Jake Sauter
Environmental Statistics
HW 4

An investigator is interested in studying the overall condition of lakes. They travel to 5 different lakes to examine a collection of different variables related to the status of the lake.

1) The investigator is interested in comparing the weight (pounds) of fish found in each of the lakes. They collect a sample of 35 fish at each of the lakes and the weights are record in the data set weightHW4. Analyze the resulting data set for the investigator and report your findings.

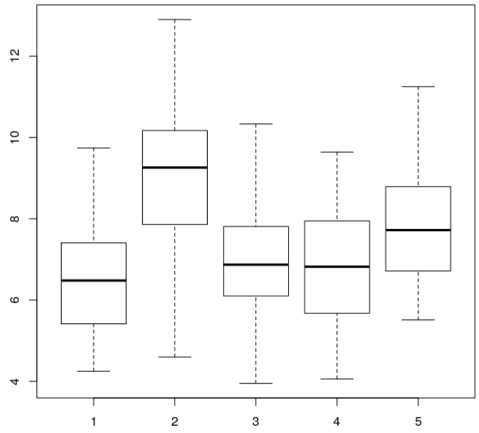
Possible tests for this analysis:

ANOVA

Kruskal-Wallis

We will first start with a boxplot of our data to get some intuition

```
>> data = read.table(”data/weightHW4.txt”)
>> attach(data)
>> boxplot(wlake1, wlake2, wlake3, wlake4, wlake5)
```



Just from observing this boxplot, it seems that a few of the groups (namely 1,3 and 4) look to be very similar. Group 2 seems ot be very different from all groups and group 5 is borderline.

ANOVA

If the assumptions for ANOVA are met, it will be an overall better test than Kruskal-Wallis as it is a parametric test that is permitted to use more representative statistics of the sample.

Assumptions for ANOVA:

1) Observations come from a random sample that represents the population

2) The residuals must be independent, normally distributed, with mean of 0 and a constant variance of σ^2

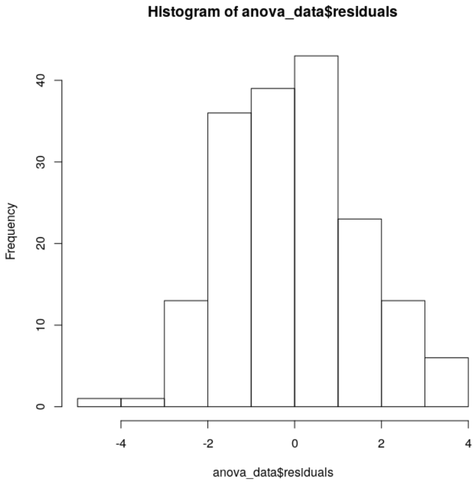
We can first test if all of our samples have the same variance, as if all of our samples have the same variance, then our residuals will also have the sampe variance

```
>> bartlett.test(list(wlake1, wlake2, wlake3, wlake4, wlake5))
Bartlett test of homogeneity of variances
data: list(wlake1, wlake2, wlake3, wlake4, wlake5)
Bartlett's K-squared = 2.8483, df = 4, p-value = 0.5835
```

A p-value of .5835 provides evidence that all population variances are equal. With this condition being met we can continue to investigate if ANOVA is valid in our current scenario. We will need to go through with the construction of ANOVA analysis in R to access the residuals, and in turn to assess if ANOVA is the correct choice of test here.

```
>> new = c(wlake1, wlake2, wlake3, wlake4, wlake5)
>> n = rep(35, 5)
>> group = rep(1:5, n)
>> data = data.frame(new, group=factor(group))
>> anova_data = aov(new group, data)
```

>> hist(anova_data$residuals)



Just from observation the distribution of the residuals seems to be normal and centered around 0, meeting the criteria for ANOVA, though we will perform a Shapiro-Wilkes test for normality to ensure that the distribution is normal

>> shapiro.test(anova_data$res)

Shapiro-Wilk normality test

data: anova_data$res W = 0.99161, p-value = 0.4015

A p-value of .4015 provides evidence that the distribution of residuals is normal, thus showing the data to meets the normality assumption. We will lastly test for centering around 0 before justifyibly analyzing ANOVA results.

>> mean(anova_data$residuals)

6.018189e-17

>> t.test(anova_data$residuals, mu=0)

One Sample t-test

data: anova_data$residuals t = 5.2227e-16, df = 174, p-value = 1 alternative hypothesis: true mean is not equal to 0 95 percent confidence interval: -0.227432 0.227432 sample estimates: mean of x 6.018189e-17

With a mean very close to 0 and a t-test of the mean of the residuals against 0 providing a p-value of 1 we have very strong evidence that the distribution of residuals meets all required assumptions. Finally we can interpret results of ANOVA.

```
>> summary(anova_data)
Df Sum Sq Mean Sq F value Pr(>F)
group 4 144.9 36.22 15.23 1.16e-10 ***
Residuals 170 404.3 2.38
—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

With a p-value very close to 0, we reject the null hypothesis that all group means are equal. This leads us into post-hoc analysis to determine which groups are different.

