

**APPROXIMATING POWER OF THE UNCONDITIONAL TEST
FOR CORRELATED BINARY PAIRS**

Grace R. Selicato	Keith E. Muller
Statistical Programming Systems	Dept. of Biostatistics, CB#7400
Quintiles, Inc., PO Box 13979	University of North Carolina
RTP, North Carolina, 27709-3979	Chapel Hill, North Carolina, 27599-7400

Key Words: 2×2 table; McNemar's Test

ABSTRACT

We provide a simple and good approximation of power of the unconditional test for two correlated binary variables. Suissa and Shuster (1991) described the exact unconditional test. The most commonly used statistical test in this setting, McNemar's test, is exact conditional on the sum of the discordant pairs. Although asymptotically the conditional and unconditional versions coincide, a long-standing debate surrounds the choice between them. Several power approximations have been studied for both methods (Miettinen, 1968; Bennett and Underwood, 1970; Connett, Smith, and McHugh, 1987; Connor, 1987; Suissa and Shuster, 1991; Lachenbruch, 1992; Lachin, 1992). For the unconditional approach most existing power approximations use the Gaussian distribution, while the accurate ("exact") method is computationally burdensome.

A new approximation uses the F statistic corresponding to a paired-data T test computed from the difference scores of the binary outcomes. Enumeration of all possible 2×2 tables for small sample sizes allowed evaluation of both test size and power. The new approximation compares favorably to others due to the combination of ease of use and accuracy.

1. INTRODUCTION

1.1 Motivation

In clinical trials one often faces the question of whether a binomial probability has changed due to treatment. If two binomial samples represent repeated measures then the resulting binomials are correlated. This situation is often referred to as the *unconditional* case in that neither row nor column marginal frequencies of the corresponding 2×2 table are fixed. Let N indicate the total number of pairs of observations. Table I summarizes notation for outcome frequencies, while Table II summarizes notation for outcome probabilities. Assuming $\mathcal{E}(c/N) = \pi_{10}$ and $\mathcal{E}(b/N) = \pi_{01}$ leads to testing $H_0: \pi_{10} = \pi_{01}$. Suissa and Shuster (1991) described how to compute p-values and power exactly for the unconditional test for two correlated binary variables.

Most of the work with 2×2 tables of correlated pairs has depended on assuming fixed marginal counts, and hence has been referred to as a *conditional* approach. As with the unconditional approach one tests $H_0: \pi_{10} = \pi_{01}$, but under the assumptions that $\mathcal{E}[c/(b+c)] = \pi_{10}$ and $\mathcal{E}[b/(b+c)] = \pi_{01}$.

A long and sometimes rancorous debate has surrounded the choice of analysis for 2×2 tables. The conditional approach requires only the counts of the discordant pairs, which allows two or more distinct configurations to provide the same statistic and same rejection regions. For example, should we treat two tables equivalently that differ only in total sample size? We believe that the debate reduces to a choice of assumptions, and that the choice should match the test to the sampling scheme used in the study. As Dozier and Muller stated (1993) "Conditional tests arise from defining the sample space in terms of the data being analyzed, while unconditional tests arise from defining the sample space in terms of hypothetical replicates of the experiment that generated the data being analyzed."

We seek a simple and accurate method for power analysis with the unconditional approach. Our interest arises from two sources. First, the approach appears appropriate for a wide range of data, including many applications in medical and behavioral science. Second, the extensive calculations needed for "exact" computations discourages more widespread use of the method.

In the unconditional case, we follow the lead of Suissa and Schuster (1991) and consider the McNemar statistic, which is the exact statistic for the conditional case. The difference between the conditional and unconditional settings lies in the description of the associated probabilities, and the corresponding computational

TABLE I

Outcome Counts

		Outcome 2		
		0	1	
Outcome 1	0	a	b	$a + b$
	1	c	d	$c + d$
		$a + c$	$b + d$	$N = a + b + c + d$

TABLE II

Outcome Probabilities

		Outcome 2		
		0	1	
Outcome 1	0	π_{00}	π_{01}	$\pi_{0\cdot}$
	1	π_{10}	π_{11}	$\pi_{1\cdot}$
		$\pi_{\cdot 0}$	$\pi_{\cdot 1}$	

difficulties. In the noncentral case the complexity arises due to the need to find the optimum of a function, with each value depending on a significant enumeration. See Suissa and Schuster (1991) for details.

