# HW 6

SDS348 Spring 2021

# Name: Santhosh Saravanan

# UTEID: sks3648

**This homework is due on Mar 29, 2021 at 8am. Submit a pdf file on Gradescope.**

*For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.*

---

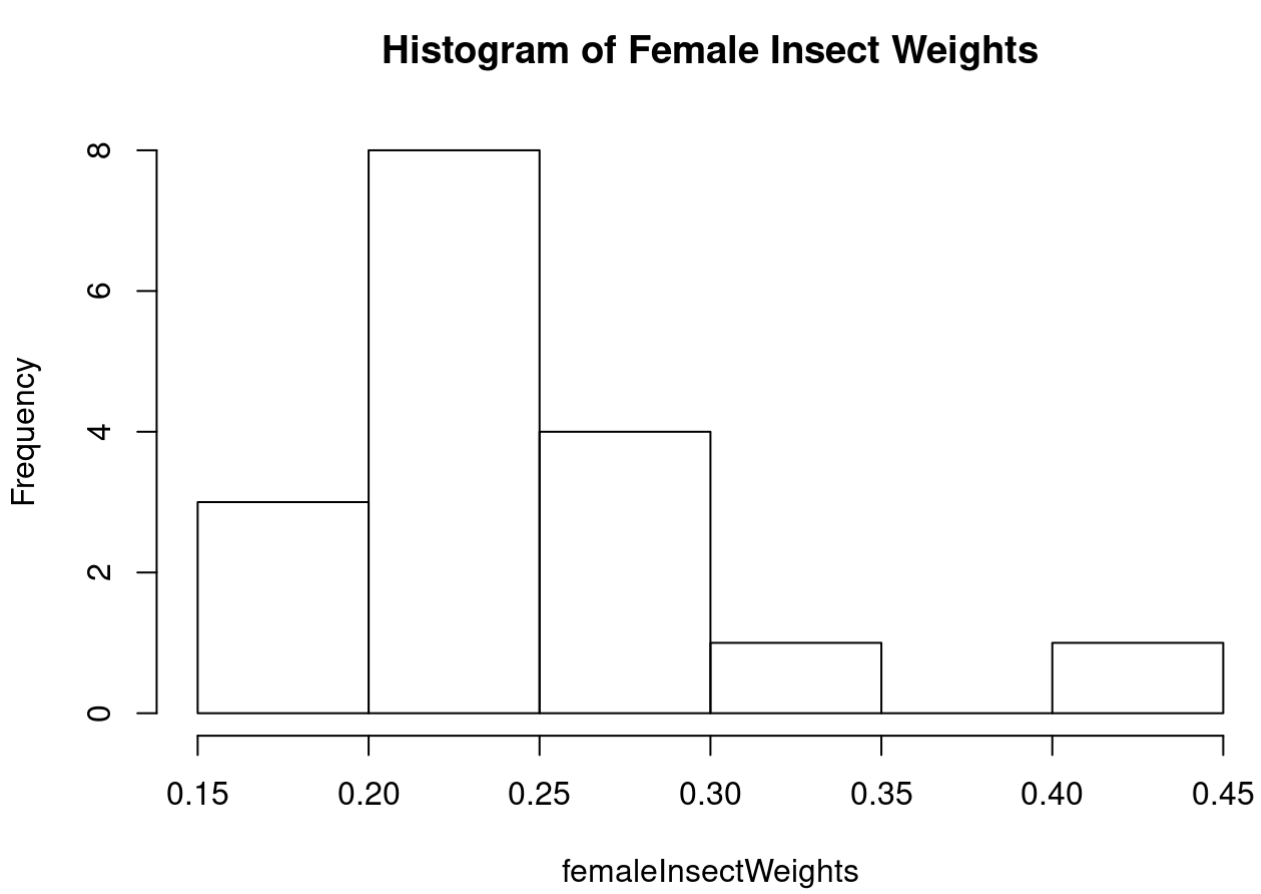
## Question 1: (14 pts)

The distribution of mosquito weight for the Aedes aegypti species is known to be log-normal (that is, weight is normally distributed if transformed with the natural log). Untransformed weights of 17 female and 15 male mosquitoes are given below (mg).

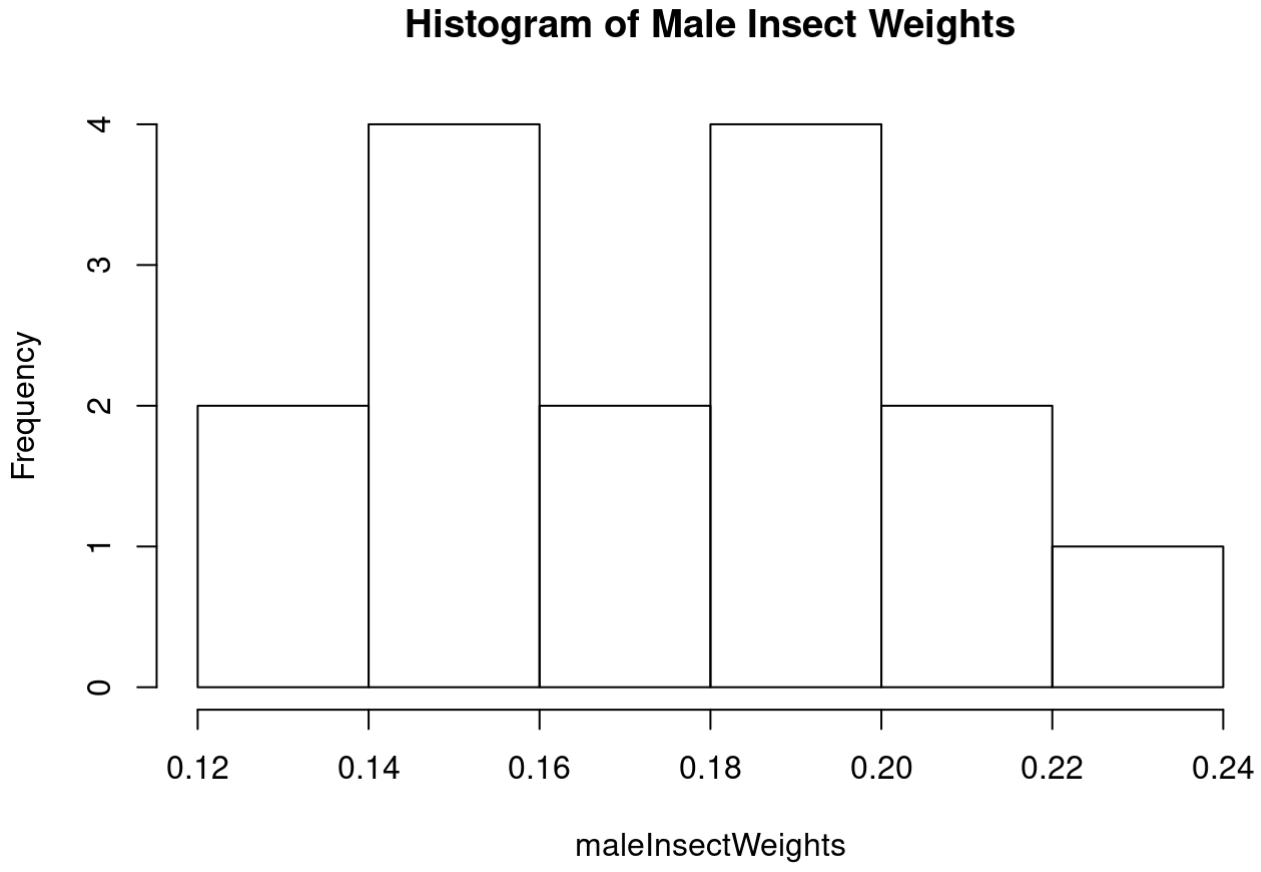
**Females:** 0.291, 0.208, 0.241, 0.437, 0.228, 0.256, 0.208, 0.234, 0.320, 0.340, 0.150
**Males:** 0.185, 0.222, 0.149, 0.187, 0.191, 0.219, 0.132, 0.144, 0.140

1.1 (2 pts) Represent the distribution of weights for females and for males in a histogram (you can use the function `hist()` to make simple histograms). Do a log transformation of weights for females and for males. Represent the transformed distributions in simple histograms. Has the log transformation improved the normality assumption?

