Demo 42: Analysis of Variance

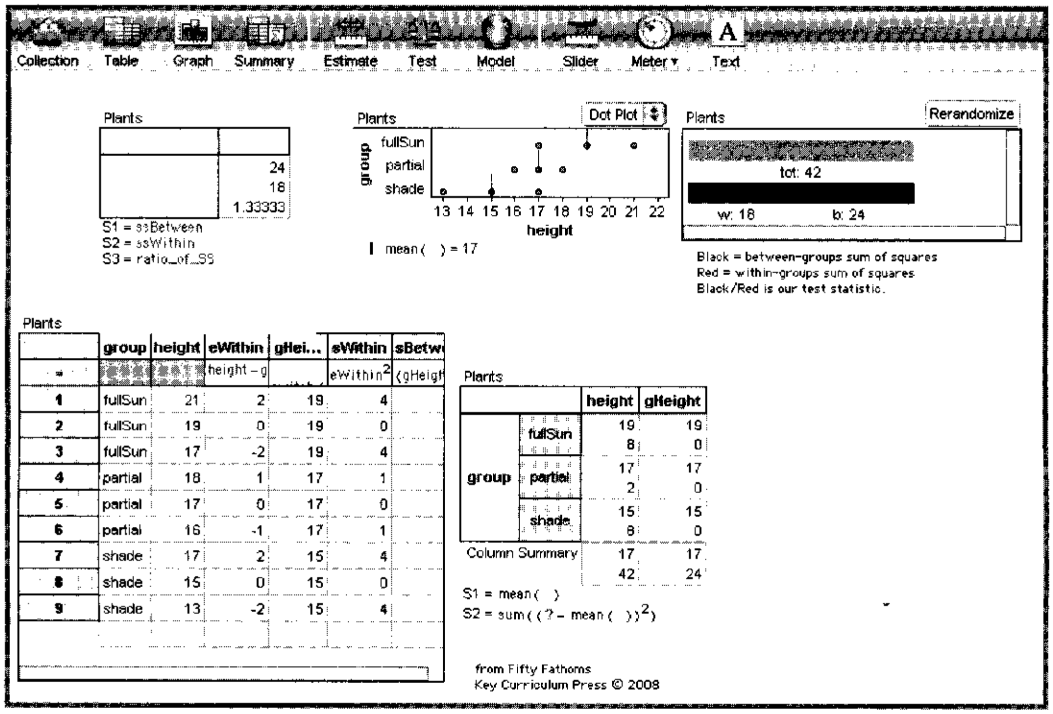
Assessing whether means are different in different groups • Introduction to ANOVA

This is an advanced topic for an introductory course, and the file for this demo is complicated, but the basic idea is simple. Suppose you're growing plants. Some are in the **sun** and some are in the **shade,** and you measure their **height.** The height is a *response* variable; we wonder if it depends on the *treatments,* **sun** and **shade.** We could use a t-test to see if the mean of the **sun** group was significantly different from the **shade** group. But what if there were *three* groups instead of two? How would you calculate t?

We need to define a new measure for how different the groups are. We'll compare the new measure's test value to its distribution under the null hypothesis—where there is no dependence at all. This is exactly what we've done with Student's t and chi-square and any other statistic we use to construct a test. We'll construct the measure using the same reasoning we used for t. There, we measured the difference *between* the groups in units of standard errors—in units of the variation *within* the groups. Generalizing this idea to more than two groups is tricky, but the idea is the same: We figure out how much of the spread is due to variation *within* the groups and how much is due to variation *between* the groups, and we make a ratio:

$$dependence = \frac{\text{variation between groups}}{\text{variation within groups}}$$

In that way, a large dependence will yield a big number—when there is little variation within groups compared to the variation between groups. And if the groups are essentially identical, the between-group variation will be small, and our statistics will be close to zero. The main thing to bear in mind from the beginning is that we will measure variation using *sums of squares* of deviations from means; that is, we'll look at numbers that are more like variances than like standard deviations.



