Biometrics 59, 441–450
June 2003

A Cautionary Note on Exact Unconditional Inference for a Difference between Two Independent Binomial Proportions

Devan V. Mehrotra,[1,*] Ivan S. F. Chan,[1] and Roger L. Berger[2]

[1]Merck Research Laboratories, UN-A102, 785 Jolly Rd., Bldg. C, Blue Bell, Pennsylvania 19422, U.S.A.
[2]Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
[*]*email:* devan_mehrotra@merck.com

Summary. Fisher's exact test for comparing response proportions in a randomized experiment can be overly conservative when the group sizes are small or when the response proportions are close to zero or one. This is primarily because the null distribution of the test statistic becomes too discrete, a partial consequence of the inference being *conditional* on the total number of responders. Accordingly, exact *unconditional* procedures have gained in popularity, on the premise that power will increase because the null distribution of the test statistic will presumably be less discrete. However, we caution researchers that a poor choice of test statistic for exact unconditional inference can actually result in a substantially less powerful analysis than Fisher's conditional test. To illustrate, we study a real example and provide exact test size and power results for several competing tests, for both balanced and unbalanced designs. Our results reveal that Fisher's test generally outperforms exact unconditional tests based on using as the test statistic either the observed difference in proportions, or the observed difference divided by its estimated standard error under the alternative hypothesis, the latter for unbalanced designs only. On the other hand, the exact unconditional test based on the observed difference divided by its estimated standard error under the null hypothesis (score statistic) outperforms Fisher's test, and is recommended. Boschloo's test, in which the p-value from Fisher's test is used as the test statistic in an exact unconditional test, is uniformly more powerful than Fisher's test, and is also recommended.

Key words: Berger and Boos confidence interval search; Boschloo's test; Conditional test; Discreteness; Fisher's exact test; Score test; 2×2 contingency table.

1. Introduction

Suppose that a trial is planned in which subjects will be randomized to receive one of two experimental treatments. At the end of the trial, each subject will be classified as being either a responder or nonresponder to the assigned treatment, based on certain criteria. A common dilemma for the researcher is deciding what method to prespecify for assessing the statistical significance of an observed difference in response proportions.

Among several existing methods, Fisher's exact test has traditionally been a popular choice in practice. This is presumably because: (a) unlike the common asymptotic theory–based tests, Fisher's test guarantees that the type I error rate will not exceed a prescribed level, and (b) it is implemented in widely used commercial software packages. However, Fisher's test involves conditioning on the group sizes as well as the observed total number of responders. While the latter conditioning eliminates the need to deal with a nuisance parameter, namely, the common true response proportion under the null hypothesis, it can result in an overly discrete null distribution of the test statistic. This makes the test arguably too conservative if the nominal significance level is fixed in advance (e.g., 5%), as is commonly done.

To circumvent the conservatism of exact conditional inference based on Fisher's test, Barnard (1945, 1947) proposed the use of exact *unconditional* inference, based on elimination of the nuisance parameter by maximization. However, invoking Fisher's principle of ancillarity (see Basu, 1977; Little, 1989), Barnard (1949) subsequently renounced his unconditional test. Despite his renouncement and other publications in support of Fisher's test (Yates, 1984; Barnard, 1989, Upton, 1992), exact unconditional inference has gained in popularity over the last few decades (Berkson, 1978; Kempthorne, 1979; Santner and Snell, 1980; Upton, 1982; Suissa and Shuster, 1985; Haber, 1986; D'Agostino, Chase, and Belanger, 1988; Rice, 1988; Haviland, 1990; Storer and Kim, 1990; Andres and Mato, 1994).

The key points of the conditional versus unconditional inference debate can be summarized as follows. Consider our aforementioned randomized comparative binomial trial. Under the null hypothesis that the probability of a subject being a responder is independent of the treatment group to which he/she is randomized, the total number of responders is predetermined. Hence, "conditionalists" argue that the statistical significance of an observed difference in response proportions should be assessed relative to its permutation

distribution in hypothetical repetitions of randomized treatment assignment after conditioning on the observed total number of responders. Moreover, they assert that the latter contains almost no information about the veracity of the null hypothesis, so little is lost by conditioning. However, "unconditionalists" maintain that by not conditioning on the observed total number of responders, the null distribution of the test statistic will presumably be less discrete, hence making it possible to construct a test with true size closer to the nominal level, and thereby resulting in a more powerful test. Moreover, they note that the observed total number of responders does contain some information about the true difference in response proportions, and is therefore not strictly an ancillary statistic (Berkson, 1978). By not conditioning, using that additional (albeit negligible) information should, in theory, increase our ability to reject a false null hypothesis.

The purpose of our article is not to contribute toward the ongoing debate on conditional versus unconditional inference. Rather, it is to caution researchers that a poor choice of test statistic for an exact unconditional test can actually result in a *substantially less powerful* analysis than Fisher's conditional test!

The rest of this article is organized as follows. In Section 2, we briefly review Fisher's exact conditional test and several unconditional tests. In Section 3, we compare the performance of the competing tests, first using a real data example that motivated this research, and then via exact test size and power results under a wide range of conditions. We conclude by making some practical recommendations in Section 4.

2. Review of Tests to be Compared

The notation in the contingency table below will be used to review the various tests that we compare. The X_i's are independent, binomial random variables with the sample sizes, N_i's, fixed by design.

	Group 1	Group 2	Total
Responders	X_1	X_2	$T = X_1 + X_2$
Nonresponders	$N_1 - X_1$	$N_2 - X_2$	$N - T$
Total	N_1	N_2	N

Let θ_i denote the true response probability for group i, and let $\theta_1 = \theta_2 = \theta$ under the null hypothesis $H_0 : \theta_1 = \theta_2$. Our alternative hypothesis is $H_a : \theta_1 \neq \theta_2$ (two-sided).

Fisher's (1935) *Exact Conditional Test (F)*

The null distribution of X_1 given T is given by

$$P_{H_0}(X_1 = x_1 \mid T = t) = \frac{\binom{N_1}{x_1}\binom{N_2}{t - x_1}}{\sum_{j \in G}\binom{N_1}{j}\binom{N_2}{t-j}}, \qquad (1)$$

where $G = \{j : \max(0, t - N_2) \leq j \leq \min(t, N_1)\}$. Note that this conditional distribution does not depend on the nuisance parameter θ. Let x_i denote the observed value of the random variable X_i. An exact two-sided p-value can be obtained in several different ways for Fisher's exact test (Agresti, 1992). Following Blaker (2000) and Agresti and Min (2001), we calculate it as

$$p_F(x_1, x_2) = \sum_{i \in H} P_{H_0}(X_1 = i \mid T = t), \qquad (2)$$

where $H = \{i : g(i, t) \leq g(x_1, t)\}$ and $g(x, t) = \min\{P_{H_0}(X_1 \leq x \mid T = t), P_{H_0}(X_1 \geq x \mid T = t)\}$.