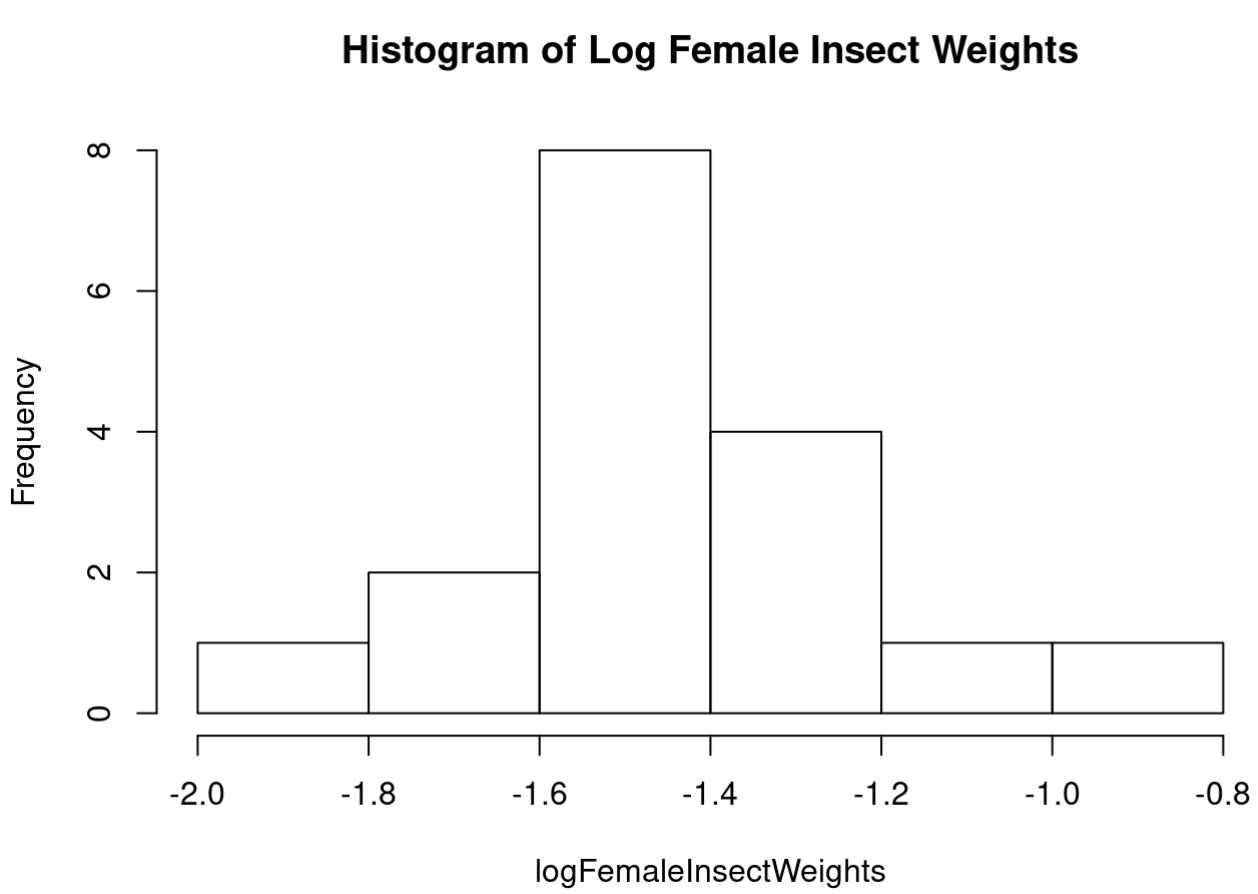
```
library(tidyverse)
# your code goes here (make sure to add comments)
femaleInsectWeights <- c(0.291, 0.208, 0.241, 0.437, 0.228, 0.256, 0.208, 0.234, 0.280, 0.340,
0.150, 0.211, 0.168, 0.221, 0.237, 0.189, 0.261)
maleInsectWeights <- c(0.185, 0.222, 0.149, 0.187, 0.191, 0.219, 0.132, 0.144, 0.140, 0.159, 0.
172, 0.198, 0.154, 0.201, 0.167)
hist(femaleInsectWeights,main="Histogram of Female Insect Weights")
```



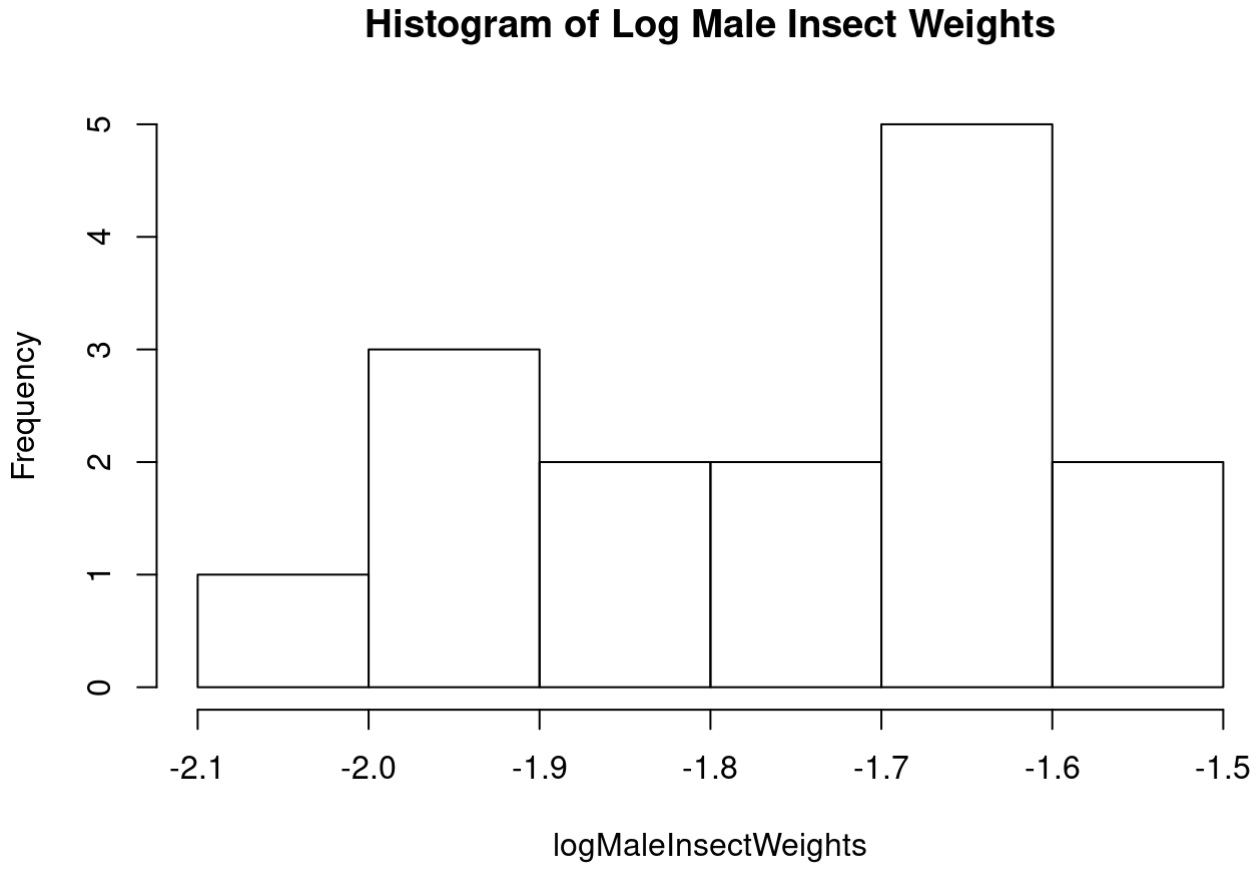

```
hist(maleInsectWeights,main="Histogram of Male Insect Weights")
```




```
logFemaleInsectWeights <- log(femaleInsectWeights)
logMaleInsectWeights <- log(maleInsectWeights)
hist(logFemaleInsectWeights,main="Histogram of Log Female Insect Weights")
```




```
hist(logMaleInsectWeights,main="Histogram of Log Male Insect Weights")
```




```
shapiro.test(femaleInsectWeights)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  femaleInsectWeights
## W = 0.88759, p-value = 0.0423
```

```
shapiro.test(maleInsectWeights)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  maleInsectWeights
## W = 0.95746, p-value = 0.6484
```

```
shapiro.test(logFemaleInsectWeights)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  logFemaleInsectWeights
## W = 0.96655, p-value = 0.7556
```

