Title
A Quasi-Exact Test for Comparing Two Binomial Proportions

Authors
Karim F. Hirji
Shu-Jane Tau
Robert Elashoff

Publication Date
2011-10-24

STATISTICS IN MEDICINE, VOL. 10, 1137–1153 (1991)

A QUASI-EXACT TEST FOR COMPARING TWO BINOMIAL PROPORTIONS

KARIM F. HIRJI, SHU-JANE TAN AND ROBERT M. ELASHOFF

Department of Biomathematics, School of Medicine, University of California, Los Angeles, CA 90024-1766, U.S.A.

SUMMARY

The use of the Fisher exact test for comparing two independent binomial proportions has spawned an extensive controversy in the statistical literature. Many critics have faulted this test for being highly conservative. Partly in response to such criticism, some statisticians have suggested the use of a modified, non-randomized version of this test, namely the mid-*P*-value test. This paper examines the actual type I error rates of this test. For both one-sided and two-sided tests, and for a wide range of sample sizes, we show that the actual levels of significance of the mid-*P*-test tend to be closer to the nominal level as compared with various classical tests. The computational effort required for the mid-*P*-test is no more than that needed for the Fisher exact test. Further, the basis for its modification is a natural adjustment for discreteness; thus the test easily generalizes to $r \times c$ contingency tables and other discrete data problems.

INTRODUCTION

Comparison of two independent binomial proportions occurs frequently in statistical practice. We use the following notation to describe it. Let A and B denote the number of successes in independent samples from two binomial populations (n_1, π_1) and (n_2, π_2), respectively. Define $n = n_1 + n_2$. Then the joint probability of a particular realization is

$$Pr(A = a, B = b) = \binom{n_1}{a}\binom{n_2}{b}\pi_1^a(1-\pi_1)^c\pi_2^b(1-\pi_2)^d \tag{1}$$

for $a = 0, 1, \ldots, n_1$ and $b = 0, 1, \ldots, n_2$, and where $c = n_1 - a$ and $d = n_2 - b$. Usually it is of interest to test $H_0: \pi_1 = \pi_2$ against $H_1: \pi_1 \neq \pi_2$, or to test $H_0: \pi_1 = \pi_2$ against $H_1: \pi_1 > \pi_2$.

A vast amount of research effort, spanning over half a century, has focused on the above problem. A voluminous literature exists, and prominent statisticians have engaged in acrimonious debates. Yet the question of the appropriateness of various approaches remains clouded by considerable controversies.[1,2] Among applied statisticians a sort of nonchalant attitude towards these controversies seems to have emerged; their practice appears to be guided by what has been described as 'conventional wisdom'.[3] Thus in practice, when the two sample sizes are large, analysts generally employ the Pearson chi-square test. With small to moderate sample sizes, they use either the Fisher exact test or the Yates continuity corrected chi-square test.

The latter two tests have, almost since their advent, faced numerous criticisms on both theoretical and empirical grounds. Yates[1] summarizes the history of and the arguments behind these critiques. Recent critiques have focused on the empirical performance of these tests.[2–6] Either from their own extensive studies, or from a review of work done by others, these authors

0277–6715/91/071137–17$08.50

Received September 1989
Revised November 1990

argue that the Fisher exact and the Yates chi-square tests are excessively conservative when used with small to moderate sample sizes. The consequence of implied loss of power then diminishes the practical utility of both these tests.

Consequently, identification of test procedures whose actual type I error rate is closer to the nominal significance level, especially when the sample sizes are not large, has become an important research issue. D'Agostino *et al.*[3] showed that even with small sample sizes the uncorrected chi-square test and the Student t-test based on binary data generally provide actual significance levels not far from the postulated levels. Upton[6] and Overall *et al.*[5] evaluated a wide variety of test procedures for this problem. From the above and related research it appears that with consideration of both ease of computation and the average or median actual significance level, one would recommend use of one of the three tests – the Pearson chi-square test, the t-test or the scaled chi-square test – for almost all sample sizes encountered in practice. The bases for all these tests, however, are asymptotic approximations. This is reflected in the tendency of the actual significance levels of these tests to vary, at times appreciably, about the nominal level. The conduct of above cited and other similar studies entailed a limited configuration of sample sizes, nominal significance levels and true common parameter values. Thus there is not yet a complete picture regarding the variability of the actual significance levels of these and other tests.

Tocher[7] presented a version of the Fisher exact test that not only attains the nominal level exactly, but that also has some optimal power characteristics. This test involves the use of an extraneous randomization procedure. Such a feature makes the test unappealing to practising statisticians, especially those who work with biomedical data.[8,9] Lancaster[8] proposed the use of a non-randomized test procedure for discrete distributions derived from a consideration of the randomized test and based on the concept of a mid-P-value. Since then various authors have suggested employment of a mid-P-based procedure in connection with the Fisher exact test.[10-13] Most recently, Barnard[14,15] advocated reporting both the traditional P-value and the mid-P-value when performing the Fisher exact test. Despite the simplicity and intuitive appeal of this concept as well as its potential for use in a wide variety of discrete data situations, apart from a limited comparison done by Miettinen,[12] no comprehensive evaluation of the mid-P-test has appeared. None of the above cited papers that critically evaluated the Fisher test mentions it, and in general statisticians appear to be unaware of its existence.

In this paper, we investigate the properties of the mid-P-test for comparing two independent binomial proportions. For a wide variety of sample and population configurations, we compare the actual significance levels of the mid-P-test (M) with those of the Fisher exact test (F), the Pearson chi-square test (X), the scaled chi-square test (S) and the Student t-test (T). The comparisons performed are for both one-sided and two-sided tests, at four commonly used levels of significance, and for a total sample size that ranges from 4 to 100. The next section describes these five test procedures.

THE TESTS

The Pearson and scaled chi-squared tests

For an observed configuration (a, b), the chi-square statistic is

$$X^2 = (ad - bc)^2 n/\{n_1 n_2 (a + b)(c + d)\}.$$

A two-sided α-level test then consists of comparing this value with a cut-off value obtained from a χ^2 distribution with one degree of freedom (d.f.) or a standard normal distribution. A one-sided

test would reject if, additionally, the difference between the observed proportions was in a specified direction.

The scaled chi-square statistic derives from use of the appropriate mean and variance of the conditional hypergeometric distribution, and is $X_s^2 = X^2(n-1)/n$, which one compares with a χ^2 statistic with 1 d.f. Both these tests are approximate tests because the distributions of X^2 and X_s^2 approach that of χ^2 with 1 d.f. only when the sample sizes are large.

The Student *t*-test

This test uses the means and the variances of the two binomial distributions to compute the classical two independent samples *t*-statistic based on a pooled estimate of the variance. The significance of the result obtains by comparison of the statistic with cut-off values derived from the Student *t* distribution with $n_1 + n_2 - 2$ d.f.

