Simonson_HW4

Marty Simonson

February 14, 2019

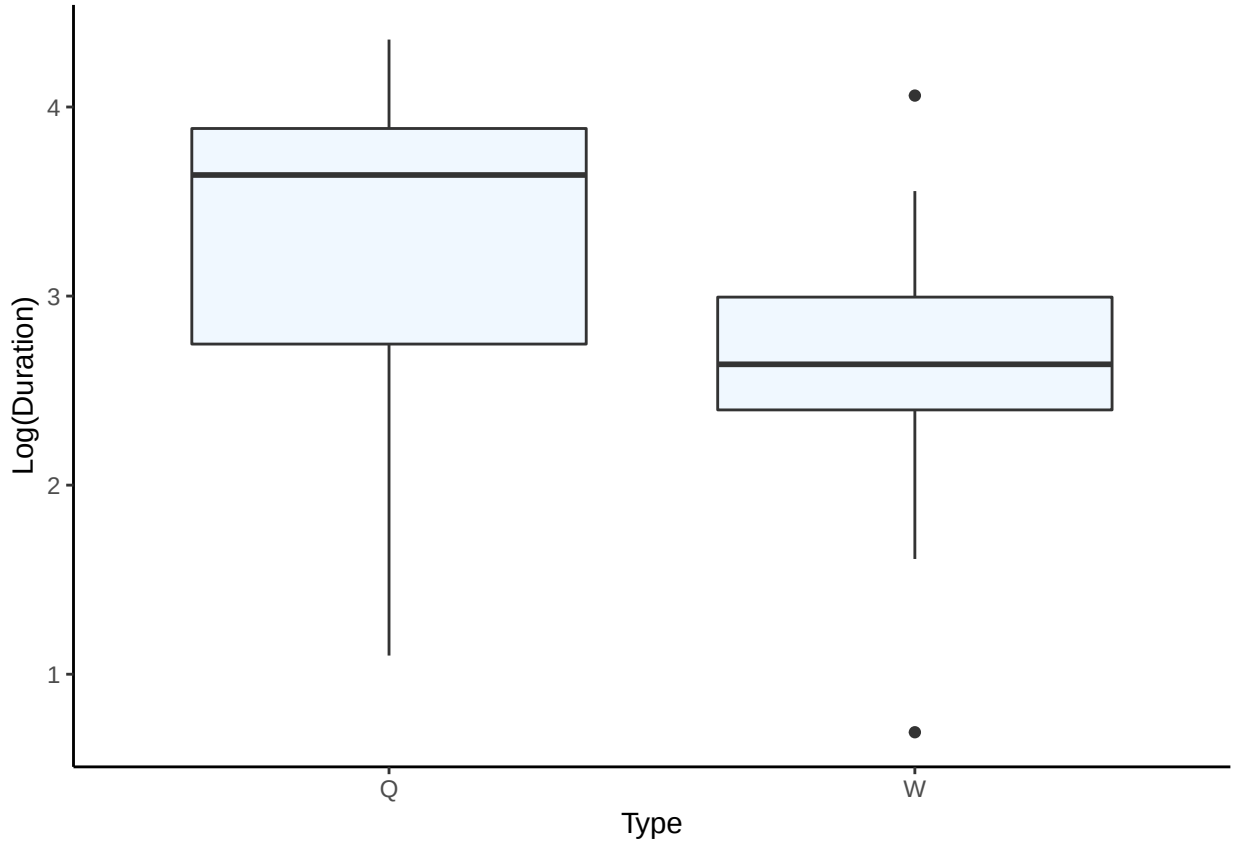
1.) Pollen Removal

Using data from *bee.csv*, test if log-transformation helps meet assumptions of pooled t-tests.

