

Exact tests based on pre-estimation and second order pivots: non-inferiority trials

Chris J. Lloyd*

Melbourne Business School, Carlton, 3053, Australia

(Received 11 July 2008; final version received 9 February 2009)

Pre-estimation is a technique for adjusting a standard approximate P -value to be close to exact. While conceptually simple, it can become computationally intensive. Second order pivots [N. Reid, *Asymptotics and the theory of inference*, Ann. Statist. 31 (2003), pp. 1695–1731] are constructed to be closer to exact than standard approximate pivots. The theory behind these pivots is complex, and their properties are unclear for discrete models. However, since they are typically given in closed form they are easy to compute. For the special case of non-inferiority trials, we investigate Wald, Score, likelihood ratio and second order pivots. Each of the basic pivots are used to generate an exact test by maximising with respect to the nuisance parameter. We also study the effect of pre-estimating the nuisance parameter, as described in Lloyd [C.J. Lloyd, *Exact P -values for discrete models obtained by estimation and maximisation*, Aust. N. Z. J. Statist. 50 (2008), pp. 329–346]. It appears that second order methods are not as close to exact as might have been hoped. On the other hand, P -values, based on pre-estimation are very close to exact, are more powerful than competitors and are hardly affected by the basic generating statistic chosen.

Keywords: nuisance parameters; higher-order asymptotics; parametric bootstrap

1. Introduction

In a clinical trial comparing a new treatment to a standard treatment, one may want to demonstrate that the new treatment is not practically inferior to the standard treatment (also called an active control). If p_1 is the probability of a positive response with the new treatment and p_0 the probability of a positive response with the active control, then we want to demonstrate that p_1 is not substantially smaller than p_0 . We consider so-called non-inferiority tests of the hypotheses

$$\mathcal{H}_0 : p_1/p_0 \leq 1 - \delta \quad \text{versus} \quad \mathcal{H}_1 : p_1/p_0 > 1 - \delta, \quad (1)$$

for some pre-chosen non-inferiority margin δ , often 0.1 or 0.2. Rejecting the null hypothesis means that the new treatment is not practically inferior to the old treatment, and may actually be better. Such tests were first considered by Chan [1]. Dunnett and Gent [2] had much earlier measured the difference in the two treatments by the difference $p_1 - p_0$ rather than p_1/p_0 , and coined the term ‘test of equivalence’. We will use the term ‘test of non-inferiority’ for the one-sided test.

*Email: c.lloyd@mbs.edu

As an introductory example, we consider the Burlington study reported in Spitzer *et al.* [3] and used by Dunnett and Gent [2]. In this case, the binary outcome was ‘adequate care’ of patients. In particular, conventional care involving an appointment with a doctor (group 0) was compared with a regime administered by nurses only (group 1). The success rates were $\hat{p}_0 = 148/225 = 65.8\%$ for conventional care and $\hat{p}_1 = 115/167 = 68.9\%$ for nurse based care. While the results for nurse based care was actually higher than for conventional care, the main interest was in establishing that nurse based care was not practically inferior to the conventional care. For this, Dunnett and Gent [2] took a margin of 0.1, though they used $p_1 - p_0$ rather than p_1/p_0 to measure the relative performance of the two types of care. Chan’s statistic, defined later, equals 2.077, which gives an approximate P -value of 0.0189. However, computation of the exact significance based on this statistic gives a considerably larger value of 0.0250. This is by no means an anomalous example, and it is easy to find data sets where the exact and approximate P -value differ by much more than this.

The issue for constructing exact and efficient tests in discrete settings is to find a test statistic that is close to pivotal, by which we mean that tail probabilities depend very little on parameters left unspecified under the null. Lloyd [4] has demonstrated numerically that this can often be achieved by simply estimating the nuisance parameter within an exact calculation. The remaining dependence on the nuisance parameter is then accounted for by maximisation.

A quite separate area of research has seen development of so-called second-order methods. This involves a local decomposition of the model into a sufficient and ancillary component. An approximate formula for the conditional distribution is available [5], and the cumulative tail probability of this distribution is further approximated [6,7]. All the approximations are very accurate, sometimes spectacularly so for continuous models. The theory is considerably simpler for models that are embedded in an exponential family (see [8]). The P -values from this theory should not only be more pivotal but, since they are conditional, may lead to more powerful unconditional tests as has been found in other contexts (see for instance [9,10]).

The plan of the article is as follows. In Section 2, we give six approximate P -values including one based on second-order theory. Section 3 explains how to generate exact tests using estimation and maximisation. In Section 4, we investigate exactness of these various approximate P -values. In Section 5, we compare the efficiency of 12 exact P -values by computing their average powers.

