

Inference for Experiments

Corey Andreasen

Inference in AP Statistics is generally introduced in the context of random samples from a large population. We are testing against (or trying to estimate) a particular population parameter. However, many questions about inference occur in the context of randomized experiments. Because an experiment involves a comparison of at least two treatments, our inference procedure in AP Statistics is generally for a Difference of Means, a Difference of Proportions, or in some cases a Chi-square test.

The purpose of this paper is not to discuss the meaning of hypothesis tests or confidence intervals in general; rather it is to see how the questions asked and, hence, the questions answered, are different for randomized experiments than for random sampling situations.

I am, correctly or not, assuming the reader has a passing familiarity with the inference procedures used in AP Statistics in the context of comparing proportions and comparing means of random samples from two populations. I summarize a bit of the reasoning here for comparison.

The inference procedures we use in AP Statistics are based on knowledge of sampling distributions. For example, say we select a simple random sample from each of two large binomial populations such that the probability of success is the same for each selection (the selections are made independently), and the sample sizes are quite large, then we subtract the proportion of successes in the first sample from the proportion of successes in the second sample. If we repeat this process many times, the distribution of the differences will be approximately normal with mean $p_1 - p_2$, where p_1 and p_2 are the proportions of successes in population 1 and population 2, respectively. The standard deviation of this distribution will be $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. Your textbook probably has a nice explanation of this.

So when we do inference, we first need to see if the situation above applies and, thus, whether we can use what we know about this sampling distribution. We check some conditions:

- We need two simple random samples taken independently from two populations.
- Each population is at least 10 times as large as its sample size (to justify treating the selections as independent)
- $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$, the number of successes and the number of failures in each sample, must all be at least 5 (or 10, depending on the textbook). This makes sure the samples are large enough to justify the use of a normal model.

If these conditions are met, we can carry out the inference procedure and make statements about the difference in proportions $p_1 - p_2$, where p_1 and p_2 are the proportions of successes in population 1 and population 2, respectively.

The problem is, with a randomized experiment, we have *one* population (the set of experimental units available for the study) from which the two ‘samples’ (treatment groups) are randomly but *not* independently selected. In fact, they are completely dependent because the selection of one group determines the other! Not only that, each ‘sample’ is frequently about *one-half* the size of the population. With such gross violations of the conditions, how can we justify using a procedure based on them at all?

Let’s consider an example. Suppose a student in your AP Statistics class believes that carrying a piece of moonstone -- a kind of stone purported to bring luck, love, and healing to its bearer -- actually works. He is convinced that moonstone is a lucky stone. A skeptic in the class proposes a study to look for evidence of this ‘luck effect.’ She suggests that, because nobody in the class of 24 students yet has a date for the upcoming dance, some students should carry moonstone in their pockets and see if those carrying moonstone are more likely to get a date. (Let’s assume that everyone would, in fact, like to have a date so that getting one would be considered good fortune.)

The class has a discussion and designs the following experiment. 24 identical pouches are prepared, and a piece of moonstone is placed into 12 of them. Small stones of similar size, found on the street, are placed into the other 12 pouches. The pouches are placed in a large bag, the bag is mixed, and each student selects one pouch. They are to carry the stone over the next two weeks leading up to the dance, but not look at it (they are not to know which stone they had). At the dance they will report whether they have a date to the teacher, who will be a chaperone for the dance. The students will also give the teacher their pouch to identify which type of stone they carried. Students are banned from taking another student from the class as a date because that might bias the results.

So, what would be the appropriate null hypothesis?

In this case, the null hypothesis would be that the stone had no effect. That means the chance of getting a date would have been the same regardless of which stone they carried. Because the student claims moonstone would *improve* the chance of getting a date, an appropriate alternative hypothesis would be that the proportion of students who get a date would be higher for those carrying moonstone than for those carrying an ordinary stone. Under the null hypothesis, the observed difference between the two groups should produce a value that can be reasonably attributed to the random assignment alone; under the alternative hypothesis, the observed difference should be larger than could be reasonably attributed to the random assignment alone.

Let us suppose that of the 12 students carrying moonstone, 9 got date, and of the others, 6 got a date. To determine whether this demonstrates a true ‘luck effect’ we need to do an inference procedure. This means we need to know what the distribution of possible differences looks like.

The following activity simulates this situation for a true null hypothesis, which implies that the students would have had the same result regardless of the type of stone they carried. 15 students would have found dates and 9 would not. Take 24 index cards to represent test results. On 15 of these, write “Date” and on the other 9 of these, write “No Date”.

Shuffle the cards and deal two piles of 12 to represent the students who were randomly selected to receive each treatment (type of stone). Use the results to fill in a table like the one below.

	Moonstone	Ordinary Stone	Total
Date			15
No Date			9
Total	12	12	24

For the results described for this class of students, the table would look like this:

	Moonstone	Ordinary Stone	Total
Date	9	6	15
No Date	3	6	9
Total	12	12	24

The difference in proportions for those getting a date for each group was:

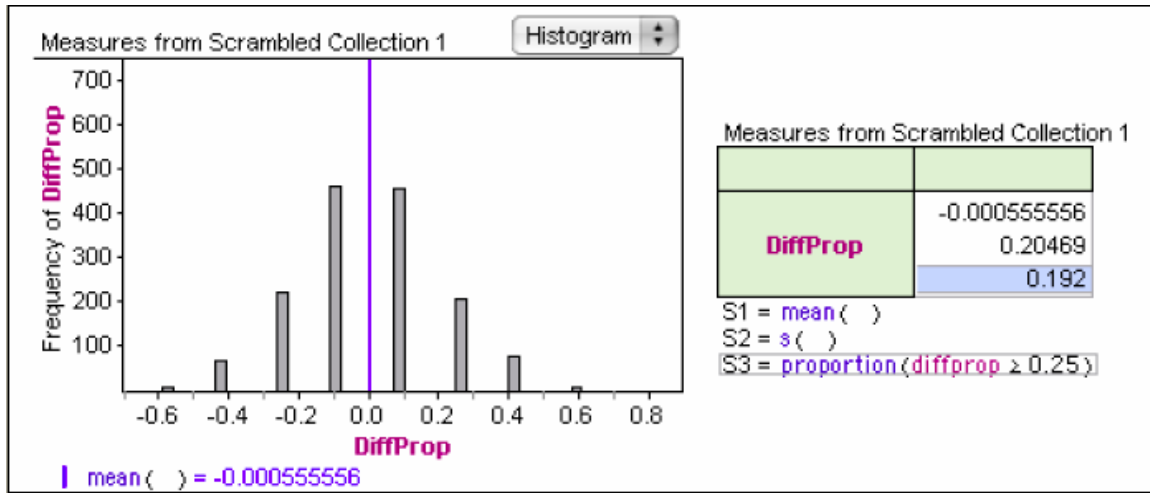
$$\frac{9}{12} - \frac{6}{12} = \frac{3}{12} = 0.25$$

Because the null hypothesis is that the type of stone has no effect, we assume the students’ scores would not change if they got a different treatment. So pick up the cards, shuffle, and redeal them. Again, fill in the table and record the difference in proportions. Repeat this several times and create a dotplot of the differences in proportions.

