

The analysis of variance (ANOVA)

Although the t-test is one of the most commonly used statistical hypothesis tests, it has limitations. The major limitation is that the t-test can be used to compare the means of only two groups at a time. Scientists often find that they need to compare the means of three or more groups. The statistical hypothesis test used to compare the means of three or more groups is the analysis of variance (ANOVA).

ANOVA: an example

ANOVA calculations are tedious enough that they are rarely done by hand, but there is no better way to really understand this particular statistical test. As an example, consider again the issue of the role of controlled fire in prairie restorations. One particularly contentious issue among restoration ecologists is the timing of prairie burns. Although natural fires may primarily have been sparked by late-summer lightning strikes, most controlled burns are done during the spring or fall. The timing of burning may strongly influence the outcome of prairie restorations because burns done at different times of year can favor dramatically different plant species. As scientists, you can collect data to help resolve such issues. For example, you could collect data to answer the following question: How does the timing of controlled burns influence the biomass of desirable prairie plant species? An example of the data you might collect is in Table 2, below.

Table 2. Total biomass (g) of *Rudbeckia hirta* (Black-eyed Susan) growing in each of 15, 0.5 m² plots. One third of the plots were burned in the spring, late summer, and fall of 1998, respectively. All plots were sampled during summer 1999.

Treatment		
Spring	Late Summer	Fall
0.10	5.56	3.85
0.61	6.97	3.01
1.91	3.01	2.13
2.99	5.33	2.50
1.06	3.53	6.10
Mean = 1.33	Mean = 4.88	Mean = 3.52

Before we go any further, we need to state null and alternative hypotheses. For example, the null hypothesis could be

H₀: The timing of controlled burning does not influence biomass of *Rudbeckia hirta*.

While the alternative hypothesis could read as follows:

H_A: The timing of controlled burns influences the biomass of *Rudbeckia hirta*.

ANOVA: calculations

To analyze these data using an ANOVA, you first need to calculate the *grand mean*. The grand mean is simply a fancy term for the mean of all of your observations, regardless of group. The grand mean can be calculated using the following formula:

$$\bar{Y} = \frac{1}{an} \sum_a \sum_n Y$$

Where a is the number of groups, n is the number of observations within each group, and Y is a single observation. Applying this formula to our prairie burn data, where $a = 3$ and $n = 5$, gives the following grand mean:

$$\bar{Y} = \frac{1}{15} (0.10 + 0.61 + 1.91 + 2.99 + 1.06 + 5.56 + \dots + 3.53 + 3.85 + 3.01 + 2.13 + 2.50 + 6.10) = 3.24$$

Once you have calculated the grand mean, you need to calculate the *sum of squares among groups* (SS_{among}). The SS_{among} is an estimate of the variation *among* your groups or, more precisely, an estimate of the deviation of the group means from the grand mean. The SS_{among} can be calculated using the following formula:

$$SS_{\text{among}} = n \sum_a (\bar{Y}_a - \bar{Y})^2$$

Where a is the number of groups, n is the number of observations within each group, and

$$\bar{Y}_a$$

is the mean of each group. Applying this formula to our prairie burn data, where $a = 3$ and $n = 5$, gives the following SS_{among} :

$$SS_{\text{among}} = 5((3.52 - 3.24)^2 + (1.33 - 3.24)^2 + (4.88 - 3.24)^2) = 32.08$$

The last thing you need to calculate is the *sum of squares within groups* (SS_{within}). The SS_{within} is an estimate of the variation among observations *within* groups, or, more precisely, an estimate of the deviation of the observations within each group from the group mean. The SS_{within} can be calculated using the following formula:

$$SS_{\text{within}} = \sum_a \sum_n (Y - \bar{Y}_a)^2$$

Where a is the number of groups, n is the number of observations within each group, Y is an individual observation, and

$$\bar{Y}_a$$

is the mean of each group. Applying this formula to our prairie burn data, where $a = 3$ and $n = 5$, gives the following SS_{within} :

$$SS_{\text{within}} = (0.10 - 1.33)^2 + (0.61 - 1.33)^2 + \dots + (2.50 - 3.52)^2 + (6.10 - 3.52)^2 = 25.55$$

ANOVA: the F-value

The test statistic for an ANOVA is called an **F-value**. The F -value for an ANOVA is calculated using the

following formula:

$$F = \frac{\left(\frac{SS_{\text{among}}}{a-1} \right)}{\left(\frac{SS_{\text{within}}}{a(n-1)} \right)}$$

Where a is the number of groups and n is the number of observations within each group. Applying this formula to our prairie burn data, where $a = 3$ and $n = 5$, gives the following F -value:

$$F = \frac{\left(\frac{32.08}{2} \right)}{\left(\frac{25.55}{12} \right)} = 7.57$$

If you look at the formula for the F -value, you can see that it is essentially the ratio of the variation *among* groups to the variation *within* groups. If the variation among groups is relatively large compared to the variation within groups, then the F -value will be relatively large. A relatively *large* F -value suggests that the variation among groups is largely caused by a given variable or experimental manipulation (in our example, the timing of burning), rather than chance variation. If the variation among groups is similar to the variation within groups, the F -value will be relatively small. A relatively *small* F -value suggests that the difference among groups is largely due to chance natural variation and measurement error, rather than to a given variable or manipulation. This interpretation of an F -value should sound familiar to you, because it is very similar to the interpretation of a t -statistic discussed above.