2. Approximate test statistics

The data comprise y_1 responses from n_1 independent individuals with treatment applied and y_0 responses from n_0 independent individuals with control. We are interested in the probabilities of response p_1, p_0 under the new treatment and control, respectively, which leads to the parameter of interest $\psi = \log(p_1/p_0)$. Then the non-inferiority hypotheses (1) are

$$\mathcal{H}_0 : \psi \leq \psi_0 \quad \text{versus} \quad \mathcal{H}_1 : \psi > \psi_0,$$

where $\psi_0 = \log(1 - \delta) \approx -\delta$. Without loss of generality, we take the nuisance parameter to be $\lambda = p_0$. The log-likelihood $\ell(\psi, \lambda; y_0, y_1)$ is

$$y_1(\psi + \log \lambda) + (n_1 - y_1) \log(1 - \lambda e^\psi) + y_0 \log \lambda + (n_0 - y_0) \log(1 - \lambda). \quad (2)$$

The maximum likelihood (MLE) estimate $\hat{\lambda}_\psi$ of λ , for fixed ψ , is obtained by solving a quadratic [11].

There are three standard test statistics to consider. The MLE of ψ has an asymptotic variance $\sigma^2(p_0, p_1) = (1 - p_0)/(n_0 p_0) + (1 - p_1)/(n_1 p_1)$ giving rise to the standard error

$\hat{\sigma} = \sigma(\hat{p}_0, \hat{p}_1)$ as well as the restricted MLE $\hat{\sigma}_\psi = \sigma(\hat{p}_{0\psi}, \hat{p}_{1\psi})$. Then, we have two Wald-type statistics

$$W_1(\psi) = \frac{\hat{\psi} - \psi}{\hat{\sigma}}, \quad W_2(\psi) = \frac{\hat{\psi} - \psi}{\hat{\sigma}_\psi}.$$

We will use the common device of adding 1/2 to all counts, to avoid problems at the boundary of the sample space. A second standard statistic is the likelihood root

$$r(\psi) = \text{sign}(\hat{\psi} - \psi)[2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\}]$$

as proposed by Munk *et al.* [12] within a more general setting. A third approach is to express the alternative hypothesis in the form $p_1 - p_0(1 - \delta) > 0$. The MLE of this parameter has variance $\tau^2(p_0, p_1) = p_1(1 - p_1)/n_1 + (1 - \delta)^2 p_0(1 - p_0)/n_0$ giving standard errors $\hat{\tau}$ and $\hat{\tau}_\psi$, which gives rise to two more statistics

$$C_1(\psi) = \frac{\hat{p}_1 - e^\psi \hat{p}_0}{\hat{\tau}}, \quad C_2(\psi) = \frac{\hat{p}_1 - e^\psi \hat{p}_0}{\hat{\tau}_\psi}$$

the second of which was considered by Chan [1]. *P*-values based on a standard normal approximation to any of these five statistics are said to be first-order accurate, and suffer errors of $O(n^{-1/2})$ for one-sided *P*-values. Second-order methods described below can give one-sided *P*-values with error $O(n^{-1})$.

Second order inference is based on the modified likelihood root

$$r^*(\psi) = r(\psi) + r(\psi)^{-1} \log \left\{ \frac{q(\psi)}{r(\psi)} \right\} \tag{3}$$

and its approximate standard normal distribution. The joint binomial model is of exponential family form with canonical parameters (φ_1, φ_0) equal to the logits of the probability parameters (p_1, p_0) . The statistic $q(\psi)$ is in general complicated, but for exponential families depends on two quantities, namely the observed information matrix in the (ψ, λ) -parameterization and the Jacobian matrix of the transformation from (ψ, λ) to (φ_1, φ_0) .

It turns out to be easier to derive an expression for $q(\psi)$ for testing a generalization of the current problem, namely, the alternative hypothesis $\psi = h(p_1) - h(p_0) > \psi_0$ for an arbitrary increasing function h . Then,

$$q(\psi) = \frac{\{\hat{w}_{1\psi}(\hat{\varphi}_1 - \hat{\varphi}_{1\psi}) - \hat{w}_{0\psi}(\hat{\varphi}_0 - \hat{\varphi}_{0\psi})\} \sqrt{\hat{V}_1 \hat{V}_0}}{\sqrt{\hat{V}_{0\psi} \hat{w}_{1\psi} + \hat{V}_{1\psi} \hat{w}_{0\psi}^2}}, \tag{4}$$

where $V_j = n_j p_j(1 - p_j)$ and $w_j = p_j(1 - p_j)h'(p_j)$. Details of this derivation are available from the author. For our present case where $h(p) = \log p$ we have $w_j = 1 - p_j$.

The formula is parameterization invariant though Davison *et al.* [13] seem to only claim invariance to linear transformations. Note that $r^*(\psi)$ breaks down in the centre of the distribution where $r(\psi) \approx 0$. This is easily corrected by setting $r^*(\psi) = r(\psi)$ when $|r(\psi)| < 0.001$. There is a further problem that $q(\psi) = 0$ whenever $\hat{V}_0 \hat{V}_1 = 0$, which occurs when (y_0, y_1) is on the boundary of the sample space. These anomalies are mentioned but not addressed in Davison *et al.* [13], and the *R*-functions in bundle *hoa* do not return an answer for boundary data sets. When $y_1 = n_1$ or $y_0 = 0$, we use $r^*(\psi)$ with a continuity correction and then divide the resulting *P*-value by 2, as recommended in Peters and Pierce [14].