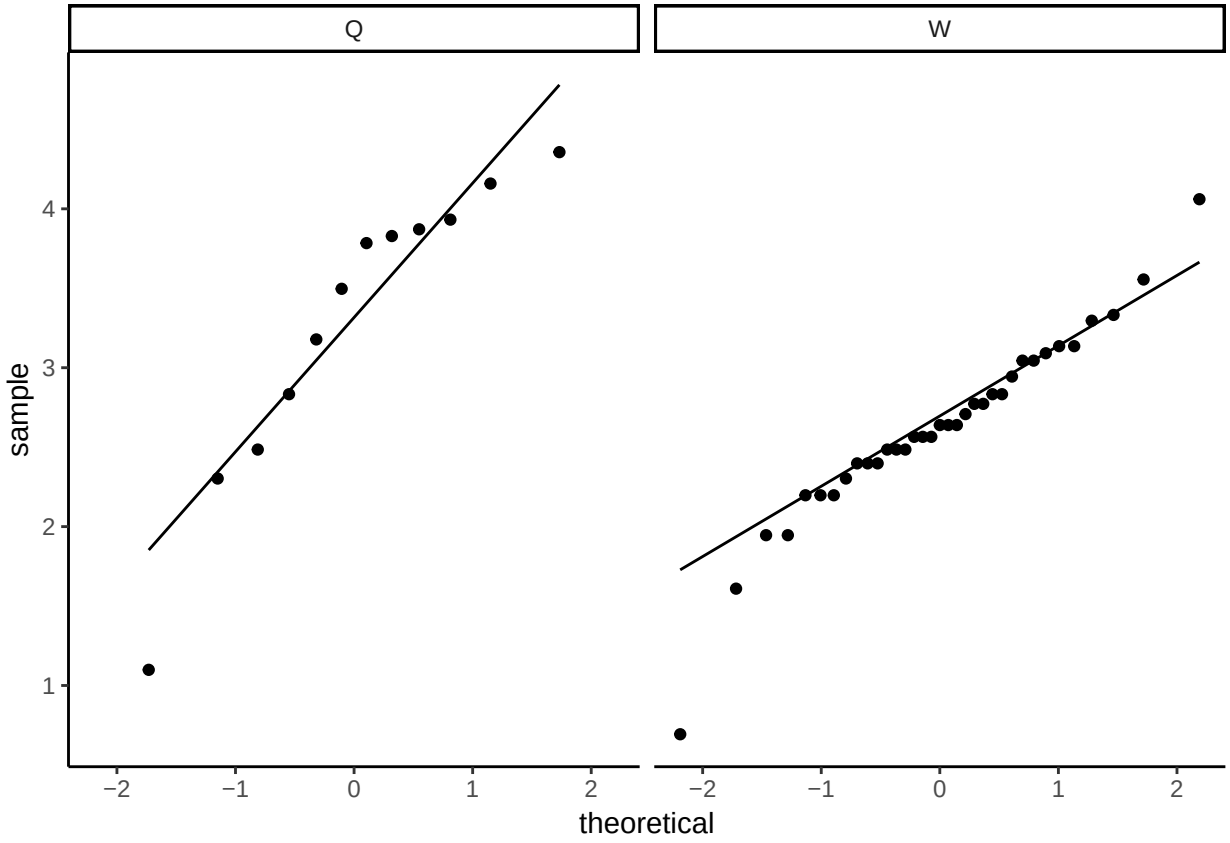
a) Construct box plots and normal quantile plots on the log-transformed duration values for both types of bees.

```
df<-read.csv("bee.csv")
df$LogDuration<-log(df$Duration)

ggplot(data=df,aes(x=Type, y=LogDuration))+
       geom_boxplot(fill="aliceblue")+
       ylab("Log(Duration)")+
       theme_classic()
```




```
ggplot(data=df, aes(sample=LogDuration)) +
        stat_qq() +
        stat_qq_line() +
        facet_wrap(facets = vars(Type))+
        theme_classic()
```



Does it appear that assuming normality for thelog-transformed duration variable is reasonable? Explain.

- **Answer:** The assumption of normality is not met after log-transformation because there is still curvature in the points along the QQ line.

Are the variabilities the same for the two types of bees on the log scale? Explain.

- **Answer:** The size of the boxes and whiskers are not equal, therefore the assumption of normality is not met.

b) Use the Welch's t-test to compare the (population) median durations. Write down the null and alternate hypotheses, provide the t-stastic and p-value.

