Matched Pairs Analysis

The term *matched pairs* is commonly used in categorical data analysis to refer to within-subjects analyses, which involve repeated measures (e.g., pre-post design), matched pairs (e.g., dyads, yoked treatment and control subjects), within-subjects experiments (e.g., participants receive both treatment and control) or multiple dependent measures for a respondent (e.g., approval of two ballot initiatives).

For a binary dependent variable, there is a form of the chi-square test for within-subjects designs called *McNemar's chi-squared*. The analogous test with a continuous measure is the dependent (paired) *t*-test or within-subjects ANOVA with two levels. The Pearson χ^2 analysis of a contingency table has the assumption that the levels of each of the variables are independent, that different individuals are in each one of the cells of the design. The matched pairs contingency table contains values for two dependent responses, such as approval of a ballot initiative at one point in time and another point in time. Any other circumstance in which the two observations being compared are not independently sampled, including two members of a dyad, same members of a household, yoked subjects in a matched control design, and so on, are considered dependent and should not be compared with the tests we have discussed up to this point (e.g., Pearson χ^2, G^2, Breslow-Day test).

Below is a 2 × 2 table of clinical depression classification of respondents to a national survey of older adults at two different time points (six months apart).

		Time 2 Depression: Not	Time 2 Depression: Depressed	
Time 1 Depression	Not	146	155	301
	Depressed	47	303	350
		193	458	651

It is important to note that this table differs from the non-matched 2 × 2 table from the Quinnipiac poll data that we examined previously. That table used party identification (independent vs. major party) by candidate choice (Biden vs. Trump) for the purpose of asking whether independent voter identification meant that the voter was more (or less) likely to support Biden—a comparison of a binary response to two different groups. The above table involves two measurements of the same variable repeated at two different time points to ask whether there is an increase or decrease in the binary response over time.

McNemar's χ^2

Introduced by quantitative psychologist Quinn McNemar in 1947 (McNemar, 1947) to examine the change in a binary repeated measurement, the formula for the *McNemar test* is nearly always stated in terms of the frequency or proportions for the discordant cells—the cells that reflect non-agreement ("no"-"yes" or "yes"-"no"). Most often, however, researchers will frame the question of change in agreement over time as pertaining to the marginal frequencies or proportions. For example, for the depression example, we would ask the question about whether there was a higher proportion of participants who were above the clinical cutoff for depression at Time 2 than at Time 1, which would compare $p_{2+} = n_{2+}/n_{++} = 350/651 = .538$ to $p_{+2} = n_{+2}/n_{++} = 458/651 = .704$. This hypothesis is often described as the *marginal homogeneity* hypothesis. For the 2 × 2 case, a little algebra can show that the marginal proportion difference is equal to the difference in discordant cells. For example, $p_{+1} - p_{1+} = (p_{11} + p_{12}) - (p_{11} + p_{21}) = p_{12} - p_{21}$.

$$McNemar's\ \chi^2 = \frac{(n_{21} - n_{12})^2}{n_{21} + n_{12}} = \frac{(c-b)^2}{c+b}$$

	Time 2: No	Time 2: Yes
Time 1: No	a	b
Time 1: Yes	c	d

The final McNemar equation on the right-hand side refers to the fourfold notation. I use lower case (*a, b, c, d*) to emphasize the difference of the matched pair and between-subjects contingency table.

Let's consider whether there was a significant change in the proportion of respondents who were depressed, .538 at Time 1 and .704 at Time 2. If there is a change overall, either decrease or increase, from one time point to another, relative to what is expected by chance, the result will be significant. Using the discordant counts from the table above,

$$\begin{aligned} McNemar's\ \chi^2 &= \frac{(n_{21} - n_{12})^2}{n_{21} + n_{12}} \\ &= \frac{(47 - 155)^2}{47 + 155} \\ &= \frac{(108)^2}{202} \\ &= 57.742 \end{aligned}$$

df in this test is 1, the critical value is 3.84 (from the chi-square table), and because calculated value of 57.742 exceeds this value, there is a significant difference in Time 1 and Time 2 depression.

The test can also be stated in terms of a *z*-statistic (a score test in this case), and is an equivalent statistical test to the chi-squared test.

$$z = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}} = \frac{p_{21} - p_{12}}{\sqrt{(p_{21} + p_{12})/n}} = \frac{p_{2+} - p_{+2}}{\sqrt{(p_{21} + p_{12})/n}}$$

where $z^2 = \chi^2$. Note that there are several equivalencies for the numerator, $p_{21} - p_{12} = p_{+2} - p_{2+} = p_{1+} - p_{+1}$. For the depression data, $z = .166 / \sqrt{(.072 + .238)/651} = .166 / .0218 = 7.599$, which is significant because it exceeds 1.96.

