual data is not normal. The p-value of the test is 3.647e-08, which means there is a 0% chance that getting W = 0.89044 through pure random chance. Because the p-value is much smaller than p = 0.05. I reject the hypothesis and can confidently say that the residual data is not normal.

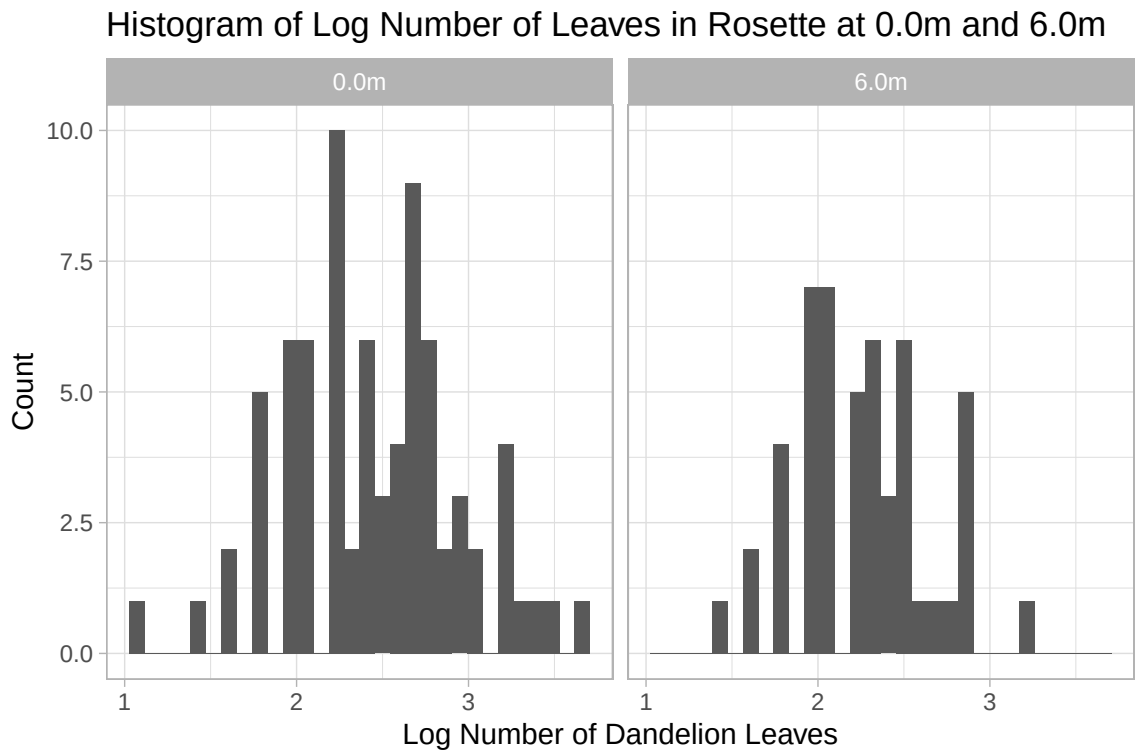
In addition, the qqPlot shows areas where many datapoints were not within the confidence bands which is also an indication of non-normal data. We could assume that the data is normal based on the Central Limit Theorem because the sample size is N > 50. However I'm going to log transform the data to double check normality.

Specify what transformation you used (you can try a couple, but just report one!) and retest your normality assumptions and check again for outliers. Show me the QQplot and Shapiro-Wilk statistic for the residuals of a transformed variable. Do you feel confident assuming normality?

```
# log transform data
sub_plant_data$log_num_leaves_in_rosette <- log(sub_plant_data$num_leaves_in_rosette + 1)

ggplot(sub_plant_data, aes(x = log_num_leaves_in_rosette)) +
  geom_histogram() +
  facet_wrap(~ dist_from_edge_m) +
  theme_light() +
  stat_bin(bins = 30) +
```

```
  labs(x = "Log Number of Dandelion Leaves", y = "Count") +
  ggtitle("Histogram of Log Number of Leaves in Rosette at 0.0m and 6.0m")
```