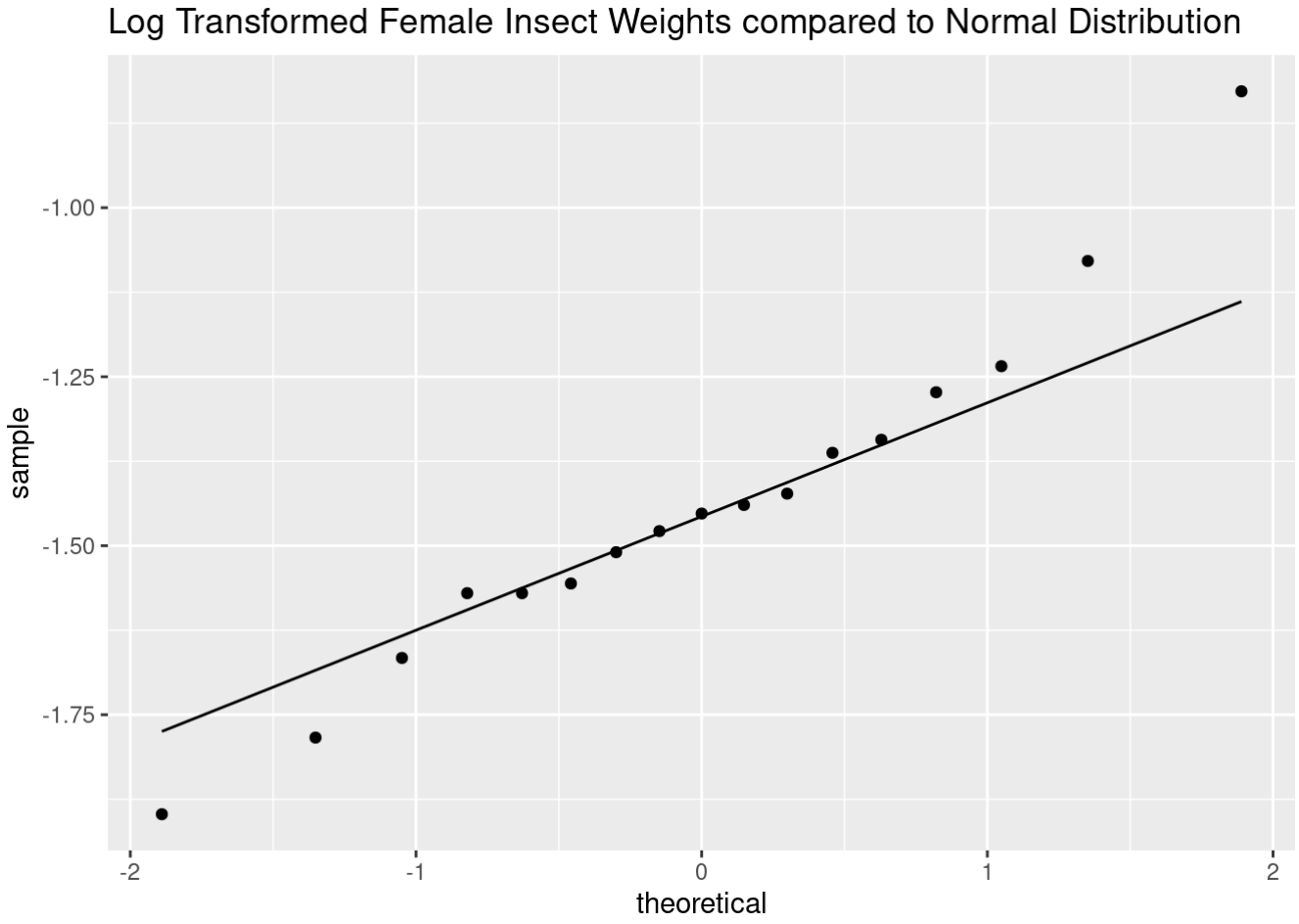
```
shapiro.test(logMaleInsectWeights)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  logMaleInsectWeights
## W = 0.96151, p-value = 0.7186
```

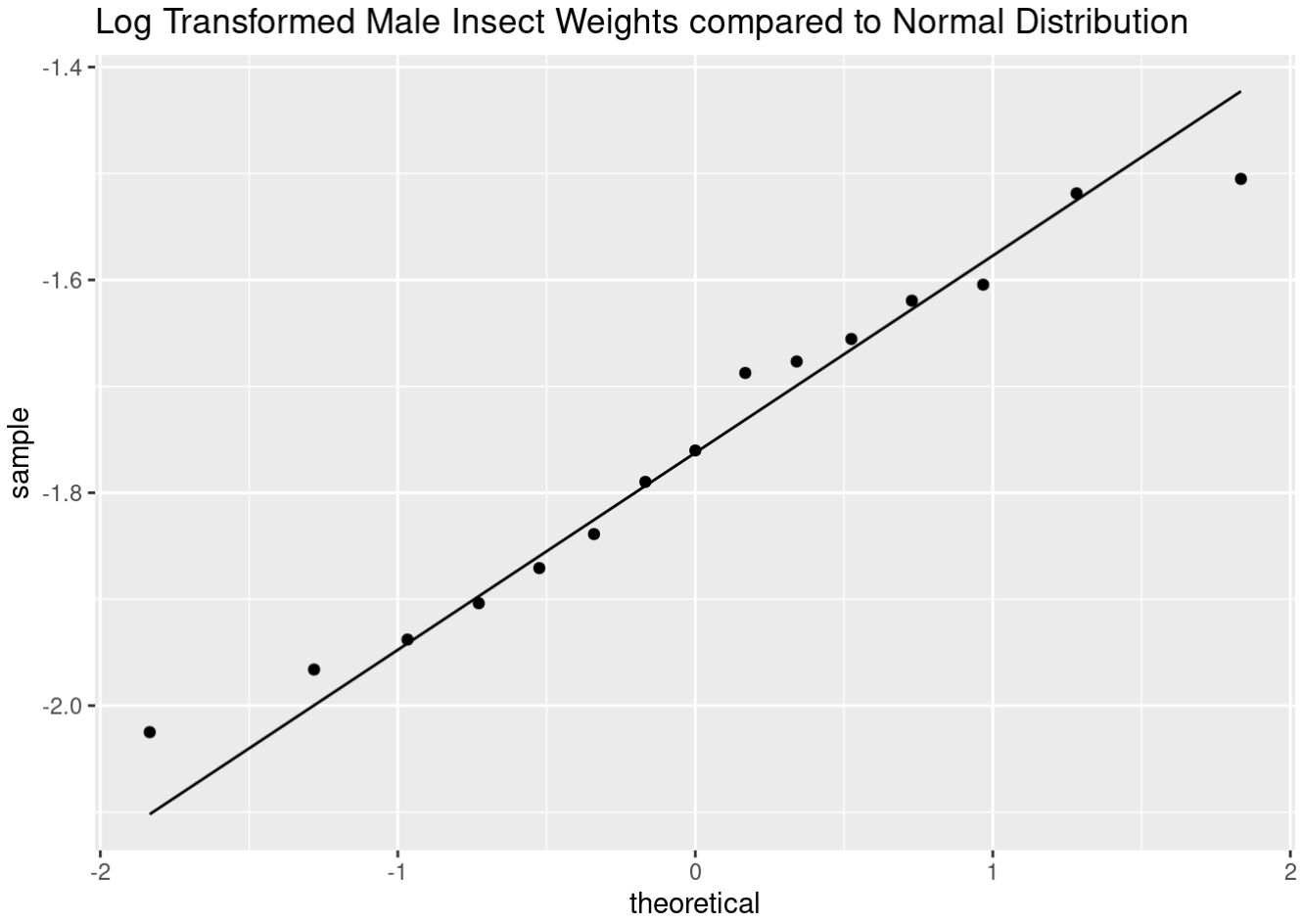
*The distributions look much more approximately normal after applying the log distributions for femaleInsectWeights and maleInsectWeights. To further validate my claim, I ran the Shapiro-Wilk test on the femaleInsectWeights and maleInsectWeights and their log counterparts. Log transforming the femaleInsectWeights increased the p-value from 0.4924 to 0.9477, a massive improvement in proving that our transformed variable is now normally distributed since the p-value is greater than 0.05. Log transforming the maleInsectWeights increased the p-value from 0.2284 to 0.2412, a slight improvement in proving that our transformed variable is now normally distributed because the p-value is greater than 0.05.*

1.2 (2 pts) Do the two groups weigh the same on average? We would like to perform an independent t-test. Assuming the samples were random and the observations were independent, check the rest of the assumptions (construct QQ-plots, conduct appropriate tests).

