EEMB146 Lab 5 Homework

Written by Rafael Romero

2025-02-08

Contents

Homework 5 – Comparing Means with Non-Parametric Tests

This homework will apply your data visualization skills (Lab 3), ability to determine normality and appropriate t-test (Lab 4 & 5). If you are having trouble with RStudio or knitting your .Rmd file please speak with a TA before the due date. **All of the information and code needed to answer homework questions can be found in Lab 4 & 5 Exercise files.** You will be graded on completeness and correctness.

Homework Questions

Question 1

Please answer **True or False** for 1a-b.

a) You have to add +1 to all log transformations. **FALSE**
b) If you log transform data that contain 0 and don't add +1, the minimum of the data will be equal to -Inf. **TRUE**

Question 2

Apply the correct statistical test (**one sample t-test, paired t-test, two sample t-test, Welch's t-test, Mann-Whitney U, Wilcoxon rank sum**) to the situation.

a) two unrelated groups with equal variance **Two-sample t-test**
b) has normal distribution, but unequal variances **Welch's T-test**
c) $H_0 : \mu_d = 0$ **Paired t-test**
d) if two groups don't have similar shapes, compares the medians of two groups **Wilcoxon Rank Sum**
e) sample compared to theoretical mean **One-sample t-test**
f) $H_0 : \mu = \mu_0$ **One-sample t-test**
g) not normal, unequal variances, but the groups have similar shapes **Mann-Whitney U**

For Question 3 & 4, it will be very important that you choose the most appropriate t-test for each data set. If you are unsure which t-test to perform please look through Lab 5 Exercise

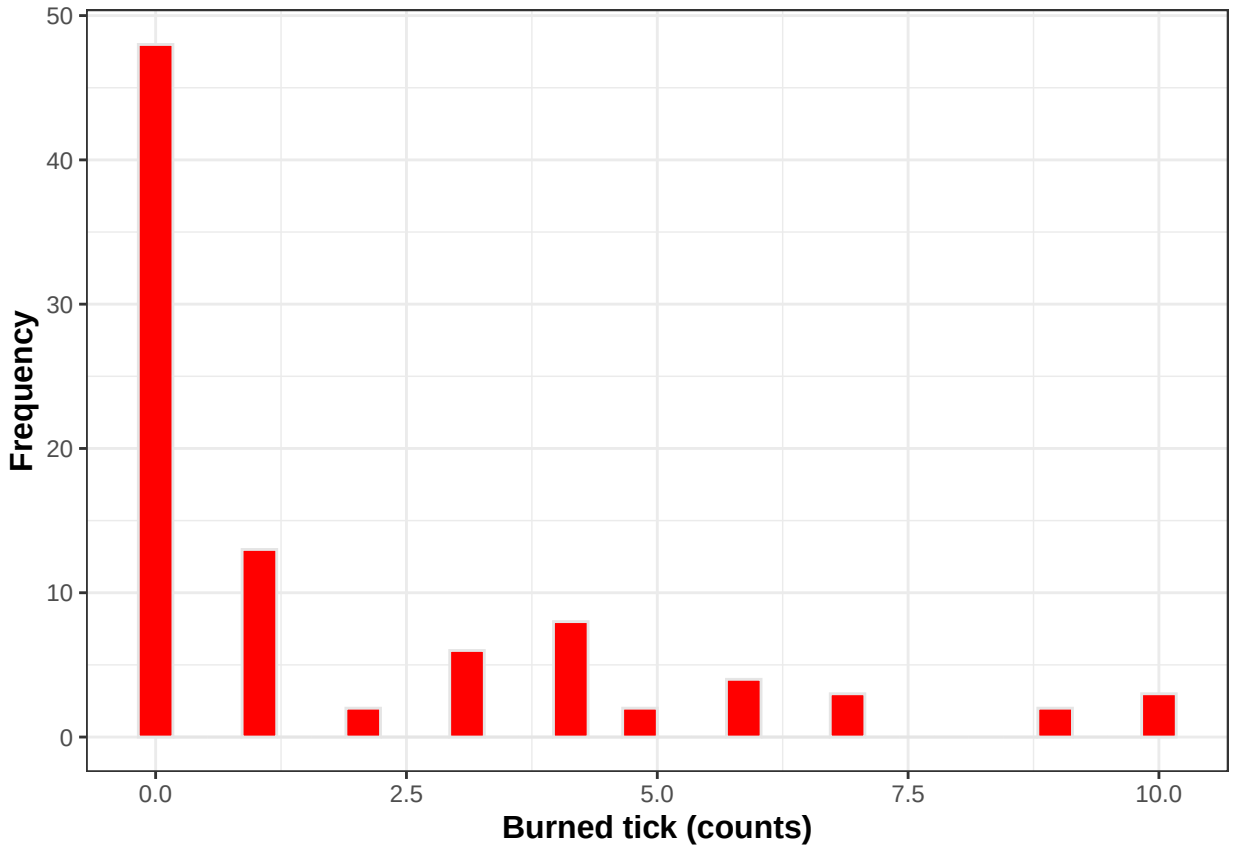
Question 3

For Question 3, you will be using the data set `ticks`. This data is part of EEMB student Sam Sambado's PhD work looking at the effect of wildfire on tick populations throughout the UC Natural Reserve System.