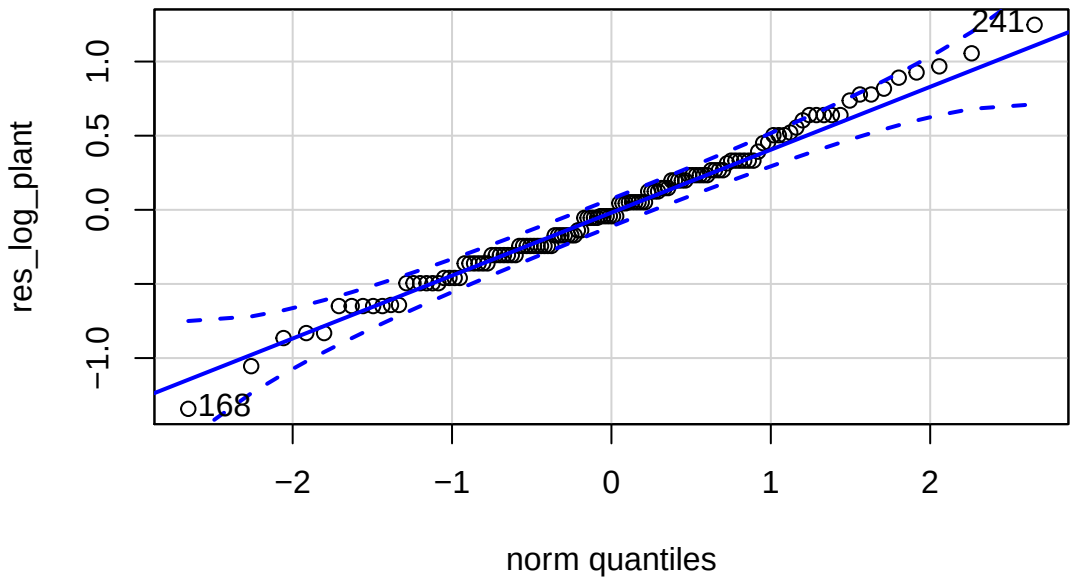

```
#Both histograms look normal
```

```
# residuals of log plant data
fit_log_plant <- lm(log_num_leaves_in_rosette~dist_from_edge_m, data = sub_plant_data)
res_log_plant = fit_log_plant$residuals

shapiro.test(res_log_plant) # p-value = 0.798
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_log_plant
## W = 0.99312, p-value = 0.798
```

```
qqPlot(res_log_plant)
```




```
## 168 241
##  79 121
```

I used log transformation to make my dataset more normal. A Shapiro-Wilk test was done on the residuals of dist_from edge_m vs. log_num_leaves_in_rosette. The null hypothesis (H_0) is that the residual of transformed data is normal and the alternative hypothesis (H_A) is that the residual of transformed data is not normal. The p-value of the test is 0.798, which means there is a 79.8% chance of getting a W-value of 0.99312 purely through random chance. Because the p-value is much larger than p = 0.05, I fail to reject the hypothesis and can confidently say that the residual of transformed data is normal.

In addition, the qqPlot shows many datapoints were within the confidence bands which is also an indication that the transformed data is normal.

A. Clearly state your null and alternative hypotheses. Remember, these hypotheses will change based on transformations and whether or not you are running a non-parametric test so you may have to update them!