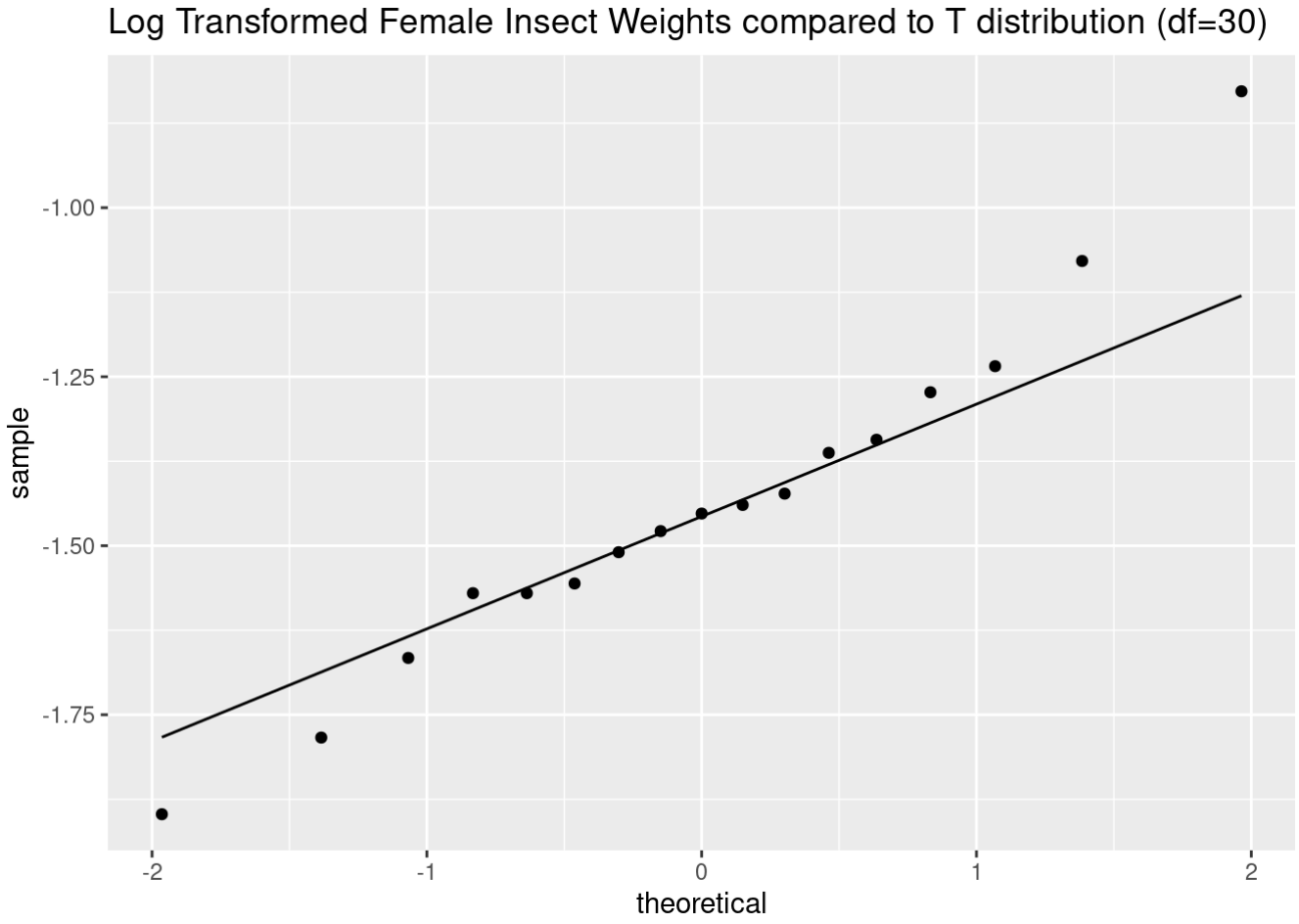
```
# Compare to a standard normal distribution
#Computer degrees of freedom: (total sample size -2) (17+15 - 2) = 30
degreesOfFreedom <- 17+15 -2
data.frame(logFemaleInsectWeights) %>%
  ggplot(aes(sample = logFemaleInsectWeights)) +
  # Compare quantiles for a normal distribution
  stat_qq() +
  # Reference qq line for a normal distribution
  stat_qq_line() +
 ggtitle("Log Transformed Female Insect Weights compared to Normal Distribution")
```