- Let μ_1 be the median duration of visit for Queen Bees, and μ_2 be the median duration of visit for Worker Bees. Therefore:
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```
attach(df)
t.test(LogDuration~Type,
       mu=0,
       alt = "two.sided",
       conf=0.95,
       var.equal=F)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  LogDuration by Type
## t = 2.2385, df = 14.025, p-value = 0.04192
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02736723 1.27496322
## sample estimates:
## mean in group Q mean in group W
##        3.277111        2.625945
```

- **Answer:** The t-statistic is 2.2385 with 14 degrees of freedom and a p-value of 0.04192.

c) Interpret the results and provide a conclusion from the context of this problem. Include an estimate of the effect size.

```
exp(3.277111-2.625945)
```

```
## [1] 1.917776
```

- **Answer:** We have weak evidence to suffest that the population median duration of visit is not the same between Worker bees and Queen bees; data suggests that foraging durations for Queen bees is longer than foraging durations for Worker bees. After applying an inverse log-transformation to the difference of mean durations, we estimate that Queen bees spend about 1.92 more time units foraging for pollen than Worker bees.

d) Construct a 95% confidence interval for the ratio of median durations between groups. Interpret the confidence interval within the context of the problem.

- **Answer:** The 95% confidence interval for the ratio of median durations between groups is (0.027, 1.27). Therefore, we are 95% sure that the true difference in median duration of flower visits for Queen bees between exp(0.027) and exp(1.27) times larger than for Worker bees.

2.) Supplementing Hog Diet

Suppose 6 pairs of hogs are used in an experiment to determine if a dietary supplement can increase weight gain. Each pair of hogs are full siblings. One sibling is randomly assigned to receive the dietary supplement; the other sibling receives a supplement that does not contain the active ingredient believed to affect weight gain. The gains in pounds are provided in *hogs.csv*.

a) Test for a difference between the population mean weight gains. write down the null and alternate hypotheses, provide a test statistic and p-value, and a conclusion in the context of this problem.

- Let μ_1 be the mean weight gain for pigs that received the supplement, and μ_2 be the median weight gain of hogs that received the placebo supplement. Therefore:
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```
df<-read.csv("Data/hogs.csv")
str(df)
```

```
## 'data.frame':    6 obs. of  3 variables:
##  $ Sib_Pair      : int  1 2 3 4 5 6
##  $ Treatment_Gain: num  70.4 38.2 59.7 40.3 93.5 50.7
##  $ Placebo_Gain  : num  69.3 37.2 58.7 39.3 92.5 49.6
```

```
df$Sib_Pair<-as.factor(df$Sib_Pair)
df.m<-melt(df)
```

```
## Using Sib_Pair as id variables
```

```
t.test(value~variable,
       data = df.m,
       paired = T,
       conf = 0.90)
```

```
##
##  Paired t-test
##
## data:  value by variable
## t = 49.015, df = 5, p-value = 6.679e-08
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  0.9908524 1.0758143
## sample estimates:
## mean of the differences
##                1.033333
```

- **Answer:** The t-statistic is 49.015 with 5df. The p-value is 6.679e-08 so we have very strong evidence against the null hypothesis and conclude that the true difference in mean weight gain is greater for hogs who received the supplement compared to hogs that received a placebo.

b) Compute a 90% confidence interval for the mean difference in weight gain, and given an interpretation in the context of these data