Null Hypothesis: there is no difference in the mean number of dandelion leaves per rosette between 0m from the habitat edge and 6m from the habitat edge (H_0: mu1 = mu2). Alternative Hypothesis: there is a difference in the mean number of leaves per rosette between 0m from the edge and 6m from the edge. (H_A: mu1 =/= mu2)

B. Check your homogeneity of variance assumption and report the p-value from your Levene's Test. Remember, you have to test this assumption even if you are using a Mann-Whitney U-test. If you reject this assumption you can still do a Mann-Whitney U-test, but you have to be careful whether you null hypothesis is that the medians are different between the two groups or that the shapes of the distributions are different between the two groups.

```
# levene test for plant data
leveneTest(sub_plant_data$log_num_leaves_in_rosette, sub_plant_data$dist_from_edge_m)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  4.1152 0.04464 *
##       124
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p-value = 0.04464
```

A levene test was done on the transformed dandelion data in which the null hypothesis (H_0) is that the variances of the logged data are equal and the alternative hypothesis (H_A) is that the variances of the logged data are not equal. The test produced a p-value of 0.04464, which means there is a 44.64% chance of getting an F-value of 4.1152 through pure random chance. Because the p-value is less than 0.05, I reject the null hypothesis that the variances of transformed data are equal.

Since the variances are not equal I decided to do a Welch's t-test on the transformed data. I'm choosing to use this instead of the Mann-Whitney U-Test because the Welch's t-test is a parametric test and therefore is more powerful than the Mann-Whitney U-test.

C. Show your test statistic and your p-value for your test.

```
# two sample t-test
dist_from_edge_0m <- subset(sub_plant_data, dist_from_edge_m == "0.0m")
dist_from_edge_6m <- subset(sub_plant_data, dist_from_edge_m == "6.0m")
t.test(dist_from_edge_0m$log_num_leaves_in_rosette, dist_from_edge_6m$log_num_leaves_in_rosette,
       var.equal = FALSE) # p-value = 0.02555
```

```
##
##  Welch Two Sample t-test
##
## data:  dist_from_edge_0m$log_num_leaves_in_rosette and dist_from_edge_6m$log_num_leaves_in_rosette
## t = 2.3875, df = 121.17, p-value = 0.01851
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03237402 0.34671592
## sample estimates:
## mean of x mean of y
##  2.440960  2.251415
```

In this Welch's t-test:

Null Hypothesis: there is no difference in the mean of log transformed number of dandelion leaves per rosette between 0m from the habitat edge and 6m from the habitat edge (H_0: mu1 = mu2). Alternative Hypothesis: there is a difference in the mean of log transformed number of leaves per rosette between 0m from the edge and 6m from the edge. (H_A: mu1 =/= mu2)

The test statistic is t = 2.3875 and the p-value is p = 0.01851, which means there is a 1.851% probability that I got this t-statistic through pure random chance.

D. Clearly state your conclusion regarding how the number of leaves in a rosette differs between 0m from the habitat edge and 6.0m from the habitat edge. Provide a one to two sentence biological interpretation of your conclusion (no right answer here, just make sure it is logical and complete).

Because the the Welch's t-test calculated a p-value < 0.05, I can reject the null hypothesis that there is no difference in the log number of rosette leaves between 0m and 6.0m from edge. Since there is a difference between the two distances from the edge this tells me that there might be environmental differences between 0.0m and 6.0m from edge such as amount of water, dirt composition, and/or amount of herbivory that affects the number of rosette leaves.

Question 2: Starving Crickets

Is there a difference in the mean waiting time to mating between female crickets who were fed and those who were starved?