1.2 Specification of the Problem

For correlated binary outcomes one usually seeks to test the difference between two proportions, each proportion representing the probability for which

the matched pairs disagree. Often the two dichotomous outcomes differ only by recording time, such as pre- and post-treatment measurements. The hypothesis of interest centers on the probabilities of discordant pairs. Under H_0 $\mathcal{E}(c/N) = \pi_{10}$ and $\mathcal{E}(b/N) = \pi_{01}$. Notation in Tables I and II allows stating $H_0: \pi_{10} - \pi_{01} = \delta = 0$, or $H_0: \pi_{1.} - \pi_{.1} = \delta = 0$. The second form arises because $\pi_{1.} = \pi_{10} + \pi_{11}$ and $\pi_{.1} = \pi_{01} + \pi_{11}$.

1.3 Related Work

McNemar's test, the exact conditional test for binary correlated pairs, converges asymptotically to the χ^2 test. The exact statistic depends only on the discordant pairs: $Q_m = (b - c)^2 / (b + c)$. The conditional method requires a larger sample size to achieve a fixed power for a fixed difference than does the unconditional method. Hence applying the conditional method to the unconditional setting yields a conservative test.

Suissa and Shuster (1985) studied the test size and power of unconditional tests. They suggested applying a maximization method to a conditional test to provide a least conservative test. Maximizing the null power function over the domain of a nuisance parameter gives the worst possible configuration, and yields a test which is never liberal. Suissa and Shuster (1991) supported Frisen's (1980) recommendation that the exact unconditional test be used, based on some unappealing properties of power for the exact conditional test. Frisen noted that under the conditional assumptions, the null hypothesis can be stated in terms of equivalent marginal probabilities ($H_0: \pi_{1.} = \pi_{.1}$) or diagonals ($H_0: \pi_{10} = \pi_{01}$). However, regarding power, "...influence of $\pi_{1.}$, $\pi_{.1}$ under H_0 on conditional power indicates that conditional power is not a suitable measure."

Most existing power approximations for the unconditional approach depend on the Gaussian distribution (Connett, Smith, and McHugh, 1987; Connor, 1987; Lachenbruch, 1992; Lachin, 1992; Miettinen, 1968; Suissa and Shuster, 1991). Bennett (1970) used the χ^2 goodness-of-fit statistic in terms of the multinomial likelihood. In addition, many unconditional sample size approximations are based on the conditional distribution under the null, and the unconditional under the alternative (Bennett and Underwood, 1970; Connett, Smith, and McHugh, 1987; Connor, 1987; Lachenbruch, 1992; Lachin, 1992; Miettinen, 1968). In contrast, both our approach (described in §2) and that of Suissa and Shuster (1991) use the unconditional distribution for both cases.

Suissa and Shuster (1991) described the exact unconditional test and computed p-values and power. As typically happens when starting with discrete random variables, the “exact” test merely guarantees test size no higher than the desired level. The test usually does not reach the target level test size of exactly α . Achieving a test with size as close to α as possible requires substantial computations. Naturally, power computations increase the burden. Hence power approximations have great appeal. The power of the best asymptotic method examined by Suissa and Shuster (1991) fluctuated as much as 14% in either direction from their computed “exact” power.