```
>> pairwise.t.test(new, group, pooled.sd=T, p.adjust=”bonf”, alternative=”two.sided”)
Pairwise comparisons using t tests with pooled SD
data: new and group
         1        2       3      4
 2  1.9e-09      -       -      -
 3   1.000   9.4e-07     -      -
 4   1.000   1.5e-07  1.000     -
 5   0.003    0.024   0.134  0.047
```

From the table of p-values above, we can see that our intuition is supported by evidence in that groups 1,3 and 4 are all similar. We also find evidence to support that groups 3 and 5 are also similar.

2) They are also interested in comparing the amount of a certain heavy metal (μ_g/g) found in seaweed in the lakes. They are using instrumentation that has a limit of detection of 0.1 μ_g/g. The investigator takes a sample of seaweed from 20 different locations at each lake resulting in the data set metalHW4. Any measured value that were below 0.1 μ_g/g were recorded as the reporting limit 0.1 μ_g/g. Analyze the resulting data set for the investigator and report your findings.

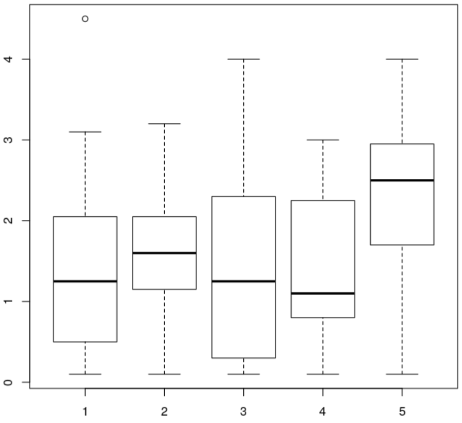
Possible tests for this analysis:

Kruskal-Wallis

For this data set we have to take into account that there is a limit of detection, and thus Non-Detects may be present, as well as ties in the data. Since we cannot make use of the magnitude of the data, we should use a non-magnitude based statistic, such as available in Kruskal-Wallis which is a test of medians.

Again we will begin our analysis with a boxplot of all of the groups we are examining.

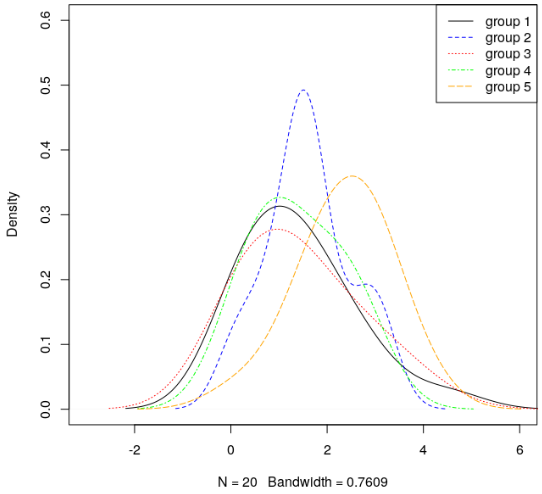
```
>> boxplot(tgroup1, tgroup2, group3, tgroup4)
```



From the boxplot above we can see that all of the groups are pretty similar, with group 5 showing the most difference from all groups.

In order to make use of the Kruskal-Wallis test, we need to make sure that all of our data has the same shape and variance. To examine this we will overlay the density distributions of the data.

```
>> dens_1 = density(mlake1, adjust=1.5)

>> dens_2 = density(mlake2, adjust=1.5)

>> dens_3 = density(mlake3, adjust=1.5)

>> dens_4 = density(mlake4, adjust=1.5)

>> dens_5 = density(mlake5, adjust=1.5)

>> plot(dens_1, lty=1, xlim=c(-3,6), ylim=c(0,.6))

>> lines(dens_2, lty=2, col=”blue”)

>> lines(dens_3, lty=3, col=”red”)

>> lines(dens_4, lty=4, col=”green”)

>> lines(dens_5, lty=5, col=”orange”)

>> legend(”topright”, c(”group 1”, ”group 2”, ”group 3”, ”group 4”, ”group 5”),
lty=c(1,2,3,4, 5), col=c(”black”, ”blue”, ”red”, ”green”, ”orange”))
```



From the plot above, we can see that all of the distributions have about the same shape and variance, so we can continue with our Kruskal-Wallis test.

```
>> tnew = c(mlake1, mlake2, mlake3, mlake4, mlake5)

>> n = rep(20,5)

>> group = rep(1:5, n)

>> data = data.frame(tnew, group=factor(group))

>> kruskal_data = kruskal.test(tnew group, data)
>> kruskal_data

Kruskal-Wallis rank sum test

data: tnew by group

Kruskal-Wallis chi-squared = 13.539, df = 4, p-value = 0.008923
```

With a p-value of .008923, we reject the null hypothesis that the medians of all of the groups are the same, and thus should continue with post-hoc analysis to analyze the similarity structure of the groups. For this post-hoc analysis we cannot use t-tests as we do not hvae the necessary normality assumption; instead we can use the Wilcoxn Rank Sum test.