The Fisher exact test

Derivation of the Fisher test involves consideration of the conditional distribution of A give $A + B$, which depends only on the parameter $\phi = \pi_1(1-\pi_2)/\{\pi_1(1-\pi_1)\}$. A test of $\pi_1 = \pi_2$ against $\pi_1 \neq \pi_2$ is equivalent to a test of $\phi = 1$ against $\phi \neq 1$, and a test of $\pi_1 = \pi_2$ against $\pi_1 > \pi_2$ is equivalent to a test of $\phi = 1$ against $\phi > 1$. Hence one can construct a test based on the conditional distribution for these hypotheses. Define

$$f(a, s, \phi) = Pr(A = a \mid A + B = s; \phi) \quad \text{and} \quad S(a, s, \phi) = Pr(A > a \mid A + B = s; \phi).$$

Then for the one-sided α-level test, we would reject if the $f(a, s, 1) + S(a, s, 1) \leqslant \alpha$.

For the three asymptotic tests described above, owing to the symmetric nature of the reference distributions, the two-sided tests are uniquely defined. But that is not the case for the hypergeometric distribution. For asymmetric distributions, there are many different ways to obtain a two-sided test. These are based either on various measures of deviation from variously defined central points of the distribution, or upon consideration of the area in the two tails, or on the principle of minimum likelihood.[16] Upton[6] computed two-sided *P*-values for the Fisher exact test in terms of the absolute deviation from the mean of the null conditional hypergeometric distribution. *P*-values based on deviation from the median may also be appropriate. Hill and Pike[17] suggested two methods for computing two-tailed *P*-values; the first was based on the odds ratio, while the second included, in addition to the conditional probability of the observed tail, all terms of the other tail such that the sum of their conditional probabilities did not exceed that of the observed tail. Yates,[1] quoting R. A. Fisher, considers doubling the conditional probability of the observed tail as the appropriate two-sided *P*-value. Cox[18] also used this method in connection with general conditional exact tests for discrete data. Pratt and Gibbons[19] suggest the use of a technique, apparently originated by Neyman and Pearson, that orders all observations on the basis of their probability. For the F test, one can use this method by ordering all observations in the conditional space in terms of their null conditional probability.

Since this paper does not seek to compare all methods for computation of a two-tailed *P*-value, we selected to study only the two-sided tests constructed with the methods of Cox[18] and Pratt and Gibbons[19] because of their generality, as well as their adaptability to the concept of mid-*P*-value described below. Hence the two methods that we used to construct a two-sided test, respectively called the minimum likelihood method and twice the smallest tail method, were: (i) F1: reject if $Pr[\{y: f(y, s, 1) \leqslant f(a, s, 1)\} \mid A + B = s] \leqslant \alpha$, or (ii) F2: reject if minimum $\{f(a, s, 1) + S(a, s, 1), 1 - S(a, s, 1)\} \leqslant \alpha/2$.

We call the above tests exact because they are based on an exact, albeit conditional, distribution for the problem. Tests based on exact conditional or unconditional distributions lead to some form of control over the actual significance levels. Thus all these three tests, F, F1 and F2, guarantee non-exceedance of the nominal significance level. Further, these are conditional tests where the conditioning has served to eliminate a nuisance parameter.

The mid-*P*-test

Lancaster[8] described the concept of a mid-*P*-value for univariate discrete distributions. Traditionally, the definition of a *P*-value is the probability of obtaining the observed or a more extreme configuration if the null hypothesis is true. In a continuous distribution, inclusion or exclusion of the observed point from the critical region is immaterial in *P*-value computation. In a discrete distribution that is not the case, and its inclusion is what leads to the conservativeness of exact test procedures with discrete data. Lancaster[8] proposed computation of the mean of the two probabilities obtained by inclusion and exclusion of the observed point. This is equivalent to inclusion of half the probability of the observed point in each tail. This quantity, called the mid-*P*-value, then forms the basis for accepting or rejecting the null hypothesis.

We can then apply the concept of a mid-*P*-value to the conditional hypergeometric distribution used for the Fisher exact test. With this modification, the above one-sided procedure becomes: M: reject if the $0{\cdot}5f(a, s, 1) + S(a, s, 1) \leqslant \alpha$. The two-sided test can now be carried out in the following two ways:

(i) M1: reject if

$$0{\cdot}5\ Pr[\{y{:}f(y, s, 1) = f(a, s, 1)\} \mid A + B = s] + Pr[\{y{:}f(y, s, 1) < f(a, s, 1)\} \mid A + B = s] \leqslant \alpha,$$

and

(ii) M2: reject if

$$0{\cdot}5f(a, s, 1) + \text{minimum}\{S(a, s, 1),\ 1 - S(a, s, 1) - f(a, s, 1)\} \leqslant \alpha/2.$$

The general formulation of M2 is due to Vollset.[20]

A mid-*P*-based test is a *quasi-exact* test because, although it is based on an exact (conditional) distribution, it does not guarantee non-exceedance of the nominal significance level. The extent to which the mid-*P* approach reduces the conservative bias of the Fisher exact test, and its performance relative to other tests for comparing two binomial proportions, need to be assessed empirically. In the next section, we describe the design of such an empirical study.

METHODS

We conducted a study to compare the actual significance levels of the five tests (X, S, T, F and M) described above. The two sample sizes n_1 and n_2 were varied among all possible values in the set $\{2, 4, 6, \ldots, 50\}$, giving a total of 625 configurations ranging from equal to the highly unequal, and from very small to moderately large sample sizes. The nominal significance levels studied were the commonly encountered levels, 0·01, 0·02, 0·05 and 0·10. For each of these combinations of sample sizes and significance levels, we computed the actual level of significance for each test when the common binomial parameter $\pi = \pi_1 = \pi_2$ took the values 0·1, 0·2, 0·3, 0·4 and 0·5. This was done as follows. For sample sizes n_1 and n_2, the sample space $\Omega(n_1, n_2)$ is the set of all integer valued pairs (a, b) with $0 \leqslant a \leqslant n_1$ and $0 \leqslant b \leqslant n_2$.[21,22] Suppose that for a given test performed at a nominal level α, $R(\alpha, n_1, n_2)$ is the subset of $\Omega(n_1, n_2)$ over which the null hypothesis is

rejected. Then, when the null is true with the binomial parameter equal to π, the actual significance level of the test is

$$\sum_{R(\alpha, n_1, n_2)} \binom{n_1}{a}\binom{n_2}{b} \pi^{a+b}(1-\pi)^{c+d}.$$

Before presenting the results, we mention some salient features of our study. We did not study the situation when $\pi > 0{\cdot}5$ as the results for it are almost equivalent to those for $\pi < 0{\cdot}5$. They are not exactly identical. See the exchange between Schawe[23] and Garside[24] for a discussion of this issue. For comparative purposes, however, not much additional information would accrue by also presenting results for $\pi > 0{\cdot}5$ if the two sample sizes are varied in a symmetric fashion so as to include balanced as well as unbalanced sample size configurations. This is what we have done in the present study. D'Agostino *et al.*,[3] who also studied only the cases with $\pi \leqslant 0{\cdot}5$, criticized earlier comparative studies for the limited number of configurations of significance levels, sample sizes and binomial parameter values considered. Their study represented an improvement over the earlier ones in this regard. Thus, for each α level, they studied 660 configurations of sample sizes and π. They gave results, however, for a two-sided test only, without explicitly stating how they performed the two-sided test. In our study, for each specified significance level, we looked at a total of 3125 sample size and binomial parameter configurations for both one-sided and two-sided tests. Thus not only did we study the same α levels as they did, but for each level we give results for almost five times the number of configurations for both one-sided and two-sided tests.

