Testing Change Over Two Measurements in Two Independent Groups

David C. Howell, Univ. of Vermont

This document began as a short answer to what I thought was a simple question from Stacey Freedenthal at the University of Denver. I have expanded it well beyond her question, and I have slightly distorted her question to simplify the problem, but I think that it addresses questions that others may have. After a lot of work it turns out that the question actually was simple, though I didn't recognize it at the time.

This entry is going to be long and rambling because I want to explain several things about contingency tables, chi-square, dependent samples, and research design. Besides, I wrote it primarily for myself and have considered several ancillary questions.

The Problem

Stacey collected data on help-seeking behavior in school children. The data were collected in the fall and again in the spring after a possible intervention. There were two groups, one of which received an intervention and the other served as a control. She found that help-seeking didn't increase significantly in either group (going from 60% to 64.9% in the treatment group and 54.4% to 55.4% in the control group. But she also wanted to make between-group comparisons in the amount of improvement in help-

seeking. Her question was complicated by different sample sizes both within and between groups, but I am going to ignore that.

To broaden the usefulness of this document I am going to leave the specifics of Stacey's study somewhat to the side and discuss a number of ways that the study *might* be run and the ways that it might be analyzed. Any clumsiness in the designs is my fault, not hers.

Measuring Independent Groups Once

Let's consider one of the simplest ways that this study might be run with independent measurements. We will measure a group of children in the fall and a different group of children in the spring after they have received some intervention. Notice that these measurements are independent because they come from different children. We might find that the fall group, which has not received an intervention, requested help 55 times while the spring group, which had received the intervention, asked for help 65 times. Is that difference significant? I don't know and I don't know how to know other than to say that 65 is a bigger number than 55. I don't think that there is a statistical test that can lead us any further because we can not compute a standard error on which to base a test.

So let's modify the design to count not only the number of children asked for help one or more times, but also the number of children that never asked for help. Now we can calculate the *proportion* of children who sought help. Suppose that in each case there were 70 children and in the fall group 42 children sought help and in the spring 45 sought help. Then the proportions are 42/70 = .60 and 45/70 = .643. We can test the

difference in at least two ways—and I am going to drag in a third way to set up what follows.

Because the groups are independent we can use a standard test to compare the two percentages.

$$z = \frac{(p_1 - p_2)}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \qquad \text{where } p = (p_1 + p_2)/2 \quad \text{and } q = 1 - p \text{ if } n_1 = n_2$$

$$st.err = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} = \sqrt{2(\frac{.6214 * .3786}{70})} = .0820$$

$$z = \frac{p_1 - p_2}{st.error} = \frac{.0429}{.0820} = .5232$$

$$p = .601$$

However we could also set this up as a chi-square test, which will be useful with respect to what follows.

	Sought Help		
	Yes	**No**	
Fall	42	28	70
Spring	45	25	70
	87	53	140

Chi-square = 0.2733, $p = .601$

The effect is not significant and you get the same probability whether you used chi-square or the test on proportions, as you should. Remember that I have independent groups in the fall and spring.

Just to make sure that everything is correct, and to lead to what follows, I will set up a simple sampling study. I will draw a sample of 70 cases from a population where $p = .60$ and another 70 cases from a population where $p = .6429$. I do this by drawing 70 random numbers between 0 and 1 and counting the proportion of times the random number is less than .60 (or less than .6429.) For each draw these proportions should hover somewhere in the vicinity of .60 and .6429. I then compute the difference in proportions, repeat this process 10,000 times, and then find the standard deviation of the resulting distribution of differences. When I do this I get

>The mean of the differences is 0.04331143

>The standard error of the differences is 0.08193536

>z = .528

>p = .597

Which should be close enough in anybody's book. So at least I am not completely off in left field.

One Sample Measured Twice—Dependent Measurements

What we have just done is nice for independent samples, but that is probably not the way that most people would run the study. If I were doing it I would probably count the proportion of children seeking help in the fall, the proportion of children who sought help

in the spring after an intervention, then test the difference in those two proportions. That sounds reasonable. But now the samples are not independent and that really messes things up. What it does is to seriously underestimate the standard error of the differences in proportions. It took me a while to figure out why this would be, but I think that I have now figured it out, which is why I am going through this convoluted discussion..