Z-Pooled Exact Unconditional Test (Z_P)

If we do not condition on T, then under H_0 and for a given θ,

$$P_{H_0}(X_1 = x_1, X_2 = x_2 \mid \theta) = \binom{N_1}{x_1}\binom{N_2}{x_2}\theta^{x_1+x_2}(1-\theta)^{N_1+N_2-x_1-x_2}. \qquad (3)$$

Note that this unconditional distribution depends on the nuisance parameter θ. Suissa and Shuster's (1985) "Z-pooled" statistic, obtained by dividing the observed difference in response proportions by its estimated standard error under the null hypothesis is

$$Z_P(x_1, x_2) = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\tilde{\theta}(1-\tilde{\theta})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}, \qquad (4)$$

where $\hat{\theta}_i = x_i/N_i$, and $\tilde{\theta} = (x_1 + x_2)/N$ is the maximum likelihood estimate of θ under H_0 based on pooling the data together from the two groups. Note that (4) is a score statistic. We calculate an exact two-sided p-value based on (4) as

$$p_{Z_p}(x_1, x_2) = \sup_{0 \leq \theta \leq 1}\left\{\sum_{i=0}^{N_1}\sum_{j=0}^{N_2} P_{H_0}(X_1 = i, X_2 = j \mid \theta) \times I_{|Z_P(i,j)| \geq |Z_P(x_1,x_2)|}\right\}, \qquad (5)$$

where $I_E = 1$ if the event E is true, and 0 otherwise. Note that while the nuisance parameter is eliminated in Fisher's exact test by conditioning on the sufficient statistic T, the above test considers all possible values of θ in its domain and selects the one that produces the largest p-value. This approach guarantees that the type I error rate is less than or equal to the nominal level.

Modified Z-pooled Exact Unconditional Test (Z_P^*)

Since the p-value in (5) is obtained by searching over the entire domain of the nuisance parameter and catering to the worst case, it can result in a potentially conservative unconditional test (Mehta and Hilton, 1993). To reduce the conservatism in situations such as this, Berger and Boos (1994) proposed that the search be restricted only to "plausible" values of θ given the observed data, with an appropriate adjustment to ensure control of the type I error

rate. Their proposal is used to modify the Z_P test as follows. If $(\theta_L^\gamma, \theta_U^\gamma)$ denotes an exact $100\,(1-\gamma)\%$ confidence interval for θ, then an exact two-sided p-value for Z_P^*, the modified Z_P test, is calculated as

$$p_{Z_P^*}(x_1, x_2) = \sup_{\theta_L^\gamma \le \theta \le \theta_U^\gamma} \left\{ \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} P_{H_0}\left(X_1 = i, X_2 = j \mid \theta\right) \times I_{|Z_p(i,j)| \ge |Z_p(x_1,x_2)|} \right\} + \gamma, \tag{6}$$

where γ is a very small positive number, typically 0.001 or 0.0001. For simplicity, we obtain $(\theta_L^\gamma, \theta_U^\gamma)$ using the Clopper and Pearson (1934) approach:

$$\theta_L^\gamma = \frac{(x_1 + x_2)}{(x_1 + x_2) + (N - x_1 - x_2 + 1)\, F\left\langle 2\,(N - x_1 - x_2 + 1)\,, 2(x_1 + x_2); \frac{\gamma}{2} \right\rangle}$$

and

$$\theta_U^\gamma = \frac{(x_1 + x_2 + 1)\, F\left\langle 2\,(x_1 + x_2 + 1)\,, 2(N - x_1 - x_2); .5\gamma \right\rangle}{(x_1 + x_2 + 1)\, F\left\langle 2\,(x_1 + x_2 + 1)\,, 2(N - x_1 - x_2); \frac{\gamma}{2} \right\rangle + (N - x_1 - x_2)},$$

where $F\langle a, b; c\rangle$ is the $100\,(1-c)$ percentile of a central $F\,(a, b)$ distribution.

Z-Unpooled Exact Unconditional Test (Z_U)

This test was also studied by Suissa and Shuster (1985). Here, the observed difference in response proportions is divided by its standard error under the alternative hypothesis. The test statistic is

$$Z_U(x_1, x_2) = \frac{\hat\theta_1 - \hat\theta_2}{\sqrt{\dfrac{\hat\theta_1(1-\hat\theta_1)}{N_1} + \dfrac{\hat\theta_2(1-\hat\theta_2)}{N_2}}}. \tag{7}$$

Analogous to (5), we calculate the exact two-sided p-value based on (7) as

$$p_{Z_U}(x_1, x_2) = \sup_{0 \le \theta \le 1} \left\{ \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} P_{H_0}(X_1 = i, X_2 = j \mid \theta) \times I_{|Z_U(i,j)| \ge |Z_U(x_1,x_2)|} \right\}. \tag{8}$$

Modified Z-Unpooled Exact Unconditional Test (Z_U^*)

Analogous to (6), we calculate an exact two-sided p-value for the modified Z_U test (Z_U^*) as

$$p_{Z_U^*}(x_1, x_2) = \sup_{\theta_L^\gamma \le \theta \le \theta_U^\gamma} \left\{ \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} P_{H_0}(X_1 = i, X_2 = j \mid \theta) \times I_{|Z_U(i,j)| \ge |Z_U(x_1,x_2)|} \right\} + \gamma. \tag{9}$$

Santner and Snell's Exact Unconditional Test (D)

Santner and Snell (1980) proposed a method for constructing an exact unconditional confidence interval for $\theta_1 - \theta_2$ based on the test statistic

$$D\,(x_1, x_2) = \hat\theta_1 - \hat\theta_2. \tag{10}$$

Although they did not explicitly propose this statistic for hypothesis testing, in practice, the resulting confidence interval could be used to reject $H_0 : \theta_1 = \theta_2$ if it does not contain zero. We included this in our comparison because, until recently (August 2002), the Santner and Snell exact confidence interval for $\theta_1 - \theta_2$ was the only option available in the widely used StatXact-4 software. (StatXact-5 now provides other options.) We calculate an exact two-sided p-value based on (10) as

$$p_D(x_1, x_2) = \sup_{0 \le \theta \le 1} \left\{ \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} P_{H_0}(X_1 = i, X_2 = j \mid \theta) \times I_{|D(i,j)| \ge |D(x_1,x_2)|} \right\}. \tag{11}$$