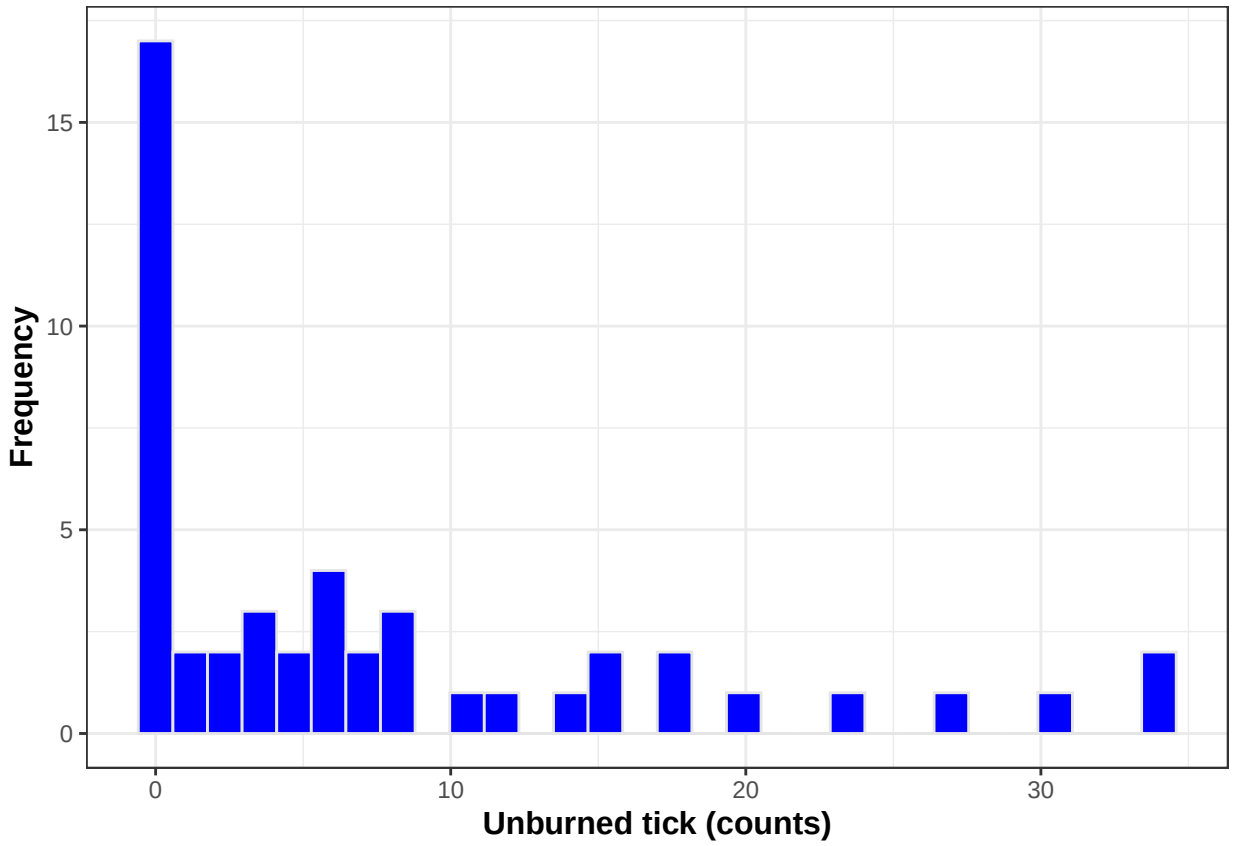
We want to know if there is a statistical difference in tick `counts` by burned and unburned `treatment` plots. To do this, you will need to answer the following questions.

a) Determine the normality of tick counts in burned and unburned plots using a histogram and qqPlot. Based on those 2 tools, what would you conclude about the normality of tick counts for burned and unburned plots?

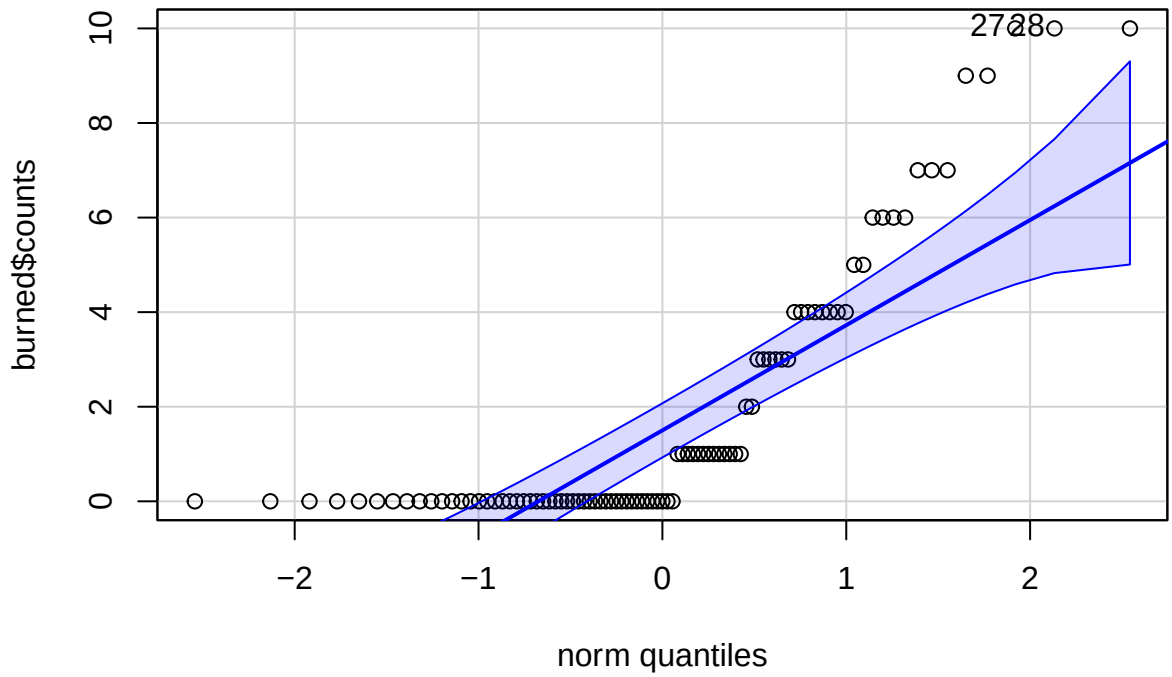
```
burned <- ticks %>% filter(treatment == "burn")
unburned <- ticks %>% filter(treatment == "unburn")
ggplot(burned, aes(x=counts)) +
   geom_histogram(fill = "red", color = "grey90")+
  theme_bw() +
  labs(x = "Burned tick (counts)", y = "Frequency") +
  theme(axis.title = element_text(face = "bold", size = 12))# right skewed distribution
```




```
ggplot(unburned, aes(x=counts)) +
   geom_histogram(fill = "blue", color = "grey90")+
  theme_bw() +
  labs(x = "Unburned tick (counts)", y = "Frequency") +
  theme(axis.title = element_text(face = "bold", size = 12))# right skewed distribution
```