```
>> pairwise.wilcox.test(tnew, group, p.adjust=”bonf”, exact=F)

Pairwise comparisons using Wilcoxon rank sum test

data: tnew and group

     1      2      3      4
2  1.000    -      -      -
3  1.000  1.000    -      -
4  1.000  1.000  1.000    -
5  0.036  0.089  0.092  0.031
```

Again, our intuition served us well as we see from the p-values above, there is evidence to suggest that groups 1,2,3 and 4 are all similar and group 5 is statistically significantly different from groups 1 and 4 at that α = .05 level.

3) The investigator is interested in comparing the turbidity (clarity) of the water (NTU). At each of the 5 different lakes they select 25 locations and measure the turbidity resulting in the data set turbidHW4. Analyze the resulting data set for the investigator and report your findings.

Possible tests for this analysis:

ANOVA

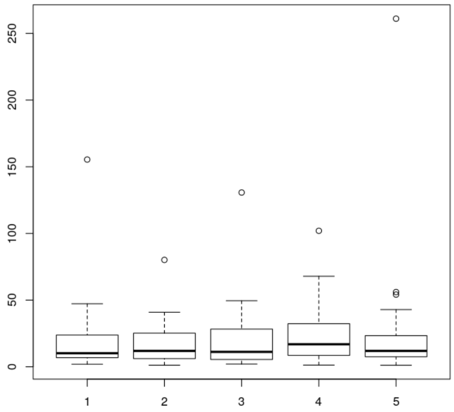
Kruskal-Wallis

We will first briefly examine our data as before with side by side boxplots.

>> data = read.table(”data/turbidHW4.txt”)

>> attach(data)

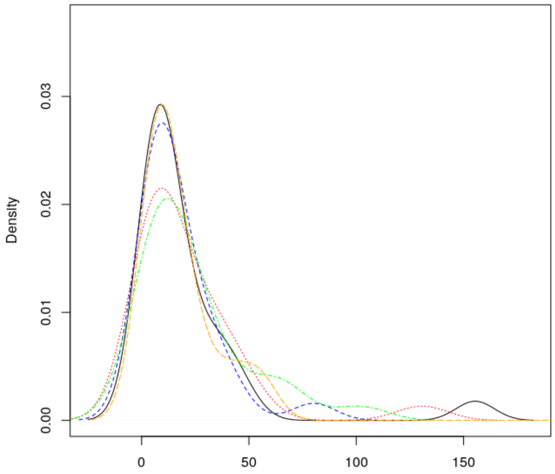
>> boxplot(tlake1, tlake2, tlake3, tlake4, tlake5)



From the boxplot above we can see that the groups look fairly similar and they all contain outliers on their right tails. Because of this, we should implement a test that makes use of a more robust statistic, so we choose to use the Kruskal-Wallis test. It is important to note that even if we did decide to implement ANOVA, we would need to check if normality of residuals is met.

In order to use the Kruskal-Wallis test, we need to make sure that all of the groups have the same shape and variance, so we will overlay the distribtuions to further examine these criteria.

```
>> dens_1 = density(tlake1, adjust=1.5)
>> dens_2 = density(tlake2, adjust=1.5)
>> dens_3 = density(tlake3, adjust=1.5)
>> dens_4 = density(tlake4, adjust=1.5)
>> dens_5 = density(tlake5, adjust=1.5)
plot(dens_1, lty=1, ylim=c(0,.037))
>> lines(dens_2, lty=2, col=”blue”)
>> lines(dens_3, lty=3, col=”red”)
>> lines(dens_4, lty=4, col=”green”)
>> lines(dens_5, lty=5, col=”orange”)
>> legend(”topright”, c(”group 1”, ”group 2”, ”group 3”, ”group 4”, ”group 5”), lty=c(1,2,3,4, 5), col=c(”black”, ”blue”, ”red”, ”green”, ”orange”))
```



As seen from the overlain density plots above, the groups most certainly have the same shape and variance, thus our assumptions for the Kruskal-Wallis test are met and we can perform the test.

```
>> tnew = c(tlake1, tlake2, tlake3, tlake4, tlake5)
>> n = rep(25,5)
>> group = rep(1:5, n)
>> data = data.frame(tnew, group=factor(group))
>> kruskal_data = kruskal.test(tnew group, data)
>> kruskal_data

Kruskal-Wallis rank sum test

data: tnew by group

Kruskal-Wallis chi-squared = 1.3503, df = 4, p-value = 0.8528
```

With a p-value of .8528, we fail to reject the null hypothesis that all groups share the same median, and since all groups share the same median there is no need for post-hoc analysis. This concludes our analysis of this data set.