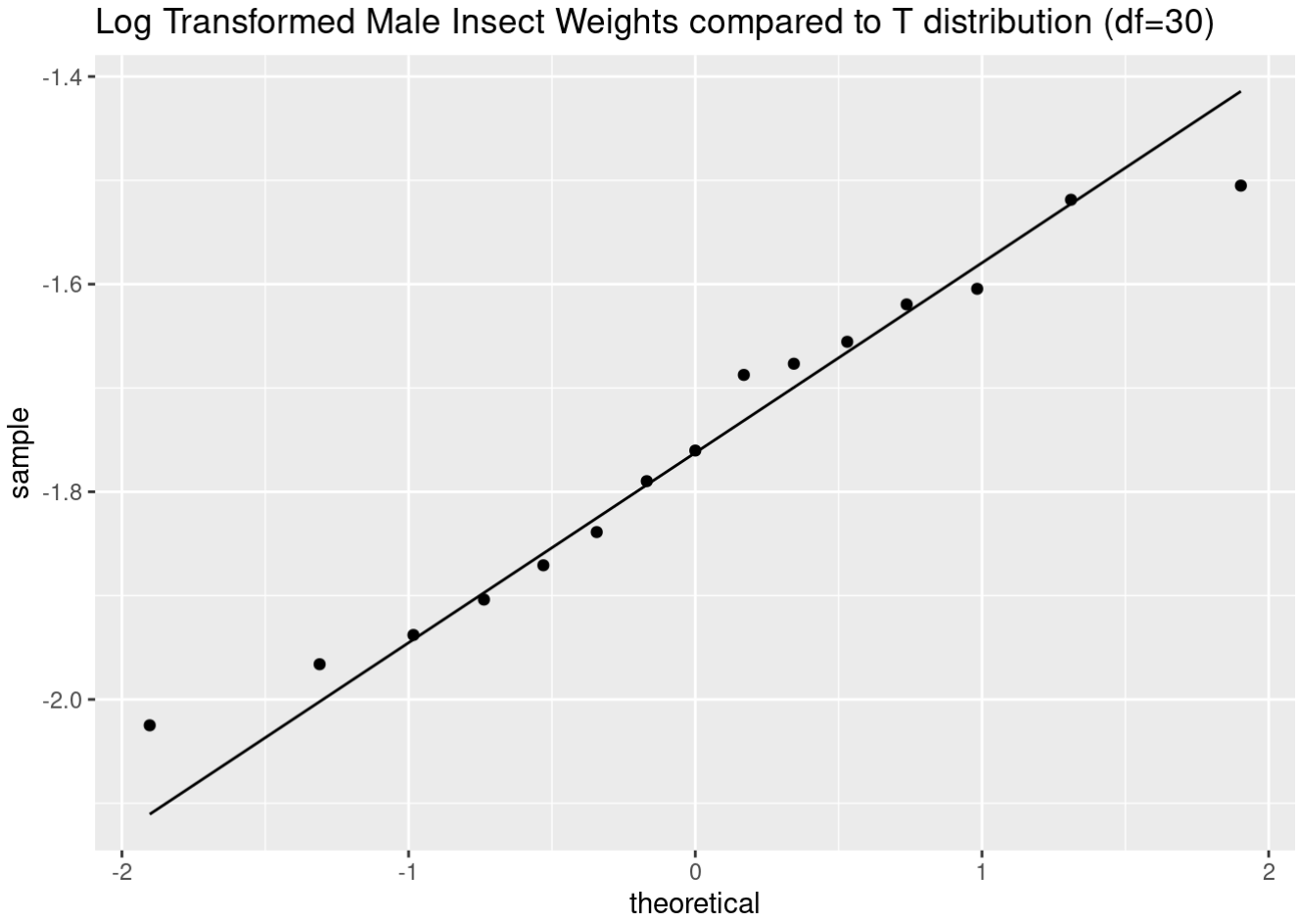

```
data.frame(logMaleInsectWeights) %>%
  ggplot(aes(sample = logMaleInsectWeights)) +
  # Compare quantiles for a normal distribution
  stat_qq() +
  # Reference qq line for a normal distribution
  stat_qq_line()+ ggtitle("Log Transformed Male Insect Weights compared to Normal Distribution"
)
```




```
# Compare to a t-distribution with df = 30
data.frame(logFemaleInsectWeights) %>%
  ggplot(aes(sample = logFemaleInsectWeights)) +
  # Compare quantiles to a t-distribution
  stat_qq(distribution = qt, dparams = list(df=degreesOfFreedom)) +
  # Reference qq line for a t-distribution
  stat_qq_line(distribution = qt, dparams = list(df=degreesOfFreedom))+ ggtitle("Log Transforme
d Female Insect Weights compared to T distribution (df=30)")
```




```
# Compare to a t-distribution with df = 30
data.frame(logMaleInsectWeights) %>%
  ggplot(aes(sample = logMaleInsectWeights)) +
  # Compare quantiles to a t-distribution
  stat_qq(distribution = qt, dparams = list(df=degreesOfFreedom)) +
  # Reference qq line for a t-distribution
  stat_qq_line(distribution = qt, dparams = list(df=degreesOfFreedom)) + ggtitle("Log Transform
ed Male Insect Weights compared to T distribution (df=30)")
```




```
#Null Hypothesis: The variances of the log-transformed female and male insect weights are equa
l.
## Alternative Hypothesis: The variances of the log-transformed female and male insect weights
 are not equal.
var.test(logFemaleInsectWeights, logMaleInsectWeights, alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  logFemaleInsectWeights and logMaleInsectWeights
## F = 2.3897, num df = 16, denom df = 14, p-value = 0.1086
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8174399 6.7318231
## sample estimates:
## ratio of variances
##           2.389698
```

*The sample size is barely greater than 30, (32 to be exact), so the sample size assumptions has been satisfied. From looking at the results of the Shapiro-Wilk test results from the previous section, the transformed data is approximately normally distributed. With regards to the QQplot, the points on the transformed variable are very close to the reference lines for the normal distribution and the t distribution with the corresponding degrees of freedom. We still need to check the equal variance assumption. With the results of the F-test, the p-value is greater than 0.05, so we fail to reject the null hypothesis of equal variances. All assumptions are met.*

1.3 (4 pts) After verifying the assumptions, perform the appropriate t-test. Write the hypotheses and write a conclusion in context, citing the appropriate statistics.

```
t.test(logFemaleInsectWeights, logMaleInsectWeights, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  logFemaleInsectWeights and logMaleInsectWeights
## t = 4.1606, df = 30, p-value = 0.0002452
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1619712 0.4742838
## sample estimates:
## mean of x mean of y
## -1.439187 -1.757315
```

*The null hypothesis is that the mean log of insect weights is the same for female vs male insects. The alternative hypothesis is that the mean log of insect weights is not the same for female vs male insects. With a p-value < 0.05, we reject the null hypothesis and state that there is strong evidence that the mean log of insect weights is not the same for female vs male insects.*

1.4 (1 pt) Build a dataframe called `mosquitoes` with a column for `weight`, a column for `logweight`, and a column for `sex`. What are the observed difference of the mean weights for untransformed and transformed data? Call them `obs_diff` and `obs_logdiff` respectively.

```
# your code goes here (make sure to add comments)
totalLogWeights <- c(logFemaleInsectWeights,logMaleInsectWeights)
logWeights <- "logweight"
totalWeights <- c(femaleInsectWeights,maleInsectWeights)
weights <- "weight"
insectSex <- "sex" #haha got em
genders <- character()
for(i in 1:32)
{
  if(i>17)
  {
    genders <- c(genders,'M')
  }
  else
  {
    genders <- c(genders,'F')
  }
}
mosquitoes <- data.frame(totalWeights,totalLogWeights,genders)
colnames(mosquitoes) <- c(weights, logWeights,insectSex)
print(mosquitoes)
```

```
##    weight  logweight sex
## 1   0.291 -1.2344320   F
## 2   0.208 -1.5702172   F
## 3   0.241 -1.4229583   F
## 4   0.437 -0.8278221   F
## 5   0.228 -1.4784097   F
## 6   0.256 -1.3625778   F
## 7   0.208 -1.5702172   F
## 8   0.234 -1.4524342   F
## 9   0.280 -1.2729657   F
## 10  0.340 -1.0788097   F
## 11  0.150 -1.8971200   F
## 12  0.211 -1.5558971   F
## 13  0.168 -1.7837913   F
## 14  0.221 -1.5095926   F
## 15  0.237 -1.4396951   F
## 16  0.189 -1.6660083   F
## 17  0.261 -1.3432349   F
## 18  0.185 -1.6873995   M
## 19  0.222 -1.5050779   M
## 20  0.149 -1.9038090   M
## 21  0.187 -1.6766467   M
## 22  0.191 -1.6554819   M
## 23  0.219 -1.5186835   M
## 24  0.132 -2.0249534   M
## 25  0.144 -1.9379420   M
## 26  0.140 -1.9661129   M
## 27  0.159 -1.8388511   M
## 28  0.172 -1.7602608   M
## 29  0.198 -1.6194882   M
## 30  0.154 -1.8708027   M
## 31  0.201 -1.6044504   M
## 32  0.167 -1.7897615   M
```