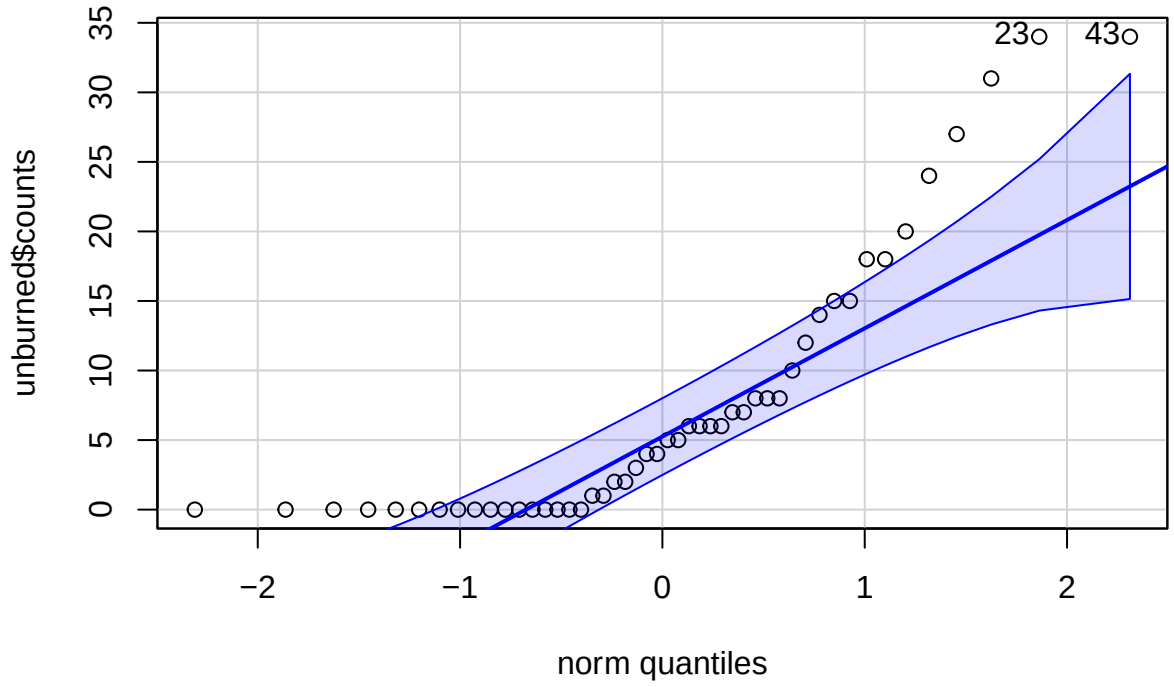



```
qqPlot(burned$counts) #Not normal
```




```
## [1] 27 28
```

```
qqPlot(unburned$counts) #Not Normal
```

