140.616.01.HW9.JeremyDelahanty

Jeremy Delahanty

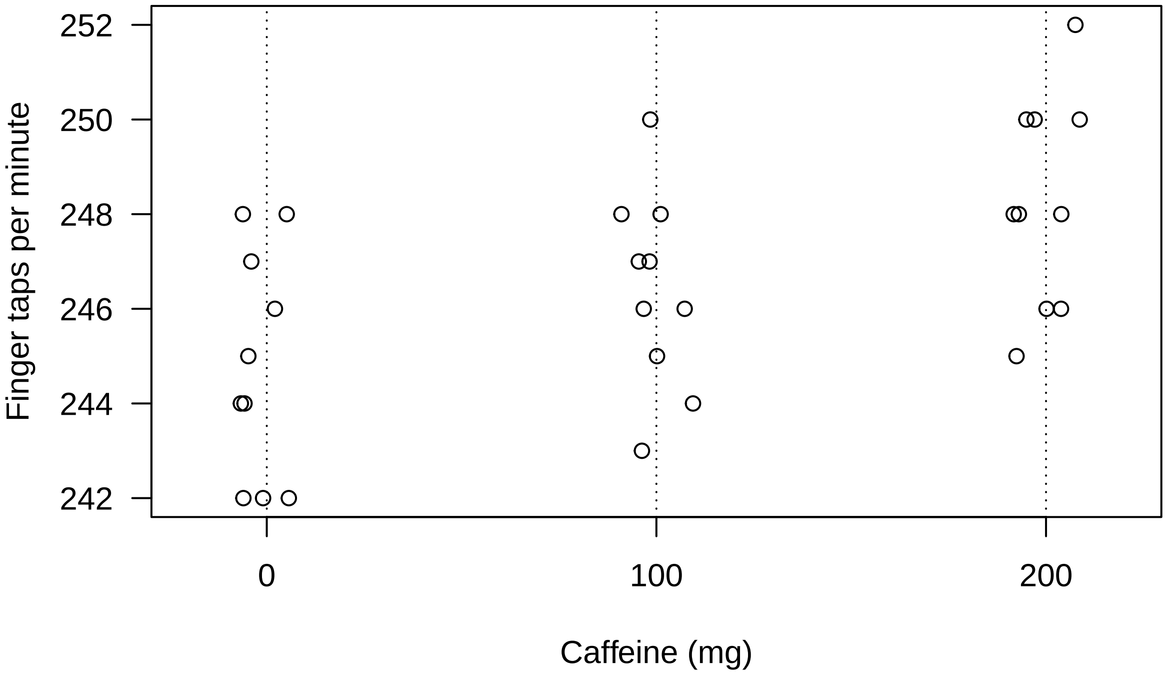
2024-04-24

```r
devtools::install_github("bllfrg/SPH.140.615", quiet = TRUE)
library(SPH.140.615)
```

#1a

Our null hypothesis is that there will be no difference on task performance between doses while our alternative hypothesis is that there is a difference between doses on caffeine doses on task performance. We will use $\alpha = 0.05$.

```r
caffeine_dose <- c(0,0,0,0,0,0,0,0,0,0,100,100,100,100,100,100,100,100,100,100,200,200,200,200,200,
200,200,200,200,200)
performance <- c(242,245,244,248,247,248,242,244,246,242,248,246,245,247,248,250,247,246,243,
244,246,248,250,252,248,250,246,248,245,250)
experiment_data <- data.frame(
  dose = factor(caffeine_dose),
  response = performance)
# Plot using stripchart
par(las=1)
stripchart(experiment_data$response~experiment_data$dose, vertical=TRUE,
  method='jitter', pch=1, xlab="Caffeine (mg)", ylab="Finger taps per minute")
abline(v=1:3, lty=3) #this will create a vertical line through each dose
```