Next you may wish to have students do some runs of the simulation on their own. Each student would need 24 cards, 15 labeled “Date” and 9 labeled “No Date”. Or you may wish to have them use plastic chips or beads, 15 of one color to represent “Date” and 9 of a second color to represent “No Date”. Have them mix the objects then draw out 12 to represent the students who would get moonstone and the remaining 12 would represent the students who get an ordinary stone.

This physical model is important to help students understand the randomization process and reinforce the idea that, if there is no difference in the effects of the two types of stone, we can randomly assign the treatment (type of stone) leaving the response variable (result of dating quest) unchanged. However, to do enough simulations to get a meaningful simulated randomization distribution (Because we are repeatedly randomizing the assignment of treatments only, this should be called a randomization

distribution rather than a sampling distribution.), it is helpful to use technology. But it is essential to start with the physical simulation so students can see what it is that the technology is emulating. Fathom works nicely for this. A graphing calculator can be programmed to carry out this simulation, but it works rather slowly. **See Appendices A and B for instructions for carrying this out in Fathom and the TI-84+, respectively.**



Randomization distribution for the difference in proportions if the moonstone and ordinary stone are equally effective

Note that the symmetric, mound-shaped distribution is centered near 0, as would be expected if the two types of stones are equally effective, and that the standard deviation of the distribution is about 0.20.

Also, because of the small groups, the plot above is very discrete – not all that close to normal. But how well does the approximation work?

The standard deviation given by the formula for the difference between two sample proportions is shown below. Define \hat{p}_M and \hat{p}_O as the observed proportions of students with moonstone and ordinary stone, respectively, that found a date. Then p is the pooled observed proportion $\frac{n_M \cdot \hat{p}_M + n_O \cdot \hat{p}_O}{n_M + n_O} = \frac{9 + 6}{12 + 12} = 0.625$ based on our hypothesis that the two expected proportions are equal.

$$\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.625 \cdot 0.375\left(\frac{1}{12} + \frac{1}{12}\right)} \approx 0.20$$

This is very close to the standard deviation of the randomization distribution, which suggests we can use the same calculations for this experiment that we would have used in the very different context of two independent samples. This was, in fact, shown to be true in general by R. A. Fisher. We have a randomization distribution that is approximately normal, with mean at $p_1 - p_2 = 0$, and standard deviation given by the

formula $\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$. However, the conditions and interpretation are a bit different.

The conditions:

The example above did not meet the conditions for inference for two independent samples. But the distribution has the same characteristics under the method we used to create this randomization distribution. The conditions were simply:

- The treatments must be randomly assigned to subjects
- The number of successes and number of failures under each treatment must be at least 5 (or 10, depending on the textbook).

In our example, the difference in proportions was 0.25. Our simulation showed that about 19.2% of the randomizations resulted in a difference of 0.25 or more, so the p -value based on the randomization test is 0.192. Using the techniques based on the normal distribution, we would calculate the test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{0.625 \cdot 0.375}{12} + \frac{0.625 \cdot 0.375}{12}}} \approx \frac{0.25 - 0}{0.20} = 1.25$$

P -value = $P(z \geq 1.25) \approx 0.11$

(The rather large discrepancy is due to the fact that the randomization distribution was so discrete. Understand that this randomization procedure is a more accurate procedure than the approximation method shown here, which is based on an assumption of normality. With larger experimental groups the gaps in the dotplot become smaller and the normal model more closely approximates the randomization distribution.)

The interpretation:

To interpret this P -value in the context of this experiment, recall how the randomization distribution was created. Randomization alone accounted for the variation in the differences in proportions. The P -value, then, is **the probability that $\hat{p}_1 - \hat{p}_2 \geq 0.25$ if the moonstone and the ordinary stone are equally effective, where \hat{p}_1 is the observed proportion of students with moonstone who got a date and \hat{p}_2 is the observed proportion of students with an ordinary stone who got a date.**

Notice that we *cannot* generalize to a larger population unless our experimental units are themselves a random sample from that larger population. That is generally not the case, and it is certainly not the case in the example here. Any generalization beyond the experimental group is made, not based on statistical reasoning, but on the researcher's judgment. The experimenter decides that the experimental group is enough like the rest of the population to justify such a generalization, but this can be risky.

In our example, the P -value (using either method) is too high to provide evidence that the moonstone is 'lucky.' We would not reject the null hypothesis. We do not have sufficient evidence that the observed difference in dating success is due to anything other than the random assignment of the type of stone.

Confidence Interval Estimates

For a confidence interval, it might be better to look at this in a slightly different way. If the two types of stone are equally effective, the proportion of students getting a date had all students used the moonstone would be the same as the proportion of students getting a date had all students used the ordinary stone. The expected difference in the observed proportions for the two treatments would be 0. But if the stones were not equally effective, the expected difference in the proportions would be different from 0. The confidence interval answers the question *What is the range of plausible values for the expected difference in proportions if all students could have used both treatments?*

Of course, all students cannot use both treatments, which is why we need to estimate the expected difference. Our best point estimate of the expected difference in proportions is the observed difference from our experiment, 0.25.

We saw a randomization distribution of the difference in proportions based on a randomized experiment. It was about the same as a sampling distribution for the difference in proportions for two independent samples. This is true in general and allows us to use the same standard deviation formula we would use for a confidence interval from a two-sample situation, this gives us the 95% confidence interval

$$95\% \text{ CI} = (\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{\frac{\hat{p}_1 \cdot (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \cdot (1 - \hat{p}_2)}{n_2}} = 0.25 \pm 1.96 \cdot 0.191 \approx (-0.124, 0.624)$$

Here \hat{p}_1 is the *observed* proportion of successes for students carrying moonstone. It estimates the corresponding 'population' parameter p_1 , which is the proportion of successes if *all* students in the class had carried moonstone.

