**One-way ANOVA**

**In-Class Activity #1**

Here are some data on a shell measurement (the length of the anterior adductor muscle scar, standardized by dividing by length; I'll call this "AAM length") in the mussel *Mytilus trossulus* from five locations: Tillamook, Oregon; Newport, Oregon; Petersburg, Alaska; Magadan, Russia; and Tvarminne, Finland, taken from a much larger data set used in McDonald et al. (1991). The nominal variable is location, with the five values Tillamook, Newport, Petersburg, Magadan, and Tvarminne. There are six to ten observations of the measurement variable, AAM length, from each location.

Copy and paste the data to excel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tillamook | Newport | Petersburg | Magadan | Tvarminne |
| 0.0571 | 0.0873 | 0.0974 | 0.1033 | 0.0703 |
| 0.0813 | 0.0662 | 0.1352 | 0.0915 | 0.1026 |
| 0.0831 | 0.0672 | 0.0817 | 0.0781 | 0.0956 |
| 0.0976 | 0.0819 | 0.1016 | 0.0685 | 0.0973 |
| 0.0817 | 0.0749 | 0.0968 | 0.0677 | 0.1039 |
| 0.0859 | 0.0649 | 0.1064 | 0.0697 | 0.1045 |
| 0.0735 | 0.0835 | 0.105 | 0.0764 |   |
| 0.0659 | 0.0725 |   | 0.0689 |   |
| 0.0923 |   |   |   |   |
| 0.0836 |   |   |   |   |

**Hypothesis**

The statistical null hypothesis is that the means of the measurement variable are the same for the different categories of data; the alternative hypothesis is that they are not all the same. For the example data set, the null hypothesis is that the mean AAM length is the same at each location, and the alternative hypothesis is that the mean AAM lengths are not all the same.

1. *Using statistical notation, write your hypothesis below:*

*Null: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Alternative: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

**Assumptions to consider before conducting a one-way anova**

* The samples must be independent
* The populations from which the samples were obtained must be normally or approximately normally distributed
* The variances of the populations must be equal

One-way anova assumes that the observations within each group are normally distributed. It is not particularly sensitive to deviations from this assumption; if you apply one-way anova to data that are non-normal, your chance of getting a P value less than 0.05, if the null hypothesis is true, is still pretty close to 0.05. You should calculate the residuals (the difference between each observation and the mean of its group) and plot them. If the residuals look severely non-normal, we might want to consider a different type of analysis.

*Using the AAM data from above, copy and paste your 5 residual normal probability plots here:*

One-way anova also assumes that your data are homoscedastic, meaning the standard deviations are equal in the groups. You should examine the standard deviations in the different groups and see if there are big differences among them.

1. *Using the AAM data from above, what do you observe about the standard deviations:*

**What happens if I don’t meet the assumptions to conduct a one-way anova?**

You should try really hard to have equal sample sizes in all of your groups. With a balanced design, you can safely use a one-way anova unless the sample sizes per group are less than 10 and the standard deviations vary by threefold or more. If you have a balanced design with small sample sizes and very large variation in the standard deviations, you should use Welch's anova instead. We’ll save this for graduate school. If you have an unbalanced design, you should carefully examine the standard deviations. Unless the standard deviations are very similar, you should probably use Welch's anova. It is less powerful than one-way anova for homoscedastic data, but it can be much more accurate for heteroscedastic data from an unbalanced design.

**One-way anova guided practice**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Tillamook | Newport | Petersburg | Magadan | Tvarminne |
|  | 0.0571 | 0.0873 | 0.0974 | 0.1033 | 0.0703 |
|  | 0.0813 | 0.0662 | 0.1352 | 0.0915 | 0.1026 |
|  | 0.0831 | 0.0672 | 0.0817 | 0.0781 | 0.0956 |
|  | 0.0976 | 0.0819 | 0.1016 | 0.0685 | 0.0973 |
|  | 0.0817 | 0.0749 | 0.0968 | 0.0677 | 0.1039 |
|  | 0.0859 | 0.0649 | 0.1064 | 0.0697 | 0.1045 |
|  | 0.0735 | 0.0835 | 0.105 | 0.0764 |   |
|  | 0.0659 | 0.0725 |   | 0.0689 |   |
|  | 0.0923 |   |   |   |   |
|  | 0.0836 |   |   |   |   |

**Step 1: Calculate the Grand Mean**

The grand mean of a set of samples is the total of all the data values divided by the total sample size.



1. *Using the AAM data from above, calculate the grand mean:*

**Step 2: Calculate the Total Variation**

The total variation (not variance) is comprised the TOTAL sum of the squares of the differences of each mean with the grand mean.