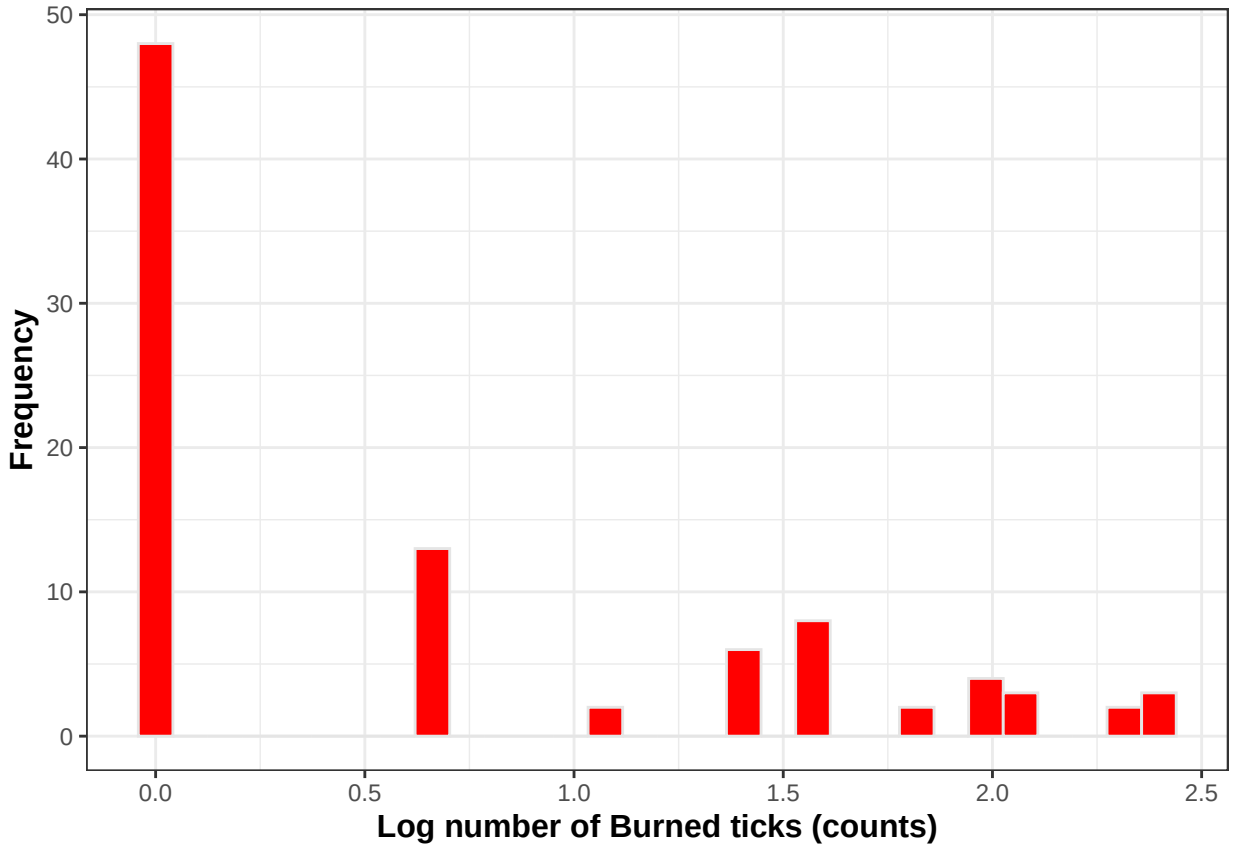



```
## [1] 23 43
```

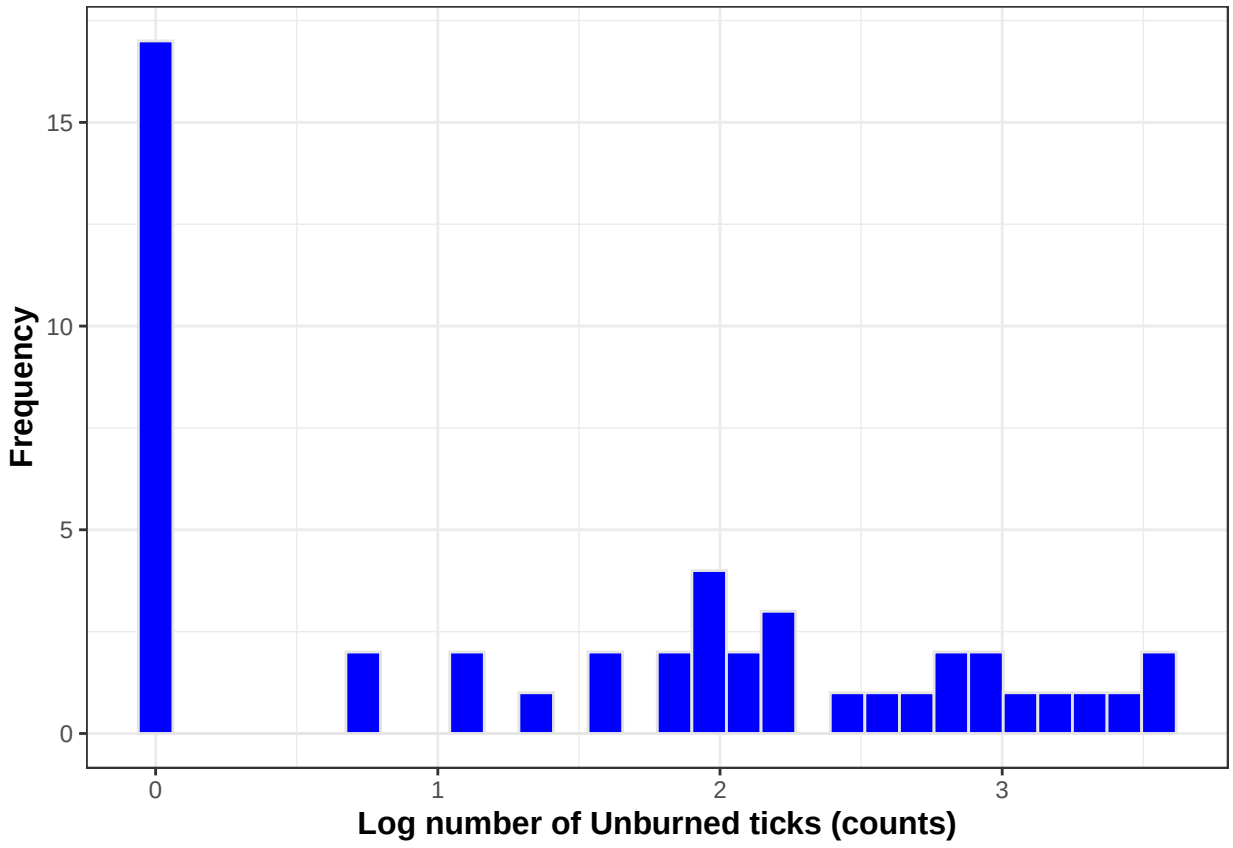
CONCLUSION: I can conclude from the histogram that tick counts for burned and unburned is right skewed and not normal. From the QQplots I can also conclude that they are not normal since the real observations do not accurately follow the CI.

b) Regardless if you think tick counts are normally distributed or not, apply a log transformation and then re-check your assumptions of normality with a histogram & qqPlot. Based on those 2 tools, report on the normality of your transformed tick counts.