A. Clearly state your null and alternative hypotheses.

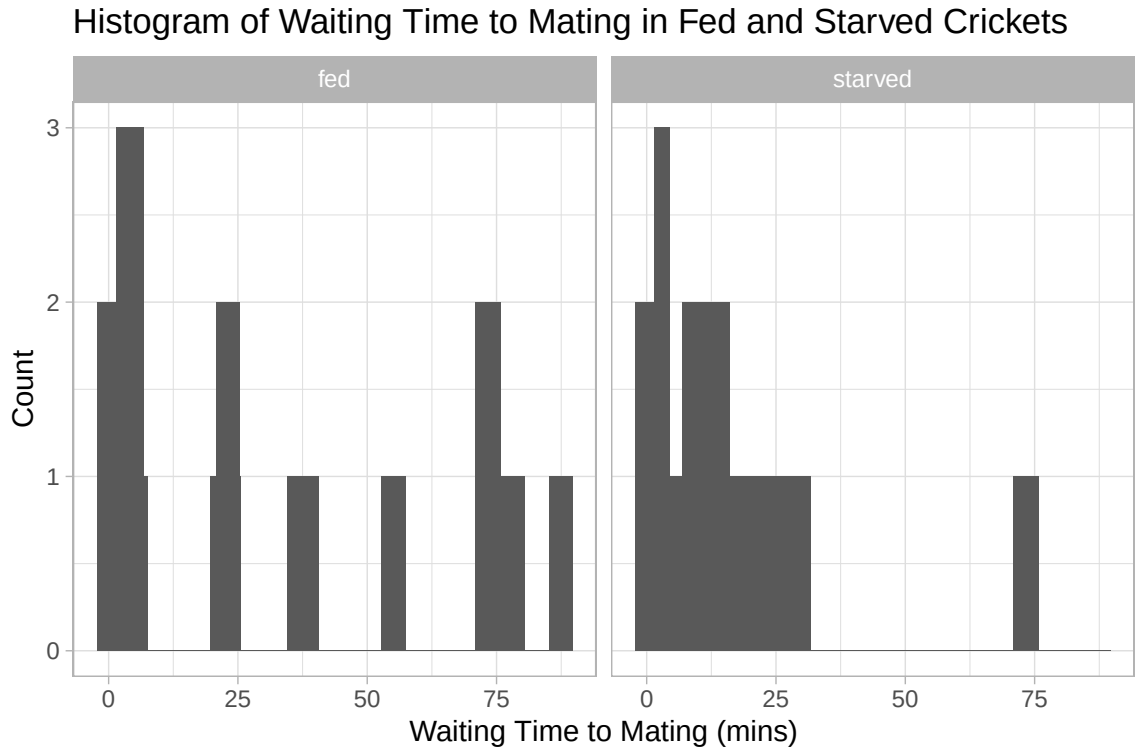
Null Hypothesis: The mean waiting time to mating in fed female crickets is equal to the mean waiting time to mating in starved female crickets (H_0: mu_1 = mu_2).

Alternative Hypothesis: The mean waiting time to mating in fed female crickets is not equal to the mean waiting time to mating in starved female crickets (H_A: mu_1 =/= mu_2).

