

# A New Exact and More Powerful Unconditional Test of no Treatment Effect from Binary Matched Pairs

BY CHRIS J. LLOYD

*Melbourne Business School, Carlton, 3053, AUSTRALIA*

c.lloyd@mbs.edu

## SUMMARY

We consider the problem of testing for a difference in the probability of success from matched binary pairs. Starting with three standard inexact tests, the nuisance parameter is first estimated and then the residual dependence is eliminated by maximisation, producing what I call an E+M P-value. The E+M P-value based on McNemar's statistic is shown numerically to dominate previous suggestions, including partially maximised P-values as described in Berger and Sidik (2003). The latter method however may have computational advantages for large samples.

# 1 Introduction

The data that arise from medical trials are often expensive. Extracting the maximum information from the data is therefore important. Since the data are scarce, asymptotic methods will behave poorly. Moreover, the chance of error will often need to be strictly controlled so that the Bayesian paradigm is not appropriate. For all these reasons, exact frequentist inference from small samples of counts has become a rich area of research over recent years, especially in this journal, see for instance Chan et al. (2003), Mehrotra et al. (2003), Tang et al. (2004) and Agresti and Min (2005ab).

The specific design of matched binary pairs arises in many statistical contexts. In case control studies, we stratify a population with respect to some control variable and randomly choose one individual per stratum to be given the treatment and one the control. In twin studies, one twin is randomly chosen for treatment and the other control. In repeated measure studies, we measure binary traits before and after some event or treatment of interest, on the same individual. In each case, the basic measurement is a bivariate variable  $Y = (Y_1, Y_2)$  measured on each stratum, or twin pair or individual. We take the binary response to be a positive outcome, such as survival. The parameter of interest  $\theta$  is the probability  $p_2$  of response with treatment 2 applied minus the probability  $p_1$  of response with treatment 1 (often no treatment or control).

Table 1 displays data used by Berger and Sidik (2003), which we use for illustration. The notation will be more fully explained in the next section. From  $n = 30$  subjects, 13 responded positively with treatment 1 which increased to 20 with treatment 2. The 15 individuals who give identical response on both occasions would seem to tell us nothing about relative merits of the two treatments, and we will see later that standard test statistics depend only on the number  $t = 15$  whose responses are “discordant” and the

number of these  $x_{01} = 11$  which are in favour of the treatment 2.

There is a large literature on testing equality of  $p_1$  and  $p_2$ , beginning with the test of McNemar (1947). An obvious alternative is the likelihood ratio (LR) statistic. The conditional likelihood is free of the main nuisance parameter in this problem, namely the probability of a discordant pair  $\phi$ . This leads to the sign test which has been criticised as being conservative (Suissa and Shuster, 1991).

When these tests are viewed unconditionally, the null distributions used to calculate the P-value necessarily involve the nuisance parameter  $\phi$ . We consider several methods of controlling this dependence. For matched pairs, Suissa and Shuster (1991) were the first to study unconditional tests and used the standard method of maximising the P-value over  $\phi$ . More recently, Berger and Sidik (2003) have used the alternative device of partial maximisation over  $\phi$  with penalty, as described in Berger and Boos (1994). This requires ad-hoc choice for the range of partial maximisation. Our preferred method is estimation followed by maximisation. The idea behind this methodology is to remove most of the dependence on the nuisance parameter by estimating it, and to control the remaining dependence by maximisation.

The aim of the paper is to construct a test which is exact and more powerful than other exact tests. The definition of exactness used is different from the traditional weak definition which merely requires the size of the test to be bounded above by the nominal size. In contrast, our definition of exactness specifies a property of a P-value  $P(Y)$ . Specifically we look at the significance  $\Pr(P(Y) \leq P(y) : \theta_0, \phi)$  as a function of  $\phi$  and we say that the P-value (and implied test) is exact if the supremum of this function exactly equals  $P(y)$ . In simple words, the P-value has the property of a uniform random variable for each possible  $y$  and for some value of  $\phi$ , see Lloyd (2007) for further details. It is not essential to accept this definition of exactness in order to

accept the new test proposed in the paper.

The paper is organised as follows. In Section 2, we establish the notation and discuss some important issues of conditionality and in Section 3 we present the three standard tests mentioned above. In Section 4, we describe in detail three methods for dealing with nuisance parameters as well as computational issues. In Section 5, we combine these three methods with the three standard tests and compare the properties of the resulting exact tests by complete enumeration rather than simulation. We evaluate the efficiency of these exact tests by the mean value of the P-value (smaller is better) and the power of the implied test for fixed size. For each basic test statistic, estimation followed by maximisation dominates maximisation clearly and partial maximisation by a lesser amount. While there is an extra cost in computation, this is not a practical issue for the kinds of sample sizes that are common in practice. Of the three basic statistics, McNemar's statistic is slightly preferred. In Section 6, we investigate the tests in the context of two examples.

## 2 Notation and model

The bivariate random variable  $Y = (Y_1, Y_2)$  takes values in  $\{00, 01, 10, 11\}$  and we denote by  $X_{jk}$  the number of individuals out of  $n$  for whom  $Y = (jk)$  and by  $\pi_{jk}$  the corresponding probabilities. The distribution of  $X_{jk}$  is multinomial and the full likelihood for the unknown probabilities  $\pi_{jk}$  is

$$L(\pi|x) \propto \pi_{11}^{x_{11}} \pi_{10}^{x_{10}} \pi_{01}^{x_{01}} \pi_{00}^{x_{00}}.$$

We are interested in comparing the probability  $p_2 = \pi_{11} + \pi_{10}$  of response under treatment 2 with the probability  $p_1 = \pi_{11} + \pi_{01}$  of response under treatment 1. The difference  $\theta = p_2 - p_1$  is the interest parameter and can also be expressed  $\pi_{01} - \pi_{10}$ .