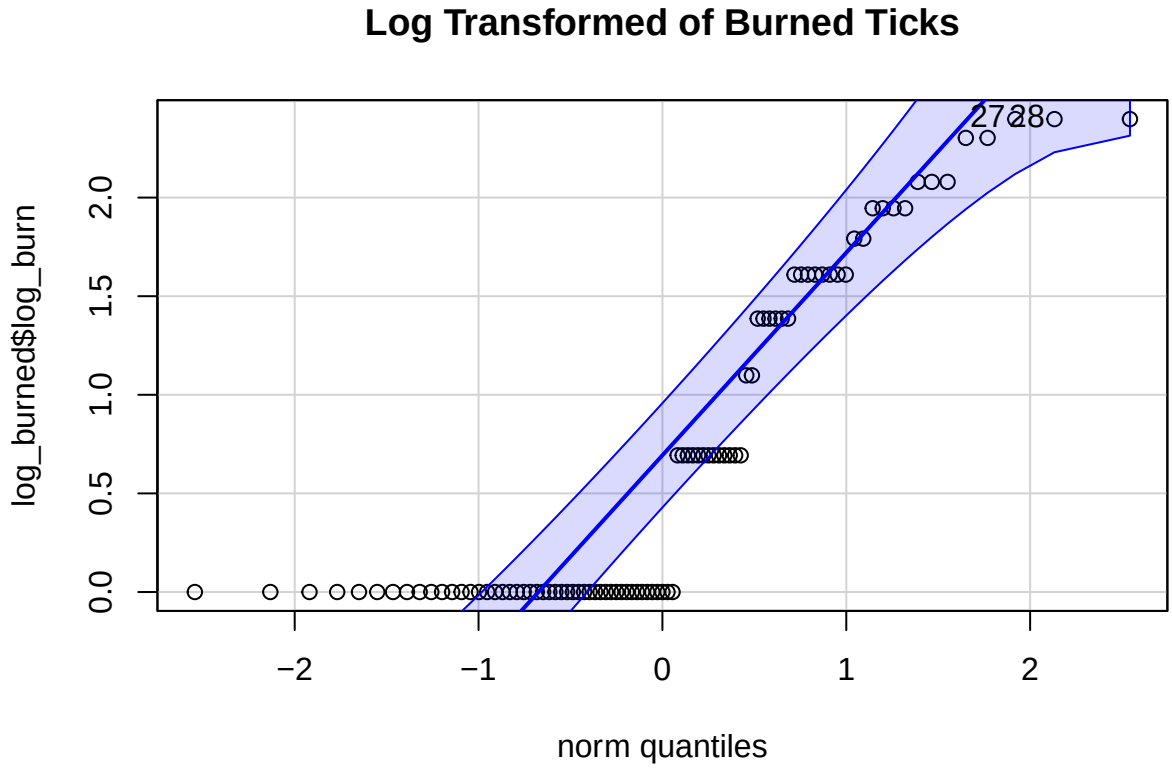
```
log_burned <- burned %>%
  mutate(log_burn = log(counts+1))
log_unburned <- unburned %>%
  mutate(log_unburn =log(counts+1))

ggplot(log_burned, aes(x=log_burn)) +
   geom_histogram(fill = "red", color = "grey90")+
  theme_bw() +
  labs(x = "Log number of Burned ticks (counts)", y = "Frequency") +
  theme(axis.title = element_text(face = "bold", size = 12))# Right skewed
```




```
ggplot(log_unburned, aes(x=log_unburn)) +
   geom_histogram(fill = "blue", color = "grey90")+
  theme_bw() +
  labs(x = "Log number of Unburned ticks (counts)", y = "Frequency") +
  theme(axis.title = element_text(face = "bold", size = 12))# right skewed
```