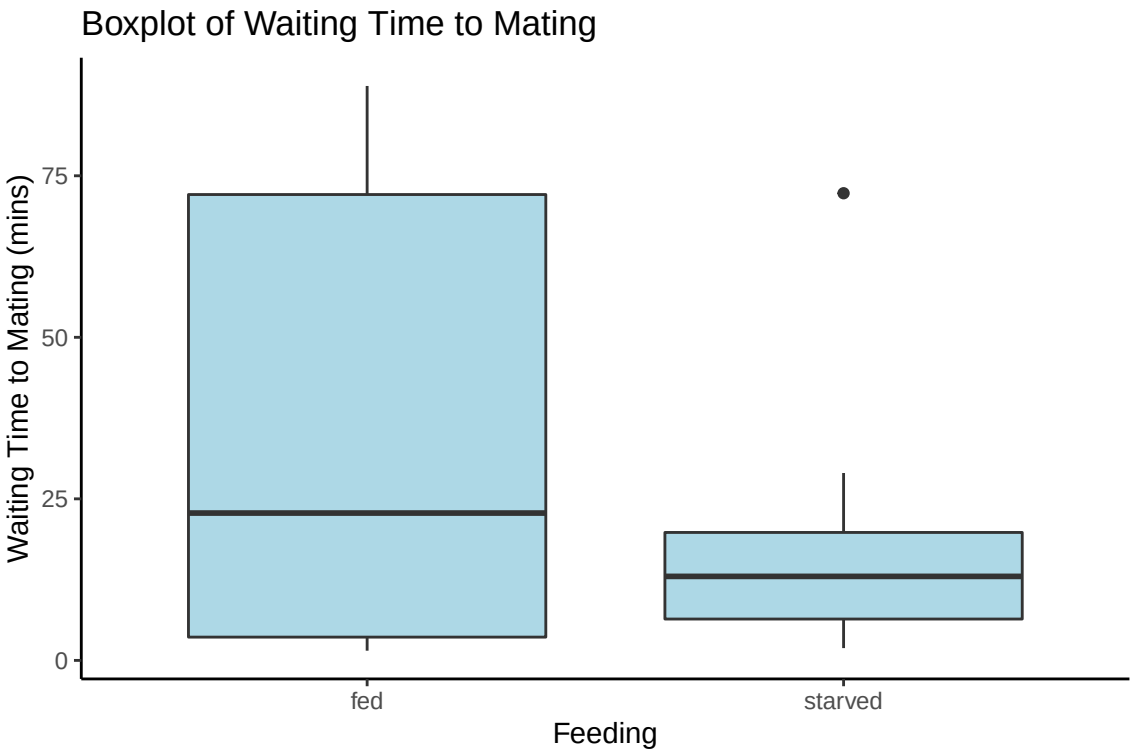
B. Visualize your data in some meaningful way and show this plot in your report.

```
#loading in cricket data
cricket_data <- read.csv("~/EEMB 146 Lab Files/Lab 5 Data/starving_cricket.csv")

# visualizing data --histogram
ggplot(cricket_data, aes(x = time_to_mating)) +
  geom_histogram() +
  facet_wrap(~ feeding) +  #separates histogram based on a certain grouping
  theme_light() +
  stat_bin(bins = 20) +
  labs(x = "Waiting Time to Mating (mins)", y = "Count") +
  ggtitle("Histogram of Waiting Time to Mating in Fed and Starved Crickets")
```




```
#boxplot
ggplot(cricket_data, aes(x = feeding, y = time_to_mating)) +
  geom_boxplot(fill = "lightblue") +
  theme_classic() +
  labs(x = "Feeding", y = "Waiting Time to Mating (mins)") +
  ggtitle("Boxplot of Waiting Time to Mating ")
```



The histogram for both fed and starved plots do not look normal and shows a large spread as well as a slight right skew. This is also prevalent in the boxplot in which fed cricket data has a larger spread compared to that of the starved cricket data. Fed cricket data has a higher median compared to that of the starved cricket data. Only the starved data has an outlier at around 75 mins.

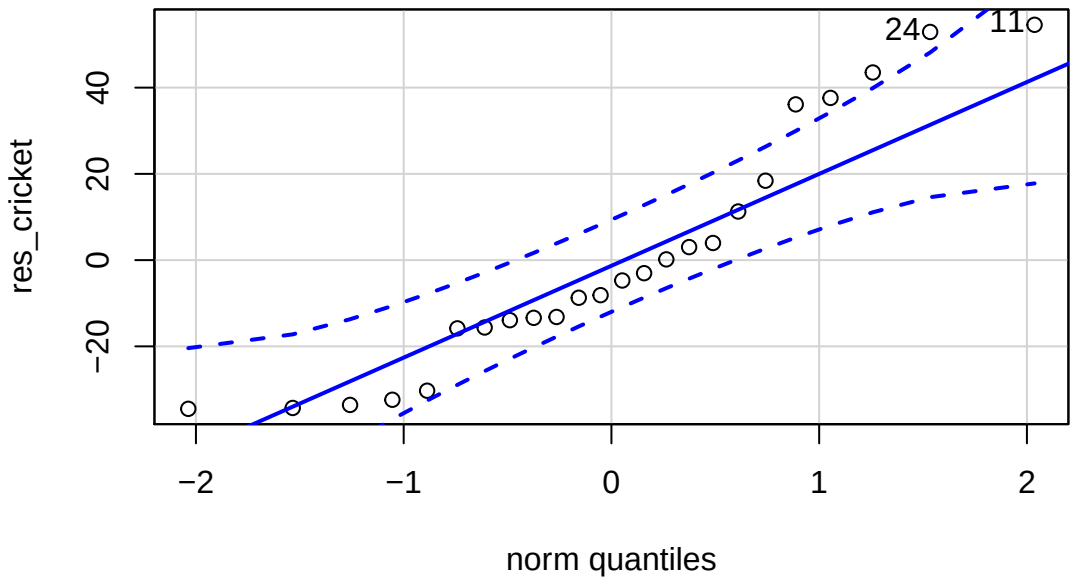
C. Test whether the two groups of crickets (starved and fed) follow a normal distribution by testing the normality of the residuals (see appendix for how to do this). Are the residuals normal or not normal? Show a QQ-plot and a Shapiro-Wilk statistic.

```
# first will be testing normality of non-transformed data
fit_cricket <- lm(time_to_mating~feeding, data = cricket_data)
res_cricket = fit_cricket$residuals

shapiro.test(res_cricket)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_cricket
## W = 0.90871, p-value = 0.03308
```

```
# p-value = 0.03308
qqPlot(res_cricket)
```




```
## [1] 11 24
```