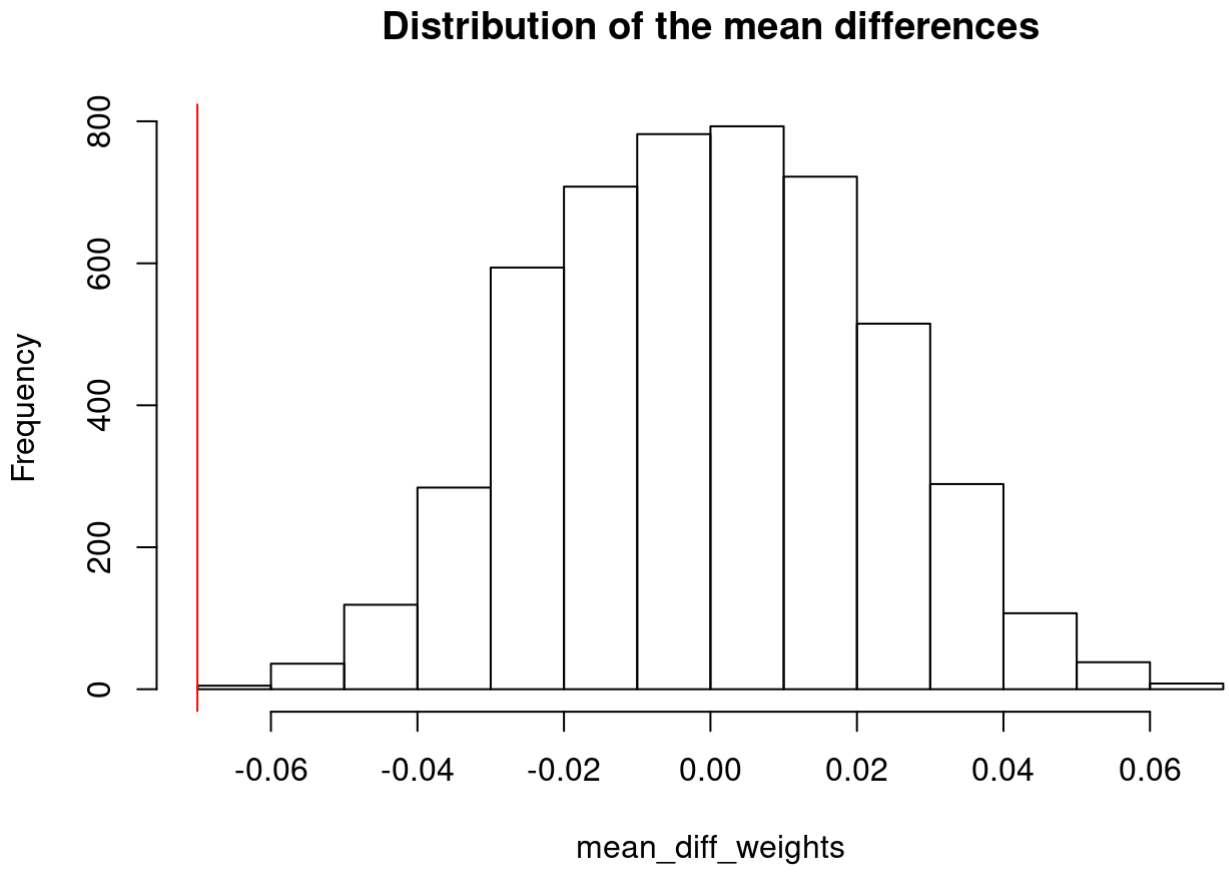
```
obs_diff <- mean(mosquitoes$weight[mosquitoes$sex == 'F']) - mean(mosquitoes$weight[mosquitoes
$sex == 'M'])
obs_logdiff <- mean(mosquitoes$logweight[mosquitoes$sex == 'F']) - mean(mosquitoes$logweight[mo
squitoes$sex == 'M'])
```

*The observed difference of the mean weights for untransformed and transformed data are 0.070039mg units and 0.3181275 log(mg) units.*

1.5 (3 pts) After setting the seed as specified below, perform a randomization test on the original weight data *then* on the log weight data. That is, for both, generate a distribution of 5000 mean differences on randomized data (with a `for` loop, *note: it might take some time to run*). Compute and report two-tailed p-values in both cases. Do both randomization tests agree? What does it mean? Are your conclusions the same as they were above for the parametric t-test?

```
set.seed(348)
# 5000 Randomizations finding mean with original weight data
# Find the new mean difference
mean_diff_weights <- vector()
# Create many randomizations with a for loop
for(i in 1:5000){
  temp <- data.frame(sex = mosquitoes$sex, weight = sample(mosquitoes$weight))

  mean_diff_weights[i] <- temp %>%
    group_by(sex) %>%
    summarize(means = mean(weight)) %>%
    summarize(mean_diff = diff(means)) %>%
    pull
}
```

```
# Represent the distribution of the mean differences with a vertical line showing the true diff
erence
{hist(mean_diff_weights, main="Distribution of the mean differences"); abline(v = -0.070039, co
l="red")}
```

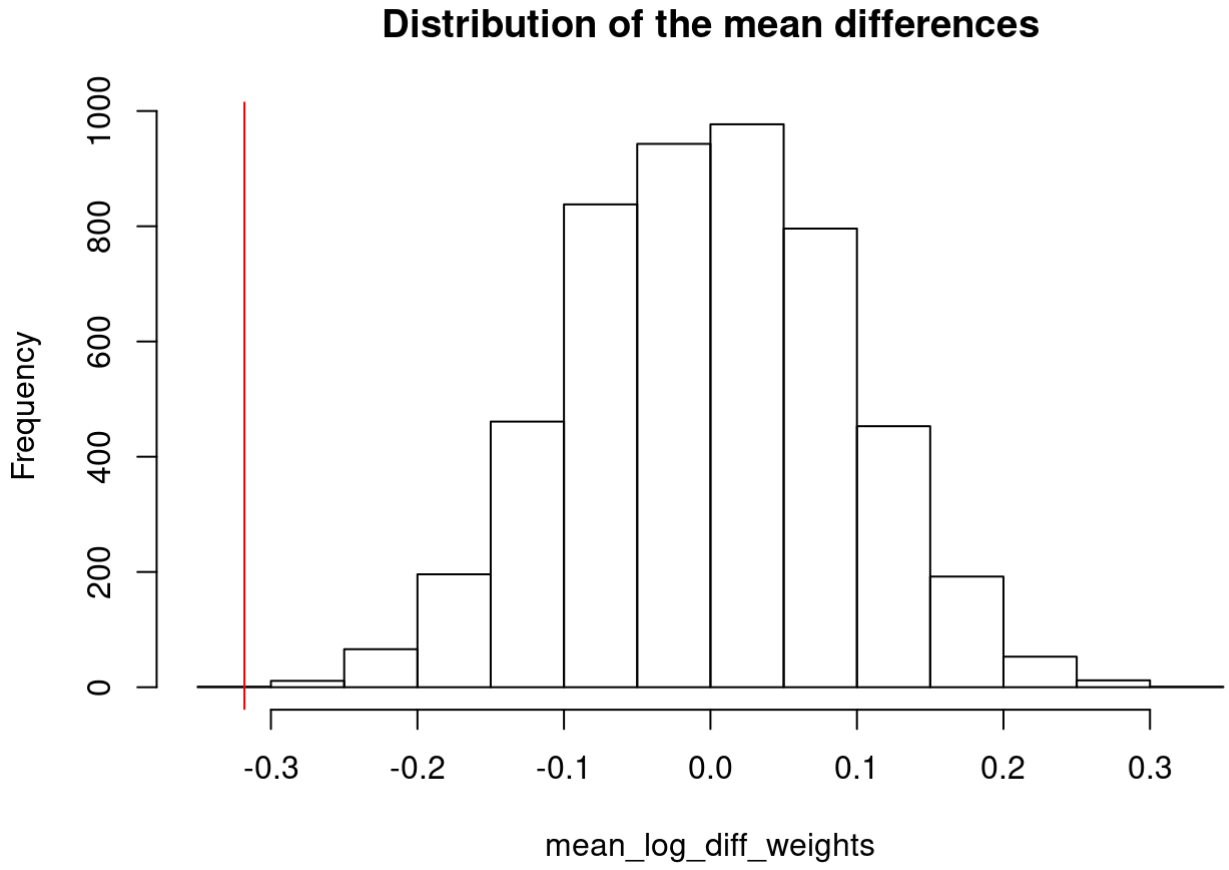



```
mean(mean_diff_weights > obs_diff | mean_diff_weights < -obs_diff)
```

```
## [1] 0
```

```
set.seed(348)
# 5000 Randomizations finding mean with original weight data
# Find the new mean difference
mean_log_diff_weights <- vector()
# Create many randomizations with a for loop
for(i in 1:5000){
  temp <- data.frame(sex = mosquitoes$sex, weight = sample(mosquitoes$logweight))

  mean_log_diff_weights[i] <- temp %>%
    group_by(sex) %>%
    summarise(means = mean(weight)) %>%
    summarise(mean_diff = diff(means)) %>%
    pull
}
```

```
# Represent the distribution of the mean differences with a vertical line showing the true diff
erence
{hist(mean_log_diff_weights, main="Distribution of the mean differences"); abline(v = -0.318127
, col="red")}
```




```
mean(mean_log_diff_weights > obs_logdiff | mean_log_diff_weights < -obs_logdiff)
```

```
## [1] 0
```

```
t.test(weight~sex,data=mosquitoes,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  weight by sex
## t = 3.7368, df = 30, p-value = 0.0007829
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.0317604 0.1083180
## sample estimates:
## mean in group F mean in group M
##       0.2447059       0.1746667
```

```
t.test(logweight~sex,data=mosquitoes,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  logweight by sex
## t = 4.1606, df = 30, p-value = 0.0002452
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1619712 0.4742838
## sample estimates:
## mean in group F mean in group M
##       -1.439187       -1.757315
```