Modified Santner and Snell's Exact Unconditional Test (D^*)

Analogous to (6), we calculate an exact two-sided p-value for the modified D test (D^*) as

$$p_{D^*}(x_1, x_2) = \sup_{\theta_L^\gamma \le \theta \le \theta_U^\gamma} \left\{ \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} P_{H_0}\left(X_1 = i, X_2 = j \,|\, \theta\right) \times I_{|D(i,j)| \ge |D(x_1,x_2)|} \right\} + \gamma. \tag{12}$$

Boschloo's (1970) *Exact Unconditional Test (B)*

This test was also proposed by McDonald, Davis, and Milliken (1977). The p-value from Fisher's conditional test is used as the test statistic in this unconditional test. We calculate an exact two-sided p-value for this test as

$$p_B\,(x_1, x_2) = \sup_{0 \le \theta \le 1} \left\{ P_{H_0}\{p_F(X_1, X_2) \le p_F(x_1, x_2)\} \,|\, \theta \right\}. \tag{13}$$

Modified Boschloo's Exact Unconditional Test (B^*)

Analogous to (6), we calculate an exact two-sided p-value for the modified B test (B^*) as

$$p_{B^*}(x_1,x_2)=\sup_{\theta_L^{\gamma}\le\theta\le\theta_U^{\gamma}}\left\{P_{H_0}\{p_F(X_1,X_2)\le P_F(x_1,x_2)\}\mid\theta\right\}+\gamma. \tag{14}$$

Z-Pooled Asymptotic Unconditional Test ($\tilde{Z}_P$)

We included this because it is a widely used nonexact test for comparing two independent binomial proportions. The test statistic is the same as (4), but instead of calculating the p-value using (5), it is obtained using a normal approximation. Under H_0 and certain regularity conditions, the distribution of $Z_P(x_1,x_2)$ converges to a standard normal distribution. Accordingly, an approximate two-sided p-value can be calculated as

$$p_{\tilde{Z}_P}=2[1-\Phi\{|Z_P(x_1,x_2)|\}], \tag{15}$$

where $\Phi(\cdot)$ denotes the c.d.f. of a standard normal distribution. Note that (15) is the same as the Pearson chi-squared p-value. It is well known that, unlike the exact tests, this asymptotic test can yield type I error rates that exceed the nominal level.

Z-Unpooled Asymptotic Unconditional Test ($\tilde{Z}_U$)

The null distribution of $Z_U(x_1,x_2)$ also converges to a standard normal distribution under certain regularity conditions. Accordingly, an approximate two-sided p-value can be calculated as

$$p_{\tilde{Z}_U}=2\left[1-\Phi\left\{|Z_U(x_1,x_2)|\right\}\right]. \tag{16}$$

As with the asymptotic $\tilde{Z}_P$ test, using $\tilde{Z}_U$ can also result in type I error rates that exceed the nominal level, sometimes by a substantial amount, as we will illustrate in Section 3.2.

3. Comparison of Tests

We first compare the different tests using a numerical example with real data, and then via exact test size and power results under a wide variety of conditions.

3.1 *Illustrative Example*

In a recent clinical trial conducted by Merck Research Laboratories, there was interest in assessing whether the observed between-group difference in the proportion of subjects experiencing a certain type of rash was statistically significant. The observed proportions were $\hat{\theta}_1=8/148=5.41\%$ for group 1, and $\hat{\theta}_2=1/132=0.76\%$ for group 2, resulting in an observed difference of 4.65%. For these data, the two-sided p-values based on the exact tests described earlier are shown below. Note that, throughout this article, we use $\gamma=.001$ when doing the confidence interval search for the nuisance parameter, as suggested by Berger and Boos (1994). As a result, for tests D^*, B^*, Z_P^*, and Z_U^*, the search for the maximum p-value in this example is restricted to (0.0080, 0.0826), the exact 99.9% confidence interval for θ based on the Clopper and Pearson (1934) method.

Conditional	*Unconditional*							
F	D	D^*	B	B^*	Z_P	Z_P^*	Z_U	Z_U^*
.0388	.4386	.1603	.0347	.0325	.0291	.0282	.0229	.0215

This real data example vividly illustrates that the choice of test statistic for exact unconditional inference can make a big difference even when the sample sizes are quite large. Note that the p-values for the unconditional tests D and D^* are both considerably larger than the p-value for Fisher's conditional test. However, the p-values for the other unconditional tests are each slightly smaller than the p-value for F. Also, observe that the restricted search for the nuisance parameter results in a smaller p-value for each test, i.e., the p-values for D^*, B^*, Z_P^*, and Z_U^* are smaller than the p-values for D, B, Z_P and Z_U, respectively. This is often (but not always) the case, for reasons elucidated by Berger (1996).

To appreciate why the p-value for the D test is so much larger than, say, the Z_P test, it is instructive to note that there are a total of $149\times133=19{,}817$ possible 2×2 tables under the unconditional sampling space. A closer look into this sampling space reveals that 18,034 tables are considered at least as extreme as that observed based on $|D(x_1,x_2)|$, compared with only 15,776 tables based on $|Z_P(x_1,x_2)|$. Unlike the latter statistic, the former focuses only on the observed difference and does not take into account its associated variability. This results in fewer distinct values for $|D(x_1,x_2)|$, i.e., more discreteness, thereby reducing the ability of the D test to detect differences in response proportions.

3.2 *Exact Type I Error Rate and Test Size*

The rejection region of a level-α test A is given by

$$R_A=\{(a,b):(a,b)\in\Omega \text{ and } p_A(a,b)\le\alpha\}, \tag{17}$$

where $\Omega=\{0,1,\ldots,N_1\}\times\{0,1,\ldots,N_2\}$ is the sample space of (X_1,X_2), and $p_A(a,b)$ is the two-sided p-value, calculated as described in Section 2. When $\theta_1=\theta_2=\theta$, the exact type I error rate of test A is

$$\alpha_A(\theta)=P_{H_0}\{(X_1,X_2)\in R_A\}=\sum_{(a,b)\in R_A}P_{H_0}(X_1=a,X_2=b\mid\theta), \tag{18}$$

where $P_{H_0}(X_1=a,X_2=b\mid\theta)$ is the point null probability given by (3). The size of test A is given by

$$\alpha_A^{\max}=\max_{0\le\theta\le1}\alpha_A(\theta). \tag{19}$$

In Table 1, for each test, we report the exact type I error rate at selected values of $\theta_1=\theta_2=\theta$, as well as the maximum type I error rate over the entire range of θ (test size). We investigated balanced as well as several unbalanced designs, with total sample sizes (N_1+N_2) of 20, 50, 100, and 300. Results are shown for $N_1=N_2$ and $N_1=4N_2$, for $\alpha=5\%$. Results for $\alpha=1\%$ and 10% for the other unbalanced designs were qualitatively similar, and are omitted for brevity (but available upon request).