We would interpret this interval as follows:

We are 95% confident that the expected difference in proportions $p_1 - p_2$ is between -0.124 and 0.624, where p_1 is the proportion of students that would have found a date for the dance had all students carried a moonstone and p_2 is the proportion of students that would have found a date for the dance had all students carried an ordinary stone.

Difference of Means

A student suggests an experiment to determine whether moonstone would help them perform better on a test, and the class designs the following experiment. The same 24 identical pouches are prepared, 12 with a piece of moonstone and 12 with an ordinary stone. The pouches are placed in a large bag and as the students walk into the room on

test day they will reach into the bag, select a pouch, and put it in their pocket. When they turn in their exam, they will also turn in their pouch so that the type of stone they carried can be identified.

The familiar conditions for a test (or interval) for a Difference of Means of two sample proportions are:

- Two samples are randomly and independently selected from two separate populations.
- Either the two samples appear consistent with a normal population or the samples are large enough that the sampling distribution of the mean will be approximately normal.
- The population size should be at least ten times as large as the sample (unless a correction factor, called the finite population factor, is used).

Of course, here we have a randomized experiment. Let's see how this works.

The scores and assignment of stones turned out as follows:

Student	Type of Stone	Score
1	Moonstone	98
2	Moonstone	80
3	Moonstone	98
4	Moonstone	74
5	Ordinary Stone	87
6	Moonstone	64
7	Ordinary Stone	71
8	Moonstone	90
9	Ordinary Stone	51
10	Ordinary Stone	76
11	Ordinary Stone	68
12	Ordinary Stone	60
13	Ordinary Stone	94
14	Ordinary Stone	95
15	Moonstone	75
16	Moonstone	92
17	Moonstone	86
18	Ordinary Stone	87
19	Moonstone	66
20	Ordinary Stone	91
21	Ordinary Stone	61
22	Moonstone	100
23	Moonstone	99
24	Ordinary Stone	99

A couple quick calculations show that observed mean score of those bearing the moonstone, \bar{x}_M , was 85.17 and the observed mean score of those bearing the ordinary stone, \bar{x}_O , was 78.33. The observed difference in mean scores is $\bar{x}_M - \bar{x}_O = 6.83$.

Again, we would like to know whether the difference in means could be reasonably attributed to the random assignment alone. To investigate this, we will create a randomization distribution by assuming the scores would have been the same regardless of the type of stone, and randomly reassigning the type of stone to the existing scores. A simulation of this could be done similarly to the previous simulation activity.

Write the 24 scores on identical index cards, shuffle the cards, and deal out two piles of 12 to represent the 12 students selected to receive each treatment. Have students calculate the mean of each pile, subtract $\bar{x}_M - \bar{x}_O$ and record the difference, then add the result to a class dotplot.

Next, use technology to do the simulation with many trials. The plot below shows the randomization distribution for 1000 trials, along with a summary of the results.

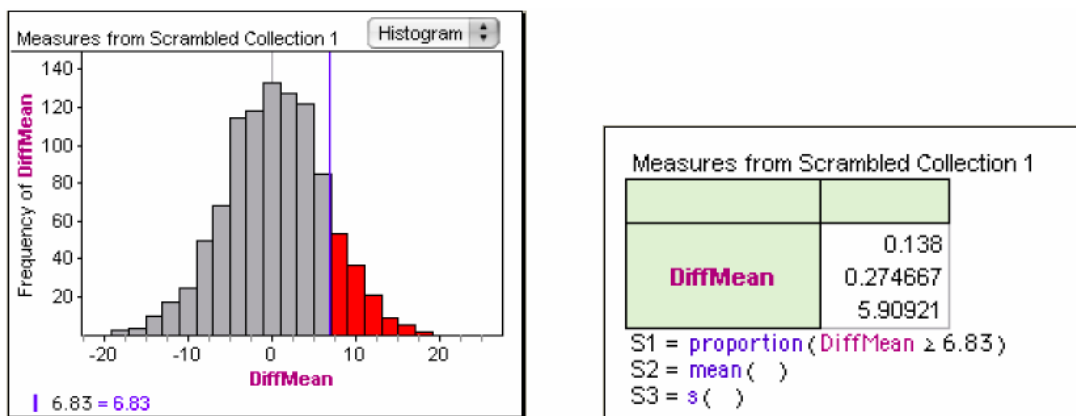


Figure 2 and 3

We can see that 13.8% of the observed differences in means are at least 6.83. This P -value is high enough that we would not reject the null hypothesis. There is not sufficient evidence to support the claim that the moonstone improved the test scores.

But notice that the distribution seems to be approximately normal and the standard deviation is approximately 6. The original scores were, indeed, a random sample selected from a normal population, then adjusted to give integer values. The standard deviation of the experimental group is 14.7

I would like to point out that we are once again blessed with good fortune! If you consider the sampling distribution of, say, \bar{x}_M by itself, the usual standard deviation formula for the sampling distribution of a mean will not work; the actual standard deviation will be much smaller than the standard formula would predict because of the relatively large sample size. (Remember that $10n \leq N$ condition!)

So, why does the formula work for the standard deviation of the difference? It works because, under the null hypothesis, the two treatment means are negatively correlated (if one happens to be relatively large, that forces the other to be relatively small). The result is that the differences are forced to be a little more extreme (larger or smaller) than would be the case for two independent sample means -- and, magically, just enough more extreme to allow the standard formula for independence to work again. That moonstone might be lucky after all!

If we had been selecting two independent samples from two identical populations and calculating the mean difference, we would expect a distribution of differences that is approximately normal, centered at 0, with standard deviation

$$\sqrt{\frac{\sigma_M^2}{n_M} + \frac{\sigma_O^2}{n_O}} = \sqrt{\frac{14.7^2}{12} + \frac{14.7^2}{12}} = \sqrt{36.015} \approx 6.00$$

This is approximately what we *did* get, though we didn't have two independent samples from two populations. We had a set of randomizations from this one group of experimental units.