RESULTS

The overall results appear in Tables I and II respectively for one-sided tests and two-sided tests. These tables give some selected percentiles of the distributions of the true α-levels for the 3125 configurations studied. The 0th, 50th and 100th percentiles correspond, respectively, to the minimum, median and maximum actual significance levels. First we note that for each nominal α-level, our results confirm the excessive conservativeness of the exact tests, namely, F for the one-sided, and F1 and F2 for the two-sided. For the two-sided exact tests, however, the minimum likelihood method (F1) is not as conservative as the method of computation of twice the observed tail probability (F2). This observation is important as some consider the Fisher exact test more conservative in a two-sided than in a one-sided situation.[24] Our results show that before one can reach a firm conclusion one must ensure that the basis is not the more conservative of the various methods for computation of a two-sided *P*-value.

Further, our results also support earlier observations regarding the Pearson chi-square (X), the scaled chi-square (S) and the t (T) tests. These tests generally provide true significance levels not far from the nominal ones. The median actual levels of these three tests are quite close to the nominal levels, with those of T closest when compared with all the other four tests. S performs somewhat better than X or T in terms of extent of exceedance of the nominal level. For two-sided tests, T has a much higher maximum actual significance level compared with the other two when the nominal level is 0·01 and 0·02. In general, S is slightly more conservative than X or T. The basic problem with these three tests is the wide range over which the actual α-level varies. In some situations they are as conservative as the Fisher exact test, while in other situations their actual levels are more than one and a half times the nominal level. Thus although the frequency of a very low, true α-value relative to the nominal value is low, and that of a relatively very high value also low, the mere simultaneous existence of these possibilities is disquieting. This is even more so if such a possibility depends, as we shall show below, on the value of the unknown nuisance parameter.

Table I. Percentiles of the distributions of ASL, one-sided test

Percentile	X	T	S	F	M
NSL = 0·01					
0	0·000	0·000	0·000	0·000	0·000
5	0·000	0·000	0·000	0·000	0·000
10	0·000	0·000	0·000	0·000	0·000
25	0·007	0·007	0·006	0·001	0·004
50	0·010	0·010	0·009	0·004	0·007
75	0·011	0·012	0·010	0·005	0·008
90	0·015	0·015	0·013	0·005	0·009
95	0·020	0·020	0·017	0·006	0·009
100	0·070	0·070	0·060	0·007	0·011
NSL = 0·02					
0	0·000	0·000	0·000	0·000	0·000
5	0·000	0·000	0·000	0·000	0·000
10	0·003	0·002	0·001	0·000	0·001
25	0·017	0·016	0·015	0·004	0·011
50	0·020	0·020	0·019	0·008	0·015
75	0·022	0·022	0·021	0·010	0·017
90	0·027	0·027	0·026	0·011	0·018
95	0·034	0·034	0·031	0·012	0·019
100	0·086	0·086	0·080	0·015	0·023
NSL = 0·05					
0	0·000	0·000	0·000	0·000	0·000
5	0·004	0·003	0·003	0·000	0·001
10	0·029	0·024	0·025	0·001	0·009
25	0·048	0·045	0·045	0·015	0·035
50	0·052	0·049	0·050	0·023	0·042
75	0·056	0·053	0·054	0·028	0·046
90	0·064	0·061	0·061	0·031	0·048
95	0·073	0·069	0·069	0·032	0·049
100	0·131	0·128	0·128	0·039	0·057
NSL = 0·10					
0	0·000	0·000	0·000	0·000	0·000
5	0·058	0·041	0·046	0·001	0·012
10	0·093	0·086	0·089	0·009	0·050
25	0·099	0·095	0·096	0·038	0·079
50	0·104	0·100	0·101	0·051	0·089
75	0·110	0·105	0·107	0·060	0·095
90	0·121	0·114	0·117	0·065	0·098
95	0·134	0·122	0·128	0·067	0·100
100	0·190	0·172	0·179	0·078	0·117

ASL is actual significance level; NSL is nominal significance level; X is chi-square; T is *t*; S is scaled chi-square; F is Fisher exact; M is mid-*P*

The performance of the mid-*P*-test falls between that of the F test on the one hand, and the S, X and T tests on the other. It is considerably less conservative than F; its median actual level is much higher than that of F but still somewhat below the nominal level. But it does not exhibit the tendency to overshoot the desired α-level by a large amount that characterized X, T and S. This is

Table II. Percentiles of the distributions of ASL, two-sided test

Percentile	X	T	S	F1	F2	M1	M2
NSL = 0·01							
0	0·000	0·000	0·000	0·000	0·000	0·000	0·000
5	0·003	0·004	0·002	0·000	0·000	0·000	0·000
10	0·005	0·006	0·004	0·001	0·000	0·003	0·001
25	0·008	0·009	0·007	0·003	0·001	0·006	0·003
50	0·009	0·010	0·009	0·005	0·003	0·008	0·006
75	0·010	0·011	0·010	0·006	0·004	0·009	0·008
90	0·012	0·013	0·011	0·007	0·005	0·010	0·009
95	0·015	0·117	0·014	0·007	0·005	0·011	0·009
100	0·050	0·125	0·047	0·009	0·007	0·014	0·011
NSL = 0·02							
0	0·000	0·000	0·000	0·000	0·000	0·000	0·000
5	0·005	0·009	0·004	0·000	0·000	0·004	0·000
10	0·009	0·014	0·007	0·003	0·001	0·009	0·003
25	0·012	0·017	0·011	0·007	0·003	0·014	0·007
50	0·014	0·020	0·013	0·011	0·007	0·017	0·014
75	0·015	0·021	0·014	0·013	0·009	0·019	0·017
90	0·016	0·024	0·015	0·014	0·011	0·020	0·018
95	0·020	0·028	0·019	0·015	0·011	0·022	0·018
100	0·057	0·125	0·057	0·018	0·014	0·026	0·023
NSL = 0·05							
0	0·000	0·003	0·000	0·000	0·000	0·000	0·000
5	0·032	0·030	0·028	0·006	0·001	0·019	0·007
10	0·037	0·036	0·035	0·012	0·005	0·029	0·013
25	0·045	0·044	0·044	0·020	0·010	0·039	0·024
50	0·050	0·049	0·048	0·030	0·020	0·045	0·039
75	0·053	0·052	0·051	0·035	0·025	0·049	0·044
90	0·056	0·055	0·054	0·038	0·029	0·052	0·047
95	0·059	0·057	0·056	0·039	0·030	0·055	0·048
100	0·125	0·125	0·086	0·046	0·038	0·075	0·057
NSL = 0·10							
0	0·005	0·005	0·004	0·000	0·000	0·000	0·000
5	0·070	0·064	0·064	0·019	0·007	0·061	0·020
10	0·079	0·073	0·074	0·031	0·014	0·073	0·033
25	0·096	0·091	0·091	0·047	0·025	0·085	0·063
50	0·102	0·098	0·098	0·064	0·045	0·093	0·084
75	0·107	0·103	0·103	0·073	0·055	0·099	0·090
90	0·114	0·107	0·107	0·078	0·061	0·106	0·095
95	0·119	0·110	0·111	0·082	0·063	0·113	0·097
100	0·156	0·146	0·146	0·094	0·076	0·134	0·114

ASL is actual significance level; NSL is nominal significance level; X is chi-square; T is t; S is scaled chi-square; F1 is Fisher exact (minimum likelihood); F2 is Fisher exact (twice smallest tail); M1 is mid-P (minimum likelihood); M2 is mid-P (twice smallest tail)

true for all nominal levels, and for both one-sided and two-sided tests. For the two-sided test, M1 corrects the conservativeness of the Fisher test by a greater degree than does M2.