*Both randomization tests agree. It means that there is a signifcant amount of evidence suggesting that there is a difference between the mean weights of female and male mosquitoes. My conclusions are the same as they were above for the parametric t-test.*

1.6 (2 pts) Recall the observed mean difference in mosquito weights between the two groups (calculated for question 4.4). Now you will create a 95% CI for this difference in means using bootstrapping. Resample from the original male mosquito data with replacement using `sample(..., replace=T)`, resample from the original female mosquito data with replacement with `sample(..., replace=T)`, take the mean difference of these samples, save it, and repeat this process 5000 times (with a `for` loop). What is the mean of the resulting distribution? Report the 95% CI of this distribution by reporting the .025 and the 0.975 percentiles of mosquito weight differences. Interpret it in a sentence.

```
set.seed(348)
# Get the mean differences in one vector
means_difference<-vector()
for(i in 1:5000){
  # Draw a bootstrap sample
  samp <- sample(femaleInsectWeights,replace = TRUE)
  # Calculate and save the sample mean
  samp1 <- sample(maleInsectWeights,replace = TRUE)
  means_difference[i] <- mean(samp) - mean(samp1)
}
mean_diff = mean(means_difference)
```

```
quantile(means_difference,c(.025,.975))
```

```
##       2.5%      97.5%
## 0.03838343 0.10621980
```

*The mean of the resulting distribution is 0.07023019. The 2.5% percentile is 0.03838343 and the 97.5% percentile is 0.10621980. We are 95% confident that the true mean difference between the weights of female and male mosquitoes lie between 0.0383838343 and 0.10621980 mg.*

# Question 2: (11 pts)

For this question, we will use the pottery data set which contains the chemical composition (the percentage of metal oxide or abundance) of ancient pottery found at four sites in Great Britain.

2.1 (0.5 pt) Import the dataset from an online resource. How many rows and how many columns are in this dataset? What does a row represent? What does a column represent?

```
library(tidyverse)
pottery <- read_csv("https://wilkelab.org/classes/SDS348/data_sets/pottery.csv")
nrow(pottery)
```

```
## [1] 26
```

```
ncol(pottery)
```

```
## [1] 6
```

*There are 26 rows and 6 columns in this dataset. Each row represents the site where the pottery was found in Great Britain. Each column represents the percentage of metal oxide in each piece of pottery found in a region in Britain.*

2.2 (2.5 pts) Let's compare the chemical composition of aluminium (Al) across the different sites. Compute the SSB (sum of squares between groups) and SSW (sum of squares within groups) for a one-way ANOVA, manually (use `dplyr` functions to get group means, finding the sum of the differences squared, …). Use the calculated values of SSB and SSW to compute an F statistic. Use `pf(..., df1=, df2=, lower.tail=F)` on the F statistic you calculated to determine the p-value. Compare your results to the output from `summary(aov())`. What is your conclusion about the chemical composition of aluminium across sites?

```
# Compute variation within groups
SSW <- pottery %>%
  group_by(Site) %>%
  summarize(SSW = sum((Al - mean(Al))^2)) %>%
  summarize(sum(SSW))

# Compute variation between groups
SSB <- pottery %>%
  mutate(mean = mean(Al)) %>%
  group_by(Site) %>%
  mutate(groupmean = mean(Al)) %>%
  summarize(SSB = sum((mean - groupmean)^2)) %>%
  summarize(sum(SSB))
# Compute the F-statistic (ratio of MSB and MSW)
# df for SSB is 4 groups - 1 = 3
# df for SSW is 26 observations - 4 sites = 22
MSB = SSB/3
MSW = SSW/22
Fstat =  MSB/MSW
pf(Fstat[,1],df1=3,df2=22,lower.tail = F)
```

```
## [1] 1.62687e-07
```

```
summary(aov(Al ~ Site,data=pottery))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Site         3 175.61   58.54   26.67 1.63e-07 ***
## Residuals   22  48.29    2.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*We have statistically significant evidence that there is a difference in mean chemical composition of aluminium (Al) across the different sites (p-value < 0.05).*

2.3 (1 pt) Let's conduct a MANOVA test to investigate whether at least one of these five response variables (chemical compositions) differ by site. Use `manova(cbind(Y1,Y2,Y3...)~X,data=data)` and report the results in writing. *Don't worry about the assumptions (there are lots).*

```
# Perform MANOVA with 2 response variables listed in cbind()
manova <- manova(cbind(Al,Fe,Ca,Mg,Na) ~ Site, data = pottery)
# OUtput of MANOVA
summary(manova)
```

```
##           Df Pillai approx F num Df den Df    Pr(>F)
## Site       3 1.5539   4.2984     15     60 2.413e-05 ***
## Residuals 22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Significant differences were found among the 4 sites for at least one of the chemical compositions(Pillai's trace = 1.5539 ,pseudo F(15,60) = 4.2984,p<0.0001).*

2.4 (2 pts) Now, let's investigate which of the elements differ by site. Report full ANOVA results for each metal variable. For the ones that differ, explore which sites are different, that is, perform posthoc t-tests for all significant ANOVAs using `pairwise.t.test(...,p.adj="none")` (*you do not have to write anything up about the post hoc tests for now*).

```
  summary.aov(manova)
```

```
##  Response Al :
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Site         3 175.610  58.537  26.669 1.627e-07 ***
## Residuals   22  48.288   2.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Fe :
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Site         3 134.222  44.741  89.883 1.679e-12 ***
## Residuals   22  10.951   0.498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Ca :
##             Df   Sum Sq  Mean Sq F value    Pr(>F)
## Site         3 0.204703 0.068234  29.157 7.546e-08 ***
## Residuals   22 0.051486 0.002340
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Mg :
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Site         3 103.35  34.450   49.12 6.452e-10 ***
## Residuals   22  15.43   0.701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Na :
##             Df  Sum Sq  Mean Sq F value    Pr(>F)
## Site         3 0.25825 0.086082  9.5026 0.0003209 ***
## Residuals   22 0.19929 0.009059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*All the sites were found to differ significantly from each other in terms of Aluminum ( F(3,22) = 26.669, p <0.05). All the sites were found to differ significantly from each other in terms of Iron ( F(3,22) = 89.883, p <0.05). All the sites were found to differ significantly from each other in terms of Magnesium ( F(3,22) = 29.157, p <0.05). All the sites were found to differ significantly from each other in terms of Calcium ( F(3,22) = 49.12, p <0.05). All the sites were not found to differ significantly from each other in terms of Sodium ( F(3,22) = 9.5026, p < 0.05).*

```
    # For Aluminum
    pairwise.t.test(pottery$Al,pottery$Site, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  pottery$Al and pottery$Site
##
##            AshleyRails Caldicot IsleThorns
## Caldicot   0.00016     -        -
## IsleThorns 0.36866     3.0e-05  -
## Llanedyrn  3.3e-06     0.44848  2.7e-07
##
## P value adjustment method: none
```

```
    # For Iron
    pairwise.t.test(pottery$Fe,pottery$Site, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  pottery$Fe and pottery$Site
##
##            AshleyRails Caldicot IsleThorns
## Caldicot   1.2e-06     -        -
## IsleThorns 0.658       2.6e-06  -
## Llanedyrn  6.0e-12     0.086    1.4e-11
##
## P value adjustment method: none
```

```
    # For Magnesium
    pairwise.t.test(pottery$Mg,pottery$Site, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  pottery$Mg and pottery$Site
##
##            AshleyRails Caldicot IsleThorns
## Caldicot   0.00013     -        -
## IsleThorns 0.89901     0.00016  -
## Llanedyrn  2.2e-09     0.13917  2.9e-09
##
## P value adjustment method: none
```

```
    # For Calcium
    pairwise.t.test(pottery$Ca,pottery$Site, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  pottery$Ca and pottery$Site
##
##            AshleyRails Caldicot IsleThorns
## Caldicot   4.8e-06     -        -
## IsleThorns 0.405       1.1e-06  -
## Llanedyrn  5.4e-06     0.019    5.2e-07
##
## P value adjustment method: none
```

```
    # For Sodium
    pairwise.t.test(pottery$Na,pottery$Site, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  pottery$Na and pottery$Site
##
##            AshleyRails Caldicot IsleThorns
## Caldicot   0.98019     -        -
## IsleThorns 0.92150     0.96039  -
## Llanedyrn  0.00049     0.01068  0.00065
##
## P value adjustment method: none
```

*At least one of the group means for chemical compositions differ by site.*

2.5 (2 pts) Between 2.3 and 2.4, how many hypothesis tests have you done in total? What is the probability that you have made at least one type I error (i.e., what is the overall type-I error rate)? What (Bonferroni adjusted) significance level should you use if you want to keep the overall type I error rate at .05? Which of your post hoc tests that were significant before the adjustment are no longer significant?