*Likelihood factorisation.*

The multinomial model can be re-expressed in terms of three binomial factors. The first factor corresponds to observing  $t = x_{10} + x_{01}$  discordant pairs out of  $n$ . The probability of a discordant pair is denoted by  $\phi = \pi_{01} + \pi_{10}$ . Conditional on this, we observe the break-up of the  $t$  discordant pairs and the  $n - t$  concordant pairs. The number of outcomes  $x_{01}$  favourable to treatment 2 amongst the  $t$  discordant pairs is binomial with parameters  $t$  and probability  $\eta = \pi_{01}/(\pi_{01} + \pi_{10})$ . Note that  $\eta$  can be expressed as  $\eta = (\theta + \phi)/(2\phi)$  and the null hypothesis  $\theta = 0$  corresponds to  $\eta = 0.5$ . The number of favourable outcomes  $x_{11}$  amongst the  $n - t$  concordant pairs is binomial with parameters  $n - t$  and probability  $\psi = \pi_{11}/(\pi_{00} + \pi_{11})$ . The likelihood is

$$L(\theta, \phi, \psi | x_{01}, t, x_{11}) \propto B(t; n, \phi) B(x_{01}; t, (\theta + \phi)/(2\phi)) B(x_{11}; n - t, \psi) \quad (1)$$

where  $B(x; n, p)$  is the probability of a binomial with parameters  $(n, p)$  equaling  $x$ .

*Model reduction.*

It seems to be universally accepted, typically without comment, that this three dimensional model be reduced to a two-dimensional model for  $(X_{01}, T)$ , by ignoring the last factor  $B(x_{11}; n - t, \psi)$ . There are several arguments in favour of this, the most common being an appeal to intuition that the breakup of the  $n - t$  concordant individuals into those who responded 00 and 11 is irrelevant to whether or not the two treatments are the same. More formal arguments involve looking at the factorisation of the likelihood in (1). Some thirty years ago, there were attempts to specify when a likelihood factor can and cannot be ignored, see Sandved (1967), Spratt (1975), Barndoff-Nielsen (1973), Kalbfleisch (1975). These papers sought to establish necessary and sufficient conditions to reduce a model. No single set of agreed conditions emerged, though there are several agreed necessary conditions.

In our case, the third factor of the likelihood satisfies the main necessary conditions: The parameter  $\psi$  is variation independent of the remaining parameters  $(\theta, \phi)$  which means that fixing a value of one parameter does not restrict the range of the other/s. So even knowing the value of  $\psi$  exactly gives no information about  $(\theta, \phi)$ . The statistic  $(X_{01}, T)$  is sufficient for  $(\theta, \phi)$  and  $X_{11}$  is sufficient for  $\psi$ . This means that  $X_{11}$  contains no information about  $(\theta, \phi)$  and that conditioning on it (which here is equivalent to ignoring it) loses no information about  $(\theta, \phi)$ . The author is not aware of any test of  $\theta = 0$  that uses the third factor of the likelihood.

*Further model reduction.*

The first factor of (1) depends on  $\phi$  only and the second factor on  $\eta$  only, raising the possibility of basing inference on the second factor alone, which is the model for  $X_{01}$  conditional on  $T$ . This leads to the sign test described in the next section. The arguments for this further conditioning are less clear. Certainly, confidence limits for  $\theta$ , or tests of non-zero values, must necessarily be based on the unconditional model, see Lloyd and Moldovan (2007) for methods of constructing exact confidence bounds for  $\theta$ . It is only for testing the specific hypothesis  $\theta = 0$  that conditioning is even contemplated. A heuristic argument against ignoring the first factor of (1) is that the number of individuals  $t$  for whom the response under treatment and control is different is surely not irrelevant to judging whether the treatment and control are different.

Clearly,  $\theta$  and  $\phi$  are not variation independent since  $|\theta| \leq \phi$  though  $\theta$  and  $\eta$  are. Neither is  $X_{01}$  sufficient for  $\theta$  or  $\eta$ . The issue of conditioning on  $T$  in this context has been addressed by Frisen (1980) who derived some undesirable properties of the conditional power function and its consequences on inference and concluded that “unconditional inference should be used” by which we mean the model for  $(X_{01}, T)$  rather than the model for  $X_{01}$  conditional on  $T$ .

### 3 Three standard inexact tests

In the sequel, we use  $X$  for the number of discordant pairs that are favourable to the treatment rather than  $X_{01}$ . A very simple test follows from treating  $T$  as fixed, in which case  $X$  has binomial  $(t, \eta)$  distribution. To test  $\eta = 1/2$ , which is identical to  $\theta = 0$ , we use the standard sign test with one and two-sided P-values

$$P_{\text{one-sided}}(x, t) = \sum_{i=x}^t B(x; t, 1/2), \quad P_{\text{two-sided}}(x, t) = 2 \sum_{i=0}^{\min(x, t-x)} B(x; t, 1/2).$$

If it is agreed that inference should be carried out conditional on  $T = t$  then this test has well-known exact properties and has no serious competitors. However, if  $T = t$  is not considered fixed then the conditional sign test is one of several competing tests.

We henceforth assume the joint model for  $(X, T)$  and call this the unconditional model. All tests, including the conditional sign (CS) test, will be evaluated within this framework. The unconditional likelihood

$$L(\theta, \phi|x, t, n) \propto \phi^t (1 - \phi)^{n-t} \eta^x (1 - \eta)^{t-x} \propto (1 - \phi)^{n-t} (\theta + \phi)^x (\phi - \theta)^{t-x}$$

generates two further tests. The score statistic  $(2X - T)/\sqrt{T}$  was proposed by McNemar (1947). We refer to tests based on this statistic with the label MC. The generalised likelihood ratio is  $2(\log L(\hat{\theta}, \hat{\phi}) - \log L(0, \hat{\phi}_0))$  where  $\hat{\phi}_0$  is the ML estimator of  $\phi$  when  $\theta = \theta_0$ , which for the case  $\theta_0 = 0$  is identical to the unrestricted ML estimator  $\hat{\phi} = t/n$ . The LR statistic simplifies to

$$\text{LR}(x, t) = t \log 2 - 2t \log t + 2x \log x + 2(t - x) \log(t - x)$$

with  $0 \log 0 := 0$ . The one-sided test statistic is  $\text{sign}(2X - T)\text{LR}(X, T)^{1/2}$ .