1b

We can use the built in `aov` function for this portion of the question.

```
aov_caffeine <- aov(response~dose, data = experiment_data)

# Show summary of AOV
summary(aov_caffeine)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## dose         2   61.4  30.700   6.181 0.00616 **
## Residuals   27  134.1   4.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that our p-value is less than 0.05! However, to be sure that we should reject the null hypothesis, we must consider different statistical corrections.

1c) Bonferroni Correction

We want to do Bonferroni's method because we are considering multiplle comparisons!

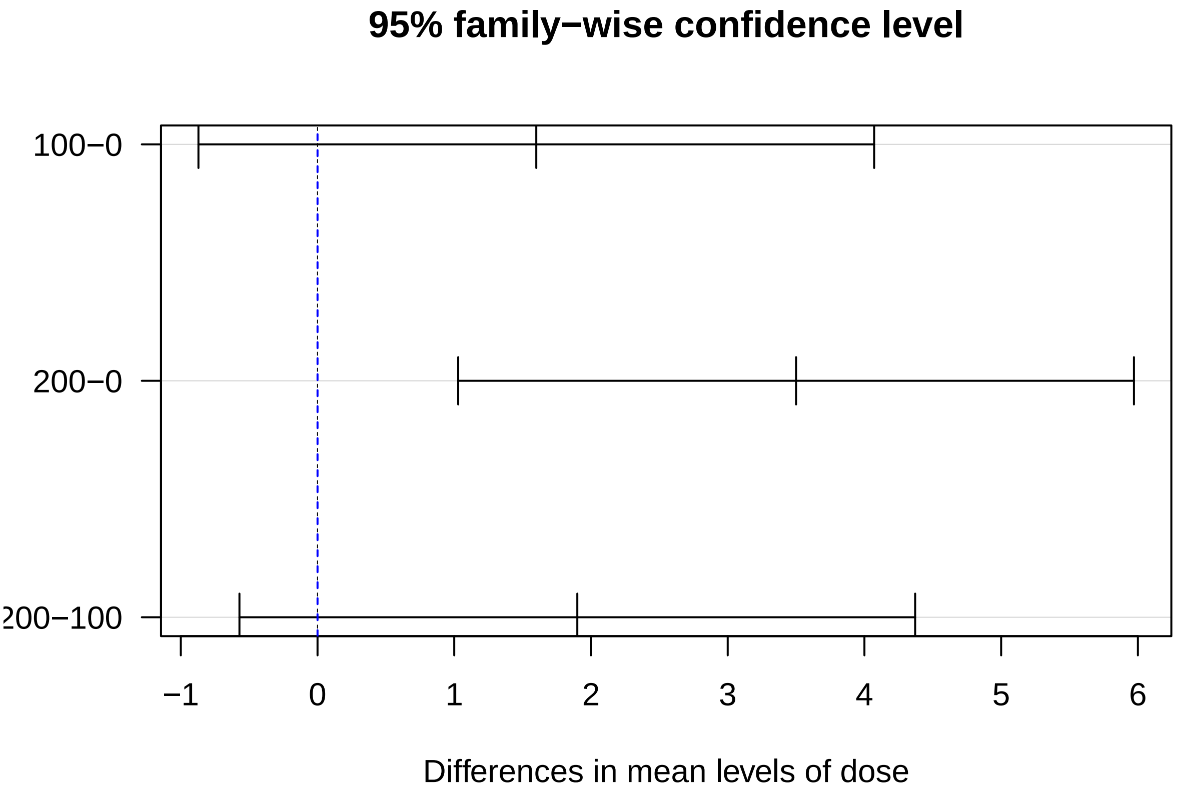
```r
# Use Built-in Bonferroni Method
pairwise.t.test(experiment_data$response, experiment_data$dose, p.adjust = 'bonf')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  experiment_data$response and experiment_data$dose
##
##     0      100
## 100 0.3601 -
## 200 0.0048 0.2019
##
## P value adjustment method: bonferroni
```

Our Bonferroni correction demonstrates that it is only the 0mg vs 200mg condition that gives a significant p-value of 0.048 while the rest are well above 0.05! We will perform additional tests according to the homework next.

1d) Tukey's Confidence Intervals (HSD)

```r
experiment_data.tukey <- TukeyHSD(aov_caffeine)
par(las = 1)
plot(experiment_data.tukey)
abline(v = 0, lty = 2, col = 'blue')
```



```
# Show the result of Tukey Intervals
experiment_data.tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = response ~ dose, data = experiment_data)
##
## $dose
##         diff        lwr      upr     p adj
## 100-0    1.6 -0.8711391 4.071139 0.2606999
## 200-0    3.5  1.0288609 5.971139 0.0043753
## 200-100  1.9 -0.5711391 4.371139 0.1562593
```

We can see again that it is only the 0mg vs 200mg conditions that has a confidence interval that does not include 0! This offers further evidence of the difference in means between these two groups is statistically significant at an alpha of 0.05. The `p-adj` value also further demonstrates that it is just the 0mg vs 200mg that is below our level of alpha as well.

We thus have evidence that we cannot reject our null hypothesis that caffeine generally influences task performance. We would only reject the null hypothesis when comparing zero caffeine against a large caffeine dose and therefore conclude that caffeine has an effect on task performance when a large 200mg dose is taken compared to no caffeine intake.

2

We will again use an alpha of 0.05. Our null hypothesis is that there is no difference in stem length across locations of the building while our alternative is that there is as difference in stem length across locations.

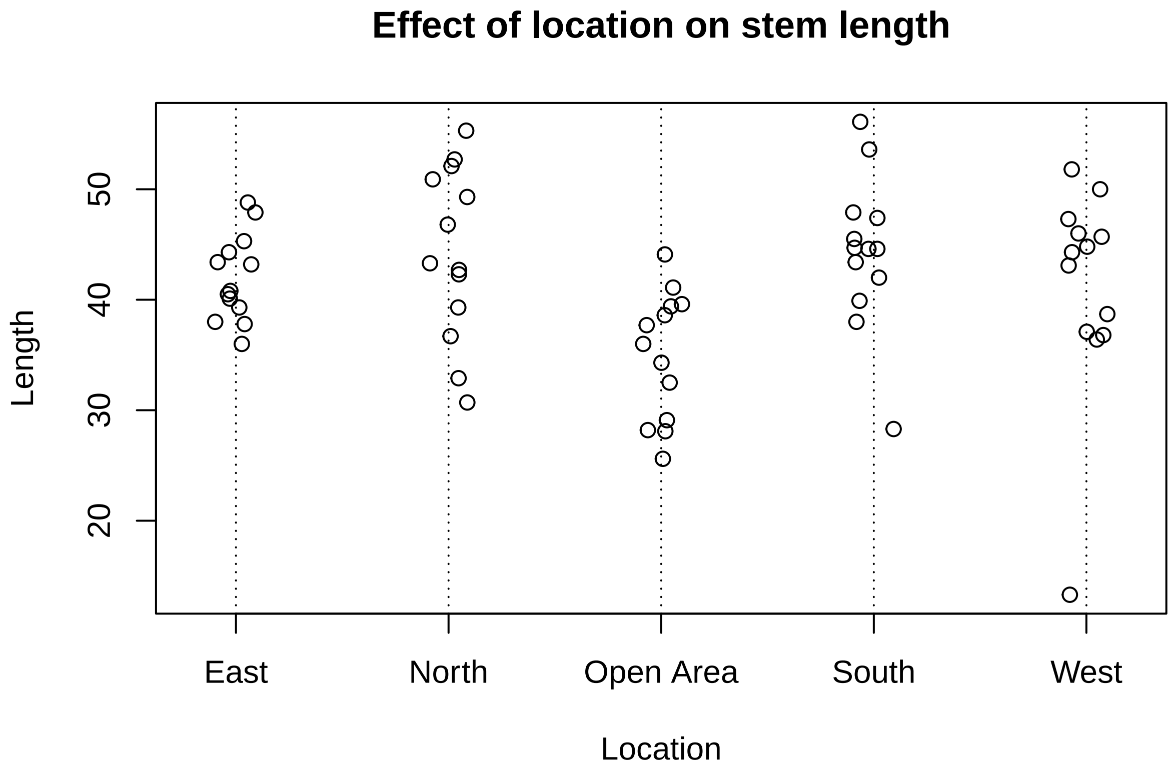
2a

We can create a `data.frame` and use the stripchart method again for plotting.

