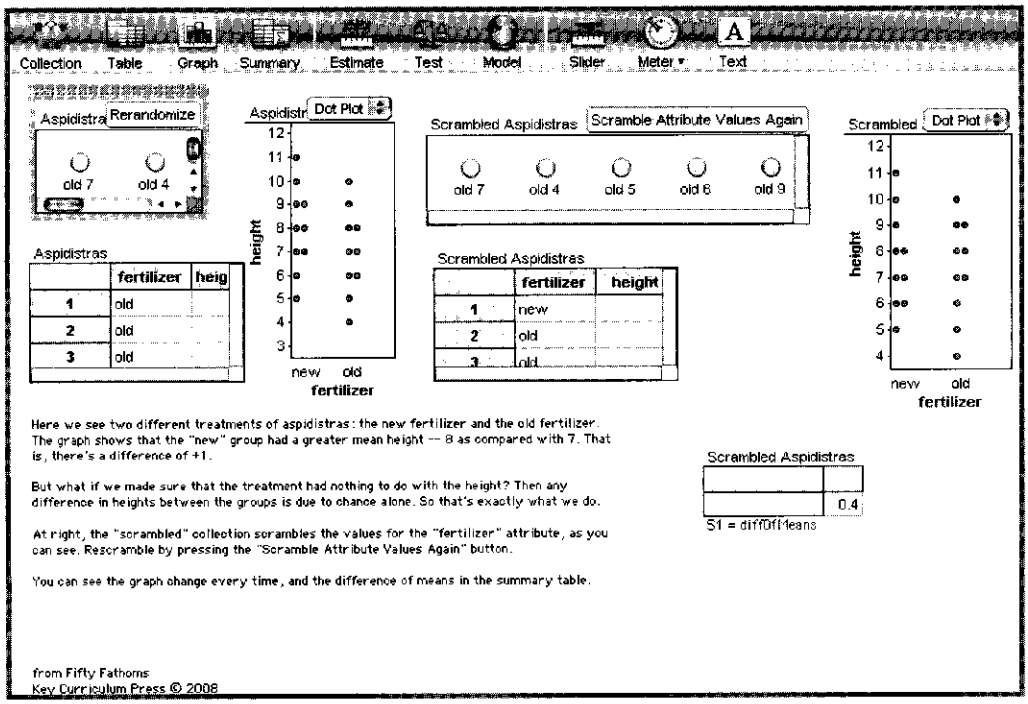
Demo 37: Scrambling to Compare Means

Randomization test • Using scrambling to simulate the null hypothesis • Generating a sampling distribution

Suppose we want to test a new fertilizer. We make an experiment in which we assign aspidistras to two groups randomly: The experimental group gets the new fertilizer and the control group gets the old one. Four weeks later, we measure the plants and compare the two groups. If the average experimental aspidistra is taller, the fertilizer is good, right?

It's not so easy. Even if we treated the plants identically, there would be some variation, and if we assigned the plants to different groups—and treated the groups the same—one group would be taller on the average than the other. So the question becomes, If the fertilizer didn't do anything (we call this the *null hypothesis*), how likely is it that the intergroup difference we measured—or a larger difference—would arise by chance? If it is very unlikely, we conclude that the fertilizer did indeed make a difference. Often, researchers define "very unlikely" as being a probability of less than 5%, but, as with so many things in statistics, you can decide for yourself what probability you want to use.

In this demo, we'll test the fertilizer using *scrambling*—officially known as a randomization test. The idea is to simulate the null hypothesis, that is, to alter the data so that any difference between the experimental group and the control is due only to chance. We'll do this repeatedly to make a *sampling distribution* of the statistic we need to look at; we compare our *test statistic* to the sampling distribution to decide whether to reject the null hypothesis.



What To Do

▷ Open **Scrambling.ftm.** It should look like the illustration.

The left side of the screen has the original (made-up) data and a graph showing the difference between the two groups. The right-hand side shows the data with

the **fertilizer** attribute scrambled. You can see this by comparing the gold balls or the case tables.

In the summary table, you can see a value for **diffOfMeans**. For this case, it's −0.2. This is the difference between the means of the experimental and control groups *in the scrambled collection*. In our original data, the difference of means is the *test statistic,* and it has a value of +1.0. Speaking symbolically, the formula is

diffOfMeans
= mean(height, fertilizer = "new")
− mean(height, fertilizer = "old")

▷ Press the **Scramble Attribute Values Again** button and see what happens. Do so repeatedly.

Questions

1 How can you tell that the **fertilizer** attribute, not **height**, was scrambled?

2 Looking informally at the summary table for **diffOfMeans**, what's the range of values you get for the scrambled statistic?