Under the null hypothesis, MC and LR statistics are asymptotically equal for large  $n$  using standard likelihood arguments. They generate approximate one and two-sided P-values based on a standard normal approximation to their null distributions. A normal

approximation to the binomial  $(t, 1/2)$  distribution without continuity correction gives the MC statistic so MC and CS are also asymptotically equal under the null hypothesis, but only for large  $t$ . The three statistics can differ substantially for small samples and even more so under the alternative hypothesis.

Any reasonable test must declare no evidence against the null for the data set  $(x, t) = (0, 0)$ . For the tests given above, we impose the standard convention that  $P(0, 0) = 0.5$  for one-sided tests and  $P(0, 0) = 1$  for two-sided tests. Consequently, all reasonable (non-randomised) tests are biased since their power converges to zero as  $\phi \rightarrow 0$ . This should be kept in mind for some of the later power and mean value comparisons. All three tests listed above depend on the data only through  $(x, t)$  and do not depend on  $n$  at all, even though the likelihood depends on  $(x, t, n)$ . The exact tests derived later all depend on  $n$  and we will consider the sometimes erratic dependence in the final section. Our recommended test depends on  $n$  only slightly and also smoothly.

Any sensible one-sided P-value for testing  $\theta > 0$  should also be a non-increasing function of  $X$  and a non-decreasing function of  $T$ . This property is sometimes called Barnard convexity, see Berger and Sidik (2003). It is simple but tedious to show that each of the one-sided P-values for tests CS, MC and LR satisfy this property. Apart from logic, the monotonicity property has two important computational consequences described in the next section.

## 4 Dealing with nuisance parameters

All three inexact tests in the previous section give rise to a P-value. Using  $Y$  now, quite generally to denote the entire data vector, any hypothesis test can be expressed in terms of a P-value  $P(Y)$  and a target size  $\alpha$ . For continuous models,  $P(Y)$  should be uniformly distributed under the null and the target size  $\alpha$  can be achieved exactly.

For discrete models,  $P(Y)$  cannot be uniform and typically no test will achieve the target size  $\alpha$ . In this sense, all discrete tests are inexact. A more realistic aim then is that the P-value have the property that  $\Pr(P(Y) \leq P(y)|H_0) = P(y)$ . This gives the P-value its interpretation. If this property holds for all  $y$  then we might still call the test exact. The one-sided CS test has this property when viewed conditionally, even though a pre-specified target size cannot be achieved exactly.

When nuisance parameters are present, even this weaker kind of exactness is usually unachievable. This is because, for all but trivial examples, the null tail probability

$$\pi(y, \phi) := \Pr(P(Y) \leq P(y); \theta_0, \phi) \tag{2}$$

depends on  $\phi$  and so cannot equal  $P(y)$ . Dependence on  $\phi$  can be extreme and  $\pi(y, \phi)$  can violate the target value  $P(y)$  grossly in some examples, see Berger & Boos (1994). I call  $\pi(y, \phi)$  the significance profile. If  $\pi(y, \phi) \leq P(y)$  for all  $y$  then the implied test will be guaranteed not to exceed the nominal size, and I will call the P-value ‘guaranteed’. Other authors, for instance Berger & Boos (1994), call this property ”validity” meaning that the implied test does not violate the size requirement. If the supremum of the profile equals  $P(y)$  then I will say that the P-value is exact, and more loosely that the test is exact. This weaker definition of exactness is always achievable and it is what we will always mean when we use the term exact. Our aim is to construct exact P-values that are as efficient as possible, where efficiency is defined in section 6.

Consideration of the significance profile leads to three different P-values. The theory is identical for one or two-sided tests. Henceforth, the data  $Y = (X, T)$ . Computational issues are addressed in the following section.

The first approach is *maximisation*. We calculate  $P^*(x, t) := \sup_{\phi} \pi(x, t, \phi)$  and use this as the P-value, see p168 of Bickel and Doksum (1977). Lloyd (2005a) points out

that  $P^*(X, T)$  is as small as possible amongst valid P-values that are non-decreasing functions of the original statistic  $P(X, T)$ . I call this the M (for maximisation) P-value.

The second approach is *partial maximisation*, motivated by examples where maximised P-values appear to be inefficient. Berger and Boos (1994) have suggested

$$P_\gamma(x, t) = \sup\{\pi(x, t, \phi) : \phi \in C_\gamma\} + \gamma,$$

where  $C_\gamma$  is a  $100(1-\gamma)\%$  confidence interval for  $\phi$ . This methodology has been applied in several recent papers, see for instance Berger and Sidik (2003) and Mehrotra, Chan and Berger (2003). I will call this the B P-value. Berger and Boos acknowledged that dependence of results on the choice of  $\gamma$  can be extreme, notwithstanding the general recommendation by Berger and Sidik that  $\gamma$  be small. Partially maximised P-values are not exact but are slightly conservative in all but trivial cases, see again Lloyd (2005a).

The third and much older approach is *estimation*. This involves replacing  $\phi$  by its maximum likelihood estimate  $\hat{\phi}_0$  under the null. This estimated P-value is denoted by  $\hat{P}(x, t) = \pi(x, t, \hat{\phi}_0)$ . While  $\hat{P}(X, T)$  is not exact, it becomes exact after application of the maximisation step. This involves maximising the profile (2) but with  $\hat{P}(x, t)$  replacing  $P(x, t)$ . The profile of  $\hat{P}(X, T)$  tends to be flatter than the profile of the original  $P(X, T)$ , see Lloyd (2005a). So our third method is estimation followed by maximisation. This produces what I denote the E+M P-value.