3. Exact P -values

A much cleaner theory of ‘exact’ P -values can be given than of ‘exact’ tests, a term that can cause some confusion for discrete models. The ideas here are explained more fully in Lloyd [4,15]. For a given approximate P -value $P(Y)$, the exact significance

$$\Pi_P(y, \lambda) := \Pr(P(Y) \leq P(y); \lambda, \psi_0) \tag{5}$$

considered as a function of λ , is called the profile of $P(y)$. To calculate this curve for an observed value $y = (y_0, y_1)$, one must first identify the subset of the sample space for which $P(Y) \leq P(y)$. For large sample spaces, this becomes more computationally intensive and would clearly become infeasible for sufficiently large sample sizes. Call this tailset $\mathcal{T}(y)$. To compute Equation (5), we calculate

$$\sum_{y' \in \mathcal{T}(y)} p_{BI}(n_0, y'_0, \lambda) p_{BI}(n_1, y'_1, \lambda e^{\psi_0}),$$

where p_{BI} denotes binomial probability. This must be done at a grid of nuisance parameter values λ .

3.1. Maximization

Define $P^*(y) := \sup_{\lambda} \Pi_P(y, \lambda)$. The maximized P -value $P^*(Y)$ only depends on the way the generating P -value $P(Y)$ orders the sample space. Subject to this ordering, it possesses many optimality properties. It is exact, in the sense described by Lloyd [4] and minimal amongst P -values that order the sample space in the same order as $P(Y)$. Moreover, any exact P -value must be expressible as the supremum of the profile of some $P(Y)$ [4,16]. In principle then, the only way to obtain an exact test is to begin with a generator P -value $P(Y)$ and apply the maximization or M -step. We can measure how close a given P -value is to exact, independent of nominal test size, by measuring the closeness of $P(Y)$ to $P^*(Y)$. We will utilize this idea later.

We illustrate with an example due to Berger and Boos [17] where there are $(y_0, y_1) = (48, 14)$ successes from $(n_0, n_1) = (283, 47)$ trials. We have $(\hat{p}_0, \hat{p}_1) = (0.170, 0.298)$ and $\hat{\psi} = 0.563$ with standard error $\hat{\sigma} = 0.260$. Consider testing $\delta = 0.1$ that corresponds to $\psi_0 = \log(0.9) = -0.1053$. The restricted MLR of (p_0, p_1) is $(\hat{p}_{0\psi}, \hat{p}_{1\psi}) = (0.190, 0.171)$ giving $\hat{\sigma}_{\psi} = 0.344$. The observed values of our various test statistics are $W_1 = 2.623$, $W_2 = 1.978$, $C_1 = 2.084$, $C_2 = 2.469$, $r = 2.316$ and $r^* = 2.331$. The differences in these statistics suggest that the first-order results are inaccurate though the closeness of r and r^* suggests the opposite, if one believes the folklore that has arisen in the higher-order asymptotics literature.

For discrete data, it is not sufficient that r and r^* be close to have confidence in r . Figure 1 plots profiles for four generating P -values. For W_2 , C_2 and r , the profiles contain a strong local maximum compared with which the quoted P -value, indicated by the solid horizontal line, strongly exaggerates the evidence against the null. The spikes are due to the differing accuracy of normal approximation across the parameter space. The profile for r^* is much better behaved.

3.2. Pre-estimation

Since the first-order statistics appear to perform badly in the previous example, we seek a simple modification prior to maximization. A very simple and successful modification is to replace $P(Y)$