A Shapiro-Wilk test was done on the residuals of time_to_mating vs. feeding. The null hypothesis (H_0) is that the residual data is normal and the alternative hypothesis (H_A) is that the residual data is not normal. The residual data got a p-value of 0.03308, which means there is 3.3% chance of getting a W-value of 0.90871 purely through random chance. Since the p-value is smaller than p = 0.05, I reject the hypothesis and can confidently say that the residual data is not normal.

In addition, in the qqPlot, many of the datapoints were not within the confidence bands which is also an indication of non-normal data.

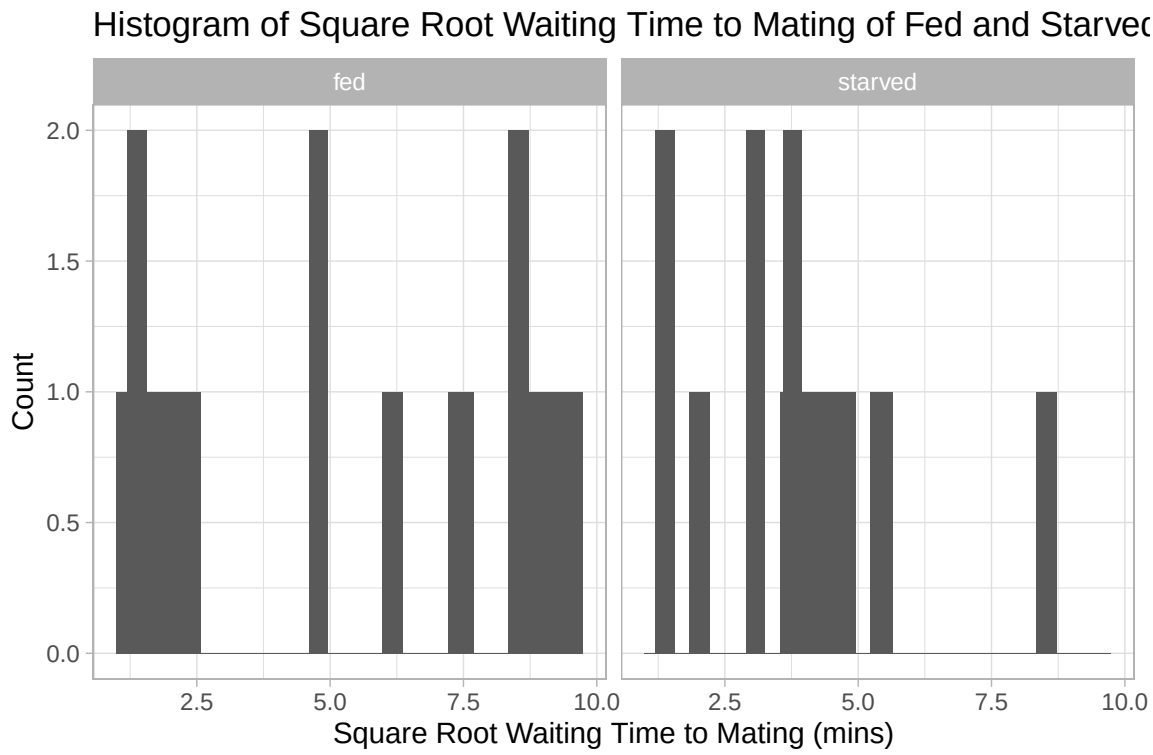
We cannot assume that the data is normal based on the Central Limit Theorem because the sample size is smaller than 50, So I will try transforming the data.