```
numTests <- 5 +5 + 1
probTypeIError <- 1 - (0.95)^11
BonferroniPepperonniLevel <- probTypeIError/numTests
```

*I have done 11 hypothesis tests (5 Multi-ANOVA tests, 5 post-hoc tests, and one MANOVA test). The probability that I have made at least one type I error is 0.431199%. The Bonferroni significance level is 0.0391999. None of my post hoc tests were no longer significant when they were significant before the adjustment.*

2.6 (1 pt) Let's now conduct a PERMANOVA test. Calculate the distances between each metal and each pot in the pottery dataset, using the function `dist`. Use the `adonis()` function from the `vegan` package to conduct PERMANOVA. Is the p-value larger or smaller than in the parametric MANOVA? Why might that be?

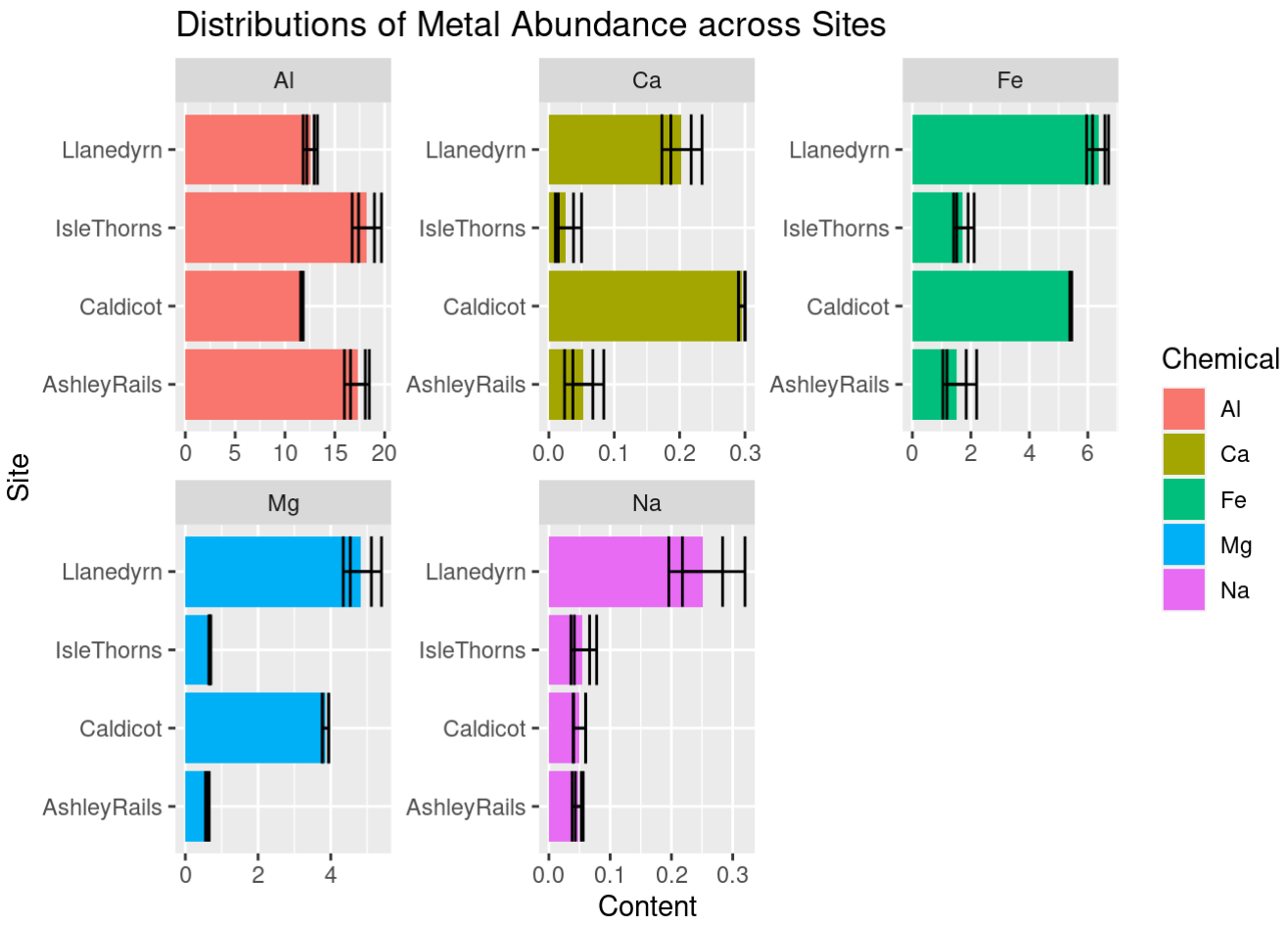
```
library(vegan)
dists <- pottery %>%
  select(Al,Na,Ca,Fe,Mg) %>%
  dist

# Perform PERMANOVA on the distance matrix
adonis(dists ~ Site, data = pottery)
```

```
##
## Call:
## adonis(formula = dists ~ Site, data = pottery)
##
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## Site       3    413.65 137.882  40.489 0.84665  0.001 ***
## Residuals 22     74.92   3.405         0.15335
## Total     25    488.56                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*The p-value is greater than the parametric MANOVA because there are no distributional assumptions, differences in variance and covariance are allowed, non-sensitivity to multicollineraity and outliers (0s), and allows more variables than samples.*

2.7 (2 pts) Make the pottery dataset long by pivoting all of the element names into a column with all of the corresponding values into another column. Use that data to make a plot (mapping abundance to x, site to y), showing the average abundance of each element at each site with `geom_bar()` (using `stat = "summary", fun = "mean"`), adding standard errors (with `geom_errorbar(stat = "summary", fun.data = "mean_se")`) and then faceting by element (set scales='free'). (Add bootstrapped with `geom_errorbar(stat="summary",fun.data=mean_cl_boot)`, or by computing them manually. *Hint: refer to HW2 for similar graphs.* For which element there is the most noticeable difference between one location compared to the others?

```
potteryDefined <- pottery %>% pivot_longer(c(`Al`,`Fe`,`Mg`,`Ca`,`Na`),names_to = "Chemical", values_to = "Content")
ggplot(potteryDefined,aes(x=Content,y=Site,fill = Chemical)) + geom_bar(stat = "summary",fun = "mean") + geom_errorbar(stat = "summary", fun.data = "mean_se") + facet_wrap(vars(Chemical),scales = "free") + geom_errorbar(stat="summary",fun.data=mean_cl_boot) + ggtitle("Distributions of Metal Abundance across Sites ")
```




*There is the most noticeable difference in aluminum between Caldicot and AshleyRails compared to the other metals.*

---

```
##                                     sysname
##                                     "Linux"
##                                     release
##                           "5.4.0-70-generic"
##                                     version
## "#78-Ubuntu SMP Fri Mar 19 13:29:52 UTC 2021"
##                                    nodename
##                       "MechaChungus-linux64"
##                                     machine
##                                    "x86_64"
##                                       login
##                                "OmniLordSanta"
##                                        user
##                                "OmniLordSanta"
##                                effective_user
##                                "OmniLordSanta"
```

/