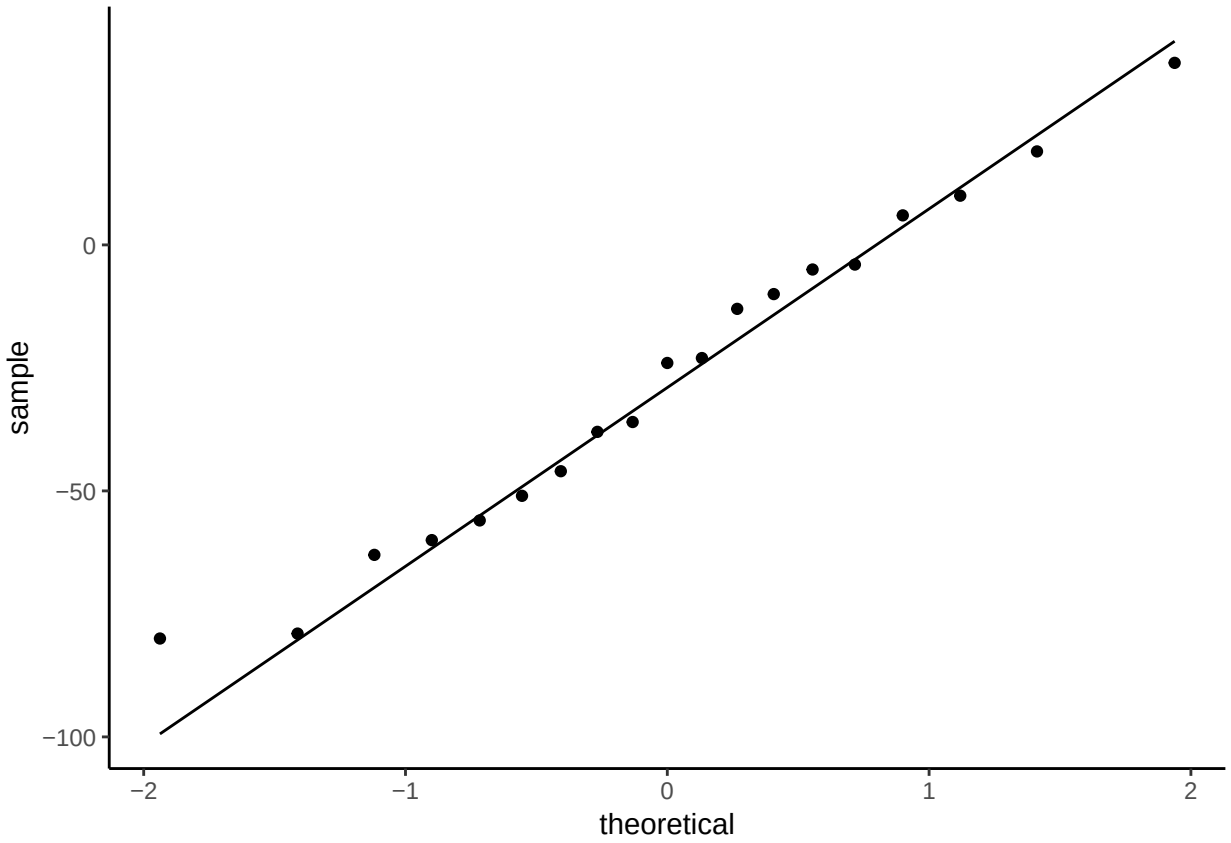
- **Answer:** The 90% confidence interval for the mean differences in weight gain between hogs given the two supplements is between 0.99 and 1.08, indicating that the hogs given the supplement experienced weight gain of 0.99 to 1.08 lbs above hogs that received a placebo.