We also looked at the influence of the total sample size and the common binomial parameter value on the actual α-level. In the interest of economy we present, in Tables III, IV and V, the

Table III. Effect of total sample size on ASL (NSL = 0·05)

Percentile	$N = 2$–20 S	$N = 2$–20 M	$N = 22$–40 S	$N = 22$–40 M	$N = 42$–60 S	$N = 42$–60 M	$N = 62$–80 S	$N = 62$–80 M	$N = 82$–100 S	$N = 82$–100 M
One-sided test										
0	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000
5	0·000	0·000	0·000	0·000	0·002	0·000	0·042	0·035	0·045	0·039
10	0·001	0·000	0·004	0·001	0·024	0·014	0·045	0·038	0·046	0·040
25	0·019	0·004	0·041	0·022	0·045	0·036	0·048	0·042	0·049	0·044
50	0·043	0·020	0·049	0·037	0·050	0·042	0·050	0·045	0·050	0·046
75	0·055	0·033	0·055	0·042	0·055	0·045	0·052	0·047	0·052	0·048
90	0·071	0·040	0·067	0·045	0·062	0·047	0·055	0·049	0·054	0·050
95	0·087	0·042	0·078	0·047	0·069	0·048	0·059	0·050	0·056	0·052
100	0·118	0·055	0·118	0·055	0·128	0·057	0·076	0·057	0·066	0·057
Two-sided test										
0	0·000	0·000	0·003	0·001	0·002	0·001	0·032	0·021	0·040	0·032
5	0·007	0·004	0·020	0·015	0·033	0·029	0·041	0·029	0·044	0·043
10	0·015	0·008	0·028	0·022	0·036	0·033	0·044	0·033	0·046	0·044
25	0·028	0·015	0·038	0·032	0·043	0·040	0·047	0·041	0·048	0·047
50	0·041	0·028	0·045	0·042	0·048	0·045	0·049	0·044	0·050	0·048
75	0·052	0·039	0·050	0·046	0·050	0·048	0·051	0·046	0·051	0·049
90	0·060	0·047	0·054	0·051	0·053	0·052	0·052	0·048	0·053	0·052
95	0·066	0·054	0·058	0·055	0·055	0·055	0·054	0·050	0·056	0·055
100	0·085	0·075	0·084	0·062	0·086	0·062	0·060	0·057	0·060	0·059

ASL is actual significance level; NSL is nominal significance level; S is scaled chi-square; M is mid-P

results for the S and M1 tests only. The results for the other tests do not add to a comparative evaluation of the mid-P-test other than what we have already observed from the overall results. Further, we restrict these results to the 0·05 nominal level, the most commonly used level in practice. The differences in the relative performance of S and M1 between the 0·05 and the other three levels appear in the text.

Table III shows the influence of total sample size on S and M1 at the 0·05 level. At all sample size levels, and for both one-sided and two-sided tests, we see that M1 provides actual levels that are better (in the sense of not exceeding the nominal level) than those of S (or X or T), at the same time as not being as conservative as the F or F1, F2, or M2. At larger sample sizes the performance of S is similar to that of M1. We observed a similar picture at the 0·01 level. At the 0·02 and 0·10 levels, the two-sided M1 test showed a tendency to have slightly larger actual α-levels than the two-sided S test when the total sample size exceeded 60. This is due to the fact that the sample sizes we selected used $2 \leqslant n_1 \leqslant 50$ and $2 \leqslant n_2 \leqslant 50$, and not $4 \leqslant n_1 + n_2 \leqslant 100$. Hence, samples with a greater degree of imbalance at the larger total sample sizes are not represented in the configurations we studied. In other words, with unbalanced samples, the S (and X and T) tests can have true α-levels quite above the nominal level even if the total sample size is large. When the sample sizes are equal, the disparity between S and M1 is not as wide. This can be seen from Table IV. We further discuss this issue below.

In Table V we show the effect of the value of the common binomial parameter on the S and M1 at the 0·05 significance level. Here we see that for one-sided tests, the relationship between S and M1 at all points in the parameter space is the same as what we observed overall. A similar picture prevailed for one sided tests at the 0·01, 0·02 and 0·10 levels. For two-sided tests, while M1

Table IV. Effect of total sample size on ASL (NSL=0·05), equal sample sizes

Percentile	$N1$ = 2–24 S	$N1$ = 2–24 M	$N1$ = 26–50 S	$N1$ = 26–50 M
One-sided test				
0	0·002	0·000	0·042	0·027
5	0·010	0·000	0·044	0·033
10	0·024	0·004	0·045	0·038
25	0·040	0·023	0·048	0·042
50	0·047	0·037	0·051	0·045
75	0·052	0·043	0·053	0·046
90	0·057	0·047	0·056	0·050
95	0·064	0·052	0·057	0·051
100	0·073	0·055	0·063	0·057
Two-sided test				
0	0·000	0·000	0·039	0·020
5	0·000	0·000	0·043	0·029
10	0·003	0·001	0·045	0·032
25	0·027	0·011	0·048	0·039
50	0·045	0·034	0·049	0·044
75	0·051	0·041	0·052	0·047
90	0·059	0·045	0·057	0·050
95	0·064	0·048	0·058	0·052
100	0·070	0·050	0·060	0·057

ASL is actual significance level; NSL is nominal significance level; S is scaled chi-square; M is mid-P

performs better than S near the boundary of the parameter space, we observe a slight reversal near the centre. We also observe this reversal at the 0·01 and 0·02 levels but not at the 0·10 level.

The relationship between π, n_1 and n_2 on the one hand, and the actual α-level on the other, can be quite complex. We illustrate this relationship for two configurations in the case of one-sided tests: Figure 1 deals with the situation with a moderately sized, balanced sample ($n_1 = n_2 = 25$), and Figure 2 with a large, unbalanced sample ($n_1 = 90$, $n_2 = 5$). Looking at Figure 1 we see that, in balanced samples, both tests are quite conservative for values of π near 0 or 1. As π approaches 0·5, S becomes non-conservative quite rapidly, with a tendency to overshoot the nominal level. The actual α-level of M, on the other hand, increases less rapidly, tending to remain close to but below the nominal α-level. For unbalanced samples (Figure 2) we see that both tests are quite conservative over a large portion of the parameter space near the boundary. For the rest of the values of π, while S exhibits a marked tendency to overshoot the true α-level, M tends to fluctuate below or close to the nominal level. In both the above cases, we found the performance of X and T to be similar to, but somewhat worse than, that of S. Further, the picture for two-sided tests is somewhat more complex but the basic message is the same.