D. Try one transformation on your data that you think is reasonable (note: you can try more than one, but only include one in your homework). Retest your normality assumptions, give the QQ-plot and the Shapiro-Wilk statistic, and report whether this transformation made the data normal.

```
# sqrt transform cricket data
cricket_data$sqrt_time_to_mating <- sqrt(cricket_data$time_to_mating)

#visualize sqrt data
ggplot(cricket_data, aes(x = sqrt_time_to_mating)) +
  geom_histogram() +
  facet_wrap(~ feeding) +
  theme_light() +
```

```
  stat_bin(bins = 25) +
  labs(x = "Square Root Waiting Time to Mating (mins)", y = "Count") +
  ggtitle("Histogram of Square Root Waiting Time to Mating of Fed and Starved Crickets")
```

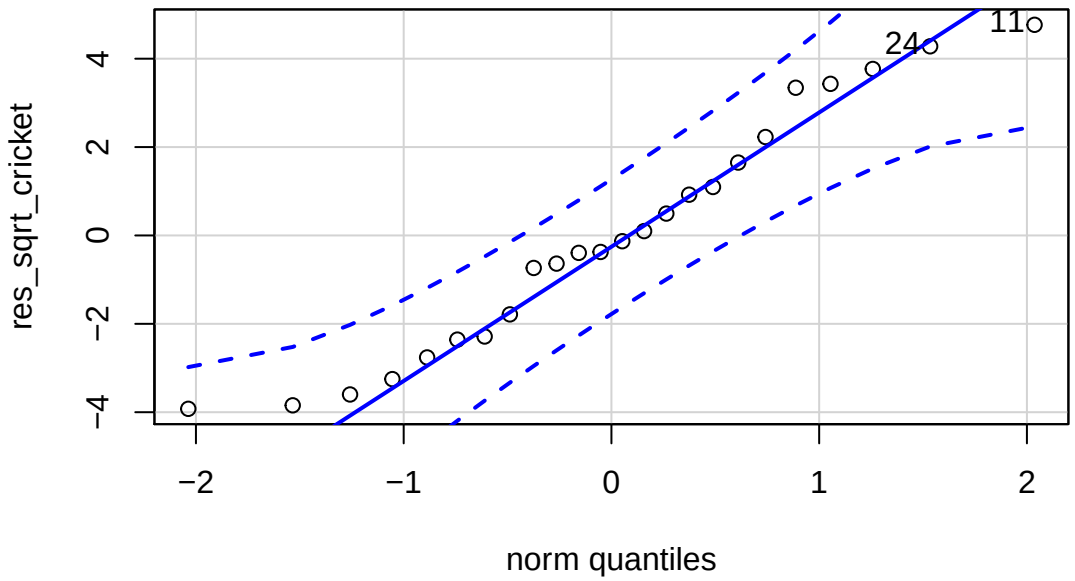



```
#test normality of sqrt data
fit_sqrt_cricket <- lm(sqrt_time_to_mating~feeding, data = cricket_data)
res_sqrt_cricket = fit_sqrt_cricket$residuals

shapiro.test(res_sqrt_cricket)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_sqrt_cricket
## W = 0.95184, p-value = 0.2968
```

```
# p-value = 0.2968
qqPlot(res_sqrt_cricket)
```




```
## [1] 11 24
```

I tested out two different transformations on the data, log transformation (not shown) and square root transformation (sqrt). The sqrt transformed data looked normal so I kept it instead of the log transformed data.