## 5 Computational issues

Computation of all three P-values in the previous section involves starting with a generating P-value (for instance one of the standard P-values in Section 3) and computing the significance profile (2). There are several computational issues. The first is to identify the subset of the sample space  $\{(x', t') : P(x', t') \leq P(x, t)\}$ . The second is to

compute the null probability of this tail set using the unconditional model

$$\Pr((X, T) = (x, t); \theta, \phi) = B(x; t, \eta)B(t; n, \phi), \quad (x, t) \in \mathcal{S}_n, (\theta, \phi) \in \Omega \quad (3)$$

where  $\mathcal{S}_n = \{(x, t) : 0 \leq x \leq t \leq n\}$  and  $\Omega = \{(\theta, \phi) : 0 \leq |\theta| \leq \phi \leq 1\}$ . This gives  $\pi(x, t, \phi)$  for a single value of  $\phi$ . The third is to maximise or partially maximise the significance profile with respect to  $\phi$ .

In the most general case, determining the tail set requires evaluating  $P(x', t')$  for all  $N = n(n+1)/2$  elements of the sample space. However, when  $P(x, t)$  is monotonic in one argument computation reduces from  $O(n^2)$  to  $O(n \log n)$ .

Computing the tail set for the E+M P-value is more computationally intensive since it requires  $O(n \log n)$  evaluations not of  $P(x, t)$  but of  $\hat{P}(x, t)$ . A single evaluation of  $\hat{P}(x, t)$  itself takes  $O(n \log n)$  evaluations of  $P(x, t)$  and so  $O(n \log n)^2$  evaluations of  $P(x, t)$  are required in total. This is the most demanding part of the computation and means that E+M P-values are computationally more demanding than M or B P-values, for large enough sample size. The good news is that the tail set only needs to be determined once.

Calculating the probability of the tail set is also  $O(n^2)$  in general but when the tail set is monotonic it can be computed as a sum of  $n$  cumulative binomial probabilities. Since cumulative binomial probabilities can be computed using the gamma function, regardless of  $n$ , the second step is  $O(n)$ .

Maximisation of  $\pi(x, t, \phi)$  requires multiple evaluations of the tail probability. The number of evaluations  $K$  depends on the degree of accuracy and confidence required. Since  $\pi(x, t, \phi)$  has no special properties such as convexity, finding the maximum will always be problematic. This issue is unacknowledged in the literature. We make the following comparative observations. For many standard examples and test statistics,

the profile can contain narrow spikes which are easy to miss and so an intense search effort is required. The B P-value requires maximisation over a restricted region of size  $O(n^{-1/2})$ , but this does not reduce the chance of missing the maximum. A schematic listing of the algorithm used for calculating the P-values compared in this paper is given in Web appendix A.

For practical sample sizes, computation times for M and B P-values are almost identical. For the E+M P-value, a much smaller number of evaluations can be used because the significance profile of  $\hat{P}(X, T)$  tends to be much better behaved than  $P(X, T)$ . Across a range of different examples, Lloyd (2005a) has found that in every case where the original profile has a problematic spike, this is not present in the profile of  $\hat{P}(X, T)$ . For the present application, I used  $k = 1000$  but in every case I the same maximum could be found with  $k = 10$  grid points. So the maximisation part of the computation is much easier for E+M P-values than for either M or B P-values.

One remaining issue is whether or not  $\hat{P}(X, T)$  has the same monotonicity properties as  $P(X, T)$ , as has been implicitly assumed in the discussion above. It is simple to show that if  $P(x, t)$  is monotone in  $x$  for fixed  $t$  then so is  $\hat{P}(x, t)$  (see Web Appendix B) which is sufficient for the computational savings above. However, for the supremum of the profile to be achieved on the boundary of the null hypothesis (i.e. when  $\pi_{01} = \pi_{10}$  rather than  $\pi_{01} < \pi_{10}$ ) monotonicity in  $t$  for fixed  $x$  is also required (Berger & Sidik, 2003). We have found numerically that  $\hat{P}(x, t)$  is monotone in  $t$  for fixed  $x$  for all sample sizes considered in the next section.

## 6 Numerical study

There are 9 possible one-sided P-values to compare, namely those based on the basic generating statistics MC, LR and CS, with the M, B or E+M transformations applied.

There are another 9 two-sided P-values. For the B P-values, we take  $\gamma = 0.0005$  as suggested in Berger and Sidik (2003). Evaluation of competing P-values is based on computing all their possible values for  $n = 20, 50, 100$ . No simulation is involved. All P-values are guaranteed and so can be meaningfully compared without worrying, for instance, that one test's extra power may be explained by its being more liberal.

Two criteria are used to evaluate the competing P-values. First, we prefer P-values to be systematically small rather than large. Certainly, if one P-value is smaller than another for all data sets and both are guaranteed we would prefer the smaller one, since it is bound to dominate the other in power for all parameter values and all nominal sizes. We measure smallness by the null mean value  $E(P(X, T); \theta = 0, \phi)$ , using the product binomial distributions (3) with  $\eta = 0.5$ . This mean value depends on  $\phi$  and tends to unity as  $\phi \rightarrow 0$  because of the earlier mentioned bias of all reasonable tests as  $\phi \rightarrow 0$ . This measure has the advantage that it does not depend on a nominal size  $\alpha$ .