FREQUENTIST APPROACHES TO THE PROBLEM

Critics of the Fisher exact test continue to produce extensive documentation of its ultra-conservativeness. The sheer scope of this evidence seems to put their position beyond challenge.

Table V. Effect of binomial parameter value on ASL (NSL = 0·05)

Percentile	$\pi = 0{\cdot}1$ S	$\pi = 0{\cdot}1$ M	$\pi = 0{\cdot}2$ S	$\pi = 0{\cdot}2$ M	$\pi = 0{\cdot}3$ S	$\pi = 0{\cdot}3$ M	$\pi = 0{\cdot}4$ S	$\pi = 0{\cdot}4$ M	$\pi = 0{\cdot}5$ S	$\pi = 0{\cdot}5$ M
One-sided test										
0	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000
5	0·000	0·000	0·001	0·000	0·013	0·004	0·040	0·012	0·035	0·004
10	0·000	0·000	0·016	0·007	0·040	0·027	0·044	0·027	0·044	0·033
25	0·026	0·015	0·044	0·034	0·046	0·039	0·047	0·040	0·048	0·040
50	0·049	0·032	0·050	0·041	0·050	0·043	0·050	0·044	0·050	0·044
75	0·058	0·040	0·055	0·045	0·054	0·046	0·053	0·047	0·052	0·047
90	0·073	0·044	0·062	0·047	0·061	0·048	0·056	0·048	0·055	0·048
95	0·085	0·046	0·068	0·050	0·065	0·049	0·060	0·049	0·057	0·049
100	0·128	0·055	0·110	0·057	0·090	0·057	0·091	0·057	0·073	0·057
Two-sided test										
0	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·000
5	0·020	0·013	0·033	0·030	0·036	0·033	0·027	0·016	0·014	0·007
10	0·028	0·019	0·036	0·033	0·039	0·037	0·040	0·029	0·043	0·028
25	0·036	0·030	0·042	0·039	0·046	0·044	0·046	0·042	0·046	0·040
50	0·044	0·039	0·047	0·046	0·048	0·048	0·049	0·046	0·049	0·045
75	0·048	0·044	0·050	0·048	0·050	0·049	0·051	0·050	0·052	0·052
90	0·055	0·047	0·052	0·049	0·053	0·050	0·053	0·055	0·055	0·056
95	0·063	0·047	0·054	0·050	0·054	0·051	0·054	0·056	0·056	0·059
100	0·086	0·053	0·074	0·053	0·062	0·056	0·065	0·070	0·070	0·075

ASL is actual significance level; NSL is nominal significance level; S is scaled chi-square; M is mid-*P*

On the other hand, the defenders of this test base their arguments on a basic imperative to use a conditional test whatever the design of the study, and also on their questioning of the validity of an unconditional evaluation of a conditional test. Excluding those who advocate a Bayesian or semi-Bayesian approach to the problem,[25,26] we have grouped the positions taken by various protagonists in this ongoing debate into four principal categories, which we describe and discuss below.

Conditional inference

The proponents of the Fisher exact test consider performance of conditional inference as the only logically sound alternative. More accurately, their position advocates performance of conditional inference embedded in the conditional sample space. Thus not only should inference be based only on information contained in the conditional sample space, but even the evaluation of this procedure should be within this restricted space. This is the position taken, for example, by Yates,[1] a number of discussants of his paper, and Hill.[27] Barnard[14] comes close to adopting this view. This approach avoids the problem of conservativeness by declaring unconditional evaluations as irrelevant, and by arguing that one should not have concern for arbitrarily fixed significance levels, but instead should report the actual attained *P*-value and, possibly, the next highest *P*-value, or even the mid-*P*-value.

There are two fundamental objections to this approach. The first relates to the level of conditioning used. Basu,[26] in a criticism of this position, argues that instead of conditioning on just the two margins, why not also condition on the difference between values on one of the

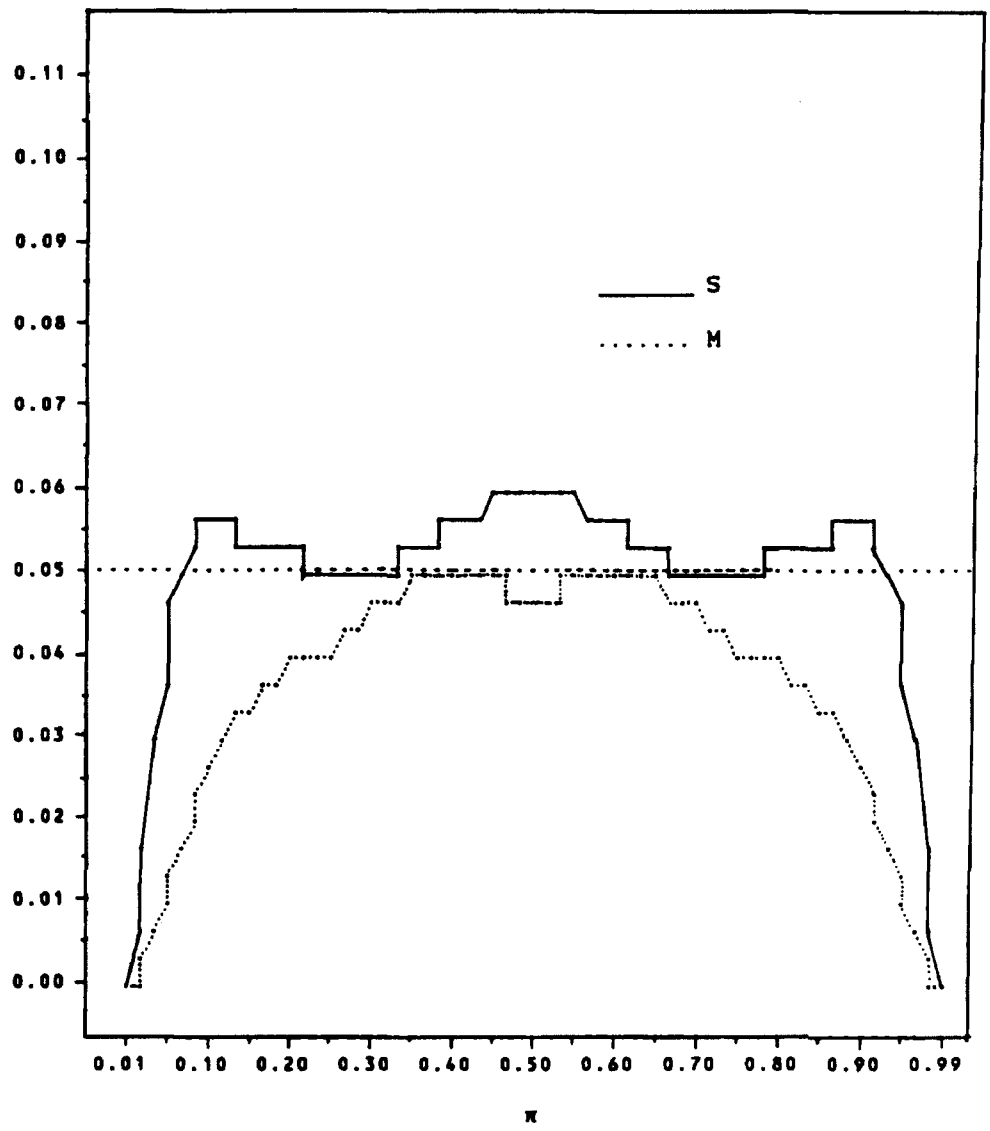

Figure 1. Actual significance levels of one-sided 0·05 nominal level S and M tests for $n_1 = n_2 = 25$

diagonals of the 2×2 table? This would yield an even more restricted sample space. Indeed, why not condition on all the data? In our opinion, the conditionalist position has not provided a satisfactory response to this critique. We note here that for the problem under study, one cannot justify the conditionalist position by resorting to the principle of ancillarity.[28] For the problem of comparing two binomial proportions, the total number of successes is not an ancillary statistic,[26] as is sometimes erroneously stated.[29,30]