A similar problem occurs in comparing two independent binomial variables. The same conditional/unconditional distinction holds, and the same computational complexity arises. D’Agostino, Chase, and Belanger (1988) demonstrated that computing a T test on the outcomes coded as 1’s and 0’s leads to a very accurate approximation of the unconditional test in small samples. Dozier and Muller (1993) extended the results to the noncentral case by demonstrating similar excellent performance for power approximation. Asymptotically the exact test and approximations converge to the same test. In the same spirit, Lachenbruch (1992) mentioned using a paired-data T for testing correlated binomials.

2. A NEW METHOD

2.1 A New Approach for Power

We suggest approximating power of the unconditional test of correlated binary outcomes by using an appropriate paired-data T test. We do not recommend analyzing data in this way. The approach parallels that of Dozier and Muller (1993), and has the same two part motivation.

First, consider the contrast between using a Z test and a T test for the hypothesis of equality of Gaussian means. Asymptotically the two coincide. The T uses an additional parameter to account for the varying impact of a finite sample. For any particular design, the critical value for the T will always be larger than for the Z test. In turn, the power of the T will never be more than for the Z . Hence using a T rather than a Z will lead to a more conservative choice for sample size. In most applications a method that has modest conservatism, and rare optimism would be preferred over a method that balances optimistic and pessimistic values.

Second, the approximation suggested here shares the same desirable asymptotic features as existing methods. See Suissa and Shuster (1991) for a

discussion of asymptotic properties of McNemar's test, the unconditional test, and approximations. Standard arguments about multivariate linear models with independent and identically distributed observation vectors apply. The central limit theorem applies to the proportions interpreted as the sample means. Consideration of the alternative hypothesis involves examining a sequence of local alternatives (Sen and Singer, 1993, p238). the approach. The availability of simple asymptotically accurate approximations of power of the unconditional test leads to examining only small sample performance.

2.2 Calculation of the Test Statistic

Additional notation allows writing simple expressions for the statistic of interest. Let $E_{ij} \in \{0, 1\}$ indicate the outcome for the i th subject and j outcome. Define $D_i = E_{i1} - E_{i2}$, with $D_i \in \{-1, 0, 1\}$. $N_1 = c$ equals the number of positive discordant pairs and $\Pr\{D_i = 1\} = \Pr\{(E_{i1} = 1) \cap (E_{i2} = 0)\}$. $N_{-1} = b$ equals the number of negative discordant pairs, with $\Pr\{D_i = -1\} = \Pr\{(E_{i1} = 0) \cap (E_{i2} = 1)\}$. $N_0 = a + d$ equals the number of concordant pairs, with $\Pr\{D_i = 0\} = \Pr\{[(E_{i1} = 1) \cap (E_{i2} = 1)] \cup [(E_{i1} = 0) \cap (E_{i2} = 0)]\}$. We need not distinguish between a and d . Note that $N_1 + N_{-1}$ equals the number of discordant pairs and $N_1 + N_{-1} + N_0 = N$, the total number of pairs. Also note the following expressions for the sample mean and variance of D :

$$\bar{D} = \sum_{i=1}^N D_i / N = \frac{N_1 - N_{-1}}{N} = \hat{\delta} = \hat{\pi}_{10} - \hat{\pi}_{01} \quad (2.1)$$

and

$$S_D^2 = (N_1 + N_{-1}) / N - \bar{D}^2. \quad (2.2)$$

The new approach for a power approximation corresponds to the simple notion of performing a paired-data T test of the difference in outcome variables coded as 1's and 0's. It will be convenient to describe the test in the equivalent form of a one sample F test of mean difference. If $S_D^2 > 0$ then express the observed statistic of interest, corresponding to the usual least squares and Gaussian theory test, as

$$F_{obs} = \frac{\bar{D}^2}{S_D^2 / (N - 1)}. \quad (2.3)$$

Only two special cases lead to $S_D^2 = 0$. The case with $a + d = N$ (and $b = c = 0$), yields no discordant pairs and $\bar{D} = \hat{\delta} = \hat{\pi}_{10} - \hat{\pi}_{01} = \hat{\pi}_1$. -- $\hat{\pi}_{.1} = 0$. Set $F_{obs} = 0$, with p-value of 1, and do not reject the null hypothesis. The case with one discordant cell count of N (either $b = N$ or $c = N$) yields all discordant pairs., Set $F_{obs} = \infty$, with p-value of zero, and reject the null hypothesis.