The second measure is more standard, namely the power surface

$$\beta_{\alpha}(\theta, \phi) = \Pr(P(X, T) \leq \alpha; \theta, \phi)$$

of the test that rejects the null when  $P(X, T) \leq \alpha$ , for the standard sizes  $\alpha = 5\%$  and  $10\%$ . For discrete models, one can find that power varies with sample size in an erratic manner, depending on whether an observable P-values falls to one side or the other of the nominal  $\alpha$ . Nevertheless, this is in some sense a bottom line measure and general patterns emerge over the three sample sizes considered.

*M, B or E+M?*

Figure 1 gives two representative plots of  $E(P(X, T); \theta = 0, \phi)$ , illustrating what are very consistent patterns across differing sample sizes, basic statistics and alternatives. Further results may be found in Lloyd (2005). Across six tests (three generating sta-

tistics and two kinds of alternative) and three samples sizes, E+M P-values are much smaller in mean than M P-values, especially in the practically interesting range where  $\phi$  is not too close to 0 and less than 0.5. B-values achieve much but not all of this improvement.

The smaller mean value associated with the E+M P-values does indeed translate into higher power which was calculated over an even  $100 \times 100$  grid of values for  $\phi$  and  $\theta/\phi$ . We chose the latter parameter since it can vary over  $[0, 1]$  for any value of  $\phi$  and also because it equals  $2\eta - 1$  which is a natural measure of deviation from the null. All powers were computed correct to seven decimal places. For sample sizes 50 and 100, and for all 18 tests, the power of the E+M test was at least as good as the B or M test over 100% of the parameter space. Table 2 gives the proportion of the 10,000 points in the parameter space where the E+M test strictly dominated either the B or M test. For the majority of the parameter space the E+M test does strictly dominate though in many regions of the parameter space the power dominance is very small. It is difficult to see systematic differences between the two sample sizes but it should be noted that the relative properties of the tests are sensitive to the nominal size chosen.

*McNemar, LR or Conditional Sign test?*

Having decided on E+M based P-values, there remains the choice between the three generating statistics. It should be noted that the generating statistics affect the E+M P-values only through the way they order the sample space. The McNemar and LR statistics in particular give almost identical orderings and it is found empirically that the E+M P-values based on these statistics are virtually identical, see Lloyd (2005) for further details. I was not able to discover any practical difference between the mean values or the powers at fixed nominal levels of 5% or 10% of tests based on these alternative statistics. Based on its relative simplicity, we would recommend E+M

P-values based on MC rather than LR.

There are some larger but still modest differences between E+M P-values based on MC and CB. Figure 2 displays plots of  $\beta_\alpha(\theta, \phi)$  for the case of 1% one-sided tests for specific values of  $\theta$  and  $\phi$ . There are many parameter constellations where the powers are identical, for instance with  $n = 50$  and  $\alpha = 10\%$  the two tests are formally identical. I could find no constellations where CS dominates MC. The MC test tends to dominate CS more clearly when both  $\alpha$  and  $\phi$  are small. It is also a fact that the mean value of the CS based P-values is higher than for the MC based P-values so that across differing nominal sizes and samples sizes we would expect MC to dominate CS more often than the reverse. Further plots and results may be found in Lloyd (2005).

## 7 Illustration and Conclusions

We illustrate on two examples. The first was given in Table 1 of the introduction. In our present notation the data is  $(x, t, n) = (11, 15, 30)$  and the ML estimate of  $\theta$  is  $7/30=0.233$ . McNemar's statistic equals 1.807 with approximate one-sided P-value 0.0353 but this test is known to be liberal and maximised P-value is larger, at 0.0494. Other P-values and equivalent normal quantiles are in Table 3. All P-values below the line are guaranteed to give valid tests. The estimated P-value is 0.0377 and the E+M P-value is virtually identical to four significant figures. The Berger and Sidik P-value here equals the E+M P-value plus the penalty 0.0005.

The second example is from a study by Brix et al. (2005) into the association of auto-immune thyroid disease (AITD) with a possible risk factor known as X-chromosome inactivation XCI. The researchers recruited 26 twin pairs, one of whom suffered from AITD and the other not. The twins form a matched pair with one each in the treatment (disease) and control (disease free) group. The subjects were then

tested for the risk factor XCI. There were 10 twin pairs where only one had the XCI risk factor (discordant in our terminology) and 16 concordant twin pairs. Out of the 10 discordant pairs, nine pairs had the risk factor XCI in the diseased twin and only one pair had the risk factor in the disease free twin. So the data is  $(x, t, n) = (9, 10, 26)$ . The study quoted the two-sided P-value 0.022 from a two-sided conditional sign test.

McNemar's test gives a Z-value of 2.589 and a two-sided P-value of 0.0114, while the LR test gives a two-sided P-value of 0.0066. Because the sample size is small, the tests are in considerable disagreement about the weight of evidence for an association and the CS test gives a less significant result, consistent with claims that it is conservative. However, the McNemar and LR P-values are not guaranteed and we have already seen that tests based on these tend to be liberal.

Various inexact and exact P-values are listed in Table 3. For heuristic purposes, the left panel of figure 3 displays the significance profile for MC. It is noteworthy that the profile is uniformly less than the asymptotic P-value 0.0114 so that 0.0114 is conservative for this data set, despite the general tendency for McNemar's test to be liberal. The M P-value is 0.0110 achieved towards the right of the plot. The E P-value is 0.0097 obtained at the indicated value  $10/26=0.386$  for  $\phi$ . The E+M P-value is 0.0098 and cannot be read off this plot but only took 0.2 seconds to compute. It is worth noting that the exact P-values obtained based on CS, MC or LR generating P-values are all very close to 0.01, consistent with our general numerical results which suggested that after the E+M transformation it makes little difference which generating statistic you started with. For CS and MC, the B P-value of 0.0103 equals the E+M P-value plus the 0.0005 penalty but for LR it is somewhat larger.