Table 1 reinforces the need for an exact test if there is a desire to keep the type I error rate strictly below the nominal level. In every case shown, the sizes of both asymptotic tests exceed 5%, even when the total sample size is as large as 300. When $N_1=N_2$, the type I error rates for $\tilde{Z}_U$ are either equal to or slightly greater than those for $\tilde{Z}_P$. However, when $N_1=4N_2$, there is no consistent trend. For example,

Table 1

Type I error rates at selected (θ_1, θ_2) and test sizes ($\alpha = 5\%$, 2-sided) of exact and asymptotic tests

(N_1, N_2)	θ_1	θ_2	F	D	D^*	B	B^*	Z_P	Z_P^*	Z_U	Z_U^*	$\tilde{Z}_P$	$\tilde{Z}_U$
						$\boldsymbol{N_1 = N_2}$							
(10, 10)	.02	.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	.10	.10	0.1	0.1	0.1	0.9	0.9	0.9	0.9	0.9	0.9	0.9	5.0
	.25	.25	1.0	1.8	1.8	3.5	3.5	3.5	3.5	3.5	3.5	3.5	9.4
	.50	.50	1.3	4.1	4.1	4.2	4.2	4.2	4.2	4.2	4.2	4.2	8.8
	Test size:[a]		1.3	4.1	4.1	4.2	4.2	4.2	4.2	4.2	4.2	4.2	9.5
(25, 25)	.02	.02	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.2	0.2
	.10	.10	0.1	0.1	0.1	1.9	1.9	3.9	3.9	3.9	3.9	4.9	4.9
	.25	.25	2.2	1.4	1.4	4.1	4.1	4.2	4.2	4.2	4.2	5.4	6.4
	.50	.50	3.3	3.3	3.3	3.7	3.7	3.7	3.7	3.7	3.7	6.5	6.5
	Test size:[a]		3.3	3.3	3.3	4.6	4.6	4.6	4.6	4.6	4.6	6.5	6.5
(50, 50)	.02	.02	0.0	0.0	0.0	0.2	0.2	1.3	1.3	1.3	1.3	1.3	1.3
	.10	.10	1.8	0.1	0.3	3.2	3.6	3.8	4.8	3.8	4.8	5.1	5.9
	.25	.25	3.0	1.5	1.6	4.1	4.1	4.5	4.7	4.5	4.7	5.1	5.7
	.50	.50	3.5	3.5	3.5	4.2	4.2	4.2	4.2	4.2	4.2	5.7	5.7
	Test size:[a]		3.5	3.5	3.5	4.9	4.9	4.9	4.9	4.9	4.9	5.7	6.1
(150, 150)	.02	.02	1.2	0.0	0.0	2.3	2.9	4.6	4.6	4.6	4.6	4.6	4.6
	.10	.10	3.3	0.1	1.0	3.9	4.5	4.8	4.8	4.8	4.8	5.0	5.2
	.25	.25	3.7	2.0	2.7	4.6	4.8	4.8	4.8	4.8	4.8	5.0	5.2
	.50	.50	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	4.3	5.7	5.7
	Test size:[a]		4.3	4.3	4.3	5.0	4.9	5.0	4.9	5.0	4.9	5.7	5.7
						$\boldsymbol{N_1 \neq N_2}$							
(16, 4)	.02	.02	0.2	0.0	0.0	0.2	0.2	0.2	0.2	0.0	0.0	5.7	0.2
	.10	.10	1.2	0.3	0.3	2.9	2.9	2.9	2.9	0.0	0.0	8.3	5.8
	.25	.25	1.5	1.3	1.3	3.7	3.7	3.7	3.7	0.4	0.4	4.3	22.4
	.50	.50	1.4	3.0	3.0	3.4	3.4	3.4	3.4	2.8	2.8	5.6	14.3
	Test size:[a]		1.5	3.0	3.0	3.9	3.9	3.9	3.9	2.8	2.8	9.0	22.8
(40, 10)	.02	.02	0.8	0.0	0.0	0.8	0.8	1.3	1.3	0.0	0.0	8.8	0.7
	.10	.10	2.4	0.2	0.4	2.6	2.6	3.9	3.9	0.6	0.6	4.5	20.8
	.25	.25	3.4	1.6	1.6	4.3	4.3	4.1	4.1	4.1	4.1	4.7	9.5
	.50	.50	2.9	3.9	3.9	3.5	3.5	4.1	4.6	1.5	1.5	5.5	8.9
	Test size:[a]		3.4	3.9	3.9	4.8	4.8	4.3	4.7	4.5	4.5	9.1	21.2
(80, 20)	.02	.02	1.5	0.0	0.0	1.6	1.6	3.3	3.3	0.0	0.0	8.7	5.2
	.10	.10	2.7	3.4	0.2	3.7	3.7	3.2	3.2	3.4	3.4	3.7	13.0
	.25	.25	3.3	1.7	2.1	4.6	4.3	4.6	4.8	1.2	2.0	5.1	6.9
	.50	.50	4.3	4.0	4.0	4.3	4.3	4.1	4.6	0.6	4.1	5.1	6.6
	Test size:[a]		4.3	4.0	4.0	5.0	4.7	4.6	4.9	3.9	4.1	9.2	21.3
(240, 60)	.02	.02	1.9	0.0	0.2	2.4	3.2	4.0	4.0	0.7	0.7	4.3	21.3
	.10	.10	3.5	0.1	1.2	3.9	4.9	4.4	4.7	1.2	2.5	4.7	7.0
	.25	.25	4.1	2.1	2.9	4.6	4.7	4.4	4.8	0.5	4.6	4.9	5.7
	.50	.50	4.3	4.6	4.6	4.3	4.3	4.6	4.6	0.4	4.9	5.3	5.5
	Test size:[a]		4.3	4.6	4.6	4.9	4.9	4.6	4.9	4.3	4.9	9.2	21.4

[a]Test size = maximum type I error rate over $0 \leq \theta\,(= \theta_1 = \theta_2) \leq 1$ for the given (N_1, N_2).

when $N_1 = 40$ and $N_2 = 10$, $\alpha_{\tilde{Z}_P}(\theta = .02) = 8.8\%$ and $\alpha_{\tilde{Z}_U}(\theta = .02) = 0.7\%$. However, when $N_1 = 240$ and $N_2 = 60$, $\alpha_{\tilde{Z}_P}(\theta = .02) = 4.3\%$ while $\alpha_{\tilde{Z}_U}(\theta = .02) = 21.3\%$! In general, $\alpha^{\max}_{\tilde{Z}_P}$ is closer to α than is $\alpha^{\max}_{\tilde{Z}_U}$, so $\tilde{Z}_P$ may be preferred over $\tilde{Z}_U$ if an asymptotic test is required.