3.) Mengovirus

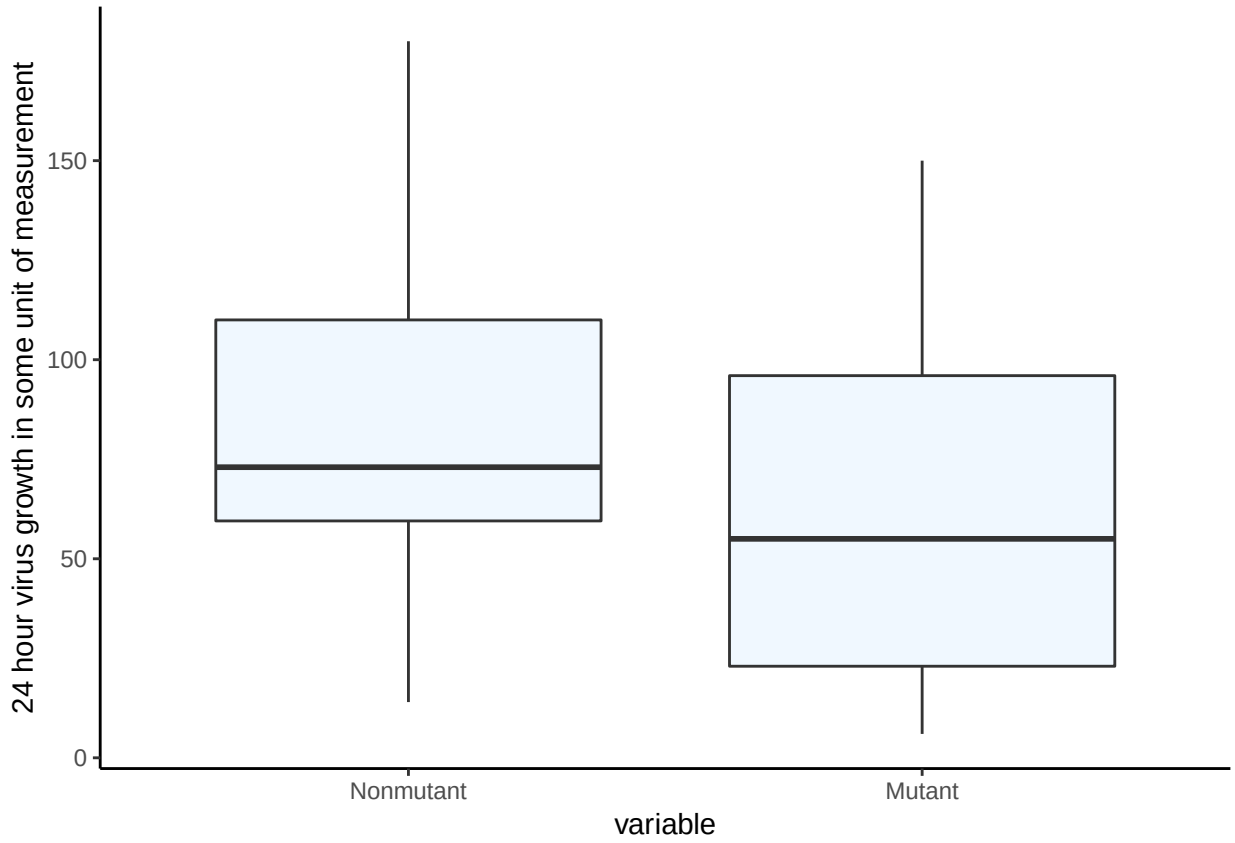
In a series of experiments on a certain virus (mengovirus), a microbiologist measured the growth of two strains of the virus - a mutant strain and a nonmutant strain - on mouse cells in petri dishes. Replicate experiments were run on 19 different days. The data are provided in *mengovirus.csv* where each number represents the total growthin 24 hours of the viruses in a single dish.

a) Should we compare the mean growths of the two strains or median growth? support your claim by providing graphical evidence.

```
df<-read.csv("Data/mengovirus.csv")
df2<-read.csv("Data/mengovirus.csv")
df2$diff<-df2$Mutant - df2$Nonmutant
ggplot(df2,aes(sample=diff)) + stat_qq() + stat_qq_line() + theme_classic()
```




```
df.m<-melt(df, id.vars = "Run")
ggplot(df.m,aes(x=variable, y=value )) +
  geom_boxplot(fill="aliceblue")+
  ylab("24 hour virus growth in some unit of measurement")+
  theme_classic()
```



- **Answer**: Given that the lines appear to fall along the straight qqline, mean virus growth should be sufficient for the assumption of normality. Variances appear to be approximately equal as well. Assuming the virus strains are independent, I will conduct a pooled t-test.

b) Based on the answer for part a), analyze the data using an appropriate method and provide a scientific conclusion.

- Let μ_1 be the mean 24 hour growth of the mutant virus strain, and μ_2 be the mean 24 hour growth of the nonmutant virus strain. Therefore:
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```
df.m<-melt(df, id.vars = "Run")
t.test(value~variable,
       data=df.m,
       var.equal = T,
       conf=0.95)
```

```
##
##  Two Sample t-test
##
## data:  value by variable
## t = 1.9272, df = 36, p-value = 0.06187
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.421631 55.737421
## sample estimates:
```