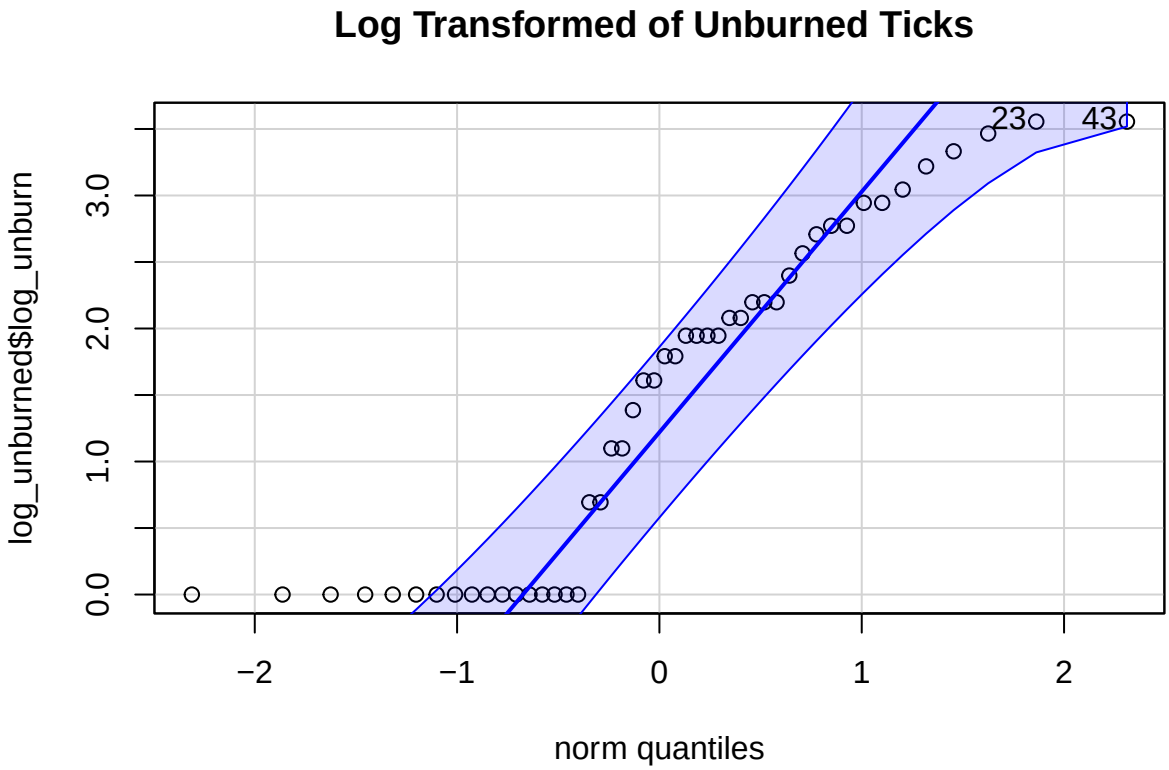



```
qqPlot(log_burned$log_burn) +
  title("Log Transformed of Burned Ticks") #More normal
```



```
## integer(0)
```

```
qqPlot(log_unburned$log_unburn) +
  title("Log Transformed of Unburned Ticks ")#More Normal
```




```
## integer(0)
```

CONCLUSION The data was more normal from both the qqPlots, but was still heavily right skewed from the histograms.

c) Test for equal variances of tick counts between burned and unburned with a Levene's test. Report the p-value from the Levene's test. What does the p-value mean in terms of the null hypothesis (ie equal or not equal variances)?

```
leveneTest(counts ~ treatment, data = ticks)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  32.537 6.914e-08 ***
##       137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion The p value was 6.914e-08 which is less than 0.05, which means we reject the null hypothesis of the variances being equal which means that there is significant evidence that both burned and unburned ticks have unequal variances.

d) Based on the results from your normality (part b) and variance (part c) assumptions tests, what is the appropriate test to run and explain why?

ANSWER Since the data was not normally distributed from part b and variances are not equal, the appropriate test to run would be a non-parametric test such as Mann-Whitney U test.

e) Run the code for your chosen test.

```
wilcox.test(counts ~ treatment, data = ticks)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  counts by treatment
## W = 1426.5, p-value = 0.0003952
## alternative hypothesis: true location shift is not equal to 0
```

f) Based on the test you ran, please report the test statistic and p-value. What can you statistically say about the tick counts at burned and unburned plots?

CONCLUSION The test statistic was 1426.5 and the p-value was 0.0003952. Therefore, since the p-value is less than our significance level of 0.05, we can say that the difference in tick counts at burned and unburned plots are significant.

Question 4

For Question 4, you will be using the data set `blue_crabs`, which is a subset of data from NOAA's Southeast Fisheries Science Center focused on The Jamaica Beach Project in Carancahua Cove, Galveston Bay estuary, Texas. This subset of data will focus on the collection of blue crabs (*Callinectes sapidus*) that were collected in `general_habitat` Marsh or Open Water. We will be focused on the temperature (`temp`) at each habitat type. Temperature (°C) can impact the amount of oxygen that is present in water and can have impacts on blue crab physiology.

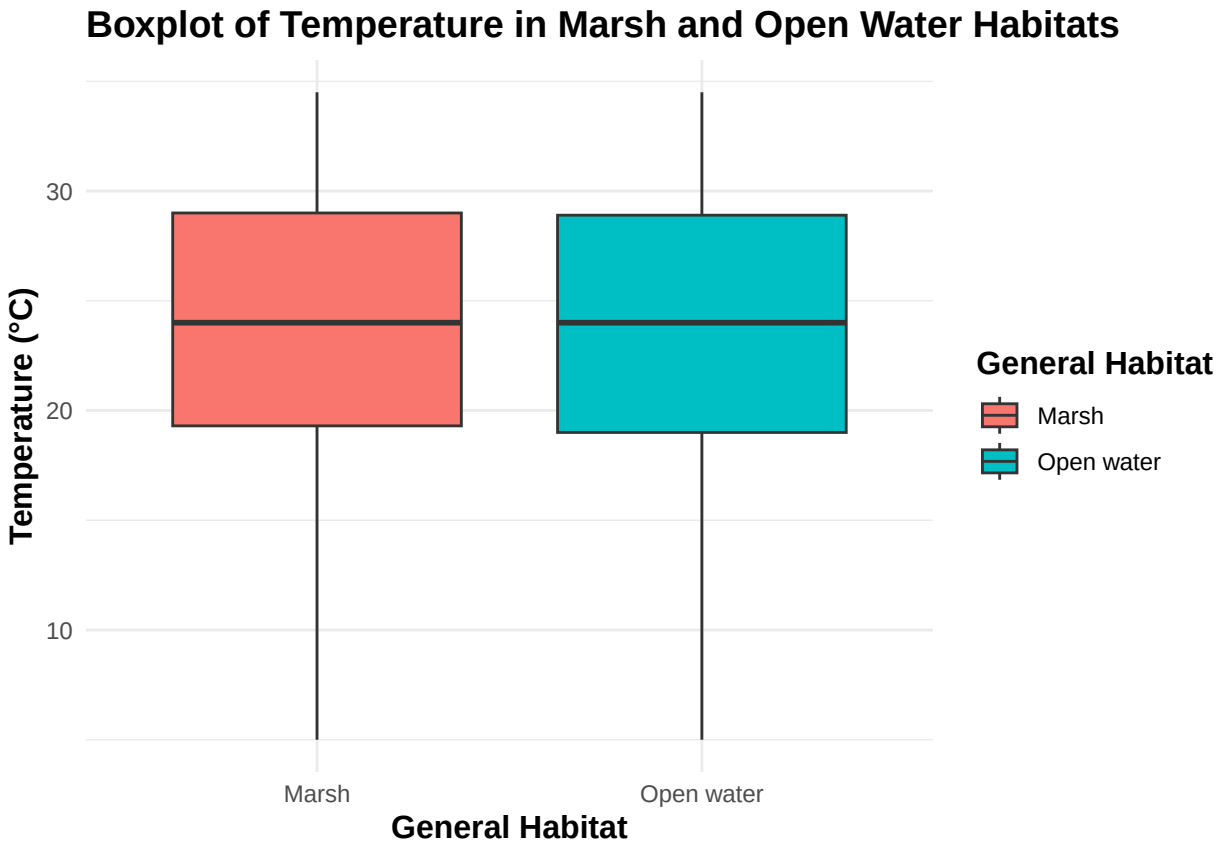
We want to know if there is a difference in temperature levels in marsh or open water. To do this, you will need to answer the following questions.

a) Format 'blue_crabs' data into two data frames `marsh` and `open_water`