The second problem with the conditionalist position concerns the choice of the *evaluative* sample space for the problem. Barnard,[21] Pearson,[22] Kempthorne,[31] Rice[32] and many others have implicitly or explicitly argued that the sample space for any problem is fixed by the study design, and hence in the case of the problem of comparing two independent binomial proportions, this space is that given by (1). We make a distinction between the *inferential* sample space, that is

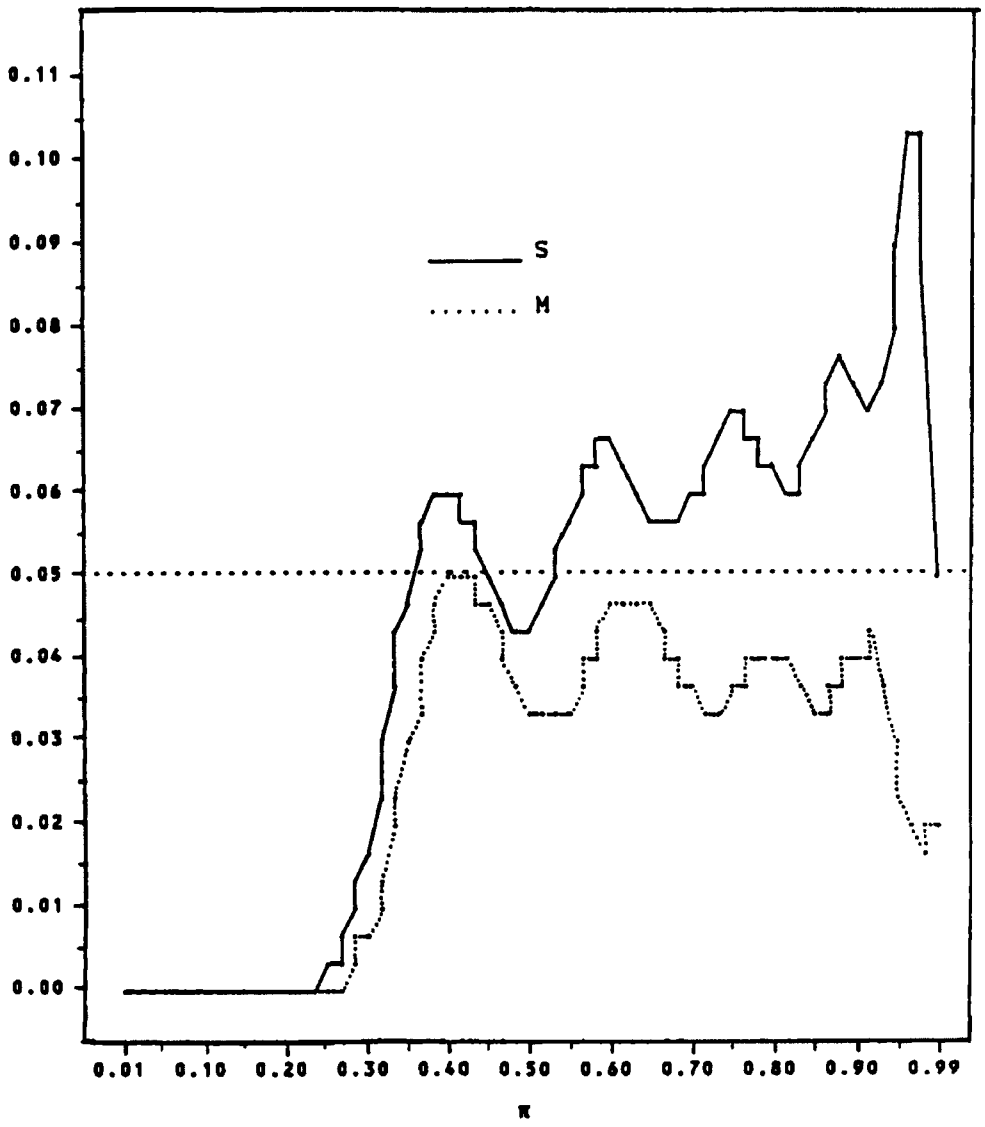

Figure 2. Actual significance levels of one-sided 0·05 nominal level S and M tests for $n_1 = 90$ and $n_2 = 5$

the probability space used for drawing inferences from the data, and the *evaluative* sample space, that is the probability space used for assessing any mode of performing inference for a problem. The former may be a subspace of the latter. Thus, while we do not rule out performance of conditional inference on an *a priori* basis, we feel that one should assess all the methods of performing inference on the basis of repeated sampling done over the probability model specified by the study design. In case of the problem under study, this model is given by (1).

Consequently, we feel that the empirical evidence of the excessive conservativeness of the classical Fisher exact test (and the Yates continuity corrected chi-square test[2]) for comparing two independent binomial proportions is valid and incontrovertible. The evidence from our study provides further support to this position. Thus a partial explanation of its continued use in practice, as noted by several authors, is the mystique associated with the word 'exact'. Undoubtedly, this test enables elimination of the nuisance parameter in a clever fashion, and the

computation of exact conditional probabilities for outcomes in the conditional sample space. These advantages, however, do not appear to outweigh its ultra-conservative performance, and the consequent loss of power for detection of realistic alternatives.

Moreover, in the context of the probability model (1), we feel that there is need for caution in interpreting the P-values given by the Fisher exact test (or, for that matter, those from the Pearson chi-square, the mid-P, and most tests for this problem). Traditionally the definition of a P-value is the probability of obtaining an observation as extreme as or more extreme than the realized one if the null hypothesis is true. If, however, any nuisance parameter is unspecified under this hypothesis, then such a probability is unknown, and the P-value loses its traditional interpretation. Thus, for the problem of comparing two binomial proportions, we *cannot* consider the P-values obtained from most tests as probabilities, under repeated sampling, of obtaining the observed or a more extreme observation. They are merely numbers that seem to behave like unconditional probabilities. We discuss below some exceptions to this rule.

Approximate inference

We describe this approach as the use of test procedures based on various approximations, derived mainly from asymptotic arguments, with the understanding that we must then evaluate such procedures in the context of the model given by (1). This is the position taken by Berkson,[4] D'Agostino *et al.*[3] and Upton,[6] for example.