```
north <- c(39.3,42.3,52.7,42.7,36.7,32.9,43.3,52.1,30.7,46.8,55.3,49.3,50.9)
east <- c(39.3,47.9,43.2,36.0,44.3,40.5,38.0,37.8,43.4,48.8,45.3,40.1,40.8)
south <- c(43.4,44.6,53.6,39.9,42.0,45.5,44.7,28.3,38.0,56.1,44.6,47.9,47.4)
west <- c(38.7,13.3,43.1,50.0,46.0,37.1,44.8,44.3,45.7,47.3,36.4,36.8,51.8)
open <- c(37.7,28.1,32.5,36.0,39.4,29.1,25.6,44.1,39.6,28.2,41.1,34.3,38.6)

daffodils <- data.frame(
  Location = factor(rep(c("North", "East", "South", "West", "Open Area"), each = 13)),
  Length = c(north, east, south, west, open))

# Plot using stripchart
stripchart(daffodils$Length~daffodils$Location, vertical=TRUE, method='jitter',
  pch=1.5, xlab='Location',
  ylab="Length", main = "Effect of location on stem length")
abline(v=1:5, lty=3)
```



By looking at the chart, we can observe that there's maybe a shorter length in the open area flowers than on the other sides of the building. There is also very likely an outlier in the West group with a very small length.

2b)

We can again use the built in `aov` function to confirm there are differences in stem lengths.

```r
daffodil_aov <- aov(Length~Location, data = daffodils)

# Show summary of AOV
summary(daffodil_aov)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Location     4  759.8  189.95   3.756 0.00859 **
## Residuals   60 3034.3   50.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our ANOVE shows a p-value of 0.009 that is less than our alpha of 0.05. It looks like we could reject the null hypothesis! However, we want to make sure that we use correction methods before we make clear conclusions.

2c)

First, I will use Bonferroni's Method.

```r
# Bonferroni's method uses pariwise.t.test, as before
pairwise.t.test(daffodils$Length, daffodils$Location, p.adjust.method = "bonf")
```