A Shapiro-Wilk test was done on the residuals of sqrt_time_to_mating vs. feeding. The null hypothesis (H_0) is that the transformed residual data is normal and the alternative hypothesis (H_A) is that the transformed residual data is not normal. The p-value of the test is 0.2968, which means there is a 29.68% chance of getting a W-value of 0.95184 purely through random chance. Because the test's p-value is larger p = 0.05, I fail to reject the hypothesis and can confidently say that the transformed residual data is normal.

In addition, in the qqPlot, most of the transformed datapoints were within the confidence bands which is also an indication of normality.

E. Test the assumption that the variances between the two groups is equal for either the transformed or untransformed data, depending on what you are going to analyze. Show me the resulting Levene's test p-value. Interpret your Levene's Test p-value.

```
# levene test for sqrt cricket data
leveneTest(cricket_data$sqrt_time_to_mating, cricket_data$feeding)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  5.1746 0.03302 *
##       22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p-value = 0.03302
```

A levene test was done on the transformed cricket data in which the null hypothesis (H_0) is that the variances of transformed data are equal and the alternative hypothesis (H_A) is that the variances of transformed data are not equal. This p-value of the levene test tells me that the chance of getting an F-score of 5.1746 by chance is 3.3%. Because the test produced a p-value of 0.03302, which is less than 0.05, I reject the null hypothesis that the sqrt variances are equal.

Since the variances are not equal I decided to do a Welch's t-test on the transformed data. I'm choosing to use this instead of the Mann-Whitney U-Test because the Welch's t-test is a parametric test and therefore is more powerful than the Mann-Whitney U-test.

F. Based on your results above, test your hypothesis using the appropriate two-sample test and give the results of the test. State your conclusions in terms of rejecting or failing to reject your null hypothesis. Based on your result, give a logical explanation on why you think female crickets eat the male crickets' wings.

```
# two sample t-test
feeding_fed <- subset(cricket_data, feeding == "fed")
feeding_starved <- subset(cricket_data, feeding == "starved")
t.test(feeding_fed$sqrt_time_to_mating, feeding_starved$sqrt_time_to_mating, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  feeding_fed$sqrt_time_to_mating and feeding_starved$sqrt_time_to_mating
## t = 1.3075, df = 20.568, p-value = 0.2055
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.837051  3.662255
## sample estimates:
## mean of x mean of y
##  5.148254  3.735652
```

```
# p-value = 0.2055
```

Because I conducted a Welch's t-test on the transformed data the null and alternative hypotheses are the following:

Null Hypothesis: The mean of sqrt transformed waiting time to mating in fed female crickets is equal to the mean of sqrt transformed waiting time to mating in starved female crickets (H_0: mu_1 = mu_2).

Alternative Hypothesis: The mean of sqrt transformed waiting time to mating in fed female crickets is not equal to mean of sqrt transformed waiting time to mating in starved female crickets (H_A: mu_1 =/= mu_2).

The test statistic is 1.3075 and the p-value is 0.2055, which means the probability of getting a t-score of 1.3075 through random chance is 20.55%. Because the p-value is much larger than p = 0.05. Therefore I fail to reject the null hypothesis and say there is no difference between the time to mating between fed and starved crickets. There is a 20.55% that any difference between time to mating is due to random chance. Based on my results, it seems that it doesn't matter whether female crickets are starved or fed and that they will mate whether hungry or not. Perhaps female crickets can detect the male's fitness through consumption of their wings which is useful deciding on a mate that will produce the best offsprings.