One might ask whether it is necessary to maximise the estimated P-value. Certainly in this example it makes no practical difference. The right panel of figure 3 describes

all possible data sets for  $n = 26$  but is typical of all sample sizes. While there are many data sets for which the E and E+M P-values are very close, there are data sets where they are practically different and the tendency is for E P-values to be smaller than E+M P-values due to their being liberal. There seems to be no method of anticipating when the E+M P-value will differ from the E P-value.

It is of interest to see how inference depends on the sample size  $n$ , recalling that the standard CS, LR and MC test depend only on  $(x, t)$  and not on  $n$  at all. For example 1, we will consider how P-values based on the MC generating P-value change when the number of subjects equals other values than  $n = 30$ . Intuitively, we would expect that as  $n$  increases the effects of discreteness might dissipate and so the P-value may slightly reduce. We would also expect and require that the P-value change smoothly with  $n$ . This issue does not seem to have been addressed at all in the literature on exact tests of matched pairs.

Figure 4 shows P-values based on the M, E+M and B methods for a range of sample sizes. The dependence of the M P-value on sample size is quite erratic. This behaviour is generated by the erratic manner in which those elements of the sample space that generate peaks in the profile enter and leave the tail set  $\{P(X, T) \leq 0.0353\}$ . In contrast, the E+M P-value more or less smoothly decreases with sample size. The B P-value largely shares this desirable behaviour except for the sample size  $n = 18$ .

The conclusion of this investigation is that E+M P-values based on MC are recommended, being virtually identical to those based on the more complicated LR statistic and tending to have a smaller mean value than CS based P-values. When computation cost becomes extreme (say  $n > 200$ ), the B P-value can be recommended. When this is computationally costly which will only be the case for sample sizes in the thousands then the E P-value is recommended, though this is not guaranteed.

## Supplementary Materials.

Web Appendices A and B referenced in Section 5 are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

## References.

- Agresti, A. and Min, Y. (2005a). Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine* **24**, 729-740.
- Agresti, A. and Min, Y. (2005b) Frequentist performance of Bayesian confidence intervals for comparing proportions in  $2 \times 2$  contingency tables. *Biometrics* **61** 515-523,.
- Barndoff-Nielsen, O. (1973). On M-ancillarity. *Biometrika* **60** 447-455.
- Berger, R.L. and Boos, D.D. (1994). P values maximised over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89**, 1012-1016.
- Berger, R.L. and Sidik, K. (2003) Exact unconditional tests for a  $2 \times 2$  matched pairs design. *Stat. Methods. Med. Res.* **12**, 91-108.
- Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics*. Holden-Day, Oakland.
- Brix T.H., Knudsen G.P.S, Kristiansen M.,Kyvik K.O., rstavik K.H. and Hegeds L. (2005) High frequency of skewed X-chromosome inactivation in females with autoimmune thyroid disease: a possible explanation for the female predisposition to thyroid autoimmunity. *J. Clin. Endocrinology and Metabolism* **90**, 5959-5953.
- Chan, I.S.F., Tang, N.S., Tang, M.L. and Chan, P.S. (2003) Statistical analysis of non-inferiority trials with a rate ratio in small-sample matched-pair designs. *Biometrics* **59** 1170-1177.
- Kalbfleisch, J.D. (1975) Sufficiency and conditionality. *Biometrika* **62** 251-259.
- Lloyd, C.J. (2005) More Powerful Unconditional Tests of no Treatment Effect from Binary Matched Pairs. Working paper 2005-24. ([http://www.mbs.edu/go/faculty-](http://www.mbs.edu/go/faculty)

and-research/faculty-publications)

- Lloyd, C.J. (2007) Exact P-values for discrete models obtained by estimation and maximisation. Submitted to *Australian and New Zealand Journal of Statistics*.
- Lloyd, C.J. and Moldovan, M. (2007) Exact one-sided confidence limits for the difference between two correlated proportions *Statistics in Medicine* **26**, 3369-3384.
- McNemar, Q. (1947) Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika* **12**, 153-157.
- Mehrotra, D. V., Chan, I. S. F. and Berger, R. L. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* **59**, 441-450.
- Sandved, E. (1967) A principle for conditioning on an ancillary statistic. *Skand. Actuar.* **50**, 23-27.
- Sprott, D. (1975) Marginal and conditional sufficiency. *Biometrika* **62**, 599-605.
- Suissa, S. and Shuster, J.J. (1991) The  $2 \times 2$  matched pairs trial: exact unconditional design and analysis. *Biometrics* **47**, 361-372.
- Tang, M-L, Tang, N-S and Carey, V. J. (2004). Confidence interval for rate ratio in a 2x2 table with structural zero: An application in assessing false-negative rate ratio when combining two diagnostic tests. *Biometrics* **60**, 550-555.

Table 1: Example of matched pairs from Berger & Sidik (2003).

Treatment 1	Treatment 2		Total
	Positive	Negative	
Positive	$9(x_{11})$	$4(x_{10})$	13
Negative	$11(x_{01})$	$6(x_{00})$	17
Total	20	10	$30(n)$

Table 2: **Power dominance of E+M over M or B based P-values for three generating P-values.** Each figure is the proportion of the parameter space where the E+M based P-value strictly dominates either B or M in power. In all cases, E+M has at least as high power as both B and M. MC(k) means k-sided McNemar test.

$G$	$\alpha = 5\%$				$\alpha = 10\%$			
	n=50		n=100		n=50		n=100	
	E+M>B	E+M>M	E+M>B	E+M>M	E+M>B	E+M>M	E+M>B	E+M>M
mc(1)	0.947	0.828	0.893	0.684	0.684	0.935	0.551	0.903
mc(2)	0.812	0.788	0.791	0.905	0.596	0.828	0.690	0.688
lr(1)	0.936	0.828	0.889	0.848	0.728	0.936	0.551	0.908
lr(2)	0.812	0.813	0.791	0.853	0.597	0.827	0.847	0.851
cb(1)	0.932	0.948	0.897	0.824	0.926	0.947	0.688	0.886
cb(2)	0.950	0.950	0.847	0.906	0.932	0.948	0.895	0.823

Table 3: Various P-values for two examples.