For all the exact methods, both conditional and unconditional, when $N_1 + N_2 \leq 50$ and $\theta = .02$, the type I error rates are substantially smaller than 5%, and range from 0% to 1.3%. However, they generally get closer to the nominal level as θ approaches 0.50 or as the sample sizes increase, the latter due to the null distributions becoming less discrete. Two notable exceptions are D and D^*: for all sample sizes shown, $\alpha_D(\theta = .02) = 0\%$ and $\alpha_{D^*}(\theta = .02) \leq 0.2\%$! When comparing Fisher's conditional test with the unconditional tests, only one comparison is definitive: since $R_F \subseteq R_B$ (Boschloo, 1970), $\alpha_F(\theta) \leq \alpha_B(\theta)$ for all θ. However, in a majority of cases shown in Table 1, especially when $\theta \leq .25$, $\alpha_F(\theta) \geq \alpha_D(\theta)$ and $\alpha_F(\theta) \geq \alpha_{D^*}(\theta)$. The comparison of F with Z_P and Z_U depends on the sample size configuration. When $N_1 = N_2$, since Z_P and Z_U are increasing functions of each other (Suissa and Shuster, 1985), they give the same ordering of the sample space, so $\alpha_{Z_P}(\theta) = \alpha_{Z_U}(\theta)$, and $\alpha_{Z_P^*}(\theta) = \alpha_{Z_U^*}(\theta)$ for all θ.

Table 1 reveals that when $N_1 = N_2$, $\alpha_{Z_P}(\theta) = \alpha_{Z_U}(\theta) > \alpha_F(\theta)$ and $\alpha_{Z_P^*}(\theta) = \alpha_{Z_U^*}(\theta) > \alpha_F(\theta)$ for all cases shown. However, when $N_1 \neq N_2$, while the same holds true almost always for Z_P and Z_P^*, that is not the case for Z_U and Z_U^*. For the latter, in most unbalanced cases shown in Table 1, $\alpha_F(\theta) > \alpha_{Z_U}(\theta)$ and $\alpha_F(\theta) > \alpha_{Z_U^*}(\theta)$. As an aside, note that the Berger and Boos restricted (confidence interval) search method usually (but not always) brings the type I error rate closer to the nominal level compared with the unrestricted search, notably so for Z_U with unequal sample sizes and θ close to 0.50.

In summary, among the exact tests considered, Z_P^* and Z_P have type I error rates that are closest to the nominal level, followed closely by B^* and B. However, the exact unconditional tests D, D^*, Z_U, and Z_U^* generally do a poor job of "spending" the allotted α, and, in that regard, are notably more conservative than Fisher's conditional test over a wide range of θ.

3.3 *Exact Power*

The exact power function of a level-α test A with rejection region (17) is given by

$$\text{power}_A(\theta_1, \theta_2) = \sum_{(a,b)\in R_A} \binom{N_1}{a}\binom{N_2}{b}\theta_1^a(1-\theta_1)^{N_1-a}\theta_2^b(1-\theta_2)^{N_2-b}. \tag{20}$$

In keeping with the Neyman-Pearson paradigm, we discuss power comparisons among the exact tests only, since the asymptotic tests do not consistently preserve the test size.

Table 2 summarizes overall pairwise power comparisons across the entire H_a parameter space ($0 \le \theta_1 \neq \theta_2 \le 1$), for equal and unequal sample size configurations, when $\alpha = 5\%$. The row and column headings indicate which pair of tests is being compared. An entry of "=" means that the power

Table 2

Power comparison of the exact tests across the entire parameter space for equal and unequal sample size configurations

	F	D	D^*	B	B^*	Z_P	Z_P^*	Z_U
Z_U^*	> 62 42	> 38 <	> 38 <	= 29 <	= 29 <	= 27 <	= 27 <	= < =
	> 51 14	> 45 19	> 47 19	> 38 12	> 47 12	= 26 12	= 26 12	= 43 =
	> 65 29	> 69 38	> 73 37	> 48 22	> 51 27	> 34 28	= 37 20	> > >
	> 71 57	> > 91	> > 56	60 42 49	> 45 49	< 46 46	= 47 40	< > >
Z_U	> 72 42	> 57 <	> 57 <	= 48 <	= 48 <	= 46 <	= 46 <	
	> 48 14	> 43 19	> 44 19	> 34 12	> 44 12	= 23 12	= 23 12	
	> 59 06	> 57 13	> 58 11	> 19 04	> 19 05	= 19 04	< 22 04	
	> 04 01	> 44 10	> 28 05	> 02 01	> 02 01	= 02 01	> 02 01	
Z_P^*	> > >	> > >	> > >	= > =	= > =	= = =		
	> > >	> 68 >	> > >	> 86 69	> > 69	= = >		
	> > >	> 80 >	> > >	> 67 76	> 70 91	> < >		
	> 85 >	> > >	> > >	60 55 87	> 49 85	< 23 >		
Z_P	> > >	> > >	> > >	= > =	= > =			
	> > >	> 68 >	> > >	> 86 61	> > 61			
	> > 83	> 85 >	> > >	> 74 44	> 78 60			
	> 88 99	> > >	> > >	> 59 54	> 55 40			
B^*	> > >	> > >	> > >	= = =				
	> > >	> 54 85	> 60 85	= < =				
	> > >	> 78 >	> 82 >	> 37 <				
	> > >	> > 82	> > 82	47 36 59				
B	> > >	> > >	> > >					
	> > >	> 58 85	> 64 85					
	> > >	> 81 >	> 85 >					
	> > >	> > 83	> > 83					
D^*	> 70 58	= = =						
	< 60 36	= < >						
	< 37 <	> 48 >						
	< < 27	> > >						
D	> 70 58							
	< 60 36							
	< 40 <							
	< < 26							

sample sizes

(10, 10)	(13, 7)	(16, 4)
(25, 25)	(33, 17)	(40, 10)
(50, 50)	(65, 35)	(80, 20)
(150, 150)	(200, 100)	(240, 60)

= indicates row and column tests are identical.