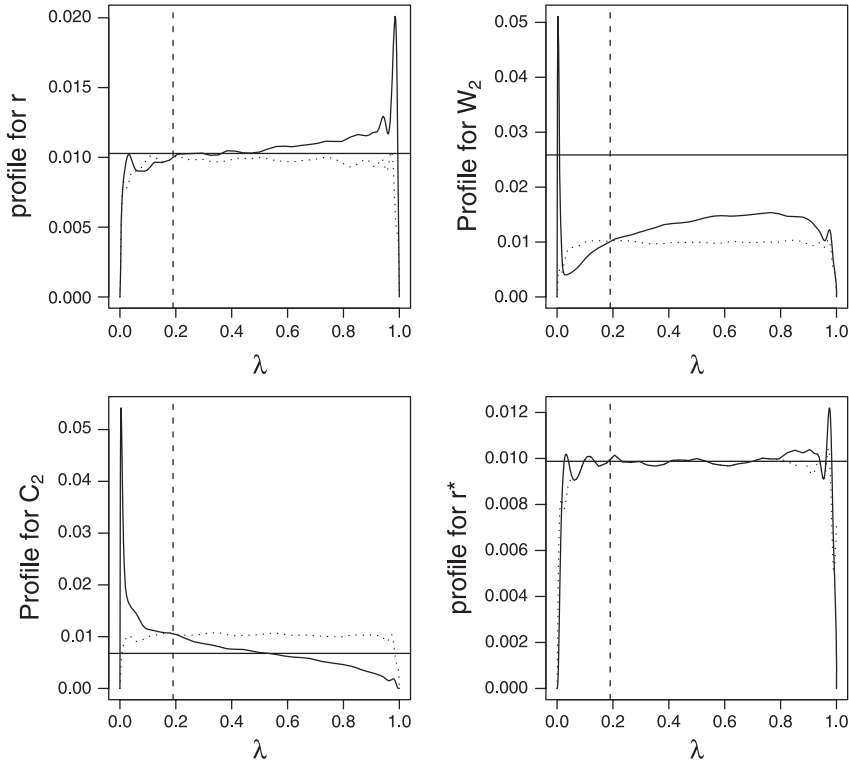


Figure 1. Profiles for four generating P -values. Data is $(y_0, y_1) = (48, 14)$, $(n_0, n_1) = (283, 47)$ and $\delta = 0.1$. Each panel shows the profile $\Pi_P(y, \lambda)$ versus λ . The horizontal line is $P(y)$ and the vertical lines is $\hat{\lambda}_\psi = 0.190$. The dotted curve is the profile function $\Pi_{\hat{P}}(y, \lambda)$ of the estimated P -value $\hat{P}(y)$.

with the pre-estimated P -value

$$\hat{P}(y) := \Pi_P(y, \hat{\lambda}_\psi). \tag{6}$$

This is nothing more than a parametric bootstrap. Estimated P -values have been around for a long time (see for instance Beran [18] for continuous models). More recently, Lee and Young [19] have shown that estimated P -values have very attractive asymptotic properties, again for continuous models. Computing this E -step is much easier than computing the full profile, because it is just a single point on the curve $\Pi_P(y, \lambda)$.

The pre-estimation step has two distinct effects. First, it seems to produce a P -value that is much closer to exact, but typically slightly liberal. It can be seen then as a computationally cheap but imperfect version of maximization. Second, it has been demonstrated in Lloyd [4] that the profile $\Pi_{\hat{P}}(y, \lambda)$ of $\hat{P}(Y)$ tends to be much flatter than that of the original generator $P(Y)$. This is presumably due to it being based on an estimate of an exact significance value rather than a normal approximation whose accuracy is inconsistent across the sample space. A detailed heuristic argument as to why pre-estimation works so well is in Lloyd [20], where it is also explained why the restricted MLE of λ is required rather than some other consistent estimator.

The dotted lines in Figure 1 show the profiles $\Pi_{\hat{P}}(y, \lambda)$ for the pre-estimated P -values generated from the four basic P -values W_2, C_2, r and r^* . Each profile is quite flat and is very close to the

observed P -value $\hat{P}(y)$, which can be read off from the original profile where it intersects the vertical line at $\lambda_\psi = 0.190$.

3.3. Estimation followed by maximization

While the estimated P -value is only approximate, it will be seen from our later numerical study that it is very close to exact. It can be made exact by the same process of maximization that is used to make the original approximate P -value exact. This involves maximizing the profile of the statistic $\hat{P}(Y)$ rather than $P(Y)$. Formally, $\hat{P}^*(y) := \sup_\lambda \Pi_{\hat{p}}(y, \lambda)$ which I call the $E + M$ P -value. Note that these will be different to the maximised P -values because the E -step induces a different ordering on the sample space that of to the original P -value.

Figure 2 illustrates the effect of the E -step for the case $(n_0, n_1) = (10, 15)$ and $\delta = 0.1$. Instead of plotting P -values, I have transformed them by a probit transform producing quasi T -values. The left panel plots the likelihood ratio T -values against the probit transformed estimated P -values. The estimated T -values tend to be smaller i.e. less significant than the first-order P -values, potentially correcting for the well-known liberalness of LR tests. Just as importantly, however, the ordering of the sample space is substantially changed. This re-ordering is just as important as the general reduction in the T -values. The right panel addresses whether the E -step improves the exactness of the P -values. Apparently, there is hardly any practical difference between the E P -values and the $E + M$ P -values. In other words, the estimated P -values are close to exact.

Munk *et al.* [12] compared the exact (i.e. maximized) test based on the LR statistic with several more exotic competitors based on explicit ordering criteria on the sample space rather than on a test statistic. They found that the LR statistic performed slightly better than other methods. Kang and Chen [21], following Chan [1], investigated the exact test based on C_2 as well as the estimated P -value based on this same statistic. They found that the estimated P -value had good power properties while the size was close to, but sometimes exceeds, nominal. They did not consider $E + M$ P -values.