Plants

	24 18 1.33333

S1 = ssBetween
S2 = ssWithin
S3 = ratio_of_SS

mean( ) = 17

Black = between-groups sum of squares
Red = within-groups sum of squares
Black/Red is our test statistic.

Plants

	group	height	eWithin	gHei...	sWithin	sBetw
			height – g		eWithin^2	(gHeig
1	fullSun	21	2	19	4	
2	fullSun	19	0	19	0	
3	fullSun	17	-2	19	4	
4	partial	18	1	17	1	
5	partial	17	0	17	0	
6	partial	16	-1	17	1	
7	shade	17	2	15	4	
8	shade	15	0	15	0	
9	shade	13	-2	15	4	

Plants

		height	gHeight
group	fullSun	19 8	19 0
	partial	17 2	17 0
	shade	15 8	15 0
Column Summary		17 42	17 24

S1 = mean( )
S2 = sum((? – mean( ))^2)

from Fifty Fathoms
Key Curriculum Press © 2008

What To Do

▹ Open **Within and Between.ftm.** It will look like the illustration.

This is all about (imaginary) plant data. The case table at lower left is a good place to start: There you see the raw data—the **group** (**fullSun, shade,** or **partial**) and **height** of each of nine plants. You also see four more attributes, calculated from the original two:

- **gHeight** is the group height—the mean of the heights of all the plants in the group. This has the same value for every plant in that group, as you can see.
- **eWithin** is the "error" of the plant's height "within the group." That is, you can think of every **height** as being **gHeight + eWithin.**
- **sWithin** is **eWithin** squared.
- **sBetween** is the squared residual of that plant's *group* (not the individual plant) within the whole data set. That is, it's the square of **gHeight – mean(height).** This will define what we mean by *between-group variation*: It's the variance of the set of groups, weighted by the number of cases in each group.

Directly above the case table is a summary table showing the sum of squares of the within-group residuals (**ssWithin**), which is simply the sum of all the numbers in the **sWithin** column. Similarly for **ssBetween.** Then **ratio_of_SS** is the ratio we talked about earlier, computed by taking **ssBetween / ssWithin.** So the value of this "dependence" measure here is 1.33; we don't know yet if that is a lot or not.

At right is a graph showing all the data; below the graph is another summary table, this time showing statistics for the **height** and **gHeight** attributes for the groups: the mean values, and the sum of the squares of their residuals *in the context of the cell they're in.*

This is a subtle but important table. At the bottom of the **height** column in that table, you can see the total sum of squares of the residuals (42) and the grand mean (17). Within the left column, you see the sums of squares of the residuals within each group (8, 2, and 8, for a total of 18). Finally, at lower right—in the column summary for **gHeight**—is yet another sum of squares of the residuals. This time, each residual is the distance of the mean of the *group* from the total mean; so this is the *between*-group sum of squares, 24. These numbers demonstrate an important identity:

$$SS_{total} = SS_{within} + SS_{between}$$

since 42 = 18 + 24. In the upper right is a display that shows how the total sum of squares (the green bar) is divided into the within and between parts (red and black).

▹ In the graph at the top, grab any point and drag it. Study its effect on the displays and numbers. Undo and try it with a different point. See if you can understand why things change the way they do.

⇨ If the red and green display does not update, click the **Rerandomize** button.

▹ Now try to grab a point and move it in a direction that will make the differences among the groups more pronounced. (For example, move the **fullSun** point at **height = 17** to the right.) Verify that the **ratio_of_SS** value increases (see the upper-left summary table) and that in the bar display, there is more and more black compared to red.