The *P*-value obtained from the usual inference procedure is based on the *t*-distribution.

$$t = \frac{(\bar{x}_M - \bar{x}_O) - 0}{\sqrt{\frac{s_M^2}{n_M} + \frac{s_O^2}{n_O}}} = \frac{85.17 - 78.33}{6.00} = 1.146$$

With 21.17 degrees of freedom, this gives a *P*-value of about 0.13.

Again, this suggests that the same approximation we used for the difference of means from independent random samples will work for a difference of means for randomized experiments. And, again, this has been shown to be true in general by Fisher.

I would like to remind you that we set this simulation up so that the treatments had no effect. This means we know, in this simulation, that the two treatment groups have the same underlying distribution and we can use the standard deviation of all the scores as σ . This is not true in practice, where we would need to use *s* from each treatment group to estimate σ .¹

Again, the focus of this paper is the conditions and interpretation when the context is a randomized experiment. Let's review the situation with which we began. We had two treatment groups whose underlying distribution was approximately normal. (The scores were a random sample from a nearly normal population, so a randomly selected subgroup

¹ If there is good reason to suspect that the two treatment groups have the same standard deviation, you can use a pooled estimate for σ , increasing the power of the test. In AP Statistics this is not generally necessary.

of the scores can also be considered a random sample from that normal population.) This is still a necessary prerequisite for the theory to hold. The condition we check is that the distributions of both groups are reasonably symmetric with no extreme outliers, or that the two treatment groups are reasonably large.

And we then randomly assigned the treatments to the experimental units. To summarize, if the context is a randomized experiment the conditions that need to be checked are

- The two treatments are randomly assigned to the experimental units
- The distributions from both treatment groups appear consistent with the assumption that the underlying distributions are approximately normal, or that the groups are “large enough” (in the same way your text describes for two independent samples).

Confidence Intervals for the Difference of Two Means

Our technique for constructing confidence intervals is based on the same randomization distributions as tests, and the standard deviation of the test statistic is the same for a confidence interval as it is for the corresponding test. This means that the technique for intervals works under the same conditions as the technique for a significance test (in the context of experiments, use the conditions listed above). As with proportions, an interval requires a different way of thinking about this difference.

If the two stones are equally effective, we would expect the mean test score had everyone carried moonstone (μ_M) to be the same as the mean score had everyone carried an ordinary stone (μ_O)². If the stones are not equally effective, we would expect a nonzero difference between these two hypothetical means. A confidence interval gives the set of expected differences that could produce the observed difference as a reasonably likely outcome.

For example, in the case above the 95% confidence interval for $\mu_M - \mu_O$ is given by the usual formula:

$$(\bar{x}_M - \bar{x}_O) \pm t^* \sqrt{\frac{s_M^2}{n_M} + \frac{s_O^2}{n_O}}$$

In our example, we have

$$(85.17 - 78.33) \pm 2.079 \sqrt{\frac{13.079^2}{12} + \frac{15.985^2}{12}} \approx (-5.56, 19.23)$$

² μ_M is the expected value of \bar{x}_M in the sense that it is the average value of \bar{x}_M in the randomization distribution.

We would interpret this as follows: We are 95% confident that the difference $\mu_1 - \mu_2$ would be between -5.56 points and 19.23 points where μ_1 is the expected mean test score had everyone carried moonstone, and μ_2 is the expected mean test score had everyone carried an ordinary stone.

Appendix A: Instructions for simulating the moonstone experiments on Fathom.

The procedure is the same for the proportion and mean situations.

1. Set up a collection with the actual class data.
2. Create a new collection with the treatments randomly scrambled.
3. Define a **Measure** on the scrambled collection to calculate the difference between the two treatments.
4. **Collect Measures** to create the randomization distribution
5. Find the proportion of randomizations that resulted in a difference as large or larger than the observed difference.

The instructions will be for both situations, but the pictures will be for proportions only unless it is necessary to show both.

Step 1: Set up a collection with the actual class data.

Drag down a new case table from the tool shelf (the row of icons).

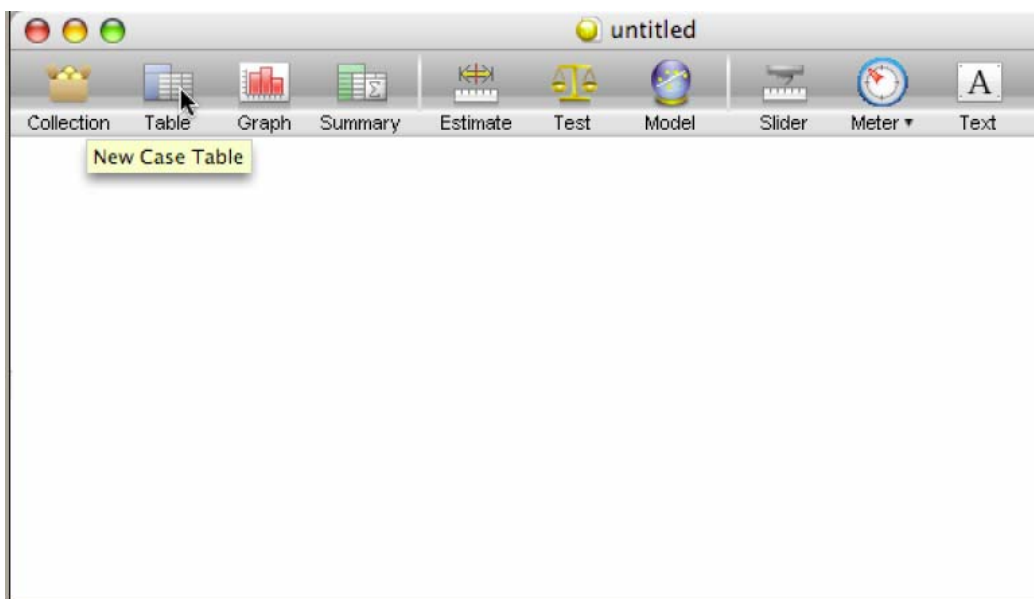


Figure A1: New Case Table

Click on the title of the column where it says <new>. For proportions, type the title *Type of Stone* (Fathom will insert underscores because spaces are not allowed in attribute names.) and enter *Moonstone* in the first 12 lines and *Ordinary Stone* in the next 12 lines. Then enter *Result* in the title of the second column, and *Date* in the first 9 lines, and *No Date* in lines 10 to 12, *Date* in lines 13 to 18, and *No Date* in lines 19 to 24. This sets up the situation described in the class. (Since we'll be randomizing the treatments, we could have simply listed the first 15 as *Date* and the last 9 as *No Date*, but students sometimes

have difficulty seeing that this is OK, especially before they have seen where this is going.) For means, type in the data as it appears in the table for that situation.