Method	Berger & Sidik (2003)	Brix et al. (2005)		
	MC	CS	MC	LR
Basic	0.0353	0.0215	0.0114	0.0066
E	0.0377	0.0099	0.0097	0.0111
M	0.0494	0.0128	0.0110	0.0120
E+M	0.0377	0.0099	0.0098	0.0112
B( $\gamma = .0005$ )	0.0382	0.0103	0.0103	0.0125

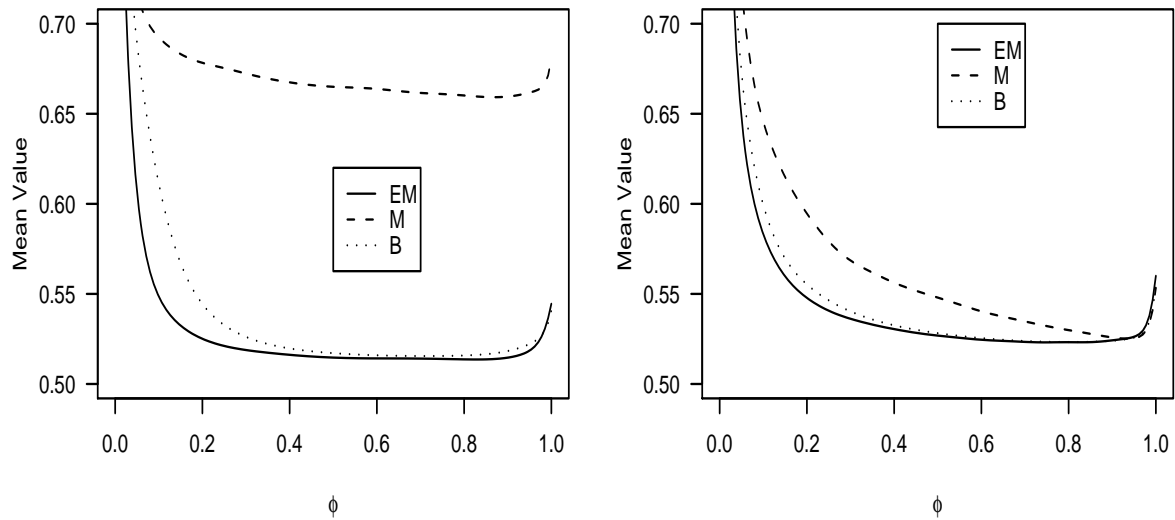


Figure 1: **Mean P-Values versus discordance probability  $\phi$ .** *Left.* LR one-sided,  $n = 50$ . *Right.* McNemar two-sided,  $n = 100$ .

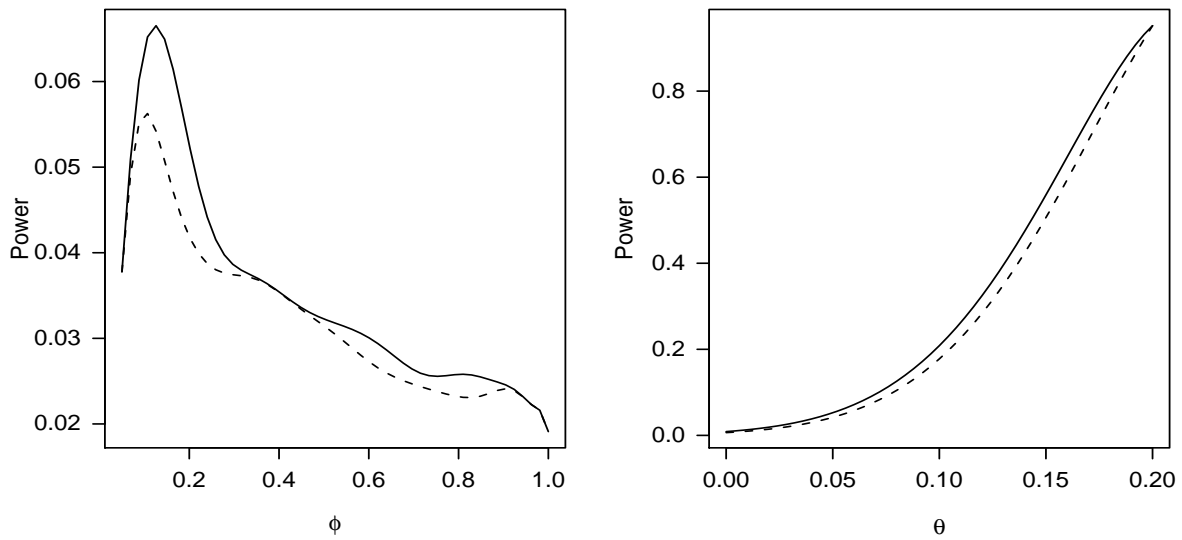


Figure 2: **Power of one-sided test based on E+M P-values derived from MC and CS.** Solid line is test based on MC, dotted line is test based on CS. *Left.* Power of 1% tests with  $n = 50$  when  $\theta = 0.05$  versus  $\phi$ . *Right.* Power of 1% tests with  $n = 50$  when  $\phi = 0.2$  versus  $\theta$ .