3 Is our test statistic of +1.0 particularly unusual, or is it the sort of thing that arises by chance? **Sol**

Moving On

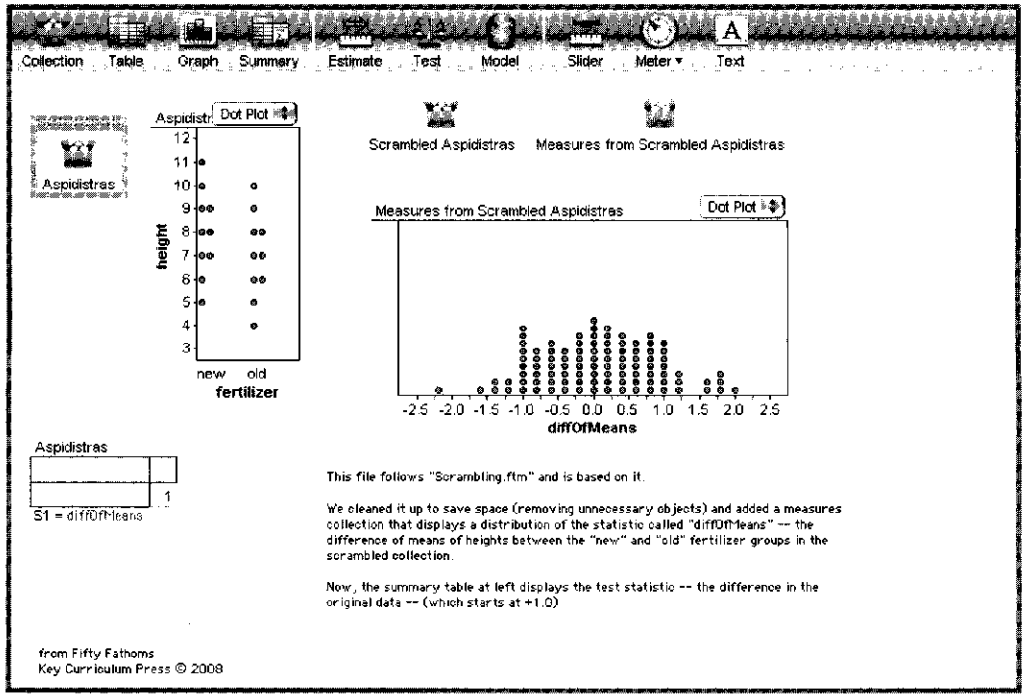
We looked at those **diffOfMeans**s informally. Let's use Fathom to record those values and look more closely at the distribution of that statistic.

▷ Open **Scrambling 2.ftm**. It should look like the illustration below.

This file compresses elements of the previous one and adds some "measures" machinery. The **Measures from Scrambled Aspidistras** collection contains 100 values from rescramblings, which are displayed in the graph. Note that the summary table here refers to the original data, not the scrambled collection. Thus, it shows the test statistic (+1.0) that we compare to the sampling distribution in the graph.

▷ Count how many cases in the distribution have values for **diffOfMeans** that are +1.0 or above.

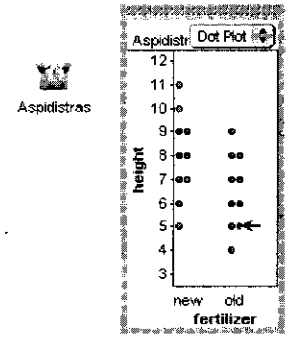
▷ Re-collect measures: Click the measures collection once to select it, then choose **Collect More Measures** from the **Collection** menu. Fathom will collect 100 more measures, rescrambling the scrambled collection every time and recomputing the statistic (that's why it takes a while). Once again see how many are +1.0 or above. Repeat as many times as you need to in order to understand what's going on.



▷ Estimate the probability that a scrambled collection has a **diffOfMeans** of +1.0 or more. Is it more than 5%?

It seems as though the two groups aren't different enough in the mean to be distinguished from chance. Let's change the original data and see if we can make it less likely for our test statistic to arise by chance:

▷ In the graph of the original data, drag down the top case in the **old** fertilizer group (value of **height = 10**) to a value of 5, as shown. The test statistic in the summary table should increase (we're increasing the difference of means) from +1.0 to +1.5.



▷ Again re-collect measures. Select the measures collection and choose **Collect More Measures** from the **Collection** menu.

▷ Now count: How many cases equal or exceed the test statistic? Note: The test statistic is now 1.5, not 1.0—so there are fewer cases than you might think.

You'll probably find that it is now less likely—but still possible—to get a **diffOfMeans** as large as 1.5. Let's make it even less likely.

▷ Repeat the above experiment, gradually altering data to make the groups more different. You might lower the same point or raise a point in the other group. Continue until the test statistic seldom occurs.

Challenges

4 In this randomization test, the *P*-value is the probability that a scrambled **diffOfMeans** is greater than or equal to the test statistic. So, if there were 7 cases out of 100 at +1.0 or greater, the *P*-value would be 0.07. At what *P*-value are you convinced that the group means really are different?

5 What result would you get if the original data had disjoint groups—that is, the highest value for one group was lower than the lowest value for the other? **Sol**

6 Explain why the logic is always "greater than or equal to" the test statistic. Why not only greater than? Why not only equal to?

7 Explain the logic of this test. What is the test statistic? Why do we scramble? What is that distribution a distribution *of*? What does it mean when we compare the test statistic to the distribution? What possible results can there be?

8 Decide whether this test should really be one-tailed or two-tailed, and explain your reasoning. That is, when we compare a test statistic of +1.0 to a distribution of **diffOfMeans**s, should we count the cases where **diffOfMeans** is greater than +1.0 or where the *absolute value* of **diffOfMeans** is greater than +1.0?

What You Should Take Away

Scrambling is genuinely useful for doing statistical inference—in this case, deciding whether the two means are different—especially if the data don't fit traditional requirements of normality. But more important for learning is that scrambling can show you what the test is really all about: comparing a test statistic to a sampling distribution. And that distribution comes from a process (scrambling) that guarantees the null hypothesis is true.