Journal of Modern Applied Statistical Methods

Volume 1 | Issue 1 Article 10

5-1-2002

An Unconditional Exact Test For Small Samples Matched Binary Pairs

Robert A. Malkin

The Joint Department of Biomedical Engineering, The University of Tennessee, Memphis and The University of Memphis

Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 69-73

1538 - 9472/02/$30.00

An Unconditional Exact Test For Small Samples Matched Binary Pairs

Robert A. Malkin
The Joint Department of Biomedical Engineering,
The University of Tennessee-Memphis and
The University of Memphis

When investigators have N pairs of binary data, a common test for an increased rate of response is McNemar's test. However, McNemar's is an approximate, conditional test. An exact, unconditional test exists, but requires restrictive assumptions. Critical values and power tables are presented for an exact, unconditional test free of these assumptions.

Keywords: Binary data, 2x2 tables, Matched pairs, Exact test

Introduction

For a 2x2 table of matched binary trials, a common large sample test is the McNemar's test [1]. For example, if blood is drawn from N patients and split into two tubes in which one is treated with a test drug and another with a control substance, then the results of each treatment would result in 1, a response, or 0, a non-response, for each tube. Each pair of tubes from a single patient would result in either a 10 (response to the test substance and a non-response to the control), 01, 00, or 11. The results from all N patients, or matched-pairs, could be summarized by the number of pairs with each type of response, m_{10}, m_{01}, m_{00} and m_{11}, respectively, where $N = m_{10} + m_{01} + m_{00} + m_{11}$. McNemar proposed to use the test statistic

$$Z = \frac{m_{10} - m_{01}}{\sqrt{m_{01} + m_{10}}}$$

to test the null hypothesis that $\Pr(10) = \Pr(01)$, or equivalently $\Pr(1\bullet) = \Pr(\bullet 1)$ (the probability of a response in the test tube is equal to that in the control tube), against an appropriate alternative, such as a greater response to the test substance $\Pr(1\bullet) > \Pr(\bullet 1)$. Z is a non-exact, conditional statistic. That is, Z is only approximately a normal deviate, and then only when $n = m_{10} + m_{01}$ is fixed before the experiment. However, the cost of a study typically depends on N, the total number of pairs, not n. Planning for small clinical trials where budgets are limited, is difficult, if not impossible, when N is unknown. Thus, the application of McNemar's test to small clinical trials is challenging.

Although there have been many proposed improvements to the McNemar's test, the first to propose an exact, unconditional test was Suissa and Shuster [2]. Their work was based on the exact formula for the power of the test, given by

$$\pi(s,q,Z) = \sum \frac{m_{10}!m_{01}!m_{00}!m_{11}!}{N!} s^{m_{10}} q^{m_{01}} (1-s-q)^{(m_{11}+m_{00})},$$

where $s = \Pr(10), q = \Pr(01)$, and the sum is taken over all $\{m_{10}, m_{01}, m_{00}, m_{11}\}$ such that $Z > z$, where z is the critical value. Under the null hypothesis $s = q =$, say, $p/2$, and therefore the p-value (α or significance) is more easily expressed as $\pi(p,z)$. However, p is unknown. To ensure that the actual p-value is less than the stated p-value, it is necessary to find the maximum of π as a function of the unknown p. This maximum has not been found analytically to date. Therefore, Suissa and Shuster suggested numerically finding a supremum of $\pi(p,z)$ over $0 \le p \le 0.995$ for $N > 10$. A supremum exists since the derivative is bounded [3]. The range $0.995 \le p \le 1.0$ was ignored by Suissa and Shuster because it required unusually long computing times and was said to represent an unlikely scenario [2]. Although unlikely, ignoring this region is somewhat arbitrary (why not 0.990 ?) and inevitably increases the possibility of a Type I error. There exists a more conservative test of the same type.

In this paper, I present new critical values for $N > 5$. A simplified approach is presented to finding a supremum, appropriate for small values of N, which allows a supremum to be found over the entire interval $0 \le p \le 1.0$.

Robert A. Malkin is Herbert Herff Associate Professor at The Joint Department of Biomedical Engineering, The University of Tennessee-Memphis and The University of Memphis. Contact him at Department of Biomedical Engineering, ET330, Memphis, TN, 38152. E-mail at ramalkin@memphis.edu.

Using the new critical values and calculation approach, new sample size (power) tables are presented. In short, I present the most conservative possible test for this problem.