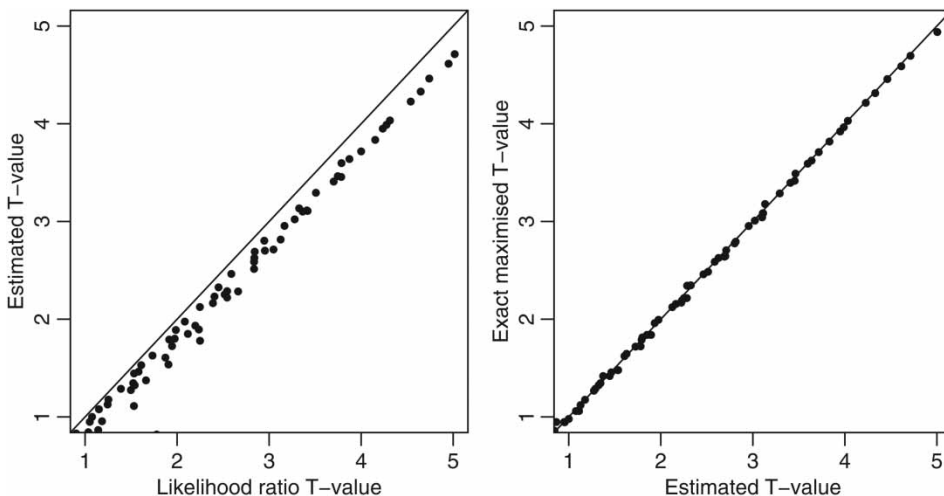


Figure 2. Effect of the E -step. The example is for $(n_0, n_1) = (10, 15)$. *Left.* Estimated T -values versus raw T -values. Estimated T -values tend to be smaller than raw T -values. *Right.* Exact maximised T -values versus estimated T -values. The estimated T -values are practically exact.

Table 1. Observed Z -values for data of Berger and Boos [17]. Each P -value is for testing alternative $\psi > -0.105$ and has been converted to the standard normal scale, to aid interpretation.

Generator	Base	M	E	$E + M$
W_2	1.945	1.632	2.321	2.317
C_2	2.469	1.598	2.305	2.297
r	2.316	2.051	2.324	2.310
r^*	2.331	2.250	2.325	2.310

A summary of the results for the Berger and Boos [17] example is in Table 1. There is considerable disagreement between the three first-order statistics. Maximization reduces the significance of the first-order statistics but has little effect on r^* . All four pre-estimated P -values agree quite closely and maximization has very little effect on these, but reduces the significance of each slightly.

Computation of E , M and $E + M$ P -values depends on the cardinality of the sample space, here $N = (n_0 + 1)(n_1 + 1)$. Computing the significance profile $\Pi_P(y, \lambda)$ at a single value of λ , which is all that is required for the E P -values, involves $O(N)$ computations. For the M P -value, the profile must be evaluated at a fine grid of λ -values. For the $E + M$ P -value, computation is $O(N^2)$ which quickly become infeasible. However, this can be reduced to $O(N \log N)$ by making use of natural monotonicity properties, for instance that any P -value should be increasing in y_1 and decreasing in y_0 . For very large sample sizes, which are not common for non-inferiority trials, some kind of simulation algorithm will be required, and is under development. Further comments about computation are in the final section.

4. Comparison of exactness of approximate P -values

The intention here is to compare six basic approximate statistics, five based on the first-order theory and one based on the second-order theory. Each can have the E -step applied giving a further six approximate statistics. The last example suggested that pre-estimation gives a P -value that is almost exact. If true, this would allow the computationally problematic M -step, which becomes more burdensome with the dimension of the nuisance parameter, to be dropped in practice.

The probability of rejection for the nominal size α test induced by $P(Y)$ is

$$\beta(\psi, \lambda, \alpha) = \Pr(P(Y) \leq \alpha; \psi, \lambda). \tag{7}$$

When $\psi = \psi_0$ this gives the size of the test. For moderate sample sizes, this can be calculated exactly, without the need for simulation. Table 2 gives results for tests of $\delta = 0.1$, with nominal size $\alpha = 0.05$. The upper figure is the supremum of $\beta(\psi_0, \lambda, \alpha)$, as recommended by Bickel and Doksum [22], while the lower figure is $\int \beta(\psi_0, \lambda, \alpha)d\lambda$. The upper section is for the tests induced by the six basic statistics.

The numbers support the notion that tests based on r^* have size closer to nominal than the first-order tests, though not in all cases. When measured by mean rather than supremum size, the Chan based tests also perform well. The flatness of the size function $\beta(\psi_0, \lambda, \alpha)$ with respect to λ can be judged by how close are the supremum and mean measures. The numbers indicate that the size of r^* is less sensitive to λ than its first-order competitors. The lower section of the table is for tests induced from the pre-estimated P -values. All of these tests have very acceptable size performance.

An alternative method of assessing exactness of a test is to see how close $P(Y)$ is to its exact version $P^*(Y)$. An advantage of this approach is that it does not depend on a nominal size α .