While concurring with the need for unconditional evaluation advocated in this approach, we feel that its proponents have yet to come to terms with two shortcomings. First, it seems that they implicitly reject conditional inference *per se* rather than conditional inference in the form given by the classical Fisher exact test. Under this view, the Fisher exact test is valid only if both marginals are fixed by design. A forthright expression of this position is expounded by Haviland[2] and Andres and Luna del Castillo,[30] and a critique of it is given by D'Agostino[33] and Mantel.[34] Thus, in their zeal to criticize the Fisher exact test, the strict anti-conditionalist position overlooks examination of other forms of conditional inference, such as that based on conditional mid-P-values. Further, even the conduct of empirical assessment of the Fisher exact test has been non-comprehensive. D'Agostino *et al.*[3] pointed out some of the shortcomings, and in this paper we mention factors such as two-sided tests, unbalanced samples and true parameter values, in relation to which a complete assessment of the performance of various tests needs to be done.

Unconditional inference

Barnard[21] pioneered this approach. He devised a test procedure from a direct consideration of the unconditional sample space (1). Subsequently various authors have given test procedures that we can regard as extensions and modifications of his basic approach.[32,35,36] This approach, which we term unconditional exact testing, allows for interpretation of a P-value as an upper bound on the unconditional probability of the critical region. Thus with this approach both the inferential space and the evaluative space are given by (1).

Upton[6] evaluated the test developed by Barnard,[21] and found it to be somewhat conservative. Another problem with this approach is its computational complexity in the general case. While in this age of omnipresent microcomputers it is quite feasible to perform such tests for two binomial proportions,[32] it is infeasible to perform such unconditional tests for the 2×2 table with no fixed margins, for $r \times c$ tables, or for general discrete data problems even with large mainframe computers. Storer and Kim[37] also point out computational problems associated with exact unconditional tests and propose the use of an approximate unconditional test. In terms of non-exceedance of nominal levels, this test appears to be similar to the mid-P-test.

Randomized inference

This approach, developed by Tocher,[7] relies upon the conduct of an auxiliary random experiment upon observation of the data and before one draws inference from the data. It is the only approach that deals satisfactorily with the problem of conservativeness. We note that the basis for this approach is a conditional inferential space. Camilli and Hopkins[38] compared the power of the Tocher test with that of the chi-square test when the total sample size is small, and Lloyd[39] provides an interpretation and justification of a two-sided version of it. To date, however, this approach has theoretical interest only, and numerous authors have criticized its use in a practical setting.[8,9]

Aconditional inference

This approach, which for the lack of a better term we call 'aconditional inference', neither considers conditional inference as the sole valid method in all settings, nor restricts conditional inference only to situations where the study design justifies it. Rather, conditioning is seen as one of the techniques, within the context of an unconditional sample space, of deriving inferential procedures. Thus for the problem at hand, this approach considers the sample space (1) to be the appropriate evaluative space. But it does not advocate use of either a conditional or an unconditional inferential procedure on the basis of some *a priori* dictum. According to this viewpoint, from the $2^{(n_1+1)(n_2+1)}$ subsets of the sample space, the Fisher exact test provides but one method selecting a critical region. There are many other subsets, such as those given by the classical approximate tests, or by various unconditional tests, that can be considered as well. Which one, if any, is to be preferred depends on the criterion used; moreover, this is a question that needs to be subject to theoretical and empirical evaluation in the context of the complete sample space.

Gleanings of this approach are found in D'Agostino[33] and Mantel,[34] and we subscribe to it as well. Here conditioning is seen as a device for deriving exact distributions, free from nuisance parameters, which can be used to perform tests of hypothesis or to derive confidence intervals for a parameter of interest.[40] Thus, unlike Barnard[14] or Williams,[29] it is in this spirit that we advocate consideration of the mid-P-test. This test provides but another method of constructing a critical region, and, if closeness to nominal levels in an important criterion for assessing a test procedure, then we have shown that it performs quite well.

CONCLUDING REMARKS

For the problem of comparing two independent binomial proportions, our study and review of the literature leads us to the following conclusions:

(i) The classical Fisher exact test and the Yates continuity corrected chi-square tests are too conservative for practical use.

(ii) When the two samples are nearly equal, and when one anticipates that the underlying true binomial value is near 0·5, one can use the scaled chi-square, the Pearson chi-square, the t or the mid-P tests for all sample sizes.

(iii) In case of quite unequal sample sizes, or when the common binomial parameter is near 0 or 1, we recommend the mid-P-test even when the total sample size is large. The approximate unconditional test of Storer and Kim[37] would also perform well in this situation.

(iv) For computation of two-sided Fisher or mid-P-values, the method of minimum likelihood is less conservative than that of doubling the one-sided P-value, and is to be preferred. This

recommendation is opposite that given earlier by Dupont[41] and Lloyd,[39] who examined the issue from the viewpoint of sensitivity of P-values to minor changes in the contingency table.

Various authors have argued for the use of a mid-P-test on intuitive and theoretical grounds. Lancaster[8] derived it from a consideration of randomized tests and investigated its properties in some discrete distributions. Here we view it as a natural adjustment, on the probability scale, for discreteness. Stone[42] provided a rationale for it in the context of a general theory of significance testing. Barnard[14] gave a justification for it in terms of approximating continuous distributions and he computed the mean and variance of the distribution of mid-P-values. Our study empirically documents its appropriateness, for the problem of comparing two binomial proportions, in a repeated sampling framework.

One may argue that the mid-P-value does not correspond to a repeated sampling probability of a defined event. But, in the presence of nuisance parameters, neither do P-values obtained from the traditional tests. All are merely numbers that provide a guide to action. Determination of which one is more appropriate requires examination of which one has actual error levels closer to postulated levels.

Further, the use of the mid-P-value generalizes to $r \times c$ contingency tables and to higher-dimensional discrete data problems. The computation of a mid-P-value does not require more work than that needed for computation of the traditional 'exact' P-value. With availability of efficient algorithms for computation of exact conditional distributions,[20, 43–46] it is feasible to use mid-P-based procedures for more complex problems. Further, we can use the concept of mid-P in an unconditional setting as well. Hirji[47] studied a multiparametric mid-P-test for matched case-control studies. Here the mid-P procedure performed better than traditional large sample procedures, while it avoided the high degree of conservativeness of the 'exact' method. Vollset[20] investigated use of mid-P-based procedures for analysis of the common odds ratio in several 2×2 tables. A more thorough assessment of the use of the mid-P-value for more complex situations, however, remains to be done.

ACKNOWLEDGEMENTS

We thank Professor Roderick A. Little for his insightful remarks on the subject that prompted us to undertake our study. Suggestions by the referees and the editor helped improve this paper. Complete results of our study are available from the authors upon request. This study was funded by USHHS grant CA 16042.