\*\*There is the between group variation and the within group variation. The whole idea behind the analysis of variance is to compare the ratio of between group variance to within group variance. If the variance caused by the interaction between the samples is much larger when compared to the variance that appears within each group, then it is because the means aren't the same.



1. *Using the AAM data from above, calculate the total variation:*

**Step 3: Calculate the Between Group Variation**

The variation due to the interaction between the samples is denoted SS(B) for Sum of Squares Between groups. If the sample means are close to each other (and therefore the Grand Mean) this will be small. There are k samples involved with one data value for each sample (the sample mean), so there are k-1 degrees of freedom.



1. *Using the AAM data from above, calculate SS (B) :*

The variance due to the interaction between the samples is denoted MS(B) for Mean Square Between groups. This is the between group variation divided by its degrees of freedom. It is also denoted by. \*\*k is the number of population means being compared

1. *Using the AAM data from above, calculate MS (B):*

 SS(B)
MS(B) = ------- =
 k-1

**Step 4: Within Group Variation**

The variation due to differences within individual samples, denoted SS(W) for Sum of Squares Within groups. Each sample is considered independently, no interaction between samples is involved. Remember $s^{2}$ is the variance.



1. *Using the AAM data from above, calculate SS (W) :*

The degrees of freedom is equal to the sum of the individual degrees of freedom for each sample. Since each sample has degrees of freedom equal to one less than their sample sizes, and there are k samples, the total degrees of freedom is k less than the total sample size: df = N - k. The variance due to the differences within individual samples is denoted MS(W) for Mean Square Within groups. This is the within group variation divided by its degrees of freedom. It is also denoted by . It is the weighted average of the variances (weighted with the degrees of freedom).

1. *Using the AAM data from above, calculate MS (W) :*

MS(W) = SS(W)

 -----------

 N-k

**Step 5: Calculate the F test statistic**

The F test statistic is found by dividing the between group variance by the within group variance. The degrees of freedom for the numerator are the degrees of freedom for the between group (k-1) and the degrees of freedom for the denominator are the degrees of freedom for the within group (N-k).

Also notated as F= $\frac{MS Between}{MS Within}$

1. *Using the AAM data from above, calculate the F test statistic:*

F =

Summary Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SS | df | MS | F |
| Between | SS(B) | k-1 | SS(B)-----------k-1 | MS(B)--------------MS(W) |
| Within | SS(W) | N-k | SS(W)-----------N-k | . |
| Total | SS(W) + SS(B) | N-1 |  |  |

The conventional way of reporting the complete results of an anova is with a table (the "sum of squares" column is often omitted). Here are the results of a one-way anova on the AAM data. Did you get the

same results?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sum ofsquares | d.f. | meansquare | Fs |
| between groups | 0.00452 | 4 | 0.001113 | 7.12 |
| within groups | 0.00539 | 34 | 0.000159 |  |
| total | 0.00991 | 38 |  |  |

1. *Using the AMD data, calculate the p value:*

=F.DIST.RT (x, d.f.1, d.f.2)

If you're not going to use the mean squares for anything, you could just report this as "The means were significantly heterogeneous (one-way anova, F4, 34=7.12, *P*=2.8×10-4)." The degrees of freedom are given as a subscript to F, with the numerator first.

**Decision Rule**

If the decision is to reject the null, then at least one of the means is different. However, the anova does not tell you where the difference lies. For this, you need another test, either the Scheffe' or Tukey-Kramer test.

1. Tukeys or Sheffe’s post-hoc analysis

Online app and Tukey’s

<http://web.mst.edu/~psyworld/tukeyscalculator.htm>

 and Tukey’s significant/probability table

1. What did you find? I found that four of the combinations were significant at 0.05. Which four?

**Good news (finally):**

Excel include an "Analysis Toolpak," which includes an "Anova: Single Factor" function that will do a one-way anova. It does not include any techniques for unplanned comparisons of means, and it does not partition the variance. Try it now.

**On your own:**

Using the Animal Genome Size Database found at <http://www.genomesize.com/> select data on the genome size (measured in picograms of DNA per haploid cell or c value) of crustaceans. The cause of variation in genome size has been a puzzle for a long time; use these data to answer the biological question of whether some families of crustaceans have different genome sizes than others. Export to excel and group by class. Following the steps outlined above, including testing the assumptions, run a one way anova to look for differences amongst families.

1. *Enter your results in the table below:*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | sum ofsquares | d.f. | meansquare | Fs | P value |
| between groups |  |  |  |  |  |
| within groups |  |  |  |  |  |
| total |  |  |  |  |  |

1. *What do you conclude?*

Gregory, T.R. 2014. [Animal genome size database](http://www.genomesize.com/).

McDonald, J.H. 2014. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, Maryland.