```
marsh <- blue_crabs %>% filter(general_habitat == "Marsh")
open_water <- blue_crabs %>% filter(general_habitat == "Open water")
```

b) Determine the normality of `blue_crab` temperature data set with a boxplot(). Must include visualization of boxplot and include informative labels to receive full credit.

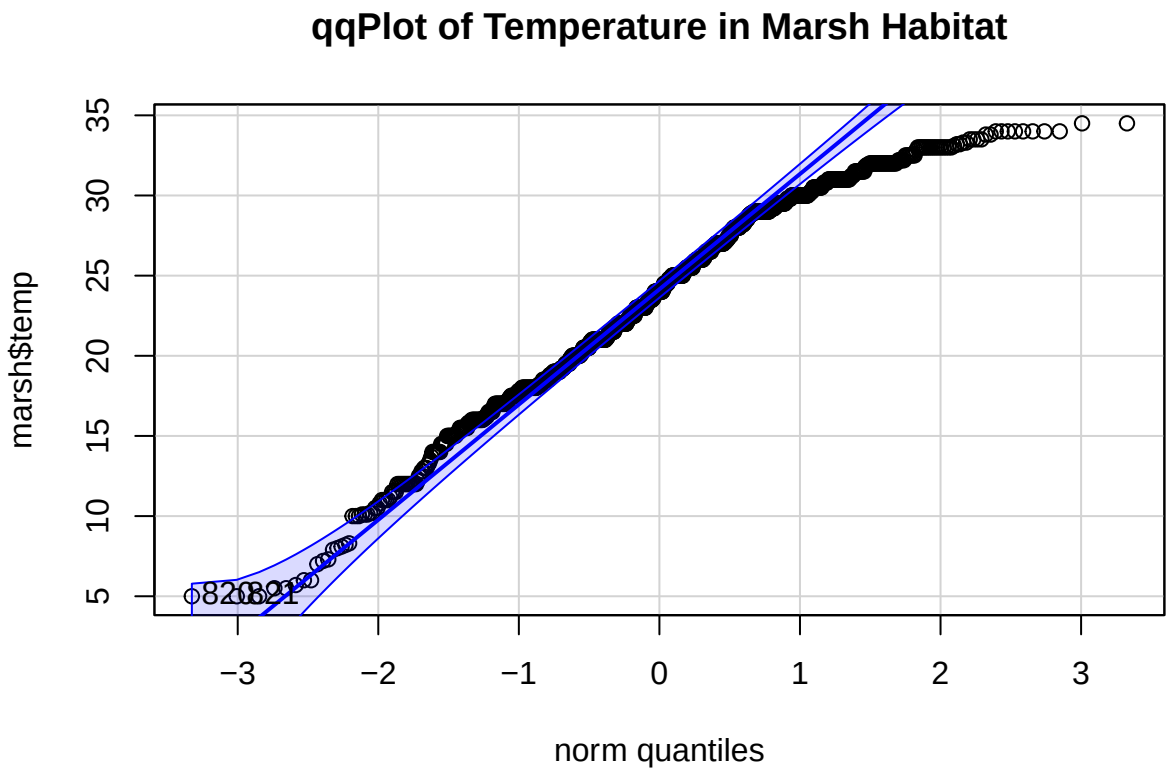
```
blue_crabs %>%
  ggplot(aes(x=general_habitat, y=temp, fill=general_habitat)) +
  geom_boxplot()+
  labs(title = "Boxplot of Temperature in Marsh and Open Water Habitats",
       x="General Habitat",
       y="Temperature (°C)", fill = "General Habitat"
         ) +
  theme_minimal() +
  theme(title = element_text(face = "bold", size = 12))
```



CONCLUSION: The boxplot shows a symmetric distribution as the median is centered within the box, so we can note that the distribution is probably normal.

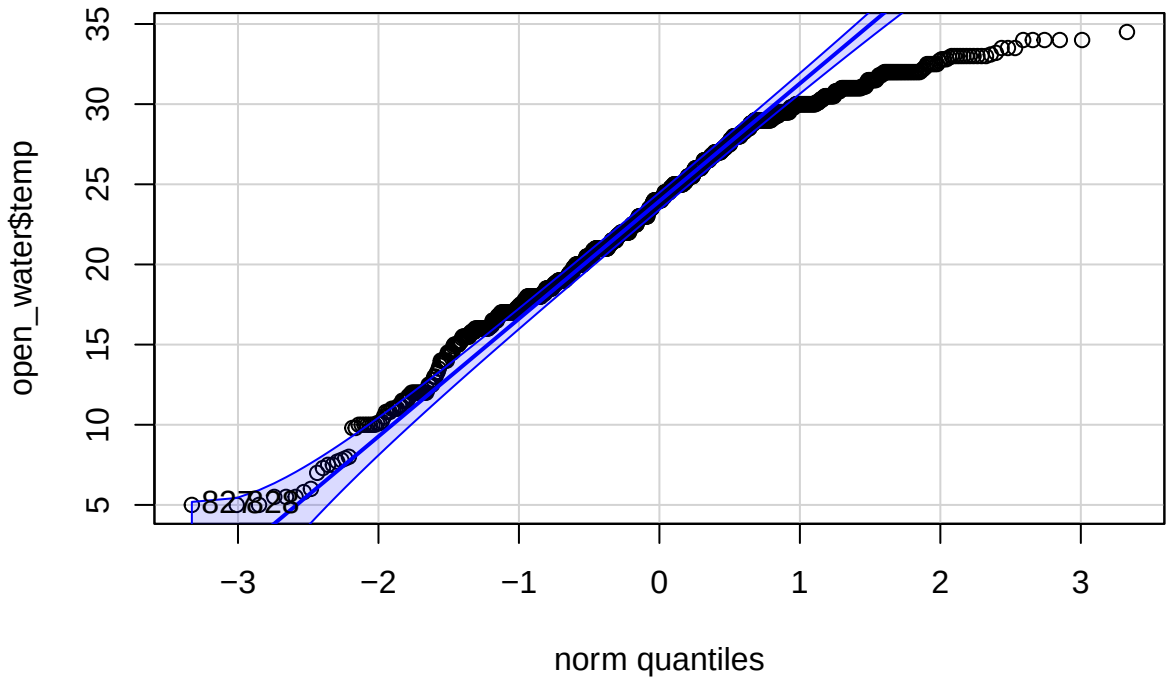
c) Determine the normality for both `marsh` and `open_water` with a qqPlot(). Must include visualization of both qqPlots to receive full credit.