REFERENCES

1. Yates, F. 'Tests of significance for 2×2 contingency tables' (with discussion), *Journal of the Royal Statistical Society, Series A*, **147**, 426–463 (1984).
2. Haviland, M. G. 'Yates's correction for continuity and the analysis of 2×2 contingency tables', *Statistics in Medicine*, **9**, 363–367 (1990).
3. D'Agostino, R. B., Chase, W. and Belanger, A. 'The appropriateness of some common procedures for testing the equality of two independent binomial proportions', *American Statistician*, **42**, 198–202 (1988).
4. Berkson, J. 'In dispraise of the exact test', *Journal of Statistical Inference and Planning*, **2**, 27–42 (1978).
5. Overall, J. E., Rhoades, H. M. and Starbuck, R. R. 'Small-sample tests for homogeneity of response probabilities in 2×2 contingency tables', *Psychological Bulletin*, **102**, 307–314 (1987).
6. Upton, G. J. G. 'A comparison of alternative tests for the 2×2 comparative trial', *Journal of the Royal Statistical Society, Series A*, **145**, 86–105 (1982).
7. Tocher, K. D. 'Extension of the Neyman–Pearson theory of tests to discontinuous variates', *Biometrika*, **37**, 130–144 (1950).

8. Lancaster, H. O. 'Significance tests in discrete distributions', *Journal of the American Statistical Association*, **56**, 223–234 (1961).
9. Liddle, D. 'Practical tests of 2 × 2 contingency tables', *Statistician*, **25**, 295–304 (1976).
10. Lancaster, H. O. *The Chisquared Distribution*, Wiley, New York, 1969.
11. Anscombe, F. J. *Computing in Statistical Science Through APL*, Springer, New York, 1981.
12. Miettinen, O. S. 'Discussion of Conover's "Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables"', *Journal of the American Statistical Association*, **69**, 374–382 (1974).
13. Plackett, R. L. 'Discussion of Yates's "Tests of significance for 2 × 2 contingency tables"', *Journal of the Royal Statistical Society, Series A*, **147**, 426–463 (1984).
14. Barnard, G. 'On the alleged gains in power from lower *P*-values', *Statistics in Medicine*, **8**, 1469–1477 (1989).
15. Barnard, G. 'Comment on the paper by Mark G. Havilland', *Statistics in Medicine*, **9**, 373–375 (1990).
16. Gibbons, J. D. and Pratt, J. W. '*P*-values: interpretation and methodology', *American Statistician*, **29**, 20–25 (1975).
17. Hill, I. D. and Pike, M. C. 'Algorithm 4: TWOBYTWO', *Computer Bulletin*, **9**, 56–63 (1965).
18. Cox, D. R. *The Analysis of Binary Data*, Methuen, London, 1970.
19. Pratt, J. W. and Gibbons, J. D. *Concepts of Nonparametric Theory*, Springer, New York, 1981.
20. Vollset, S. E. 'Exact and asymptotic inference in stratified one parameter conditional logistic model', PhD dissertation, University of California, Los Angeles, University Microfilms, Ann Arbor, 1989.
21. Barnard, G. A. 'Significance tests for 2 × 2 tables', *Biometrika*, **34**, 123–138 (1947).
22. Pearson, E. S. 'The choice of statistical tests as illustrated on the interpretation of data classed in a 2 × 2 table', *Biometrika*, **34**, 139–167 (1947).
23. Schawe, D. 'Error probabilities for the 2 × 2 contingency table', *American Statistician*, **31**, 134 (1977).
24. Garside, G. R. 'Reply to Schawe, D. "Error probabilities for the 2 × 2 contingency table"', *American Statistician*, **31**, 134 (1977).
25. Little, R. J. A. 'Testing the equality of two independent binomial proportions', *American Statistician*, **43**, 238–288 (1989).
26. Basu, D. 'Discussion of Joseph Berkson's paper "In dispraise of the exact test"', *Journal of Statistical Inference and Planning*, **3**, 189–192 (1979).
27. Hill, I. D. 'Discussion of the paper by William R. Rice', *Biometrics*, **44**, 14–16 (1988).
28. Kalbfleisch, J. D. 'Ancillary statistics', entry in Kotz, S. and Johnson, N. L. (eds), *Encyclopedia of Statistical Sciences*, vol. 1, Wiley, New York, 1982.
29. Williams, D. A. 'Tests for differences between several small proportions', *Applied Statistics*, **37**, 421–434 (1988).
30. Andres, A. M. and Luna del Castillo, J. D. 'Letter to the editor', *Statistics in Medicine*, **8**, 243–245 (1989).
31. Kempthorne, O. 'In dispraise of the exact test: reactions', *Journal of Statistical Inference and Planning*, **3**, 199–213 (1979).
32. Rice, W. 'A new probability model for determining exact *P*-values for 2 × 2 contingency tables when comparing two binomial proportions', *Biometrics.*, **44**, 1–22 (1988).
33. D'Agostino, R. B. 'Comment on the paper by Mark G. Havilland', *Statistics in Medicine*, **9**, 377–378 (1990).
34. Mantel, N. 'Comment on the paper by Mark G. Havilland', *Statistics in Medicine*, **9**, 369–370 (1990).
35. McDonald, L. L., Davis, B. M. and Milliken, G. A. 'A nonrandomised unconditional test for comparing two proportions in 2 × 2 contingency tables', *Technometrics*. **19**, 145–157 (1977).
36. Suissa, S. and Shuster, J. J. 'Exact unconditional sample sizes for the 2 × 2 binomial trial', *Journal of the Royal Statistical Society, Series A*, **148**, 317–327 (1985).
37. Storer, B. E. and Kim, C. 'Exact properties of some exact test statistics for comparing two binomial proportions', *Journal of the American Statistical Association*, **85**, 146–155 (1990).
38. Camilli, G. and Hopkins, K. D. 'Testing for association in 2 × 2 contingency tables with very small sample sizes', *Psychological Bulletin*, **86**, 1011–1014 (1979).
39. Lloyd, C. J. 'Doubling the one-sided *p*-value in testing independence in 2 × 2 tables against a two-sided alternative', *Statistics in Medicine*, **7**, 1297–1306 (1988).
40. Lehmann, E. L. *Testing Statistical Hypothesis*, 2nd edn, Wiley, New York, 1986.
41. Dupont, W. D. 'Sensitivity of Fisher's exact test to minor perturbations in 2 × 2 contingency tables', *Statistics in Medicine*, **5**, 629–635 (1986).
42. Stone, M. 'The role of significance testing: some data with a message', *Biometrika*, **56**, 485–593 (1969).

43. Hirji, K. F., Mehta, C. R. and Patel, N. R. 'Computing distributions for exact logistic regression', *Journal of the American Statistical Association*, **82**, 1110–1117 (1987).
44. Hirji, K. F., Mehta, C. R. and Patel, N. R. 'Exact inference for matched case-control studies', *Biometrics*, **44**, 803–814 (1988).
45. Mehta, C. R. and Patel, N. R. 'A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables', *Journal of the American Statistical Association*, **78**, 427–434 (1983).
46. Pagano, M. and Tritchler, D. 'Algorithms for the analysis of several 2×2 contingency tables', *SIAM Journal of Scientific and Statistical Computing*, **4**, 302–309 (1983).
47. Hirji, K. F. 'A comparison of exact, mid-*P* and score tests for matched case-control studies', *Biometrics*, (in press) (1990).