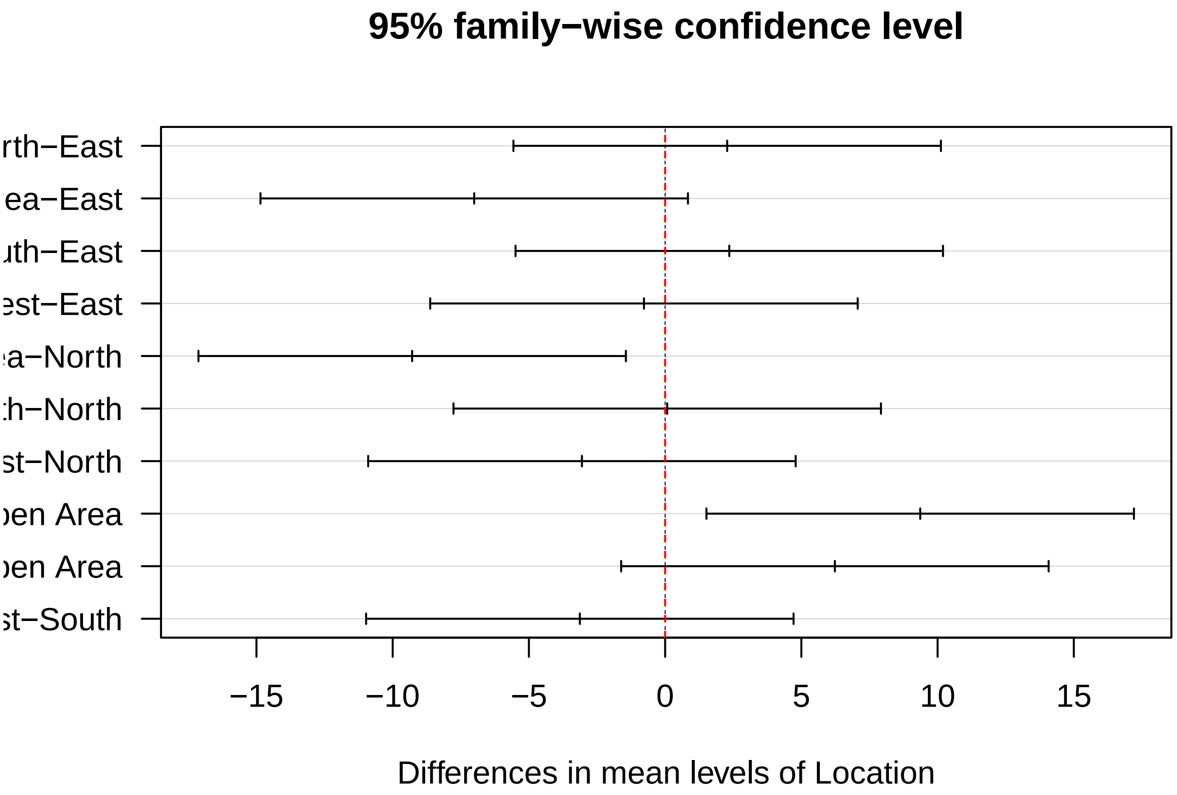
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  daffodils$Length and daffodils$Location
##
##           East  North Open Area South
## North     1.000 -     -         -
## Open Area 0.147 0.015 -         -
## South     1.000 1.000 0.014     -
## West      1.000 1.000 0.292     1.000
##
## P value adjustment method: bonferroni
```

Our Bonferroni correction demonstrates the p-value for each comparison of locations and we observe that it is just two comparisons that yield a significant p-value below 0.05! The north vs open area flowers and the south vs open area flowers.

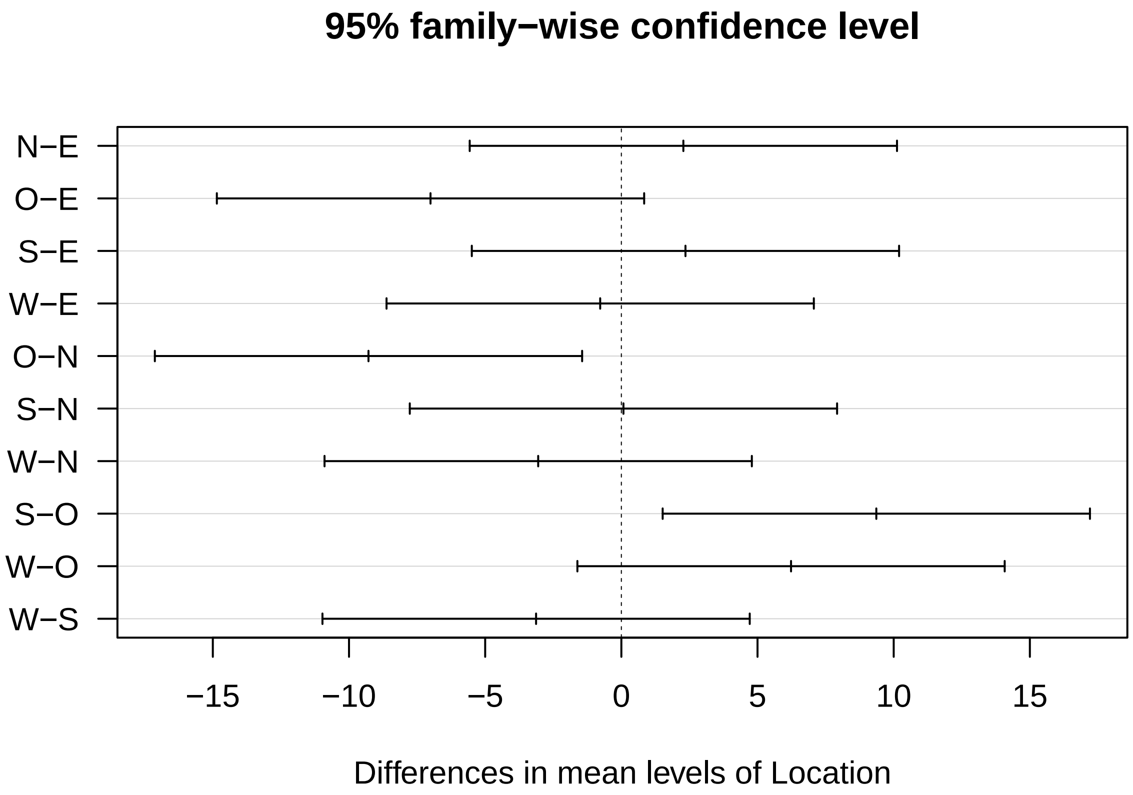
Next, I will use Tukey's HSD.

```r
daffodils.tukey <- TukeyHSD(aov(Length~Location, data = daffodils))
par(las = 1)

# Plot the data
plot(daffodils.tukey)
abline(v = 0, lty = 2, col = 'red')
```



```r
# The data doesn't plot very well like this, so we should change our factor names as so:
new_labels <- factor(rep(c("N", "E", "S", "W", "O"), each = 13))
daffodils$Location <- new_labels
daffodils.tukey <- TukeyHSD(aov(Length~Location, data = daffodils))
plot(daffodils.tukey)
```



```
# Show the Tukey results
daffodils.tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Length ~ Location, data = daffodils)
##
## $Location
##            diff        lwr        upr     p adj
## N-E  2.27692308  -5.567907 10.1217533 0.9246101
## O-E -7.00769231 -14.852522  0.8371379 0.1014116
## S-E  2.35384615  -5.490984 10.1986763 0.9157106
## W-E -0.77692308  -8.621753  7.0679071 0.9986441
## O-N -9.28461538 -17.129446 -1.4397852 0.0125298
## S-N  0.07692308  -7.767907  7.9217533 0.9999999
## W-N -3.05384615 -10.898676  4.7909840 0.8085014
## S-O  9.36153846   1.516708 17.2063686 0.0115753
## W-O  6.23076923  -1.614061 14.0755994 0.1815410
## W-S -3.13076923 -10.975599  4.7140609 0.7941028
```

Our plots from the Tukey HSD demonstrate that the open vs north locations and open vs south location show a statistically significant difference between mean levels where the confidence interval does not include zero! The Tukey HSD function's return further demonstrates this result as these two comparisons have a p-value that is less than 0.05.

2d) Kruskal-Wallis

We can use the Kruskal-Wallis test built into R.

