## 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?**

Those three readings that came from the same jar, meaning that the values they get from one chromatography reading will affect the other two readings. All of the water in that jar was taken at the same time and from the same specific area of water, so obviously what is present in the whole liter taken will be related to each other and cannot be considered independent. However, the 10 individual grab samples can be considered independent, and was how the authors avoided violating the assumption. Although they are taken from the same group area (such as bottom, middepth or surface), they were taken at different times in space, and they probably had a layout planned of where they were taking these samples from. That would be so that they have, for example, “bottom” readings from multiple areas around the river instead of all close together.

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

Is the concentration of hydrophobic organic substances uniformly vertically distributed 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 in IQR. The surface group has similar whiskers, but the mean and the median are at the top of the IQR 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 because of outlier

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

-Although there are similar patterns as mentioned before, it makes those deviations from normality less severe. –

Overall, I would say that it is still normal enough to perform parametric 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 is slightly left skewed but similar whisker lengths. The bottom group has a median slightly closer to the bottom of the IQR with a longer whisker on the bottom but no outliers. I would say the data is normal enough to still analyze parametrically

-The ratio is 1.75 so the variances are equal.

* Another way to address the assumptions is with a residual plot.
* For both the raw Aldrin and the log aldrin residuals plot, it looks as though the residuals increase as the fitted values increase, but not to the extent of an aggressive triangle shape, so after checking assumptions again we are still okay.
* For the HCB, the plot does not have that triangle shape, so the data still meets the assumption of normality

* 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** | **4.638** | **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** | **0.1379** | **0.06895** | **5.592** | **0.00929** |
| **Error** | **27** | **0.3329** | **0.01233** |  |  |
| **Total** | **29** | **0.471** |  |  |  |

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.

**-**The only pair that is significantly different is the Bottom and the surface group (with p< 0.00684)

* 1. **Report your results for log10 transformed aldrin concentration and hexachlorobenzene as you would in a scientific paper.**
* The log10 transformed aldrin concentrations varied significantly with depth (One-way ANOVA: F2,27= 5.59, p = 0.00929)
* The log10 transformed aldrin concentration were significantly different in the bottom compared to the surface group, (Tukey HSD: t=3.34, p = 0.00673). But not statistically different when comparing the middepth and surface (Tukey HSD: t=1.48, p = 0.316) and middepth and bottom (Tukey HSD: t=1.86, p = 0.171)
* The HCB concentration did not vary significantly with depth (One-way ANOVA: F2,27= 3.03, p = 0.0649)
  1. **Based on these results, would you recommend a change to the standard procedure of sampling at six-tenths of the depth? Why or why not?**
* Probably not. If the protocol had been measuring surface concentration instead of at 6/10 depth, then I would maybe advise changing because that concentration was significantly different from the bottom. However, the middepth is not different from the surface and neither from the bottom measurements. Therefore, if it is not statistically different to sample from the middle and from the bottom, then the 6/10 depth is okay, you do not miss the bulk of pollutants.