ANOVA: EXCEL output

For a data set of any size, an ANOVA is extremely tedious to calculate by hand. As such, you will be using EXCEL to do ANOVAs (see below). The EXCEL ANOVA output for our prairie burn study is inserted below:

ANOVA

Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	31.99836	2	15.99918	7.515351	0.007655	3.88529
Within Groups	25.5464	12	2.128867			
Total	57.54476	14				

By custom, the results of an ANOVA are expressed in the above format, which not surprisingly is called an *ANOVA table*. As you can see, this table contains the (now familiar to you) estimates of the SS_{among} , SS_{within} , and F -value (the values in this table are not identical to the values calculated earlier in this handout because of rounding error in the hand calculations). It also contains the associated ***P-value***. To review, a P -value ranges from 0 to 1, and is the probability of calculating a given test statistic assuming that the means of your groups are identical. The larger the test statistic is, the lower the chance (P) that that an observed difference among groups is due to chance environmental variation, and the greater the chance that a difference has biological causation. With a sample size of 15 observations, the probability that the differences we see among spring, late summer, and fall burns are due to chance alone is 0.007655. This value is bolded in the EXCEL output. That is, the p -value equals 0.007655. We can therefore reject the null hypothesis that burning season has no effect on *R. hirta* biomass, and accept the alternative

hypothesis that it burning season does affect *R. hirta* biomass. If the p-value for our *F*-value had been >0.05, then we would have accepted that null hypothesis.

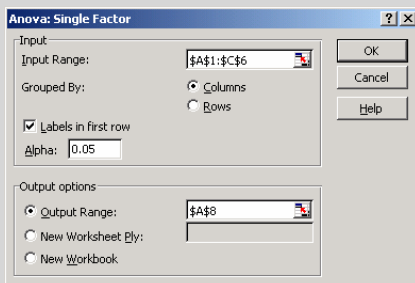
Note that the ANOVA does *not* tell you anything about pairwise differences between groups, only if there are *overall* differences among all groups. There are procedures that can be done after an ANOVA that test for pairwise differences between groups, but they are beyond the scope of this handbook.

How to do an ANOVA in EXCEL

1. Enter your data so that each group is in a separate column. Label your columns appropriately. For example, in our prairie burn example, the columns could be labeled "spring", "late summer", and "fall".

1	Treatment		
2	Spring	Late Summer	Fall
3	0.1	5.56	3.85
4	0.61	6.97	3.01
5	1.91	3.01	2.13
6	2.99	5.33	2.5
7	1.06	3.53	6.1

2. Click on Tools > Data analysis > Anova: single factor. If the Data analysis module is not available in the Tools menu, click on Add-ins and install the Analysis ToolPak.



3. Input the range of your data by highlighting all of your data (including labels).

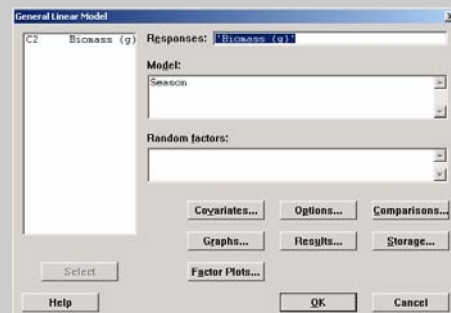
Click Labels, then click OK. The ANOVA output should appear on the screen

How to do an ANOVA in Minitab

1. Enter your data so that one labeled column contains the grouping variable (e.g., for the prairie burn example, the entries in this column could be "spring", "summer", and "fall") and a second labeled column contains the dependent variable.

	C1-T	C2
	Season	Biomass (g)
1	spring	0.10
2	spring	0.61
3	late summer	5.56
4	late summer	6.97
5	fall	3.85
6	fall	3.01
7		

2. Select Stat > ANOVA > general linear model.
3. In the dialog window, place the dependent variable in the response box. The dependent variable is what you measure. In this case biomass is the dependent variable. Place the grouping variable in the model box. The grouping variable is the same as the independent variable or treatment, so in this example it would be season.



4. Click OK to view the ANOVA output.

Example problems

Carry out ANOVAs for the following data sets:

1. Three different methods were used to measure dissolved oxygen (mg/kg) content of lake water. (Data from JH Zar's *Biostatistical Analysis*, 4th ed.)

Method		
1	2	3
10.96	10.88	10.73
10.77	10.75	10.79
10.90	10.80	10.78
10.69	10.81	10.82
10.87	10.70	10.88
10.60	10.82	10.81

2. Number of eggs laid per female per day for females from each of three lines of *Drosophila melanogaster*. The RS and SS lines were selected for resistance and susceptibility to DDT. The NS line is the nonselected control. (Data from RR Sokal and FJ Rohlf's *Biometry*, 3rd ed.)

Line		
RS	SS	NS
12.8	38.4	35.4
21.6	32.9	27.4
14.8	48.5	19.3
23.1	20.9	41.8
34.6	11.6	20.3
19.7	22.3	37.6
22.6	30.2	36.9
29.6	33.4	37.3

3. Strontium concentrations (mg/ml) in three different bodies of water. (Data from JH Zar's *Biostatistical Analysis*, 4th ed.)

Grayson's Pond	Beaver Lake	Angler's Cove
28.2	39.6	46.3
33.2	40.8	42.1
36.4	37.9	43.5
34.6	37.1	48.8
29.1	43.6	43.7
31.0	42.4	40.1