```
kruskal.test(Length~Location, data = daffodils)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Length by Location
## Kruskal-Wallis chi-squared = 15.035, df = 4, p-value = 0.004629
```

The K-W Rank Sum Test demonstrates that we have a statistically significant value of 0.0046 that is even stronger than our 1-Way ANOVA from earlier!

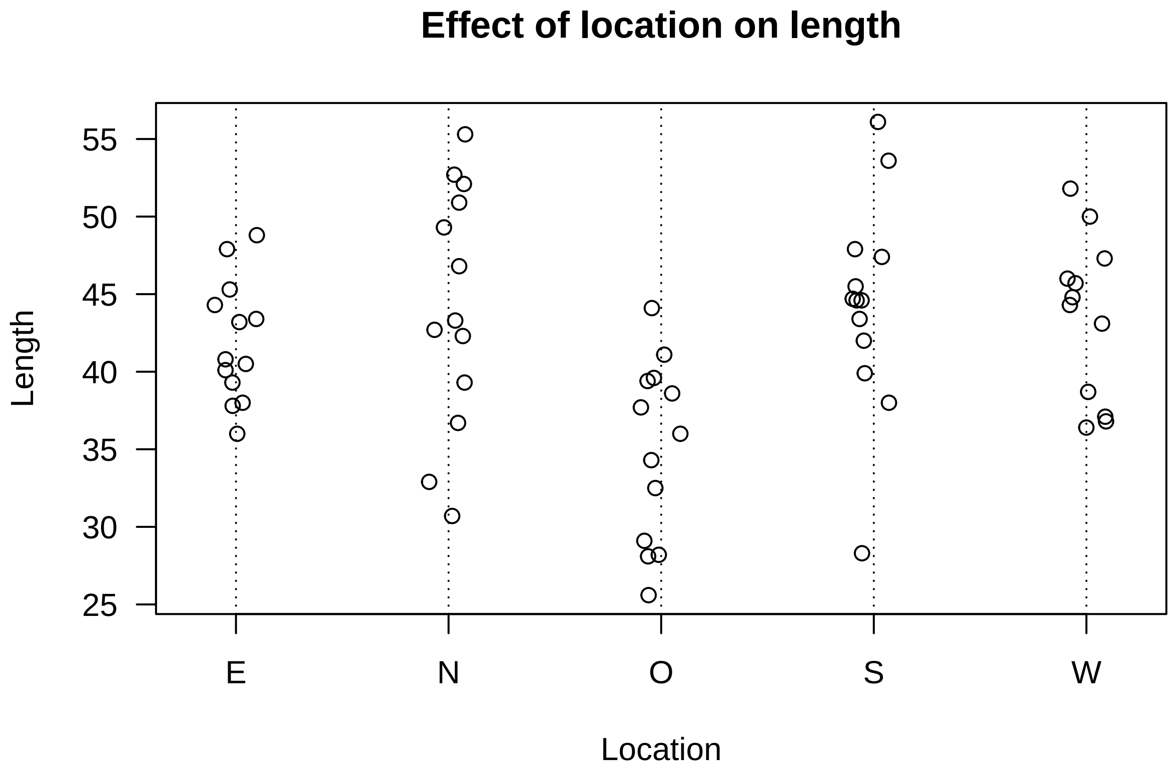
2e) Parametric one-way ANOVA without smallest data

We can find the smallest observation using the `min` function and remove it from the dataset in `R`.

```
min_value <- which.min(daffodils$Length)

# remove smallest value
new_daffodils <- daffodils[-min_value, ]

# Replot the data
par(las=1)
stripchart(new_daffodils$Length ~ new_daffodils$Location, method='jitter',pch=1,
  vertical=TRUE, xlab="Location", ylab="Length",
  main="Effect of location on length")
abline(v=1:5, lty=3)
```



```
# 1-way ANOVA
new_daffodils.aov <- aov(Length ~ Location, data = new_daffodils)
# Show summary
summary(new_daffodils.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Location     4  804.1  201.03    5.41 0.000885 ***
## Residuals   59 2192.4   37.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# K-W Test
kruskal.test(Length ~ Location, data = new_daffodils)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Length by Location
## Kruskal-Wallis chi-squared = 16.175, df = 4, p-value = 0.002793
```

Removing the smallest value here gave us a more significant p-value in both the one-way ANOVA and Kruskal-Wallis test!

2f) Pairwise Comparisons with Bonferroni