2.3 Approximating Power

The test just described allows approximating power very easily. With $\delta = \pi_{10} - \pi_{01}$ and $\psi = \pi_{10} + \pi_{01}$ define

$$\omega = \frac{N\delta^2}{\psi - \delta^2}. \tag{2.4}$$

Let $F_F(f; \nu_1, \nu_2, \omega)$ indicate the cumulative distribution function of a noncentral F random variable with degrees of freedom ν_1 for the numerator, ν_2 for the denominator, and noncentrality ω . In turn let $F_F^{-1}(1 - \alpha; \nu_1, \nu_2, 0) = f_{crit}$ indicate the $(1 - \alpha)$ central quantile. Approximate power of a two-sided test with

$$P = 1 - F_F(f_{crit}; 1, N - 1, \omega). \tag{2.5}$$

For a one-sided test replace $(1 - \alpha)$ by $(1 - 2\alpha)$ in calculating f_{crit} .

3. ENUMERATION STUDIES

3.1 Enumeration Methods

Here we describe the enumeration of small sample behavior of the proposed statistic under the null and alternative hypothesis. Enumeration was preferred to simulation because it produces exact results. We calculated probabilities based on a trinomial distribution, for every possible 2×2 configuration given a number of total pairs (N), for a range of π_{10} and π_{01} , for both the null and non-null cases.

Using a one-sided test leads to observing only whether $N_1 \geq N_{-1}$. Critical values for the nominal α from the F distribution were used to evaluate each configuration for significance. The probabilities of the significant configurations were summed to give the attained test size (Table III) and sample size approximations (Table IV). Write the probability of a particular configuration as

$$\Pr\{N_1, N_{-1}, N_0; \pi_{10}, \pi_{01}\} = \frac{N!}{N_1!N_{-1}!N_0!} \pi_{10}^{N_1} \pi_{01}^{N_{-1}} (1 - \psi)^{N_0}. \tag{3.1}$$

Table III
Maximum Attained Test Size of the F Test
Compared to “Exact” Test Actual Size from Suissa and Shuster (1991)

N	$\alpha = .01$			$\alpha = .025$			$\alpha = .05$		
	π	F	“Exact”	π	F	“Exact”	π	F	“Exact”
10	.292	.0132	.0099	.241	.0265	.0208	.463	.0652	.0265
20	.314	.0119	.0071	.347	.0287	.0246	.498	.0557	.0396
40	.161	.0116	.0080	.030	.0269	.0250	.127	.0527	.0500
80	.081	.0115	.0094	.151	.0267	.0234	.064	.0522	.0499

The special case of the null hypothesis has $\pi_{10} = \pi_{01} = \pi$ and reduces (3.1) to

$$\Pr\{N_1, N_{-1}, N_0; \pi\} = \frac{N!}{N_1!N_{-1}!N_0!} \pi^{N_1+N_{-1}} (1-2\pi)^{N_0} \quad (3.2)$$

Note that when one assumes the null hypothesis is true, the distribution is symmetric. Send e-mail to the first author (GSelicat@Quintiles.Com) for a copy of the SAS® (version 6.08) program used for the enumerations.