▹ Repeat, exploring how the relationships among the groups affect the **ratio_of_SS** statistic and the size of its two components, **ssWithin** and **ssBetween.**

Questions

1 In the lower summary table, the top number in each box is the same in the left column as in the right. Why?

2 In the lower summary table, the right-hand column has a lot of zeros in it. Why are those numbers zero? **Sol**

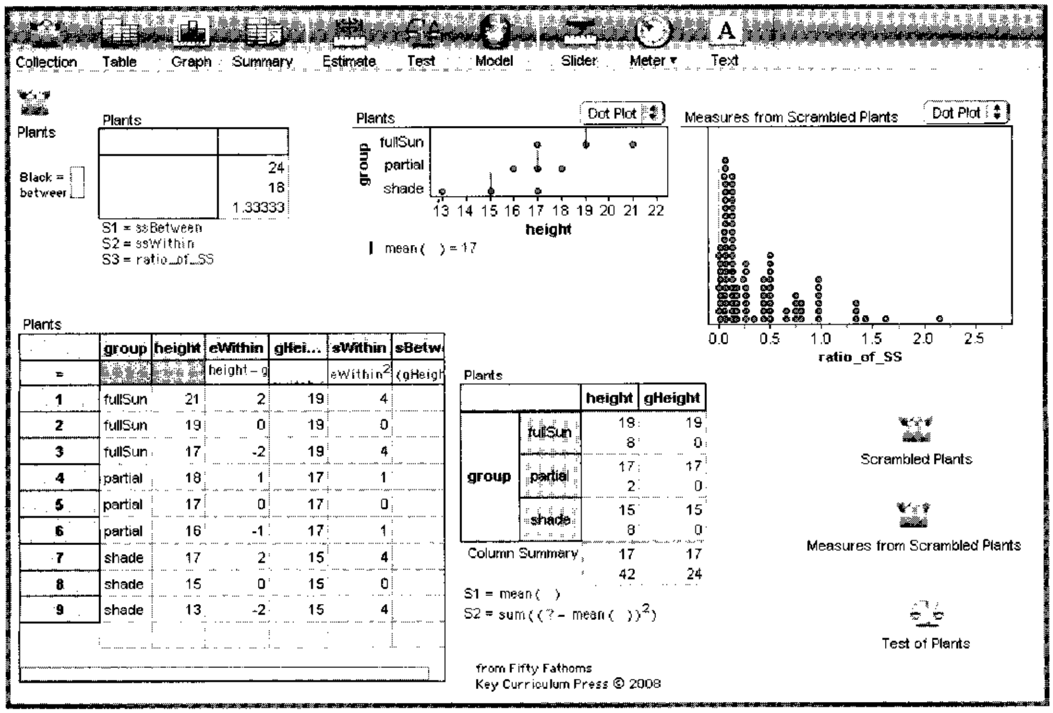
Onward!

Now let's figure out if our value for **ratio_of_SS** is unusual, or if it could arise easily by chance. We'll use the same strategy we used in Demo 37, "Scrambling to Compare Means"—we want to see what happens in the case of the null hypothesis. So we will make the null hypothesis true: We will scramble the values in the **group** attribute so that any apparent relationship between **group** and **height** is due to chance alone. Then we compute **ratio_of_SS** for those data, and repeat the process, building up a distribution of the **ratio_of_SS** statistic for the case that the null hypothesis is true.

- ▹ Put everything back the way it was.
- ▹ To save space, shrink the upper-right bar display until it turns into an icon (**plants**). Then move it to the upper-left (empty) corner of the window. Do the same with the text below it.
- ▹ Finally, choose **Show Hidden Objects** from the **Object** menu. The file should now look something like the illustration below.

Now you can see the sampling distribution on the right. It has 100 points. You can see that 6 of the 100 points have values of **ratio_of_SS** equal to or greater than our test statistic, 1.33. Thus, 0.06 is our empirical P-value.

- ▹ Drag data in the **plants** graph to make the statistic bigger and the groups more distinct.
- ▹ Select the lower-right collection, **Measures from Scrambled Plants**, by clicking on it once. Then choose **Collect More Measures** from the **Collection** menu. Fathom will construct a new graph.
- ▹ Compare the new test statistic to the new graph; repeat as necessary.