The 95% confidence interval using the score test standard error (SE_{score} = .0218) around the proportion difference, $d = p_{2+} - p_{+2} = .704 - .538 = .166$, is then

$$\begin{aligned} &d \pm SE_{score}(1.96) \\ &LCL = .166 - (.0218)(1.96) = .123 \\ &UCL = .166 - (.0218)(1.96) = .209 \end{aligned}$$

The Wald confidence interval, usually used in practice assuming large sample size, uses a bit different standard error estimate (as with the single group *z*-proportion test, the CI is a bit narrower).

$$SE_{Wald} = \sqrt{(p_{12} + p_{21}) - (p_{12} - p_{21})^2 / n} = .021$$

$$\begin{aligned} &d \pm SE_{Wald}(1.96) \\ &LCL = .166 - (.021)(1.96) = .125 \\ &UCL = .166 - (.021)(1.96) = .207 \end{aligned}$$

I × *I* Matched Pairs Table

The McNemar test was generalized by Stuart (1955) for square tables larger than the 2 × 2 case, a test usually referred to as the Stuart-Maxwell statistic for testing marginal homogeneity. With more than two

categories, tests of matched pairs differences become more complex. Using I for the number of levels of rows and columns and i as the index for each level, the marginal homogeneity hypothesis is that all $\pi_{i+} = \pi_{+i}$ in the population. The algebraic explanation usually ventures into matrix algebra at this point, with the cross-product of the vector of differences, $d = (p_{1+} - p_{+1}), (p_{2+} - p_{2+}), \ldots (p_{i+} - p_{+i})$ multiplied by the inverse of (divided by) the variance covariance matrix of expected values, which obtains the Stuart-Maxwell chi-squared test. McNemar's test is a special case of the Stuart-Maxwell statistic. The Stuart-Maxwell test is considered a score test. Bhapkar (1966) developed another test, which is a Wald test version. Although the Bhapkar has greater power, the Stuart-Maxwell test controls Type I error better (Yang et al., 2012).[1]

There are several other hypotheses that can be tested with square tables. The *symmetry* hypothesis is that every cell proportion formed by a row × column, π_{ab}, will be equal to its corresponding proportion on the other side of the diagonal, π_{ba}, with a and b any numbers that are not equal (e.g., $\pi_{12} = \pi_{21}$). One test of symmetry is the Bowker test (or sometimes McNemar-Bowker test), which is an omnibus test of all the possible 2 × 2 McNemar comparisons. There are several attempted corrections to the Bowker test (Krampe & Kuhnt, 2007). The *quasi-symmetry* hypothesis involves the comparison of odds ratios above and below the diagonal, which is less strict than symmetry. Symmetry is more stringent, because it requires that the marginal proportions are equal, which rarely occurs. Likelihood ratio tests can be specified for symmetry and quasi-symmetry using loglinear analysis (loglinear models are discussed in my univariate class). The difference in the two types of symmetry give a likelihood ratio for marginal homogeneity.

$$G^2\left(\text{marginal homogeneity}\right) = G^2\left(\text{quasi-symmetry}\right) - G^2\left(\text{symmetry}\right)$$

Three-Way Matched Tables

Cochran's Q (Cochran, 1950) is a test of three or more binary matched pairs (e.g., comparison of depression diagnosis over three time points). Cochran's Q is a generalization of the McNemar test and might be used first as an omnibus test of any change (increase or decrease) over three or more time points, for instance. Paired comparisons of two variables using the McNemar test could be used as follow-ups to a significant Cochran's Q. Cochran's Q is also equivalent to the nonparametric Friedman test.

Software Examples

The data for the examples below come from the Late Life Study of Social Exchanges (LLSSE; Sorkin & Rook, 2004). The depression diagnosis is determined by the recommended cutoff scores for the brief 9-item version (Santor & Coyne, 1997) of the Center for Epidemiologic Studies-Depression scale (Radloff, 1977).