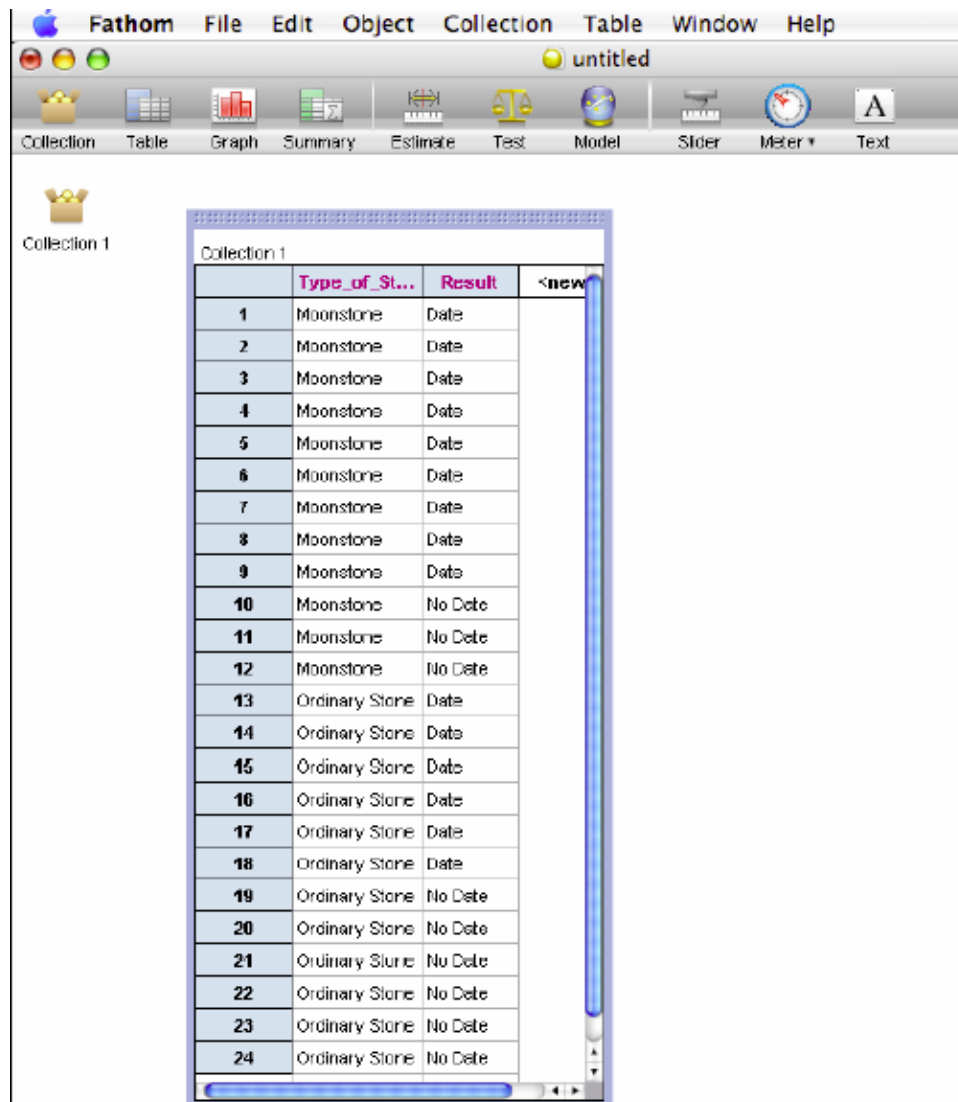


Fig A2

Step 2: Create a new collection with the treatments randomly scrambled.

Click once on the box labeled **Collection 1**. A blue frame should appear around it indicating that the collection is selected. In the **Collection Menu** select **Scramble Attribute Values**. A second collection will appear with the title **Scrambled Collection 1** and an inspector window will appear for this new collection.