3.2 Null Case Results

The test size attained by using the approximate unconditional statistic was compared with Suissa and Shuster’s (1991) results. Table III contains results for a one-sided test with $\alpha \in \{.01, .025, .05\}$. Since symmetry holds under the null case, a two-sided test with these values can be applied for a test with $\alpha \in \{0.02, .05, .10\}$. Values of π very near .50 represent “the highly unlikely and practically impossible scenario of the most negative correlation between [the two outcomes]” (Suissa and Shuster, 1991). Table III contains the attained test size with a supremum of π selected from the interval $(0, .995)$, with a precision of .001, the same method as Suissa and Shuster (1991). The columns labeled F and “Exact” contain results for the approximation proposed here and for the test described by Suissa and Shuster. Table III also contains the π at which the supremum occurred. For computational convenience, .498 was used as the π upper limit rather than the .4975 as used by Suissa and Shuster. The attained test size of the approximate test was a bit liberal, relative to the target α . Some liberality remains even with $N = 80$. The discrete nature of the data leads to the “exact” test usually being somewhat conservative. The test sizes of both tests grow closer to the nominal value as N increases.

3.3 Alternative Case Results

Table IV contains approximate sample sizes for the minimal number of pairs needed to attain power of at least .80. For a one-sided test, a nominal significance level of .01, .025, or .05 was used. The column labeled N_F contains sample sizes suggested by the approximate F approach, while the column labeled N_S contains sample sizes for the exact unconditional approach (Suisa and Shuster, 1991). Missing sample sizes have $N > 200$. Table IV results have $N_F \leq N_S$ (with the exception of some results for $\psi \leq 0.25$), which agrees with the test size results in Table III. The approximation usually provides a sample size an average of 2-3 units smaller than N_S . The new approximation appears to have less optimism than the suggestion of Miettinen (1968), and fluctuated less from the exact value than the approximation studied by Connett, Smith, and McHugh (1987) and Conner (1987). The exact conditional sample size proves very conservative as an approximation for the unconditional case (Suisa and Shuster, 1991). Overall, the accuracy of the power approximation improves with sample size and as power increases from .80 to .90 (results were computed but not tabled for .90).

4. DISCUSSION

4.1 Discussion of Results

Overall we saw a maximum of 8 and an average of 2-3 units of optimism in sample size approximations, as compared with the exact unconditional method (Suisa and Shuster, 1991). Hence we recommend merely increasing the approximate sample size by 3 units. The power optimism stems from the test size optimism. Consequently another approach would be to reduce the nominal test size, such as by using $\alpha \cdot [1 - (2/N)]$. Further research would be needed to develop and evaluate any such modification.

The algorithm used for the enumeration studies could be used to compute an exact version of the approximate test describe here. The computational burden would be modest for current desktop personal computers. The test would be exact in a similar sense as the test described by Suisa and Shuster: test size would be guaranteed to be no more than the nominal size, although usually less. Perhaps more importantly, the approach might be generalized to allow two or more groups with two correlated binary responses, or three or more repeated binary responses measures. Note that Agresti (1991) reported a method due to O'Brien for approximating power for general categorical data models.

Table IV
Minimum Sample Sizes Needed for Power of .80

δ	ψ	$\alpha = .01$		$\alpha = .025$		$\alpha = .05$	
		N_S	N_F	N_S	N_F	N_S	N_F
0.10	0.15	142	144	107	112	86	88
	0.20	196	194	152	152	118	119
	0.30					185	181
0.20	0.25	57	56	42	44	36	34
	0.30	69	68	53	53	43	42
	0.40	98	94	76	73	63	58
	0.50	126	119	97	93	76	73
	0.60	151	144	118	112	93	88
	0.70	177	169	136	132	108	104
	0.80		194	159	152	124	119
	0.90			179	171	139	135
0.30	0.35	34	32	26	25	21	20
	0.40	41	38	31	30	26	23
	0.50	53	49	42	38	34	30
	0.60	65	60	51	47	40	37
	0.70	76	71	60	56	49	44
	0.80	89	82	69	64	55	51
	0.90	101	94	78	73	65	58
	0.40	0.45	23	21	18	17	15
0.50		28	25	21	19	17	15
0.60		36	31	26	24	23	19
0.70		41	37	32	29	27	23
0.80		49	43	38	34	32	27
0.90		54	50	43	39	35	30
0.50		0.55	17	15	15	12	11
	0.60	21	17	16	14	13	11
	0.70	25	21	20	17	17	13
	0.80	30	25	23	20	19	16
	0.90	35	29	27	23	22	18
0.60	0.65	14	11	11	9	10	7
	0.70	15	13	12	10	11	8
	0.80	20	16	15	12	13	10
	0.90	23	18	19	14	16	11