```
pairwise.t.test(new_daffodils$Length, new_daffodils$Location, p.adjust.method = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  new_daffodils$Length and new_daffodils$Location
##
##   E      N      O      S
## N 1.0000 -      -      -
## O 0.0480 0.0026 -      -
## S 1.0000 1.0000 0.0024 -
## W 1.0000 1.0000 0.0088 1.0000
##
## P value adjustment method: bonferroni
```

2g)

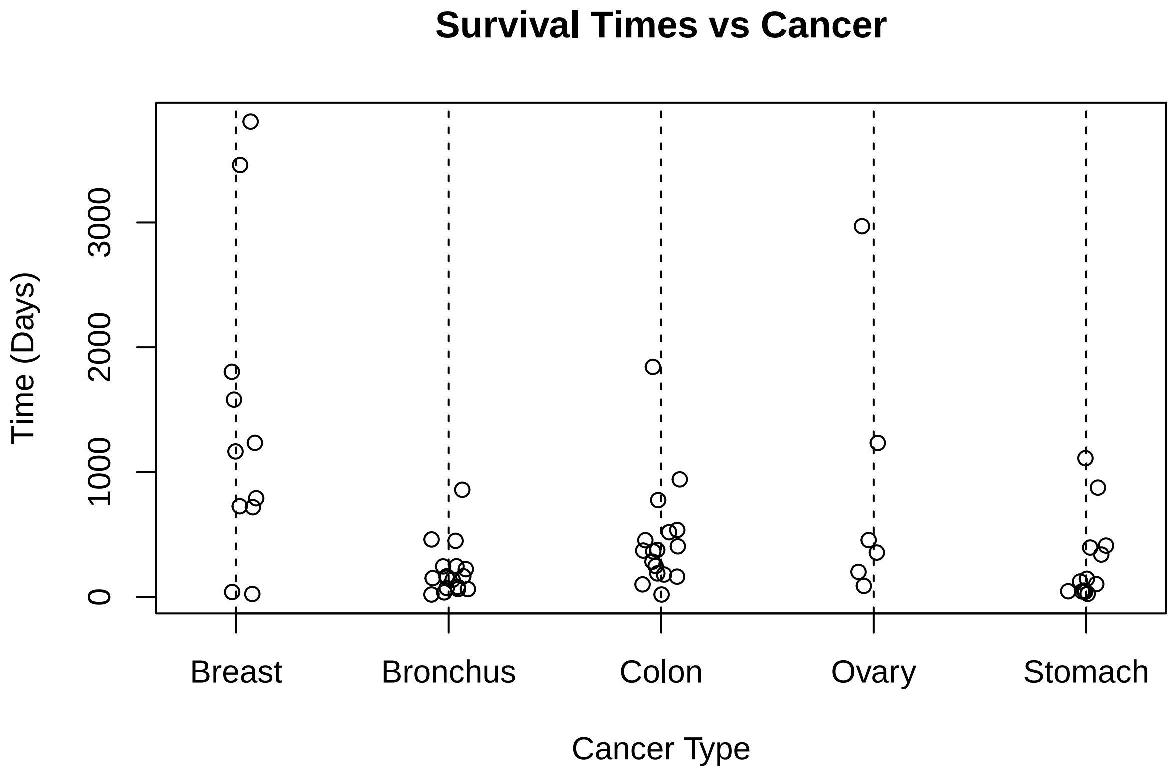
When removing the smallest value of the dataset, we found that the strength of our evidence that there are differences in stem lengths between locations increased when compared to an alpha of 0.05. We can also see that the Kruskal-Wallis test demonstrated increased significance when compared to the ANOVA which relies upon the F-statistic. The F-statistic is the ratio of between and within group variances, so if there's an extreme value in one of the datasets it is more vulnerable to being biased towards that value. The K-W test, which is dependent on rank, is more robust to extreme values. In addition, our use of multiple comparisons was appropriate in this case because we are interested in a comparison between all locations in our null hypothesis.

Once the extreme value was removed, we observed that not only is the north vs open and south vs open condition significant, but that the east vs open and west vs open comparisons were significant as well! Thus, we can rejuct the null hypothesis for these comparisons.

3

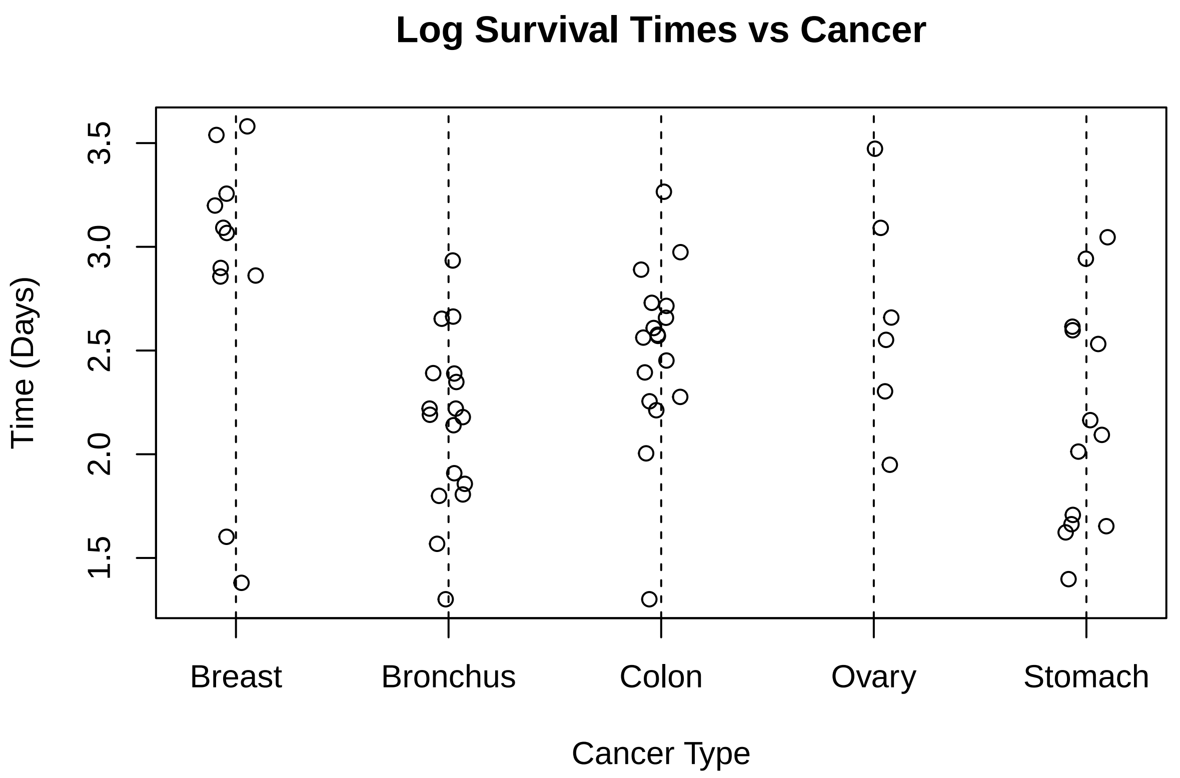
I will again use an alpha = 0.05.