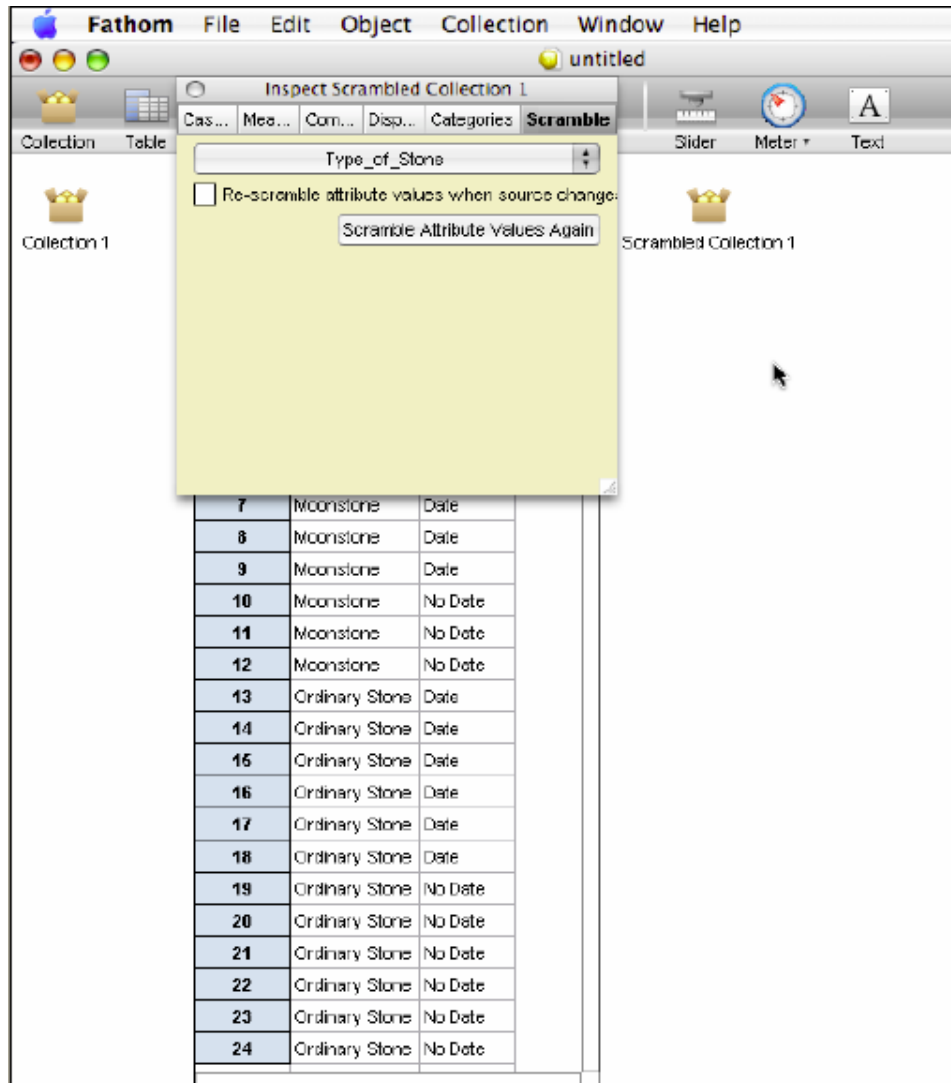


Fig A3

We no longer need to see the **Case Table** so select it by clicking on it once and delete it (**Object/Delete Case Table**). Now select the collection box **Scrambled Collection 1** and drag down a new case table. Now you can see how the treatment was randomly reassigned to each person, but the result for each person did not change.

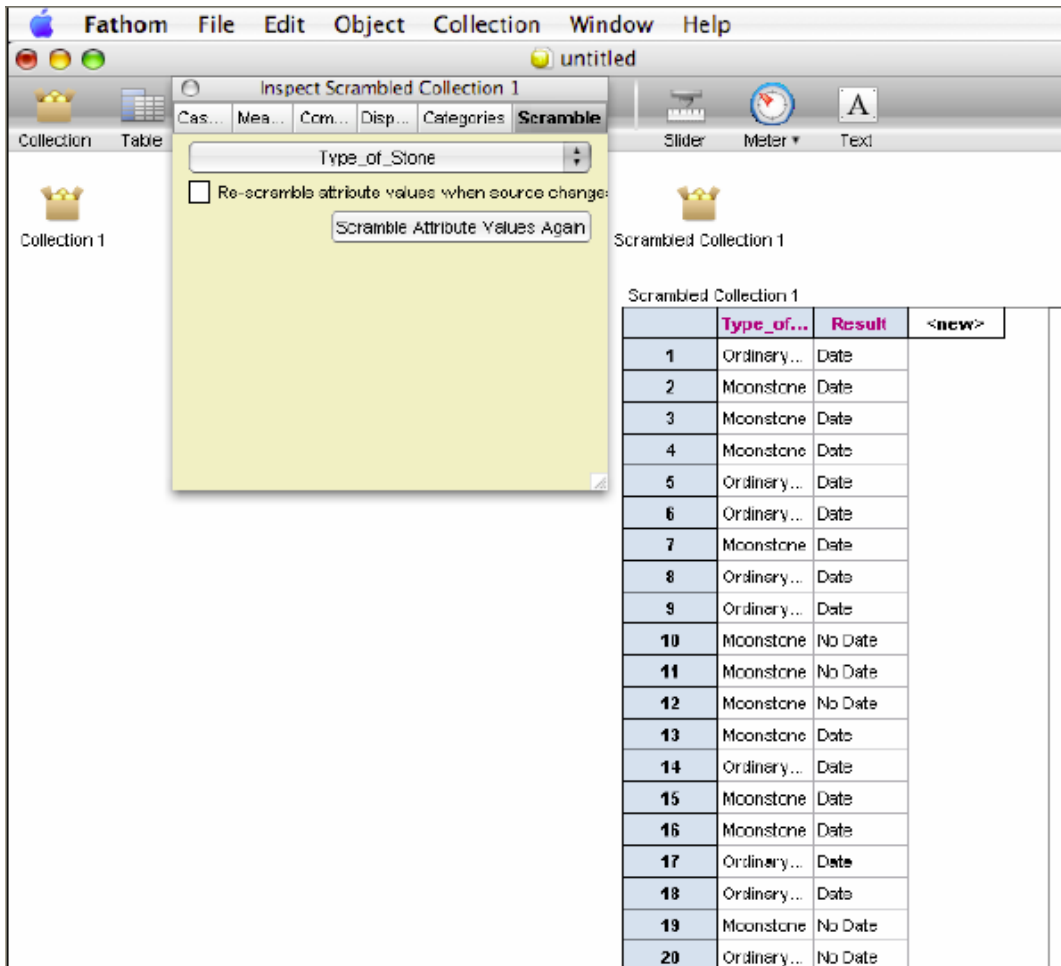


Fig A4

Step 3: Define a **Measure** on the scrambled collection to calculate the difference between the two treatments.

Click on the **Measures** tab in the inspector for **Scrambled Collection 1**. Where it says <new> type *DiffProp* for Difference in Proportions or *DiffMean* for a difference of means.

Measure	Value	Formula
DiffProp		
<new>		

Fig A5

Double-click the box under **Formula** and a formula editor will appear. Enter the appropriate formula for the difference of proportions or the difference of means.

DiffProp = `proportion (result = "Date", Type_of_Stone = "Moonstone") - proportion (Result = "Date", Type_of_Stone = "Ordinary Stone")`

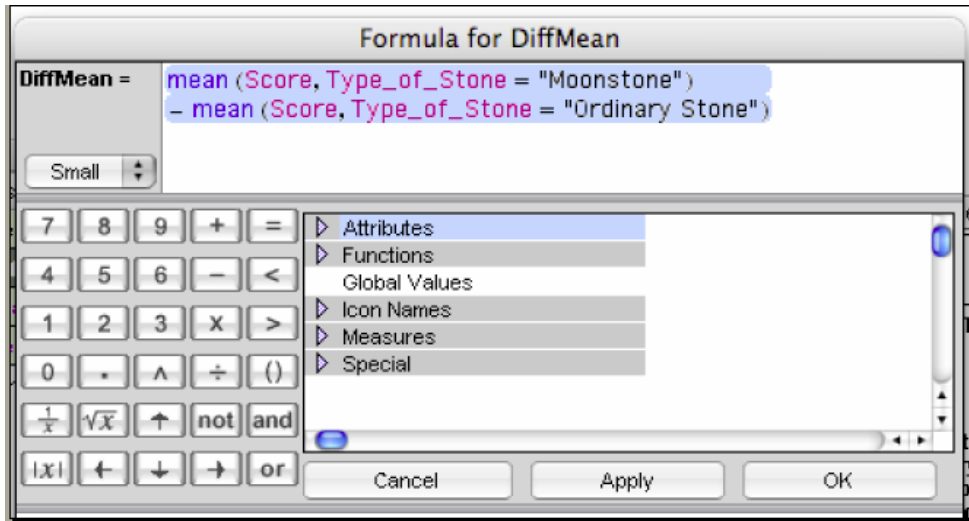
Small

- Attributes
- Functions
- Global Values
- Icon Names
- Measures
- Special

Cancel Apply OK

Attributes are the names you can use in expressions. They refer to attributes in a collection.