Table 2. Size of approximate tests. Summaries of $\beta(\psi_0, \lambda, \alpha)$ for $\psi_0 = -0.105$ and $\alpha = 0.05$. The upper value is supremum with respect to λ . The lower figure is mean with respect to Lebesgue measure on λ . The bottom half of table is for test based on pre-estimated P -values.

(n_0, n_1)	W_1	W_2	C_1	C_2	r	r^*
50 20	0.0995	0.0343	0.1123	0.1139	0.0995	0.0795
	0.0584	0.0272	0.0559	0.0536	0.0526	0.0505
80 25	0.0731	0.0323	0.1595	0.1006	0.0993	0.0561
	0.0644	0.0266	0.0575	0.0531	0.0537	0.0497
60 40	0.0740	0.0553	0.1115	0.0553	0.0609	0.0559
	0.0526	0.0414	0.0543	0.0497	0.0528	0.0499
75 50	0.0838	0.0547	0.0989	0.0551	0.0573	0.0536
	0.0525	0.0420	0.0527	0.0511	0.0517	0.0487
60 60	0.0646	0.0528	0.1093	0.0528	0.0893	0.0646
	0.0456	0.0454	0.0530	0.0484	0.0534	0.0498
200 50	0.0816	0.0361	0.1060	0.0922	0.0738	0.0563
	0.0651	0.0321	0.0526	0.0517	0.0517	0.0503
50 20	0.0524	0.0524	0.0522	0.0524	0.0524	0.0502
	0.0463	0.0456	0.0443	0.0456	0.0463	0.0452
80 25	0.0490	0.0502	0.0483	0.0502	0.0490	0.0502
	0.0449	0.0453	0.0444	0.0453	0.0449	0.0449
60 40	0.0502	0.0499	0.0502	0.0499	0.0553	0.0499
	0.0453	0.0453	0.0453	0.0453	0.0468	0.0453
75 50	0.0515	0.0508	0.0525	0.0496	0.0547	0.0496
	0.0471	0.0468	0.0471	0.0468	0.0483	0.0468
60 60	0.0499	0.0499	0.0499	0.0499	0.0499	0.0499
	0.0463	0.0463	0.0463	0.0463	0.0463	0.0463
200 50	0.0501	0.0499	0.0511	0.0501	0.0499	0.0499
	0.0482	0.0482	0.0485	0.0482	0.0482	0.0482

Let us define the inexactness of a test by

$$\epsilon(P; \lambda) = E^{1/2}\{(P^*(Y) - P(Y))^2 | R, \psi_0, \lambda\},$$

where R is a subset of the sample space considered relevant. For instance, sample points where both P -values are between 0.9 and 1.0 are of no interest. For our application, we take $R = \{\hat{\psi} \geq \psi_0\}$, which corresponds to there being some evidence against the null and any reasonable P -values being 0.5 or less. It should be noted that, for continuous models, $\epsilon(P, \lambda)$ can be interpreted as size error, averaged over alternative nominal α .

Table 3. Inexactness of approximate P -values. Values of $\int_0^1 \epsilon(P, \lambda) d\lambda$ for generating statistics given in column headings. Upper section is for basic generating statistic, lower section of table is for pre-estimated P -values.

(n_0, n_1)	W_1	W_2	C_1	C_2	r	r^*
(50, 20)	0.156	0.115	0.109	0.064	0.078	0.062
(80, 25)	0.109	0.146	0.104	0.065	0.076	0.062
(60, 40)	0.153	0.135	0.092	0.066	0.075	0.045
(75, 50)	0.152	0.134	0.082	0.063	0.068	0.041
(60, 60)	0.071	0.057	0.069	0.056	0.067	0.044
(200, 50)	0.155	0.173	0.066	0.042	0.045	0.038
(50, 20)	0.002	0.002	0.003	0.002	0.002	0.003
(80, 25)	0.002	0.002	0.003	0.002	0.002	0.004
(60, 40)	0.005	0.002	0.002	0.002	0.002	0.004
(75, 50)	0.005	0.002	0.002	0.002	0.002	0.004
(60, 60)	0.030	0.011	0.011	0.011	0.011	0.011
(200, 50)	0.001	0.001	0.002	0.001	0.001	0.004

Table 3 shows values of $\int_0^1 \epsilon(P, \lambda) d\lambda$ for the same statistics and samples sizes as in Table 2. The upper section is for the six basic P -values. For instance, for $(n_0, n_1) = (20, 50)$ P -values based on C_2 needs to be adjusted by a mean value of 0.064 on average. Approximate P -values based on r^* require the least adjustment to exactness, though the improvement is not spectacular. The lower section gives inexactness values for the pre-estimated P -values. Again, all pre-estimated P -values are close to exact. There also appears to be little difference between the six pre-estimated P -values.

The conclusion from this section is that, among the six basic statistics, r^* is closer to exact than its first-order competitors. However, pre-estimation is much more successful at producing an approximately exact test than using r^* . Indeed, the effect of the M step after the E step is, in mean at least, negligible.