4.2 Using the Power Approximation

The SAS® code below uses the F statistic to find approximate power.

```
DATA POWER;
  DELTA=0.20;
  PSI =0.45;
  ALPHA=0.05;
  N = 91 ;
  FCRIT=FINV(1-2*ALPHA,1,N-1);
  OMEGA = [N*(DELTA**2)] / [PSI - (DELTA**2)];
  POWER = 1 - PROBF(FCRIT, 1, N-1, OMEGA);
```

The code assumes a one-sided test. For a two-sided test replace “2*ALPHA” with “ALPHA”. The example code gives $FCRIT = 2.7621$, $OMEGA = 8.8781$, and $POWER = .9053$. The computational efficiency of the approximation allows conveniently examining a wide range of scenarios. For example, a plot of power for a range of alternatives provides a very informative display and one extremely well received by scientists.

4.3 Conclusions and Recommendations

When demanded by the sampling situation, we suggest testing the hypothesis for correlated binary pairs with the exact unconditional test. The F approximation described here provides a convenient and reasonably accurate method of approximating the corresponding power, with adjustments as noted above.

ACKNOWLEDGMENTS

An earlier version of this paper was submitted by the first author in partial fulfillment of the requirements for the M. S. degree in Biostatistics. Muller's work supported in part by NCI program project grant P01 CA47 982-04, NIH Clinical Research Center grant M01 RR000-46-33, and NIEHS grant N01-ES-35356. The authors gratefully acknowledge helpful comments on an earlier draft of this paper by Lisa M. LaVange and Susan Kenny.

BIBLIOGRAPHY

- Agresti, A. (1991). *Categorical Data Analysis*, New York: Wiley.
- Bennett, B. M., and Underwood, R. E. (1970). “On McNemar's test for the 2×2 table and its power function,” *Biometrics*, **26**, 339-343.

- Connett, J. E., Smith, J. A., and McHugh, R. B. (1987). "Sample size and power for pair-matched case-control studies," *Statistics in Medicine*, **6**, 53-59.
- Connor, R. J. (1987). "Size for testing differences in proportions for the paired sample design," *Biometrics*, **43**, 207-211.
- D'Agostino, R. B. Chase, W., and Belanger A. (1988) "The appropriateness of some common procedures for testing the equality of two independent binomial populations," *The American Statistician*, **42**, 198-202.
- Dozier, W. G. and Muller, K. E. (1993). "Small-sample power of uncorrected and Satterthwaite corrected t tests for comparing binomial proportions," *Communications in Statistics: Simulation and Computation*, **22**, 245-264.
- Frisen, M. (1980). "Consequences of the use of conditional inference in the analysis of a correlated contingency table," *Biometrika*, **67**, 23-30.
- Lachenbruch, P. A. (1992). "On the sample size for studies based upon McNemar's test," *Statistics in Medicine*, **11**, 1521-1525.
- Lachin, J. M. (1992). "Power and sample size evaluation for the McNemar test with application to matched case-control studies," *Statistics in Medicine*, **11**, 1239-1251.
- Miettinen, O. S. (1968). "The matched-pairs design in the case of all-or-none responses," *Biometrics*, **24**, 339-352.
- Suissa, S., and Shuster, J. J. (1985). "Exact unconditional sample sizes for the 2×2 binomial trial," *Journal of the Royal Statistical Society, A* **148**, 317-327.
- Suissa, S., and Shuster, J. J. (1991). "The 2×2 matched pairs trial: exact unconditional design and analysis," *Biometrics*, **47**, 361-372.
- Sen P. K., and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*, New York: Chapman and Hall.

Received December, 1996; Revised October, 1997.