Suppose that 60% of your 70 students (= 42 students) in the fall sought help. Because this is the dependent sample case, we use those same 70 students again in the spring. It *could* be that those same 42 students (plus a few more to get up near $p = .6429$) sought help in the spring. OR, it could be that an entirely different set of (approximately) 45 students sought help in the spring. Does this make a difference? I hope so or else I can't figure out the problem with lack of independence. The problem gets complicated because we need to change how we record out data. We need to know if Johnny sought help in the fall and in the spring. We can't just say 42 kids sought help one time and 45 kids sought help the other. That requirement may or may not be feasible experimentally, but we need it statistically.

The following are three tables that I set up to mirror what I think is going on. In the first I deliberately had mostly repeaters—if you sought help in the fall you also sought help in the spring. In the second I relaxed that somewhat but not completely. If you sought help in the fall there was a good, but not perfect, chance that you would seek help in the spring and quite a few who did not seek help in the fall did go ahead and seek it in the spring. In the third scenario I pretty much made spring help-seeking independent of fall help-

seeking. But in all three cases I had (approximately) 60% help seekers in the fall and 64.29% help seekers in the spring, so that part of the data does not change.

Mostly the same students seek help

		Spring		
		Yes	No	
Fall	Yes	42	**0**	42
	No	**3**	25	28
		45	25	70

Notice that this table is fundamentally different from the table earlier. Here I know that 3 children who did not seek help in the fall did seek it in the spring. This is not your standard chi-square contingency table. In fact it is part of what is called McNemar's test. I used to cover that test in my books, but somewhere along the line it disappeared. The critical feature of McNemar's test is to ask if more people switched from No to Yes than switched from Yes to No. (Notice that I have entered the changes—the off-diagonal entries—in bold. We ignore the data on the main diagonal. If we label the cells

A	B
C	D

then the formula for McNemar's chi-square is

$$\chi^2 = \frac{[B-(B+C)/2]^2}{(B+C)/2} + \frac{[C-(B+C)/2]^2}{(B+C)/2} =$$
$$= \frac{[0-1.5]^2}{1.5} + \frac{[3-1.5]^2 1.5}{} = 3.0$$

Using R to run the statistical test I have

>McNemar's Chi-squared test

>McNemar's chi-squared = 3, df = 1, p-value = 0.08326

Although the difference is not significant, it isn't way off. But just wait!

A lot of the same students seeking help but a bunch of new ones as well

		Spring		
		Yes	No	
Fall	Yes	35	**7**	42
	No	**10**	18	28
		45	25	70

Here R gives a different result even though 60% seek help in the fall and 64.29% seek help in the spring. The data are no where near significant.

>McNemar's Chi-squared test

>McNemar's chi-squared = 0.5294, df = 1, p-value = 0.4669

A random sample of students switch—whether you are Yes in the fall has nothing to do with whether you are Yes in the spring.

		Spring		
		Yes	No	
Fall	Yes	27	**15**	42
	No	**18**	10	28
		45	25	70

From R we get

McNemar's Chi-squared test

McNemar's chi-squared = 0.5294, df = 1, p-value = 0.4669

If we were to apply a continuity correction (analogous to Yates' correction) to these data we would get chi-square = 0.2353, which is very close to the chi-square we obtained when the data came from independent groups. That is because when I created this last set of data I made the response in the spring independent of the response in the fall.

BUT the question that I have often wondered about, but not enough to worry about it too much, was why the results changed when the data were not independent. Why does a lack of independence matter? Agresti (2002) says that the lack of independence affects the standard error, but as the test is run here we don't see a standard error—though there would be one. So I started playing with a simple sampling design in R to see if I could see what happens to the standard error.

A Simple Sampling Study

First I will try to create a model with a scheme like that behind the first set of data. I will repeat it 10,000 times and look at the standard deviation of the differences, which is the standard error of the difference. Then I will do that all over with a model with a scheme like the third set. (The results for a scheme like the second set would fall in the middle.)

To do this I created the Before data, which would have a mean proportion of .60. Then I created the After data by taking whatever number of Yeses that I got with Before and

taking, ON AVERAGE, 4.29% of 70 more cases to add to the ones I already had. In other words, the same people who said yes the first time said yes the second PLUS a couple more because 64.29 is more than 60. This is the extreme because everyone who had a yes first was forced to have it again in the spring.