```
stripchart(cancer$response~cancer$type, method='jitter',vertical=TRUE, pch=1.5,
           xlab="Cancer Type", ylab="Time (Days)", main = "Survival Times vs Cancer")
abline(v=1:5, lty=2)
```



```
# The data is quite bunched up around small values and our scale is on the order of
# thousands. A log transformation may help us.
log_responses <- log10(cancer$response)
cancer_log <- data.frame(response = log_responses, type = cancer$type)

# Plot transformed data
stripchart(cancer_log$response~cancer_log$type, method='jitter',vertical=TRUE, pch=1.5,
           xlab="Cancer Type", ylab="Time (Days)", main = "Log Survival Times vs Cancer")
abline(v=1:5, lty=2)
```



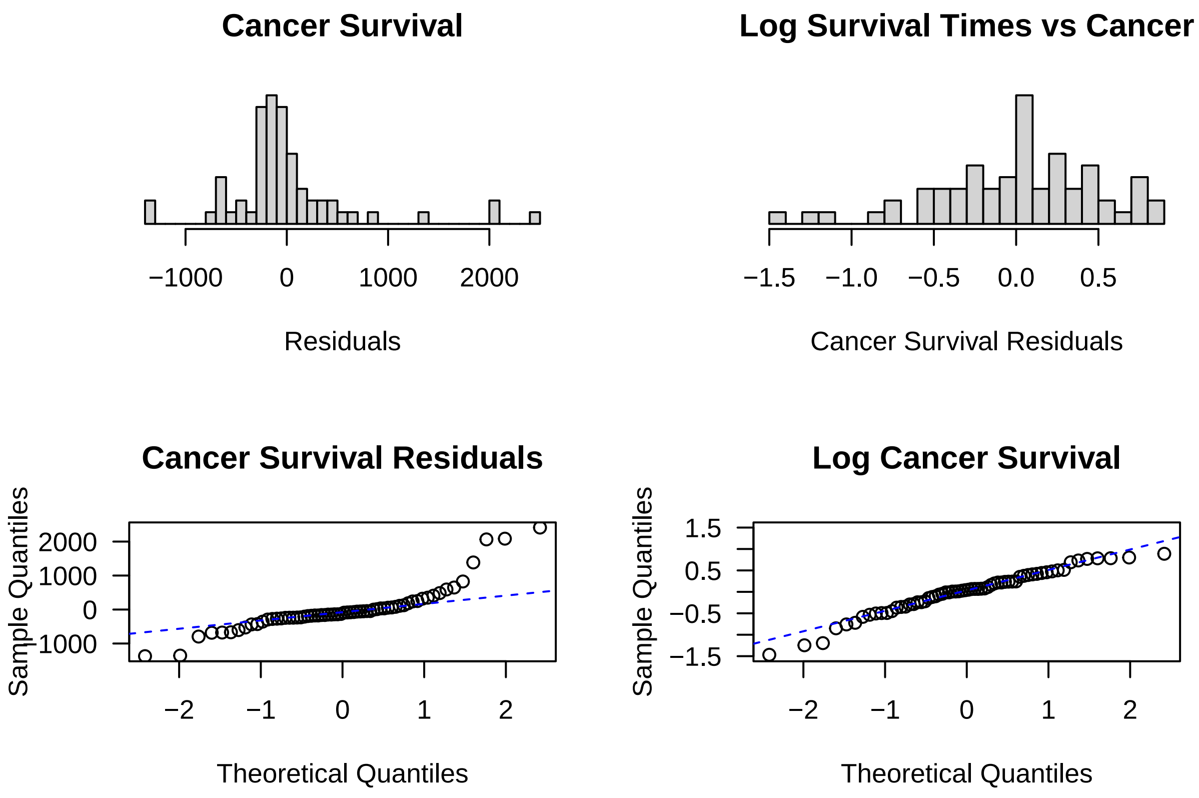
```
# Perform ANOVA Test on Original and Transformed Data to compare
cancer_aov <- aov(cancer$response~cancer$type)
summary(cancer_aov)
```

```
##              Df   Sum Sq Mean Sq F value   Pr(>F)
## cancer$type   4 11535761 2883940   6.433 0.000229 ***
## Residuals    59 26448144  448274
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