```
## mean in group Nonmutant    mean in group Mutant
##                 85.94737                 58.78947
```

- **Answer:** Using a pooled T-test we obtain a t-statistic of 1.9272 with 36 degrees of freedom and a p-value of 0.06187. Therefore, we have very weak evidence to reject the null hypothesis and cannot conclude that there is any true difference in mean 24 hour growth among mutant and non-mutant virus strains.

c) Construct a 95% confidence interval for an appropriate effect size. give an interpretation within the context of these data.

- **Answer:** The 95% confidence interval for these data lies between -1.42 and 55.7 units of growth, therefore, we cannot conclude that the true difference in mean virus growth is different than 0,

4.) Schizophrenia study

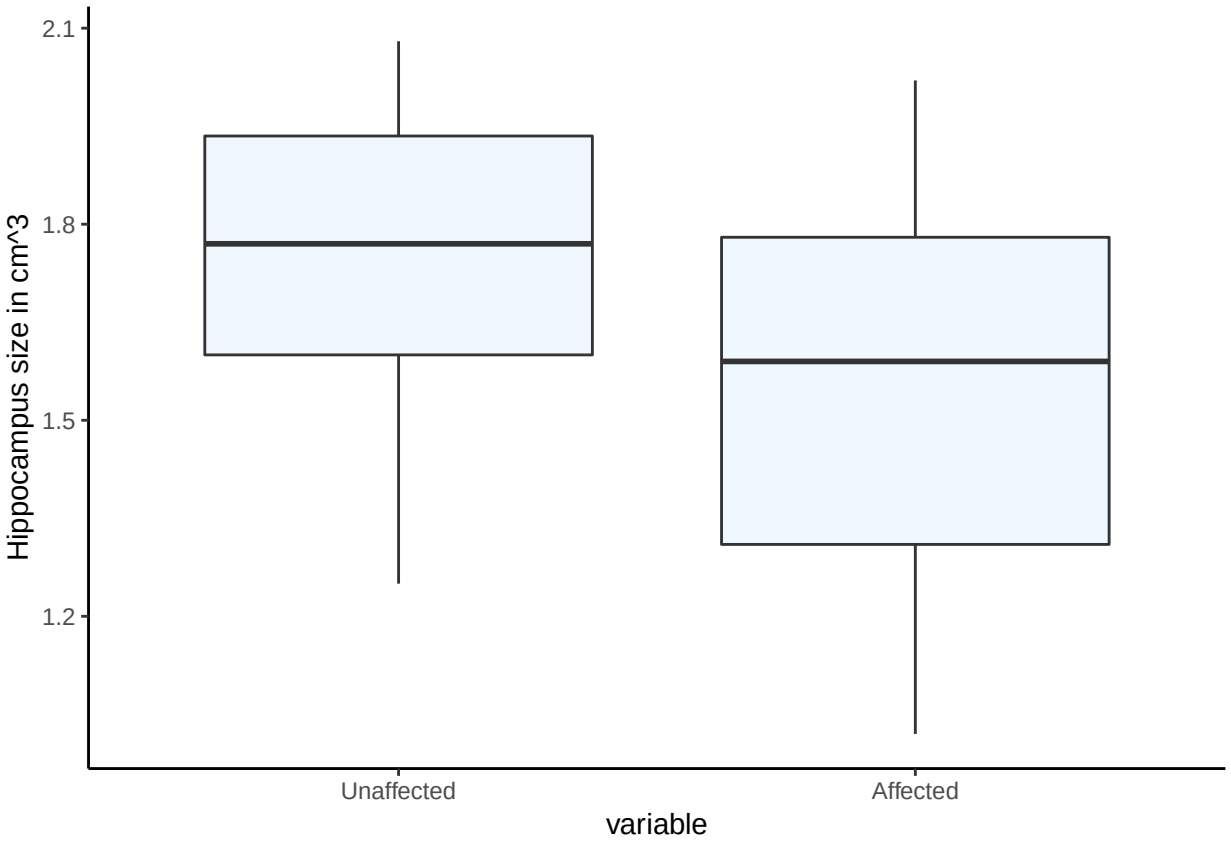
Data are provided in *Schizophrenia.csv*. Use these data to answer the following questions:

a) Draw a boxplot of the differences in hippocampus volumes. What does this plot suggest about the normality assumption of paired t-tools?