Fig A6



A7

(I'd like to point out a couple things for those who have not used Fathom. First, when Fathom recognizes a command in the formula editor, it turns color. Proportion and Mean turn blue when you hit the parentheses key because that is when Fathom recognizes the command. Attribute names turn a reddish color. Also, when you type the open parenthesis, both parentheses appear and you type between them. The editor is very efficient if you type perfectly, but if you make a mistake it is sometimes easier to delete everything and start over. In these formulas, the portion after the comma acts as a filter. Proportion(result = "Date",Type_of_Stone = "Moonstone) returns the proportion of cases for which *Date* appears in the result column, but only for the cases where the type of stone is *Moonstone*.)

Now that the measure has been defined, we want to collect a bunch of them to see what proportion of the time we get a difference as large or larger than we observed. Single-click to select the **Scrambled Collection 1** and in the **Collection** menu, select **Collect Measures**. A new collection will appear and its inspector will open. Deselect **Animation on** to speed things up, select **Replace existing cases**, and change the number of measures to 1000. Click **Collect More Measures**.

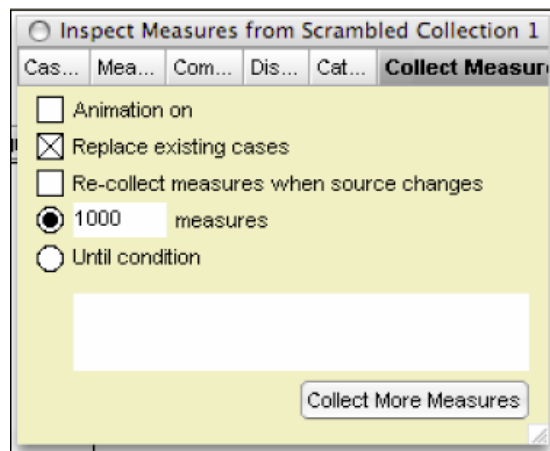


Fig A8

Then, click on the **Cases** tab in the inspector. Drag down a **New Graph** from the tool shelf, and drag the attribute name “DiffProp” or “DiffMean” to the horizontal axis.

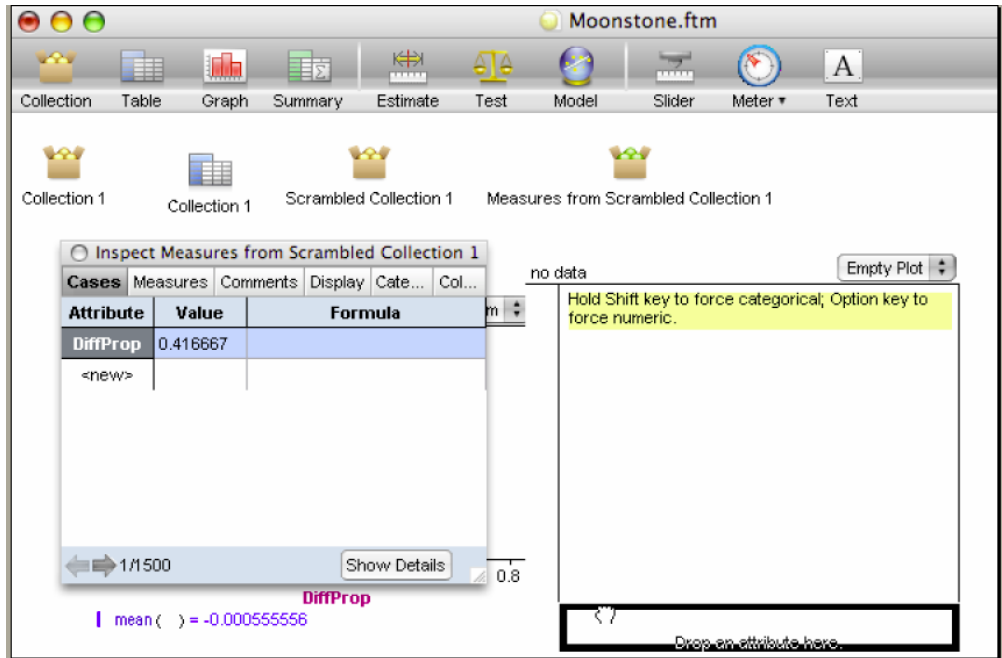


Fig A9

This will create a dotplot of the differences. The pull-down menu in the upper right corner of the graph will allow you to change this to a histogram if you prefer.

To find the *P*-value, we need the proportion of differences at least as large as that observed by the class of students. One efficient way to do this is with a summary table. Drag down a **New Summary** from the tool shelf. Drag the attribute name “DiffProp” or “DiffMean” from either the graph axis or the inspector to one of the cells (arrows will indicate where you can place it).