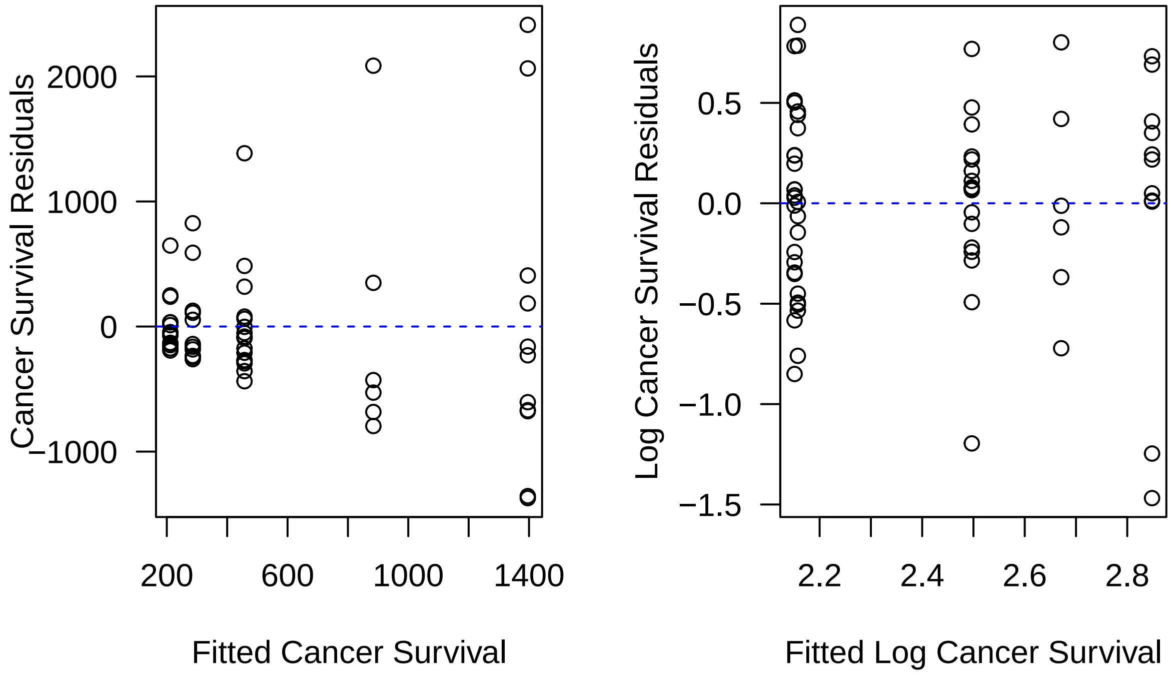
```
# AOV on transformed data
log_cancer_aov <- aov(cancer_log$response ~ cancer_log$type)
summary(log_cancer_aov)
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## cancer_log$type   4  4.618  1.1546   4.286 0.00412 **
## Residuals        59 15.894  0.2694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# We should also plot the residuals and a QQ Plot for the original and log
# transformed dataset to check for any substantial deviations in the data.
par(las = 1, mfcol=c(2,2))
hist(cancer_aov$residuals, breaks=30, yaxt='n', ylab='',xlab='Residuals',
     main="Cancer Survival", las=1)
qqnorm(cancer_aov$residuals, main="Cancer Survival Residuals", las=1)
qqline(cancer_aov$residuals, col="blue", lty=2, lwd=1)
#Plotting residuals & QQplot for transformed dataset
hist(log_cancer_aov$residuals, breaks=30, yaxt='n',ylab='',xlab='Cancer Survival Residuals',
     main="Log Survival Times vs Cancer", las=1)
qqnorm(log_cancer_aov$residuals, main="Log Cancer Survival", las=1, ylim=c(-1.5,1.5))
qqline(log_cancer_aov$residuals, col="blue", lty=2, lwd=1)
```



```
# Now that there are fitted values, we want to plot them against the residuals
par(las=1, mfrow=c(1,2))
plot(cancer_aov$fitted, cancer_aov$residuals, pch=1, xlab="Fitted Cancer Survival",
     ylab="Cancer Survival Residuals")
abline(h=0, lty=2, col="blue")
plot(log_cancer_aov$fitted, log_cancer_aov$residuals, pch=1, xlab="Fitted Log Cancer Survival",
     ylab="Log Cancer Survival Residuals")
abline(h=0, lty=2, col="blue")
```



```
# Now, perform a Kruskal-Wallis test on both datasets to compare
kruskal.test(cancer$response, cancer$type)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  cancer$response and cancer$type
## Kruskal-Wallis chi-squared = 14.954, df = 4, p-value = 0.004798
```

```
kruskal.test(cancer_log$response, cancer_log$type)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  cancer_log$response and cancer_log$type
## Kruskal-Wallis chi-squared = 14.954, df = 4, p-value = 0.004798
```

```
# Finally, we can perform Bartlett test for homogeneity of variance, another
# ANOVA assumption
bartlett.test(cancer$response ~ cancer$type)
```

```
##
##  Bartlett test of homogeneity of variances
##
```

```
## data:  cancer$response by cancer$type
## Bartlett's K-squared = 48.097, df = 4, p-value = 9.009e-10
```

```
bartlett.test(cancer_log$response, cancer_log$type)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  cancer_log$response and cancer_log$type
## Bartlett's K-squared = 4.809, df = 4, p-value = 0.3075
```

It appears from our ANOVA that we could reject the null hypothesis that there is no difference between survival times of different cancers. Indeed, the p-value is quite a lot smaller than our defined value for alpha. However, our checks for the model's assumptions demonstrates that the data is more homogeneous in its spread with the log transformed data than expected for a normal distribution visualized by the dashed lines in the QQ plots. Thus the transformed data fits better to the expected values and we should rely upon the log transformed dataset instead of the original dataset for our ANOVA.

In the K-W test, which does not assume normality from homogeneity in the dataset's spread, it can be better to use the original dataset which can protect the analysis from a Type I error (false positive). However, here we see that we get the same exact p-values for both the original and transformed datasets! We can reject the null hypothesis with this p-value and conclude that survival is indeed influenced by the type of cancer.

Finally, in Bartlett's Test for Homogeneity of Variance, we can observe that the p-value for our data having equal variances between groups in the non-transformed data is highly significant! The log-transformed data, however, is not. Thus, our transformed data would be better to use for our ANOVA.

Finally, our results demonstrate that, when using the log-transformed data which better satisfies the conditions for ANOVA, we can reject the null hypothesis that there is no difference in survival times of different cancers and conclude that cancers of different organs do indeed have different survival times.