Plants

	24 18 1.33333

S1 = ssBetween
S2 = ssWithin
S3 = ratio_of_SS

Plants

	group	height	eWithin	gHei…	sWithin	sBetw
=			height – g		eWithin²	(gHeigh
1	fullSun	21	2	19	4	
2	fullSun	19	0	19	0	
3	fullSun	17	-2	19	4	
4	partial	18	1	17	1	
5	partial	17	0	17	0	
6	partial	16	-1	17	1	
7	shade	17	2	15	4	
8	shade	15	0	15	0	
9	shade	13	-2	15	4	

Plants

		height	gHeight
group	fullSun	19 8	19 0
	partial	17 2	17 0
	shade	15 8	15 0
Column Summary		17 42	17 24

S1 = mean( )
S2 = sum((? – mean( ))^2)

from Fifty Fathoms
Key Curriculum Press © 2008

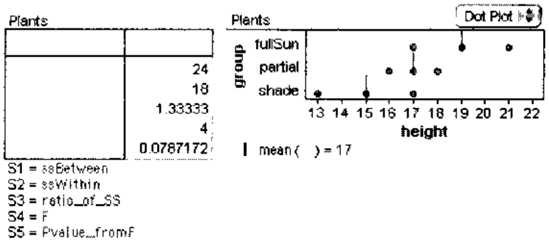
Extension: Using the *F*-Statistic

The ratio we used is similar to the traditional statistic that is used in *Analysis of Variance* (ANOVA): the *F*-statistic. We used the ratio of sums of squares; the *F*-statistic actually looks at the ratio of the variances. To get the variance—the mean square deviation—from the sum of the squares of the deviation, you have to divide by the number of *degrees of freedom*[1] (not *n*, because we are trying to infer from a sample). Let's look at *F* and see what happens:

▹ Click the top summary table to select it. Then choose **Add Formula** from the **Summary** menu. The formula editor appears.

▹ Enter the one-letter formula **F.** (We have already defined it as a measure; you may look at its formula if you like; it's a measure in **plants.**) Close the formula editor with **OK.**

▹ Now add another formula to the table, **pValue_fromF.** This is the *P*-value from an *F*-test. You may need to stretch the summary table to see all the values; they will look something like the illustration.



Plants

	24
	18
	1.33333
	4
	0.0787172

S1 = ssBetween
S2 = ssWithin
S3 = ratio_of_SS
S4 = F
S5 = Pvalue_fromF

(Here, *P* is about 0.08; if there were no relationship between **group** and **height**, we'd see this value of *F* or greater 8% of the time.)

▹ Finally, open up the inspector for the lower-right measures collection, go to the **Cases** panel, and drag **F** to the horizontal axis, replacing **ratio_of_SS.** Now you can compare the test value for **F** with the distribution.

▹ One more thing: There is a *test* in the lower-right corner. Drag it into the middle and expand it; it's a Fathom ANOVA test, which does all that we have just done in a simple display.

More Questions

3 If the null hypothesis were true, which would usually be bigger—the within-group or the between-group variance?

4 What value for **F** gives a *P*-value of about 0.05? (that is, What is the critical value for *F* at the 0.05 level?) **Sol**

5 In our example, our three groups' means are more or less evenly spaced. Is the *F*-statistic you get larger than, smaller than, or the same as if you take the one in the middle and move it to the end—so that we'd have two groups about the same and one different?

Challenges

6 All we're doing here is comparing means. Why can't we just use a *t*-test?

7 We have blithely stated—and shown empirically—that the sum of the squares of the within-group residuals, plus the sum of the squares of the between-group residuals, equals the sum of the squares of the "total" residuals (the distances of each data value from the mean of the entire data set). That is,

$$SS_{total} = SS_{within} + SS_{between}$$

Prove it.

8 Plot **F** against **ratio_of_SS.** Explain the graph you see. **Sol**

[1] The number of degrees of freedom—*df*—is a lot like the $n - 1$ we divide by to get a sample SD. So, since there are three groups, the *df* for the "between" is 2. For within-groups, we get *df* of 2 for each group, for a total of 6. Notice that the total *df* is 8—one less than the number of points.