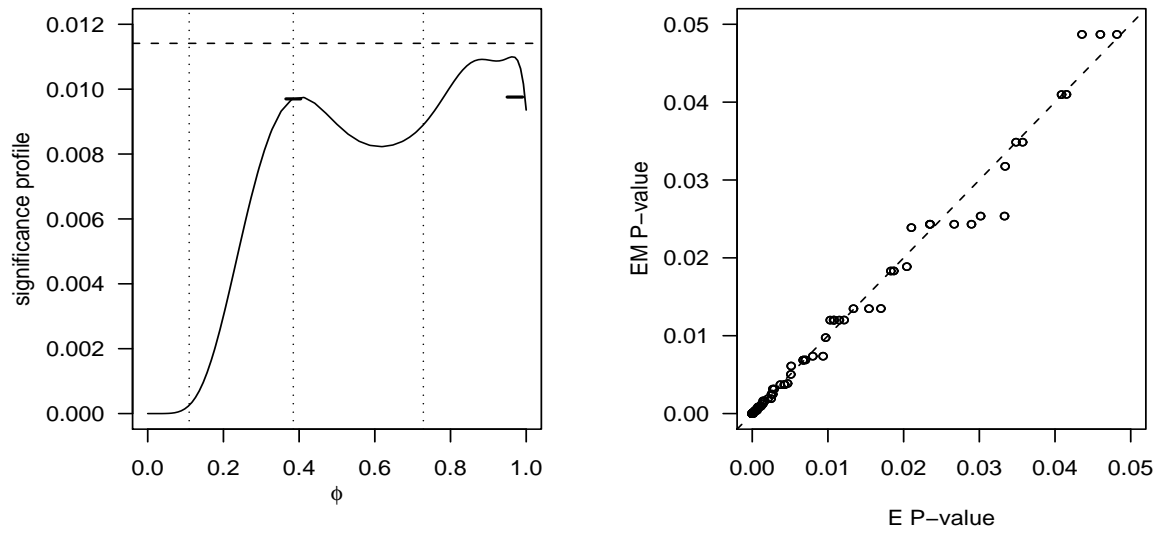


Figure 3: All plots are based on McNemar's normally approximated two-sided P-value as generating P-value. *Left.* Significance profile for data set  $(x, t, n) = (9, 10, 26)$  from Brix et al. (2005). Dashed line is the normally approximated P-value of 0.0114. E and M P-values are indicated by short thick lines. *Right.* E P-values versus E+M P-values when  $n = 26$ . Scale is limited to P-values less than 0.05.

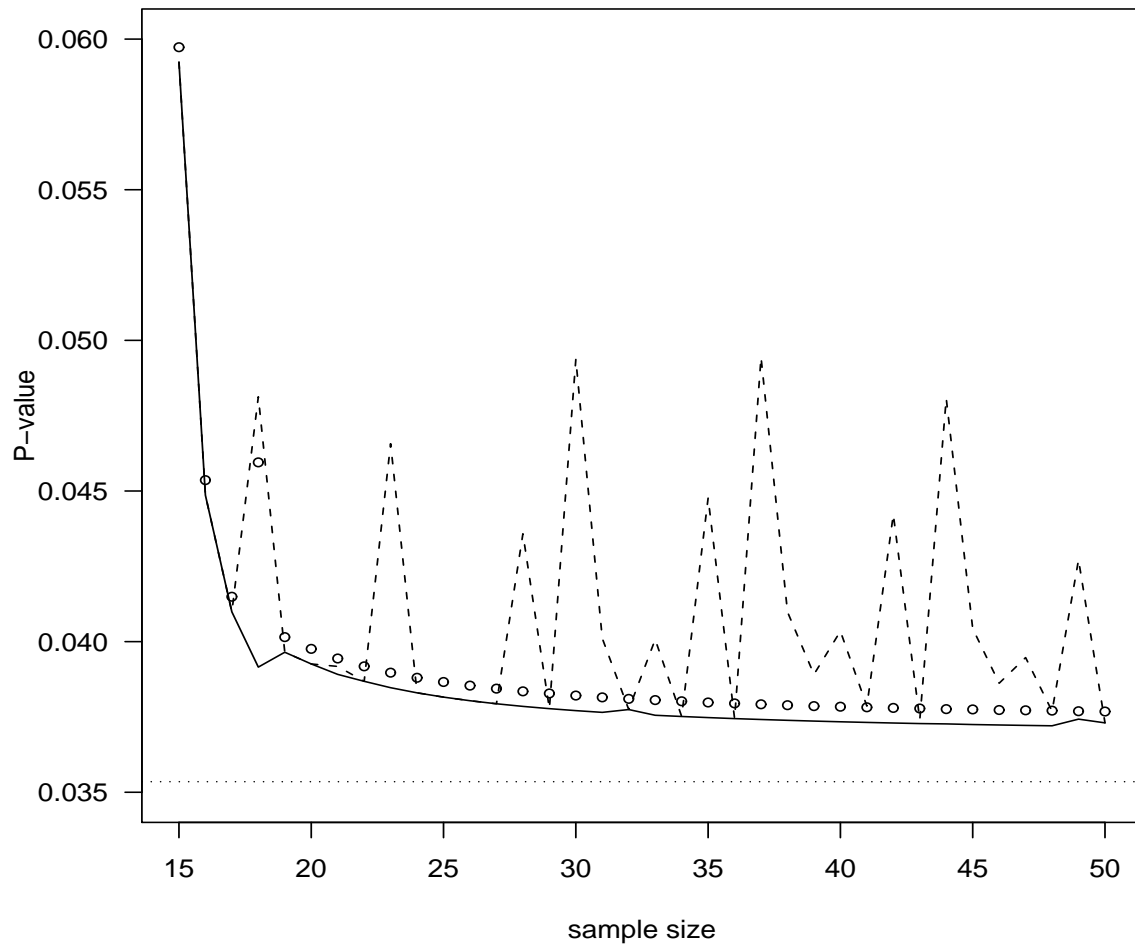


Figure 4: **Dependence of P-values on sample size.** Plot is for data  $(x, t) = (11, 15)$  of Berger & Sidik (2003). Solid line is E+M P-value, dashed line is M P-value and points are B P-values.