This time the mean difference was .04930, which is very close to what it should be. The standard deviation of the difference was 0.0243, which is much smaller than the standard error was in the independent case where I tested the difference of independent proportions (0.0820). Agresti told me that this should happen because the lack of independence underestimates the standard error. And the standard error should be small because so much of the spring data is dominated by the fall data. If in one replication there were 58% Yeses in the fall, then there are going to be very close to 58% in the spring because I would only add about 6.29% of 70. If we had a weird replication with only 50% of yeses in the fall, and I had only about 6.29% of 70 to that, I am still going to be down close to 50% and the difference will again be small. Since almost all of the differences will be small, the standard error will be small. The spring data cannot differ much from the fall data.

Ooops—4.29

Ooops—4.29

But now I want to see if that standard error rises when I change the sampling scheme. In what I just did I maximized the lack of independence by making sure that it was the same people getting Yes both times, plus a few more the second time. Now I am going to a case where I am using the same people, but I am drawing data where their responses are (almost?) forced to be independent. In other words we will have, on average 60% yes in

Before and 64.29% yes in After, but I won't impose any requirement that those who were scored Yes the first time are more likely to be Yes the second. This will make it possible for the spring data to differ considerably from the fall in some replications, so the standard deviation of differences (the standard error) will be much larger.

>The mean of the differences is 0.04325571

>The standard error of the differences is 0.08155602

This is almost exactly what we found with the case of two independent groups, and it should be because I allowed the responses to be independent even if they came from the same children. Notice how large the standard error is here.

Two Groups With Repeated Measures Fall and Spring

Now we come to the fun stuff. This is the question that Stacey originally asked. How do we test to see if the change in the Intervention group is significantly different from the change in the control group? I am sticking roughly to her data, where it is unlikely that the difference in change is significant, so don't get your hopes up. For the Intervention group the fall and spring proportions were (nearly) 60% and 64.29%. For the Control group the fall and spring proportions were (nearly) 54.4% and 55.4%. (I say "nearly" because I had to fudge things to get reasonable integer frequencies for the cells because I set my *n*s at 70 whereas hers varied. I also assume that I had 70 kids in each group and that I know who was a Yes both times, who was a Yes and then a No, etc.)

Suppose that I had the following data

Treatment

		Spring		
		Yes	**No**	
Fall	**Yes**	32	**10**	42
	No	**13**	15	28
		45	25	70

Control

		Spring		
		Yes	**No**	
Fall	**Yes**	30	**8**	38
	No	**9**	23	32
		39	31	70

What follows is based on work by Marascuilo & Serlin (1979, *British Journal of Mathematical and Statistical Psychology*). The logic of the test is much easier than you might suppose.

The important data from both of these tables are the data in the cells that represent change from fall to spring. You will note that for the treatment group 13/(10+13) = 13/23 = 56.52% of the changes were from not seeking help to seeking help. In the control group 9/17 = 52.94% of the changes were toward seeking help. All we need to ask is whether these two percentages are significantly different. Using a test on two independent proportions we have

$$z = \frac{p_1 - p_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} = \frac{.5652 + .5294}{\sqrt{\frac{.5473 * .4527}{23} + \frac{.5473 * .4527}{17}}} = \frac{.0358}{\sqrt{.025434}} = .22487$$

where p is the average of p_1 and p_2 and $q = 1 - p$.

$p(z \geq \underline{+0.22487}) = .822$

The probability under the null is .822, so we can certainly not reject the null hypothesis.

If we set the same data up as a standard contingency table we would have

		Condition		
		Treatment	**Control**	
Change	**Increase**	13	9	22
	Decrease	10	8	18
		23	17	40

From the chi-square test (without a correction for continuity) we have

>chisq.test(data, correct = F)

>Pearson's Chi-squared test

>X-squared = 0.0506, df = 1, p-value = 0.822

which is the same result.

This Doesn't Answer All of the Problems

Although each of the solutions above is correct for the particular design, It leaves out several possible designs. For one thing it might be possible to count the help-seeking behaviors but not be able to record which child sought help. Perhaps it is the same few kids asking over and over again, or perhaps it is most of the class asking relatively few times. That can make a difference.

Alternatively we might not be able to identify change at the level of the individual child. For the last several designs it was necessary to ask if Johnny switched from no help-seeking in the fall to help-seeking in the spring, or some other pattern. It is easy for a statistician to say that we need that kind of data, but it is much harder for a researcher, especially in a busy chaotic school, to collect those data.

And of course you have the common problem that differences may be due to classrooms rather than the intervention. In my first design we might find that the intervention group did better not because of the intervention but because they had a teacher that was seen as more willing to help.

dch

8/28/2008