Methodology

Under the null hypothesis, equation (1) can be simplified to [2]

$$\pi(p,z) = \sum \binom{N}{n} p^n (1-p)^{N-n} F_n(n - i_n - 1) \qquad [2]$$

where F_n is the cumulative binomial distribution function. A supremum for this function with respect to the unknown p exists, because the derivative is bounded. Specifically, the derivative is

$$\frac{\partial \pi(p,z)}{p} = \sum \binom{N}{n} np^{n-1}(1-p)^{N-n} F_n(n - i_n - 1) - \sum \binom{N}{n} p^n (N-n)(1-p)^{N-n-1} F_n(n - i_n - 1) \qquad (3)$$

A supremum for the magnitude of Formula (3) can be obtained by assuming that F_n =1 (since $F_n \le 1$), and all the negative terms are zero. An equivalent magnitude is found by assuming that all the positive terms are zero. The maximum range for the sum is $0 \le n \le N$. Thus, equation (3) is bounded by

$$\pm \sum_{n=0}^{N} \binom{N}{n} np^{n-1}(1-p)^{N-n} \qquad (4)$$

From equation (2), when $p = 0$ or $p = 1, \pi = 0$. So, a supremum need only be found over the open interval $0 < p < 1$. The maximum of $p^r(1-p)^{t-r}$, over the open interval $0 < p < 1$, is $\left(\frac{r}{s}\right)^r \left(1 - \frac{r}{s}\right)^{s-r}$ [3]. Thus, for small N, the largest possible range of the slope of π can be found by substituting the maximum value of $p^n(1-p)^{N-n}$ for each occurrence in the sum in (4). Using this substitution technique, it was discovered that the slope of π is bounded by ± 212.20 for $N \le 29$.

A supremum for π of any desired accuracy can be obtained with the knowledge of bounds for the slope of π and the values of π at appropriately selected points. For example, to find a supremum for $\pi(p,z)$ which is no more than 0.002 greater than the maximum, calculate $\pi(p,z) + 0.001$ for every value of p from (0.001/212.20) up to (1 - 0.001/212.20) in steps of (0.001/212.20). The supremum is the maximum of the calculated values rounded up, in this case, to the nearest 0.001.

Critical Values and Sample Sizes

The exact, unconditional critical values, z, for $N < 30$ are given in Table 1 for one-sided tests with p<0.05, 0.025, and 0.01. Suprema were calculated to be within 0.001 of the maxima. Symmetrically, this table can also be used for the two-sided tests with p<0.10, 0.05 and 0.02. Based on the critical values shown in Table 1, the minimum number of matched pairs for the one-sided test was calculated to attain a power of at least 80% (Table 2) or 90% (Table 3) for a test of Pr(1•) > Pr(•1). The parametric notation of Suissa and Shuster [2] has been retained: Ψ = Pr(10) + Pr(01) and Δ = Pr(10) – Pr(01). Thus, a larger value of Ψ for the same Δ indicates a smaller value of Pr(10), and therefore, requires a larger N.

Table 1 can be directly compared to Table 2 in [2]. Some values appear for the first time as Suissa and Shuster did not present the critical values for $N < 10$. In addition, seven critical values which appear in both tables are different. Five of these differences, indicated by * in Table 1, are attributable to the maximum lying outside the range $0 \le p \le 0.995$. Two additional differences are found as indicated by the ** in Table 1. In these two cases, the suprema for the significance value were approximately 0.000039 below the desired p-value, and, as with all suprema in their software, (kindly provided by Dr. Suissa) were rounded up at the fourth decimal place. For example, a suprema of 0.024961 was not considered to be sufficient to satisfy $p < 0.025$. Such rounding is required for numerical calculations, but the software presented here rounded up at the 8th decimal place. The differences in critical values, both due to rounding and ranging, are reflected in differences in the sample sizes shown in Tables 3 and 4.