5. Comparison of efficiency of exact P -values

The example in Section 3 suggested two other conjectures that are worth resolving. First, $E + M$ P -values tended to be smaller (the Z -value was larger) than the M P -value. If true generally, pre-estimated P -values will lead to a more powerful exact test than those based on the original asymptotic P -value. This was not the case, however, for P -values based on r^* , where the M and $E + M$ P -values were similar. If true generally, this would allow the E -step to be dropped for the particular generator r^* .

Power is measured by replacing ψ_0 in the formula (7) by an arbitrary alternative value $\psi_A > \psi_0$. Power tends to increase with ψ_A and approaches zero as $\lambda \rightarrow 0$, since the null is never rejected from the data $(y_0, y_1) = (0, 0)$. To summarize results, we use $\int \beta(\psi_A, \lambda, \alpha) d\lambda$, being the power averaged over λ . We select ψ_A so that the power is in the statistically interesting range, i.e. neither close to nominal size nor close to one. Table 4 lists the results for 12 exact P -values. The results in the upper section indicate that if pre-estimation is not to be employed then P -values based on r^* are slightly preferred.

The lower section is for $E + M$ P -values showing that power is distinctly and almost uniformly higher with pre-estimation than without pre-estimation. The results also indicate that pre-estimation leads to exact P -values that have almost identical performance, regardless of the basic generating P -value. Moreover, the values of five of the six $E + M$ P -values are highly correlated. A measure of the extent to which they give similar inferential results is the root mean

Table 4. Mean powers of alternative exact P -values. Values of $\int_0^1 \beta(\psi_A, \lambda, 0.05) d\lambda$ for 12 difference exact tests. The upper section is for M P -values, the lower section is for $E + M$ P -values.

(n_0, n_1)	$\exp(\psi_A)$	W_1	W_2	C_1	C_2	r	r^*
(50, 20)	1.5	0.303	0.499	0.000	0.492	0.372	0.497
(80, 25)	1.44	0.268	0.546	0.000	0.505	0.393	0.487
(60, 40)	1.4	0.414	0.530	0.284	0.539	0.493	0.535
(75, 50)	1.35	0.494	0.548	0.422	0.557	0.515	0.567
(60, 60)	1.35	0.382	0.591	0.333	0.588	0.524	0.591
(200, 50)	1.2	0.520	0.576	0.389	0.472	0.532	0.567
(50, 20)	1.5	0.497	0.497	0.497	0.497	0.497	0.510
(80, 25)	1.45	0.546	0.546	0.546	0.546	0.546	0.546
(60, 40)	1.4	0.572	0.575	0.572	0.575	0.571	0.575
(75, 50)	1.35	0.559	0.554	0.555	0.561	0.559	0.561
(60, 60)	1.35	0.595	0.595	0.595	0.595	0.595	0.595
(200, 50)	1.2	0.578	0.578	0.576	0.578	0.578	0.578

Table 5. Closeness of 6 $E + M$ P -values. Sample size $(n_0, n_1) = (75, 50)$. Root mean square differences between all pairs \hat{P}_1^* and \hat{P}_2^* , averaged over λ .

	W_1	W_2	C_1	C_2	r	r^*
W_1	0.000	0.035	0.034	0.035	0.035	0.035
W_2	0.035	0.000	0.004	0.003	0.003	0.008
C_1	0.034	0.004	0.000	0.002	0.002	0.008
C_2	0.035	0.003	0.004	0.000	0.001	0.008
r	0.035	0.003	0.002	0.001	0.000	0.008
r^*	0.035	0.008	0.008	0.008	0.008	0.000

square difference between them, averaged over the nuisance parameter λ . These numbers are in Table 5, for the special case $(n_0, n_1) = (75, 50)$. It would seem that the E -step results in P -values that order the sample space in the same way, almost regardless of the original generating P -value, apart from W_1 . The same conclusion is true for the other sample sizes (results not presented).

6. Conclusion

For discrete models, especially, testing theory faces some practical and theoretic difficulties. While there are several standard test statistics available that all lead to P -values that are first-order accurate, these can give quite different answers for a given problem. Moreover, the exact frequentist properties of standard approximate P -values are disappointing, notwithstanding their asymptotic properties.

We have investigated two methods of producing better exact tests. One is an analytic method based on second-order theory that leads to a closed form P -value known as r^* . The second is a more computationally demanding but conceptually simple method of pre-estimation, which must be applied to a pre-chosen approximate P -value.

P -values based on r^* are somewhat closer to pivotal than the standard first-order P -values. Consequently, the exact tests based on them tend to be slightly more powerful. But the dominance over first-order methods is not great. The source of the disappointing performance is that the r^* formula breaks down on the boundary. At the same time, the performance of exact tests is quite sensitive to boundary anomalies.