> indicates row test is uniformly more powerful than column test.

< indicates column test is uniformly more powerful than row test.

Numeric value is estimate of percent of alternative space over which row test has higher power than column test.

functions of the two tests are identical, while ">" ("<") means that the row test is uniformly more (less) powerful than the column test. If none of the preceding uniform conditions apply, we report the approximate proportion of points in the H_a space at which the row test has greater power than the column test for the given sample size configuration. As an example, consider Z_U versus B. For $(N_1, N_2) = (10, 10)$, Table 2 reveals that Z_U and B have identical power, but for the other equal sample size cases, Z_U is uniformly more powerful than B. However, for $(N_1, N_2) = (16, 4)$, Z_U is uniformly less powerful than B, while for the other unequal sample size cases, Z_U is more powerful than B between 1% and 48% of the time under H_a.

When comparing Fisher's conditional test with each of the exact unconditional tests, only one comparison is definitive: B is uniformly more powerful than F, by construction (Boschloo, 1970). Tests B^*, Z_P, and Z_P^* for all sample sizes, and Z_U and Z_U^* for equal sample sizes, are more powerful than F over most (but not all) points in the H_a space. Note that while B^* is more powerful than Fisher's test in every case shown in Table 2, it is not uniformly more powerful than F in all cases. For example, among cases not shown, when $(N_1, N_2) = (200, 100)$ and $\alpha = 1\%$, F is more powerful than B^* over approximately 31% of the H_a space.

It is clear from Table 2 that tests D and D^* for moderate-to-large sample sizes, and tests Z_U and Z_U^* for unequal sample sizes, are less powerful than F over most (but not all) points in the H_a space. When $(N_1, N_2) = (10, 10)$, both D and D^* (with $\gamma = .001$) are uniformly more powerful than F, but the opposite is true for all the other equal sample cases. Interestingly, for the $N_1 \neq N_2$ cases, in general, as the total sample size increases, F becomes more powerful than D, D^*, and Z_U over an increasingly greater proportion of the H_a space. For example, when $N_1 = 4N_2$, F is more powerful than Z_U over approximately 58% of the H_a space when $N_1 + N_2 = 20$, but that percentage increases to 99% when $N_1 + N_2 = 300$.

Table 2 is helpful for overall power comparisons across the entire H_a space. However, the magnitude of the power differences is not revealed. Accordingly, in Table 3, we provide the power of each of the exact tests at selected values of (N_1, N_2) and (θ_1, θ_2). The sample size pairs mimic those in Table 1, while the latter were chosen such that, when $N_1 = N_2$, the power of Z_P and Z_U is approximately 0.80 for each total sample size shown. In each row of Table 3, [a] is attached to the power of an unconditional test if it is less than or equal to the corresponding power of F. Note that there are numerous cases where Fisher's conditional test is substantially more powerful than the unconditional tests D and D^* (for both equal and unequal sample sizes), and Z_U and Z_U^* (for unequal sample sizes). For example, when $(N_1, N_2) = (150, 150)$ and $(\theta_1, \theta_2) = (.02, .09)$, F has 70.4% power, while the power of D is only 4.0%, even less than α! The corresponding power for D^* is 43.1%, a dramatic improvement over D, but still notably less than F. For the same point in the H_a space and the same total sample size, but now with $(N_1, N_2) = (60, 240)$, the power of F (64.1%) is substantially greater than that of D (3.4%), D^* (38.2%), Z_U (2.8%), and Z_U^* (2.8%), and only a bit smaller than that of B (67.3%), B^*(68.5%), Z_P (68.9%), and Z_P^* (69.2%).

When testing at $\alpha = 5\%$, the following general conclusions can be drawn collectively from Tables 2 and 3. F is often substantially more powerful than D and D^*. However, B is uniformly more powerful (Boschloo, 1970), and B^*, Z_P, Z_P^*, Z_U, and Z_U^* (last two for $N_1 = N_2$ only) are almost always more powerful than F. When the total sample size is small to moderate (∼50 or less), and the true response proportions are, say, ≤0.15 or ≥.85, the power advantage over F can be quite large, typically 8 to 12 percentage points. However, in most other cases, the power advantage (if any) is much less, typically less than five percentage points. Tests Z_U and Z_U^* behave very erratically when $N_1 \neq N_2$, and depending on the combination of (N_1, N_2) and (θ_1, θ_2), either have the best or the worst power among all the exact unconditional tests. For example, consider the cases in the bottom part of Table 3, where $N_1 + N_2 = 100$. When $(N_1, N_2) = (80, 20)$ and $(\theta_1, \theta_2) = (.02, .18)$, both Z_U and Z_U^* have 2.4% power, which is substantially less than that of F (64.6%) and all the other unconditional tests. However, at the same sample size configuration, when $(\theta_1, \theta_2) = (.50, .77)$, Z_U^* (60.7%) has slightly higher power than F (59.5%), but Z_U (32.1%) continues to have much lower power. Note that when the sample sizes are switched, i.e., $(N_1, N_2) = (20, 80)$, an opposite pattern is seen! In that case, both Z_U and Z_U^* (62.7% each) are substantially more powerful than F (32.4%) and all the other unconditional tests at $(\theta_1, \theta_2) = (.02, .18)$. However, at $(\theta_1, \theta_2) = (.50, .77)$, both Z_U (18.2%) and Z_U^* (37.8%) are notably less powerful than F (59.3%) and all the other unconditional tests.

Note that the Berger and Boos (1994) confidence interval search method usually results in greater power compared with the unrestricted search, particularly for D and Z_U. Interestingly, for D (versus D^*), regardless of the sample sizes, the gain in power becomes more pronounced as the true proportions get further way from 0.50. However, for Z_U (versus Z_U^*), with unequal sample sizes, the gain in power becomes more notable as the true proportions get closer to 0.50. For Z_P (versus Z_P^*), in about one-third of the cases, Z_P^* is more powerful than Z_P, and there is only one case in which Z_P is more powerful than Z_P^*, but the gain in power is modest. For B (versus B^*) the power of B exceeds that of B^* in about as many cases as the power of B^* exceeds B, and the differences in power are not large.

4. Concluding Remarks

Through a real data example and a detailed study of the actual test size and power of several tests for both balanced and unbalanced designs, we have shown that the choice of test statistic is important when conducting exact unconditional inference for a difference between two independent binomial proportions. Fisher's conditional test generally outperforms exact unconditional tests based on using as the test statistic either $D(x_1, x_2)$, with equal or unequal sample sizes, or $Z_U(x_1, x_2)$ with unequal sample sizes. On the other hand, the exact unconditional test based on the score statistic, $Z_P(x_1, x_2)$, is generally more powerful than Fisher's test. Moreover, Boschloo's test, in which the p-value from Fisher's test is used as the test statistic in an exact unconditional test, is uniformly more powerful than Fisher's test.

