## Lab 7. Introduction to 1-way ANOVA

The purpose of this lab is to use R to perform and interpret 1-way ANOVA and Tukey’s HSD

Question 1.

Jaffe, Parker and Wilson investigated the concentration of several hydrophobic organic substances (such as hexachlorobenzene, chlordane, heptachlor, aldrin, dieldrin, endrin) in the Wolf River in Tennessee. Measurements were taken downstream of an abandoned dump site that had previously been used by the pesticide industry to dispose of its waste products.

It was expected that these hydrophobic substances might have a nonhomogeneous vertical distribution in the river because of differences in density between these compounds and water and because of the adsorption of these compounds on sediments, which could lead to higher concentrations on the bottom. *It is important to check this hypothesis because the standard procedure of sampling at six-tenths of the depth could miss the bulk of these pollutants if the distribution were not uniform.*

Grab samples of 1L were taken at various depths of the river. Ten surface, 10 mid-depth and 10 bottom samples were collected, all within a relatively short period. Until they were analyzed the samples were stored in 1-quart mason jars at low temperature. In the analysis of the samples, a 250-mL water sample was taken from each mason jar and was extracted with 1 mL of either hexanes or petroleum ether. A sample of the extract was then injected into a gas chromatograph and the output was compared against standards of known concentrations. The test procedure was repeated two more times, injecting different samples of the extract in the gas chromatograph. The average aldrin and hexachlorobenzene (HCB) concentrations (in nanograms per liter) in these 30 samples are given in the file Jaffe.csv in datasets/demos.

* 1. **Why would it be inappropriate to treat the three gas chromatograph readings from the same 1L mason jar as independent observations? What have the authors done to avoid violating the assumption of independence?**

It would be inappropriate because taking samples from the same jar will make the concentration readings related to each other.

* 1. **What is the main research question being tested statistically?**

Is the concentration of hydrophobic organic substances uniformly distributed vertically in the Wolf River.

**To address this question statistically, it is first necessary to re-express the question from a statistical perspective. The null hypothesis tested in a single factor ANOVA can be expressed in two (functionally synonymous ways): as a relationship between group means, or a relationship between group effects.**

* 1. **Select the null hypothesis for the hydrophobic substances example, based on the relationship between group means**

H0: µb = µm = µs

* 1. **One-way ANOVA partitions variance into two categories. These categories are…**

Explained Variance (between groups) and Unexplained Variance (Within Groups)

* 1. In an earlier question we addressed the assumption of independence.  **Do the data meet the remaining assumptions of ANOVA? How can you tell? What should you do?** A complete answer will address assumptions for aldrin concentration and HCB concentration and give specific details.
* The assumptions of ANOVA are similar to two sample t test. We already talked about the assumption of independence above. The other assumptions are the normal distribution of the variables and that the groups have equal variance. To test these assumptions, you can look at qq, histogram and boxplots to test for normality in R.
* Aldrin:

-For the histogram, the plots look pretty normal except for an outlier right in the bottom group.

-The middepth boxplot is normal, with mean & median being the same and centered. The surface group has similar whiskers, but the mean and the median are right skewed in the IQR, at the top of the box, but with no outliers. For bottom group, the mean is much higher than the median, and there is a high outlier. The longer whisker on the bottom might balance out the outlier.

-The qq plot shows linear for middepth and surface but not completely for bottom.

-I will try log transforming the data for [Aldrin]

-Although there is similar patterns as mentioned before, it makes those deviations less severe. Overall, I would say that it is still normal enough to perform ANOVA.

-The ratio of variance before log transformation was a bit above 3, but then after log transformation was 2.09, so checks off the assumption of equal variance

* **HCB:**

-For HCB data, histogram has no real squew visible. For the boxplots, there are no outliers in any group. THe surface group has similar mean and median in the almost center of the IQR. The middepth has a slightly right skewed median but similar whisker lengths. THe bottom group is left skewed with a longer whisker but no outliars. I would say the data is normal enough to still analyze parametrically

-The ratio is 1.75 so the variances are equal.

* 1. **For this example, what is *p* (the number of levels)? What is N? What should your dfgroups and dferror be?**
* *p* = 3
* N= 30
* dfgroups = (p-1) = 2
* dferror = (N-p) = 30-3 = 27
  1. Use R to perform one-way ANOVAs. **Record your results in the tables below. Confirm for yourself with some quick math that the df are assigned correctly and that the MS and *F* are calculated correctly. Round to 2 decimal places.**

Table 1. Results of a one-way ANOVA testing the effect of water depth on **aldrin concentration** (ng/L).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **df** | **SS** | **MS** | ***F*** | ***P-value*** |
| **Depth** | **2** | **25.454** | **12.727** | **4.634** | **0.0186** |
| **Error** | **27** | **74.099** | **2.744** |  |  |
| **Total** | **29** | **99.553** |  |  |  |

Table 2. Results of a one-way ANOVA testing the effect of water depth on **log-transformed aldrin concentration** (ng/L).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **df** | **SS** | **MS** | ***F*** | ***P*** |
| **Depth** | **2** | **0.731** | **0.366** | **5.592** | **0.00929** |
| **Error** | **27** | **1.765** | **0.0654** |  |  |
| **Total** | **29** | **2.496** |  |  |  |

Table 3. Results of a one-way ANOVA testing the effect of water depth on hexachlorobenzene (HCB) concentration (ng/L).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **df** | **SS** | **MS** | ***F*** | ***P*** |
| **Depth** | **2** | **5.357** | **2.678** | **3.032** | **0.0649** |
| **Error** | **27** | **23.848** | **0.883** |  |  |
| **Total** | **29** | **29.205** |  |  |  |

Although we have now established that there is a statistical difference between the group means for log10 transformed aldrin concentrations, we do not yet know which group(s) are different from which other(s). For this data a Tukey multiple comparison test (to determine which depth groups are different from each other, *in terms of log10 transformed aldrin concentration*) is appropriate.

Perform a Tukey-Kramer Honestly Significant Difference (HSD) pairwise comparison of group means. **Which pairs are significantly different?** Make sure you are looking at the correct response variable.

* 1. **Report your results for log10 transformed aldrin concentration and hexachlorobenzene as you would in a scientific paper.**
  2. **Based on these results, would you recommend a change to the the standard procedure of sampling at six-tenths of the depth? Why or why not?**