Pre-estimation, which is a form of parametric bootstrap, is more successful at producing a P -value that is closer to exact, and the exact tests based on pre-estimated P -values have distinctly higher power as compared with exact tests based on the raw P -values. Just as importantly, the estimated P -values are virtually identical regardless of the initial approximate pivotal that is chosen (Table 5). For practical purposes, the estimated P -value computed from any of the six basic pivots can be treated as if it were exact (Table 3). The simplest of these is C_1 . In principle, one should quote the $E + M$ P -value which is computationally feasible for sample sizes up to several hundred.

The estimated P -value requires us to compute the probability of $\{P(Y) \leq P(y_{\text{obs}})\}$ for a single parameter value. The exact computation is of order $N = O(n^2)$, where N is the cardinality of the sample space. This can be reduced to $O(n \log n)$ for testing non-inferiority in 2×2 tables because the approximate test statistics are monotonic in both arguments. To compute the $E + M$ P -value, we must compute all possible estimated P -values that is a $O(n^2 \log n)$ calculation. For larger sample sizes and more general problems, it is not difficult to develop sampling algorithms to approximate the probability to any desired accuracy. An algorithm based on importance sampling is currently under development.

References

- [1] I.S.F. Chan, *Exact tests FPF equivalence and efficacy with a non-zero lower bound for comparative studies*, Stat. Med. 17 (1998), pp. 1403–1413.
- [2] C.W. Dunnett and M. Gent, *Significance testing to establish equivalence between treatments with special reference to data in form of 2×2 tables*, Biometrics 33 (1977), pp. 593–602.
- [3] W.O. Spitzer, D.L. Sackett, J.C. Sibley, R.S. Roberts, M. Gent, D.J. Kergin, B.C. Hackett, and A. Olynich, *The Burlington randomised trial of the nurse practitioner*, New. Engl. J. Med. 290 (1974), pp. 251–256.
- [4] C.J. Lloyd, *Exact P-values for discrete models obtained by estimation and maximisation*, Aust. N. Z. J. Stat. 50 (2008), pp. 329–346.
- [5] O.E. Barndorff-Nielsen and D.R. Cox, *Inference and Asymptotics*, Chapman and Hall, London, 1994.
- [6] R. Lugannani and S. Rice, *Saddlepoint approximation for the distribution of the sum of independent variables*, Adv. Appl. Probab. 12 (1980), pp. 475–490.
- [7] N. Reid, *Asymptotics and the theory of inference*, Ann. Statist. 31 (2003), pp. 1695–1731.
- [8] A. Brazzale, A.C. Davison, and N. Reid, *Applied Asymptotics: Case Studies in Small-Sample Statistics*, Cambridge University Press, New York, 2007.
- [9] R.D. Boschloo, *Raised conditional level of significance for the 2×2 table when testing the equalities of probabilities*, Statist. Neerlandica 24 (1970), pp. 1–35.
- [10] C.J. Lloyd and M. Moldovan, *Unconditional efficient upper limits for the odds ratio based on conditional likelihood*, Stat. Med. 26 (2007), pp. 5136–5146.
- [11] O. Miettinen and M. Nurminen, *Comparative analysis of two rates*, Stat. Med. 4 (1985), pp. 213–226.
- [12] A. Munk, G. Skipka, and B. Stratmann, *Testing general hypotheses under binomial sampling: the two sample case - asymptotic theory and exact procedures*, Comput. Statist. Data Anal. 49 (2005), pp. 723–739.
- [13] A.C. Davison, D.A.S. Fraser, and N. Reid, *Improved likelihood inference for discrete data*, J. R. Stat. Soc. Ser. B 68 (2006), pp. 495–508.
- [14] D.A. Pierce and D. Peters, *Improving on exact tests by approximate conditioning*, Biometrika 86 (1999), pp. 267–277.
- [15] C.J. Lloyd, *A new exact and more powerful unconditional test of no treatment effect from binary matched pairs*, Biometrics 64 (2008), pp. 716–723.
- [16] J. Rohmel and U. Mansmann, *Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority or superiority*, Biom. J. 41 (1999), pp. 149–170.
- [17] R.L. Berger and D.D. Boos, *P values maximised over a confidence set for the nuisance parameter*, J. Amer. Statist. Assoc. 89 (1994), pp. 1012–1016.
- [18] R. Beran, *Prepivoting test statistics: a bootstrap view of asymptotic refinements*, J. Amer. Statist. Assoc. 83 (1988), pp. 687–697.
- [19] S.M.S. Lee and G.A. Young, *Parametric bootstrapping with nuisance parameters*, Stat. Prob. Lett. 71 (2005), pp. 143–153.
- [20] C.J. Lloyd, *Estimated P-values in discrete models: asymptotic and non-asymptotic effects*, MBS working paper 2009-12.
- [21] S. Kang and J.J. Chen, Stat. Med. 19 (2000), pp. 2089–2100.
- [22] P.J. Bickel and K.A. Doksum, *Mathematical Statistics*, Holden-Day, Oakland CA, 1977.