SPSS

Note: SPSS results differ from R and SAS because it uses the continuity correction. There are three ways to call a McNemar's test in SPSS. To save space, I include the frequencies from the crosstabs approach and the test from the `npar` approach.

```
crosstabs /tables w1dep by w2dep
  /cells=count row column total
  /statistics= mcnemar.

npar tests mcnemar=w1dep w2dep.

nptests /related test (w1dep w2dep) mcnemar.
```

[1] Yang and colleagues (2012) proposed two alternatives not widely available, and show that their χ^2_{o2} test improves the Type I error problem with the Bhapkar.

Crosstabs

w1dep Time 1 depression * w2dep Time 2 depression Crosstabulation

			w2dep Time 2 depression		
			.00 not depressed	1.00 depressed	Total
w1dep Time 1 depression	.00 not depressed	Count	146	155	301
		% within w1dep Time 1 depression	48.5%	51.5%	100.0%
		% within w2dep Time 2 depression	75.6%	33.8%	46.2%
		% of Total	22.4%	23.8%	46.2%
	1.00 depressed	Count	47	303	350
		% within w1dep Time 1 depression	13.4%	86.6%	100.0%
		% within w2dep Time 2 depression	24.4%	66.2%	53.8%
		% of Total	7.2%	46.5%	53.8%
Total		Count	193	458	651
		% within w1dep Time 1 depression	29.6%	70.4%	100.0%
		% within w2dep Time 2 depression	100.0%	100.0%	100.0%
		% of Total	29.6%	70.4%	100.0%

NPar Tests

McNemar Test

Test Statistics b

	w1dep Time 1 depression & w2dep Time 2 depression
N	651
Chi-Square a	56.678
Asymp. Sig.	.000

a. Continuity Corrected
b. McNemar Test

R[2]

```
> counts <-array(
+   c(146,155,47,303),
+   dim=c(2, 2),
+   dimnames=list(w1dep=c("not", "depressed"),
+                 w2dep =c("not", "depressed"))
+ )
> mcnemar.test(counts, y=NULL, correct = FALSE)

        McNemar's Chi-squared test

data:  counts
McNemar's chi-squared = 57.743, df = 1, p-value = 0.00000000000002988
```

SAS[3]

```
*replace with your own file path;
proc import datafile="c:\jason\spsswin\cdaclass\dep.sav" out=one dbms = sav replace;
run;
```

[2] The DescTools package has functions for both the Stuart-Maxwell and Bhapkar tests.
[3] Bhapkar's test can be obtained from SAS in the CATMOD procedure. The Stuart-Maxwell test can be obtained in SAS with a macro developed by Sun and Yang (2008).

```
PROC FREQ DATA=one;
  TABLES  w1dep * w2dep / AGREE ;
RUN;
```

(frequencies omitted)

Statistics for Table of w1dep by w2dep

McNemar's Test

Statistic (S)	57.7426
DF	1
Pr > S	<.0001

Write up Example

A McNemar's test was used to investigate whether the percentage of participants who were clinically depressed changed over the six-month interval. At baseline, 53.8% of participants met the clinical cutoff for depression, whereas, at follow-up, 70.4% of participants met the clinical cutoff for depression. The difference was significant, McNemar's $\chi^2(1) = 57.74$, $p < .001$, and was a moderate-sized effect, $w = .30$.[4]

References and Further Reading

Bhapkar VP. (1966) A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association, 61*, 228-235.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, *43*(244), 572-574.

Cochran, W. G.The comparison of percentages in matched samples.*Biometrika*37256–266

Krampe, A., & Kuhnt, S. (2007). Bowker's test for symmetry and modifications within the algebraic framework. *Computational statistics & data analysis*, *51*(9), 4124-4142.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychology and Measurement, 1, 385-401.

Santor, D. A. & Coyne, J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. Psychological Assessment, 9, 233–243.

Sorkin, D. H., & Rook, K. S. (2004). Interpersonal control strivings and vulnerability to negative social exchanges in later life. *Psychology and Aging, 19*, 555–564.

Sun, X., & Yang, Z. (2008, March). Generalized McNemar's test for homogeneity of the marginal distributions. In SAS Global forum (Vol. 382, pp. 1-10).

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, *42*(3/4), 412-416.

Yang, Z., Sun, X., & Hardin, J. W. (2012). Testing marginal homogeneity in matched-pair polytomous data. Drug Information Journal, 46(4), 434-438.

[4] Though I have not seen authors report Cohen's w with McNemar's test, I see no reason we cannot compute it in this case,

$$w = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{57.74}{651}} = .30$$