Illustration

Assume that a new defibrillation waveform must be clinically tested to determine whether it is more efficient than a standard waveform at terminating ventricular fibrillation (VF), which is an acute condition and is fatal if untreated. For example, Chapman et al. [4] described a laboratory study of this type. Chapman used upwards of twenty VF inductions and terminations per laboratory

Table 1

N	p<0.05	p<0.025	p<0.01
5	1.74		
6	1.74	2.01	
7	1.89	2.01	2.24
8	1.74	2.13	2.24
9	1.74	2.01	2.34
10	1.90	2.01	2.53*
11	1.74	2.12	2.34
12	1.74	2.01	2.34
13	1.74	2.01	2.50
14	1.74	2.14	2.31
15	1.81	2.01	2.33
16	1.74	2.01	2.51*
17	1.74	2.01	2.33
18	1.74	2.01	2.36
19	1.74	2.07	2.36
20	1.79	2.01	2.36
21	1.74	2.01	2.41
22	1.74	2.14*	2.33**
23	1.74	1.97	2.33
24	1.74	2.05	2.45
25	1.80*	1.97	2.33
26	1.74	1.97**	2.36
27	1.74	2.12*	2.36
28	1.74	1.97	2.36
29	1.74	2.05	2.42

* Different from [2] due to supremum calculation; ** Different from [2] due to rounding differences

Table 2

Δ	Ψ	p<0.05	p<0.025	p<0.01	Δ	Ψ	p<0.05	p<0.025	p<0.01
0.30	0.35	21	25-	*	0.50	0.55	11	15	17
	0.40	26	*	*		0.60	13	16	21
						0.65	15	18	23
0.40	0.45	15	18	22-		0.70	17	20	25
	0.50	17	23-	28		0.75	18	23-	28
	0.55	20	24	*		0.80	19	23	*
	0.60	23	26	*		0.85	21	25-	*
	0.65	25	*	*		0.90	22	28-	*
	0.70	27	*	*		0.95	24	*	*
	0.75	29	*						
					0.60	0.65	9+	11	14
						0.70	11	12	15
						0.75	11	15	18
						0.80	13	15	20
						0.85	14	17	22
						0.90	16	19	22-
						0.95	16	20	25

* N>30, Calculate the sample size using approximation formula; - Different from [1] due to critical value difference;
+ Different from [2] due to limited sample size calculations

Table 3

Δ	Ψ	p<0.05	p<0.025	p<0.01
0.30	0.35	28	*	*
0.40	0.45	19	23·	29
	0.50	23	28·	*
	0.55	27	*	*
0.50	0.55	14	17	22
	0.60	17	20	25
	0.65	20	23	*
	0.70	22	25·	*
	0.75	24	28	*
	0.80	26	*	*
	0.85	28	*	*
0.60	0.65	11	15	17·
	0.70	13	16	20
	0.75	15	18	22·
	0.80	17	20	25
	0.85	18	23·	26
	0.90	20	23	*
	0.95	21	25·	*

*N > 30, Calculate the sample size using approximation formula
· Different from [2] due to critical value difference

subject. Clinically, using more than one or two VF inductions for research purposes is uncommon because the danger to the patient increases rapidly as the number of inductions increases. Using the paired approach described here, the study could be designed to include two inductions and termination attempts per patient, one with the test stimulus waveform and one with a control stimulus waveform. When the same stimulus strength is used for both waveforms, the results are correlated matched-pairs of the type treated here, since some patients are easier to defibrillate than others, independent of waveform. If the test stimulus strength is selected to average about 40% effective in the population with the intent of finding waveforms that are dramatic improvements, say more than 80% effective in the population, such as those in [4], then $\Delta \approx 0.4$ and $\Psi \approx 0.56$. For 80% power, Table 2 indicates that a minimum of 20 patients should be planned for a one-sided test with $p < 0.05$.

Conclusion

When the cost--in dollars or otherwise--of each sample pair is great, small samples are naturally preferred. Furthermore, some studies may be impossible when the total number of sample pairs cannot be predicted *a priori*. Under such conditions, the exact, unconditional approach presented here offers a conservative alternative to the McNemar's test.

Acknowledgments

This work was supported a faculty research grant from The University of Memphis. Editorial assistance was provided by Katara Herron.

References

1. McNemar Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*, 153-157.

2. Suissa S., & Shuster J. J. (1991). The 2x2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics, 47(2)*, 361-369.

3. Suissa S., & Shuster J. J. (1985). Exact unconditional sample sizes for the 2x2 binomial trial. *J Royal Stat Soc, A148*, 317-327.

4. Chapman P. D., Vetter J. W., Souza J. J., Wetherbee J. N., & Troup P. J. (1989). Comparison of monophasic with single and dual capacitor biphasic waveforms for nonthoracotomy canine internal defibrillation. *J Am Coll Cardiol, 14*, 242-245.