Andres and Mato (1994) have also conducted a study of several exact tests for the two independent sample binomial

Table 3
Power (%) of each exact test at selected (θ_1, θ_2)

(N_1, N_2)	θ_1	θ_2	F	D	D^*	B	B^*	Z_P^*	Z_P^*	Z_U	Z_U^*
						$\boldsymbol{N_1 = N_2}$					
(10, 10)	.02	.54	62.8	66.9	66.9	80.8	80.8	80.8	80.8	80.8	80.8
	.10	.68	62.9	77.7	77.7	79.4	79.4	79.4	79.4	79.4	79.4
	.25	.84	61.9	78.9	78.9	79.2	79.2	79.2	79.2	79.2	79.2
	.50	.99	57.9	59.9	59.9	78.4	78.4	78.4	78.4	78.4	78.4
(25, 25)	.02	.29	68.1	36.8[a]	36.8[a]	76.4	76.4	80.4	80.4	80.4	80.4
	.10	.45	72.3	66.9[a]	66.9[a]	80.9	80.9	80.9	80.9	80.9	80.9
	.25	.65	78.4	78.4[a]	78.4[a]	80.4	80.4	80.4	80.4	80.4	80.4
	.50	.86	70.8	69.3[a]	69.3[a]	79.7	79.7	79.7	79.7	79.7	79.7
(50, 50)	.02	.18	68.6	19.1[a]	39.6[a]	77.8	78.6	79.5	82.2	79.5	82.2
	.10	.33	75.7	60.0[a]	65.3[a]	80.1	80.3	80.7	82.1	80.7	82.1
	.25	.52	74.5	74.1[a]	74.1[a]	80.1	80.1	80.1	80.1	80.1	80.1
	.50	.77	75.3	74.3[a]	74.3[a]	80.7	80.7	80.8	80.8	80.8	80.8
(150, 150)	.02	.09	70.4	4.0[a]	43.1[a]	75.0	77.7	79.4	79.4	79.4	79.4
	.10	.22	77.6	53.1[a]	69.0[a]	80.6	81.5	81.5	81.5	81.5	81.5
	.25	.40	76.1	73.4[a]	73.8[a]	78.9	78.5	78.9	78.9	78.9	78.9
	.50	.66	78.0	78.0[a]	78.0[a]	80.0	79.7	80.0	79.7	80.0	79.7
						$\boldsymbol{N_1 \neq N_2}$					
(16, 4)	.02	.54	64.2	37.4[a]	37.4[a]	73.0	73.0	73.0	73.0	8.5[a]	8.5[a]
	.10	.68	58.3	53.1[a]	53.1[a]	73.5	73.5	73.5	73.5	21.4[a]	21.4[a]
	.25	.84	47.9	53.3	53.3	61.9	61.9	61.9	61.9	45.8[a]	45.8[a]
	.50	.99	10.1	21.8	21.8	21.9	21.9	21.9	21.9	21.8	21.8
(4, 16)	.02	.54	16.3	31.0	31.0	31.2	31.2	31.2	31.2	31.0	31.0
	.10	.68	41.0	52.9	52.9	56.6	56.6	56.6	56.6	50.8	50.8
	.25	.84	53.7	53.2[a]	53.2[a]	68.4	68.4	68.4	68.4	31.4[a]	31.4[a]
	.50	.99	63.2	31.2[a]	31.2[a]	68.3	68.3	68.3	68.3	6.3[a]	6.3[a]
(40, 10)	.02	.29	68.7	28.8[a]	31.5[a]	68.9	68.9	77.6	77.6	4.0[a]	4.0[a]
	.10	.45	67.3	46.6[a]	50.0[a]	71.3	71.3	71.8	71.8	16.6[a]	16.6[a]
	.25	.65	60.0	59.8[a]	59.8[a]	68.7	68.7	65.7	66.9	29.1[a]	29.1[a]
	.50	.86	50.2	51.7	51.7	54.0	54.0	56.1	56.4	42.9[a]	42.9[a]
(10, 40)	.02	.29	30.3	12.9[a]	12.9[a]	30.5	30.5	30.5	30.5	70.4	70.4
	.10	.45	51.2	48.6[a]	48.6[a]	56.1	56.1	56.8	56.8	47.1[a]	47.1[a]
	.25	.65	55.6	60.9	60.9	60.9	60.9	61.9	65.2	36.8[a]	36.8[a]
	.50	.86	64.1	49.5[a]	50.1[a]	68.7	68.7	68.5	68.5	18.5[a]	18.5[a]
(80, 20)	.02	.18	64.6	12.9[a]	38.5[a]	70.6	70.6	76.2	76.2	2.4[a]	2.4[a]
	.10	.33	65.0	39.9[a]	51.1[a]	68.1	68.1	68.4	68.4	10.6[a]	10.6[a]
	.25	.52	57.4	54.6[a]	55.7[a]	64.8	63.3	63.9	63.9	18.4[a]	41.8[a]
	.50	.77	59.5	56.2[a]	56.2[a]	59.7	59.7	57.7[a]	59.9	32.1[a]	60.7
(20, 80)	.02	.18	32.4	2.9[a]	5.0[a]	40.6	39.8	40.7	40.7	62.7	62.7
	.10	.33	49.4	39.5[a]	42.1[a]	56.5	55.7	53.2	56.9	42.6[a]	57.1
	.25	.52	58.6	56.0[a]	56.0[a]	58.6[a]	58.6[a]	56.9	59.0	30.1[a]	59.0
	.50	.77	59.3	54.5[a]	56.3[a]	65.9	64.8	65.7	65.7	18.2[a]	37.8[a]
(240, 60)	.02	.09	64.1	3.4[a]	38.2[a]	67.3	68.5	68.9	69.2	2.8[a]	2.8[a]
	.10	.22	62.7	33.2[a]	53.4[a]	65.1	67.4	67.8	67.8	13.4[a]	41.4[a]
	.25	.40	59.8	53.4[a]	56.2[a]	61.1	61.2	61.1	62.3	18.1[a]	54.5[a]
	.50	.66	59.2	59.6	59.6	59.2[a]	59.2[a]	59.7	60.0	25.3[a]	61.5
(60, 240)	.02	.09	41.1	0.1[a]	8.5[a]	41.1[a]	48.5	43.2	44.6	45.0	46.8
	.10	.22	54.1	31.5[a]	43.5[a]	57.5	57.8	55.5	58.3	37.0[a]	66.2
	.25	.40	56.0	54.5[a]	54.6[a]	58.6	58.6	57.5	58.5	27.1[a]	61.8
	.50	.66	58.7	59.0	59.0	62.7	62.0	61.3	62.2	21.0[a]	59.0