```
qqPlot(marsh$temp) + title("qqPlot of Temperature in Marsh Habitat")
```




```
## integer(0)
```

```
qqPlot(open_water$temp) + title("qqPlot of Temperature in Open Water Habitat")
```




```
## integer(0)
```

CONCLUSION: The qqplots for both marsh and open water are not normally distributed since a lot of the values towards the top lie outside of the CI.

d) Test the normality with a Shapiro-Wilk test. Report the p-value for each grouping (marsh and open water). What does each p-value mean in terms of the null hypothesis (ie normal or not)?

```
shapiro.test(marsh$temp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  marsh$temp
## W = 0.97403, p-value = 2.064e-13
```

```
shapiro.test(open_water$temp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  open_water$temp
## W = 0.96885, p-value = 5.723e-15
```

CONCLUSION: The P-value for the marsh group is 2.064e-13 and for the open water group is 5.723e-15. This means we reject the null hypothesis and conclude that they are both not normal.

e) Base on steps c-e and the Central Limit Theorem, do you consider the data to be normally distributed and why? Use your best judgement if not all of your normality tools agree with one another.

CONCLUSION: Based on the central limit theorem we can consider the data to be normally distributed since the sample size is large enough.

f) Test for equal variances of temperature between marsh and open water with a Levene's test. Report the p-value from the Levene's test. What does the p-value mean in terms of the null hypothesis (ie equal or not equal variances)?

```
leveneTest(temp ~ general_habitat, data = blue_crabs )
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  0.0786 0.7792
##       2280
```

CONCLUSION: The p-value is 0.7792 which is greater than 0.05. This means we fail to reject the null hypothesis of the variances being equal. This means the variances can be considered equal.

g) State the null and alternative hypotheses of your chosen t-test. Must state which t-test you are choosing for full credit.

ANSWER: We will be doing a two sample t-test with hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

h) Run the correct code for your chosen test

```
t.test(marsh$temp, open_water$temp, var.equal= TRUE)
```

```
##
##  Two Sample t-test
##
## data:  marsh$temp and open_water$temp
```

```
## t = 0.56974, df = 2280, p-value = 0.5689
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3469446  0.6311041
## sample estimates:
## mean of x mean of y
##  23.66016  23.51808
```

i) Based on the results from your test, what can you conclude about the difference of temperature (°C) between marsh and open water habitats? To receive full credit your response must include the p-value from your test, test statistic, degree of freedom, and the sample mean for marsh and open water.

CONCLUSION: Based on the results we know the p value was 0.5689, test statistic was 0.56974, degrees of freedom were 2280. The sample mean for marsh was 23.66016 and the sample mean for open water was 23.51808. Therefore, since our p value was greater than 0.05 we fail to reject the null hypothesis of the means being equal to each other. This means that there is not a significant difference between the temperatures in open water and marsh habitats.

End of Homework 5