```
df<-read.csv("Data/Schizophrenia.csv")
df.m<-melt(df)
```

```
## No id variables; using all as measure variables
```

```
ggplot(df.m,aes(x=variable,y=value))+
  geom_boxplot(fill="aliceblue")+
  ylab("Hippocampus size in cm^3")+
  theme_classic()
```



- **Answer:** This plot suggests that the data may not be normal, however, box and whisker plots are more effective at visualizing *variance*, not normality. Perhaps a QQ plot would be more appropriate.

b) Take the log-transformation of the volumes for each of the 30 subjects. Conduct a *paired t-test* on the log transformed values to test the hypothesis of no difference in hippocampus volume. Include the null and alternative hypotheses, report the t-ratio, degrees of freedom, and p-value. Give a conclusion within the context of this study and on the original scale.

- Let μ_1 be the median hippocampus volume for unaffected people, and μ_2 be the median hippocampus volume for affected people. This is a study of twins. Therefore, in a paired sample framework:
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```
df.m$LogValue<-log(df.m$value)
t.test(LogValue~variable,
       data=df.m,
       conf=0.95,
       paired = T)
```

```
##
##  Paired t-test
##
## data:  LogValue by variable
## t = 3.1967, df = 14, p-value = 0.006463
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04228227 0.21470597
```

```
## sample estimates:
## mean of the differences
##               0.1284941
```

```
exp(0.1284941)
```

```
## [1] 1.137115
```

- **Answer:** The results from this paired t-test provides a t-ratio of 3.1967 with 14 degrees of freedom and a p-value of 0.006463. Therefore we have strong evidence against the null hypothesis that there is no difference in median hippocampus sizes between affected and unaffected twins. We would expect that, on average, the difference in hippocampus volumes is 1.137 cm^3 greater in twins that were unaffected with schizophrenia.

c) Compute the 95% confidence interval for the median ratio of hippocampus volumes between schizophrenic and non-schizophrenic twins. Interpret the interval within the context of these data.

```
exp(c(0.04228227, 0.21470597))
```

```
## [1] 1.043189 1.239497
```

- **Answer:** We are 95% confident that the true difference in mean volume of the hippocampus in twins unaffected with schizophrenia is between 1.042 and 1.244 cm^3 larger than those twins affected with schizophrenia.

5.) Cell Regeneration

In a study involving 8 rhesus monkeys, four were randomly selected and the nerves from the right sides of their cords were cut, while nerves from the left side were kept intact. The other four monkeys have nerves severed on the left side (and assumed to be intact on the right side). During the regeneration process, the content of of creatine phosphate **(CP; mg/100g tissue)** was measured in the left and right portions of the spinal cords. The average CP level for severed nerves is 10.875 mg/100g tissue and the average CP level for intact nerves is 15.5 mg/100g tissue. The standard deviation of the difference in CP is 4.886 mg/100g tissue.

a) Write downt he null and alternative hypotheses for comparing mean CP levels, compute the paired t-statistic and an exact p-value. Provide a conclusion in the conext of the data.

- Let μ_1 be the mean CP (mg/100g tissue) for severed nerves, and μ_2 be the mean CP (mg/100g tissue) for intact nerves. Therefore, in a paired sample framework:
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu 1 \neq \mu_2$

```
se<-4.886/sqrt(8)
t_ratio<-((10.875-15.5)/se)
abs(t_ratio)
```

```
## [1] 2.677338
```

```
p<-2*(1-pt(2.667,7))
p
```

```
## [1] 0.03213918
```

- **Answer:** We have a t-statistic of 2.677 and 7 degrees of freedom with a p-value of 0.0321. Therefore, the data show some evidence to reject the null hypothesis that there is no true difference in mean CP between severed and intact nerves.

b) Compute a 95% confidence interval for the difference between mean CP levels and interpret the interval in the given context.

```
lower<-(10.875-15.5)-(2.667*se)
upper<-(10.875-15.5)+(2.667*se)

data.frame(lower,upper)
```

```
##       lower      upper
## 1 -9.232141 -0.0178592
```

- **Answer:** we are 95% confident that the true difference in mean CP is bewteen -9.232141 and -0.0178592 mg/100g tissue in severed nerves when compared to intact nerves. This indicates that, on average, the severed nerves have less creatine phosphate than intact nerves per 100 grams of tissue.