[a]power is less than or equal to the corresponding power of F; $\alpha = 5\%$, 2-sided.

problem, albeit with a different focus than ours. They conclude that Barnard's test (1945) is the most powerful test for exact unconditional inference. However, it is also clear from their report that because Barnard's test requires evaluation of the true size of the test multiple times (once for each candidate outcome to be introduced into the rejection region), it is substantially more computationally intensive compared with the best exact unconditional tests in our

investigation (which require only a single evaluation of the true size). Accordingly, based on relative performance and computational feasibility, we recommend tests B, B^*, Z_P, and Z_P^* if exact unconditional inference is desired. Both one-sided and two-sided p-values for F, Z_U, Z_U^*, and our recommend tests can be calculated using the online interactive software at `http://www.stat.ncsu.edu/~berger/tables.html`. In addition, methods F, D, D^*, Z_P, and Z_P^* are implemented in StatXact-5. An advantage of using the latter software is that, in addition to p-values, it also provides improved exact unconditional test-based confidence intervals for $(\theta_1 - \theta_2)$ based on recent articles by Chan and Zhang (1999), and Agresti and Min (2001).

Résumé

Utilisé pour la comparaison de proportions dans le cadre d'essais randomisés, le test exact de Fisher peut se révéler particulièrement conservateur, principalement lorsque les échantillons sont de petite taille ou lorsque les proportions observées sont proches de 0 ou de 1. Cette propriété indésirable s'explique surtout par une distribution de la statistique de test (sous hypothèse nulle) trop discrétisée, notamment en raison d'une inférence effectuée *conditionnellement* au nombre total des répondeurs. De ce fait, les procédures exactes *non conditionnelles* ont conquis une popularité fondée sur l'idée d'une augmentation de la puissance entraînée par des distributions de statistiques de tests (sous hypothèse nulle) nécessairement moins discrétisées. Toutefois, nous mettons en garde les chercheurs contre le choix hasardeux d'un test exact non conditionnel car, dans certains cas, leur utilisation peut s'avérer franchement moins puissante que le test de Fisher. A l'appui de nos propos, nous traitons d'un exemple réel, puis nous calculons les risques de type I et les puissances associées à différents tests, lors de comparaisons portant sur des groupes de tailles égales ou inégales. Nos résultats montrent que le test exact de Fisher fait généralement mieux que les tests exacts non conditionnels basés sur la différence des proportions, mieux aussi, dans le cas d'échantillons de taille inégales, que les tests exacts non conditionnels basés sur la différence des proportions divisée par son écart–type (estimé sous hypothèse alternative). A l'inverse, le test exact non conditionnel basé sur la différence des proportions divisée par son écart–type (estimé sous hypothèse nulle) – il s'agit de la statistique du score – fait mieux que le test de Fisher. Ce test est recommandé, tout comme le test de Boschloo, test dans lequel le test de Fisher est utilisé comme une statistique de test dans un test exact non conditionnel; ce dernier test s'avère, quant à lui, uniformément plus puissant que le test de Fisher.

References

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science* **7,** 131–177.

Agresti, A. and Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57,** 963–971.

Andres, A. M. and Mato, A. S. (1994). Choosing the optimal unconditioned test for comparing two independent proportions. *Computational Statistics and Data Analysis* **17,** 555–574.

Barnard, G. A. (1945). A new test for 2×2 tables. *Nature* **156, 177,** 783–784.

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika* **34,** 123–138.

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society, Series B* **11,** 115–139.

Barnard, G. A. (1989). On alleged gains in power from lower p-values. *Statistics in Medicine* **8,** 1469–1477.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association* **72,** 355–366.

Berger, R. L. (1996). More powerful tests from confidence interval p-values. *The American Statistician* **50,** 314–318.

Berger, R. L. and Boos, D. D. (1994). P-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89,** 1012–1016.

Berkson, J. (1978). In dispraise of the exact test. *Journal of Statistical Planning and Inference* **2,** 27–42.

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics* **28,** 783–798.

Boschloo, R. D. (1970). Raised conditional level of significance for the 2×2 table when testing the equality of probabilities. *Statistica Neerlandica* **24**, 1–35.

Chan, I. S. F. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **55,** 1202–1209.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26,** 404–413.

D'Agostino, R. B., Chase, W., and Belanger, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial proportions. *American Statistician* **42,** 198–202.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, Series A* **98,** 39–54.

Haber, M. (1986). An exact unconditional test for the 2×2 comparative trial. *Psychological Bulletin* **99,** 129–132.

Haviland, M. G. (1990). Yates' correction for continuity and the analysis of 2×2 contingency tables (with comments). *Statistics in Medicine* **9,** 363–383.

Kempthorne, O. (1979). In dispraise of the exact test: Reactions. *Journal of Statistical Planning and Inference* **3,** 199–213.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *American Statistician* **43,** 283–288.

McDonald, L. L., Davis, B. M., and Milliken, G. A. (1977). A nonrandomized unconditional test for comparing two proportions in 2×2 contingency tables. *Technometrics* **19,** 145–157.

Mehta, C. R. and Hilton, J. F. (1993). Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table. *American Statistician* **47,** 91–98.

Rice, W. R. (1988). A new probability model for determining exact p-values for 2×2 contingency tables when comparing binomial proportions. *Biometrics* **44,** 1–22.

Santner, T. J. and Snell, M. K. (1980). Small-sample confidence intervals for p1-p2 and p1/p2 in 2×2 contingency tables. *Journal of the American Statistical Association* **75,** 386–394.

Storer, B. E. and Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial

proportions. *Journal of the American Statistical Association* **85,** 146–155.

Suissa, S. and Shuster, J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *Journal of the Royal Statistical Society, Series A* **148,** 317–327.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2×2 comparative trials. *Journal of the Royal Statistical Society, Series A* **145,** 86–105.

Upton, G. J. G. (1992). Fisher's exact test. *Journal of the Royal Statistical Society, Series A* **155,** 395–402.

Yates, F. (1984). Test of significance for 2×2 contingency tables (with discussion). *Journal of the Royal Statistical Society, Series A* **147,** 426–463.

Received September 2002. *Revised* n/a.
Accepted September 2002.