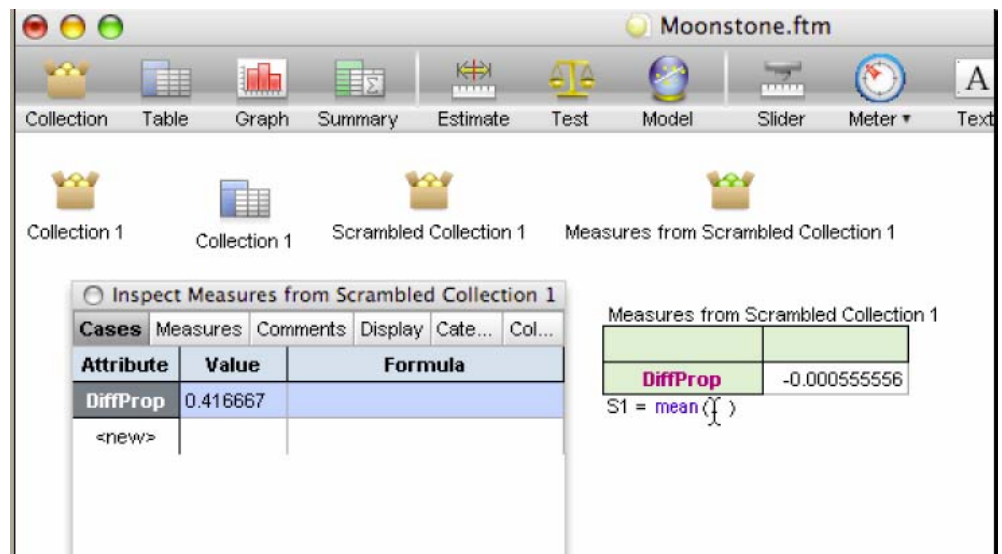


Fig A10

Double-click on the formula for S1 in the summary table to change it to the appropriate condition.

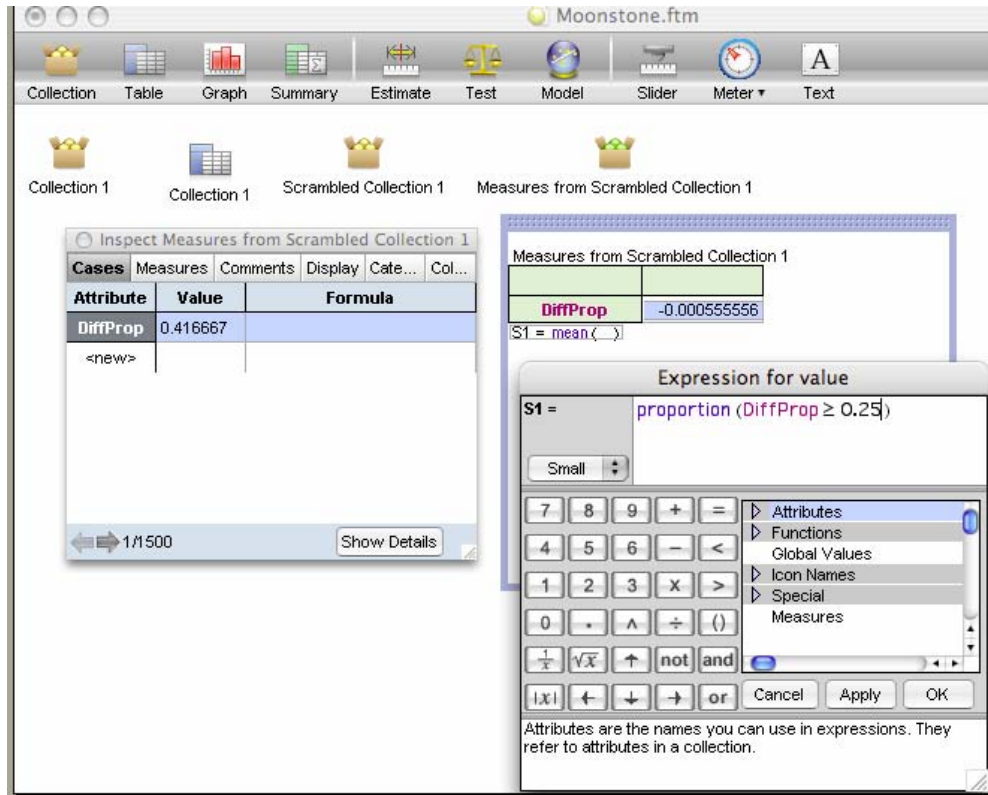


Fig A11

This will give you the *P*-value.

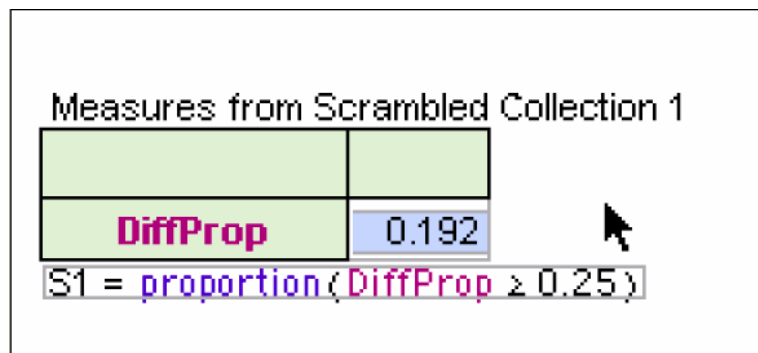


Fig A12

Appendix B: Instructions for simulating the moonstone experiment on TI-84+ Graphing Calculators

I will give the instructions for simulating the difference of proportions on the calculator, and make notes where changes are needed for the difference of means. The simulation will be done with a program, which allows you to set up a loop to repeat a procedure many times.

The calculator is much slower than a computer, so efficiency is important in a complicated simulation. We will not enter data for “Moonstone” or “Ordinary Stone”. We will randomly reorder the list of results and assign the first 12 to be “Moonstone.”

Enter the data for the results into L_1 . For proportions, we will need to recode the data as numerical data for the calculator. “Date” will be represented by the number 1, and “No Date” by the number 0. Enter the number 1 fifteen times and the number 0 nine times in L_1 . (For means, simply enter the 24 test scores.)

What follows is the program that will randomly shuffle the list, subtract the mean of the first half of the list from the mean of the second half of the list, record the difference in means, and report the proportion of scores at or above the given value. Note that by recoding successes as 1 and failures as 0, the mean does give the proportion of successes.

```
: ClrList L3
: For(I,1,100)
: rand(24) -> L2
: SortA(L1,L2)
: mean(seq(L1(J),J,1,12))-mean(seq(L1(J),J,13,24)) -> L3(I)
: End
: L3 ≥ 0.25 -> L4 (For test scores, use : L3 ≥ 6.83 -> L4)
: Disp mean(L4)
```

The second to last line converts values greater than or equal to 0.25 (or 6.83 for test scores) to 1 and any other values to 0. The mean of this new list will be the proportion of differences greater than the observed difference.