

Bootstrap and Second-Order Tests of Risk Difference

Chris J. Lloyd

University of Melbourne, Carlton, 3053, Australia

**email:* c.lloyd@mbs.edu

SUMMARY. Clinical trials data often come in the form of low-dimensional tables of small counts. Standard approximate tests such as score and likelihood ratio tests are imperfect in several respects. First, they can give quite different answers from the same data. Second, the actual type-1 error can differ significantly from nominal, even for quite large sample sizes. Third, exact inferences based on these can be strongly nonmonotonic functions of the null parameter and lead to confidence sets that are discontinuous. There are two modern approaches to small sample inference. One is to use so-called higher order asymptotics (Reid, 2003, *Annal of Statistics* **31**, 1695–1731) to provide an explicit adjustment to the likelihood ratio statistic. The theory for this is complex but the statistic is quick to compute. The second approach is to perform an exact calculation of significance assuming the nuisance parameters equal their null estimate (Lee and Young, 2005, *Statistic and Probability Letters* **71**, 143–153), which is a kind of parametric bootstrap. The purpose of this article is to explain and evaluate these two methods, for testing whether a difference in probabilities $p_2 - p_1$ exceeds a prechosen noninferiority margin δ_0 . On the basis of an extensive numerical study, we recommend bootstrap P-values as superior to all other alternatives. First, they produce practically identical answers regardless of the basic test statistic chosen. Second, they have excellent size accuracy and higher power. Third, they vary much less erratically with the null parameter value δ_0 .

KEY WORDS: Equivalence test; Exact test; Noninferiority study; Nuisance parameters; r-star statistic.

1. Introduction

In a clinical trial comparing a new treatment to a standard treatment, one may want to demonstrate that there is at most a practically irrelevant difference between the failure rates of the two treatments. So-called noninferiority studies seek to place a limit on the possible inferiority of a new treatment, which may have safety and tolerability advantages compared to the standard treatment. If p_2 is the probability of a positive end-point with the new treatment and p_1 the probability with the old treatment, then the hypotheses under test are

$$\mathcal{H}_1 : p_2 - p_1 \leq \delta_0 \quad \text{versus} \quad \mathcal{H}_2 : p_2 - p_1 > \delta_0,$$

for some prechosen noninferiority margin δ_0 , often -0.1 or -0.2 . The data will typically comprise y_1 responses from n_1 independent individuals with the old treatment, and y_2 responses from n_2 independent individuals with the new treatment. The same hypotheses can be tested for matched binary data (Liu et al., 2002; Lloyd, 2008d) but this design will not be considered here.

An early approach to testing these hypotheses (Dunnett and Gent, 1977) was based on a chi-square goodness-of-fit statistic with expected values estimated under the null by a method of moments estimator. Obviously there is a huge literature on testing a difference of binomial probabilities, which will not be reviewed here. The most recent improvements and refinements are described in Röhmel and Mansmann (1999); Mehrotra, Chan, and Berger (2003); and Munk, Skipka, and Stratmann (2005).

Example. A standard illustrative example is from the seminal paper of Dunnett and Gent (1977). They report results

from the Burlington study (Spitzer et al., 1974) comparing patients given doctor-based care (regime 1) with those given care administered by nurses only (regime 2). The success rates, appropriately defined, were $\hat{p}_1 = 148/225 = 65.8\%$ for conventional care and $\hat{p}_2 = 115/167 = 68.9\%$ for nurse-based care providing weak evidence that nurse-based care was better than doctor-based care. However, the main interest was in establishing that nurse-based care was not practically inferior to the doctor-based care. For instance, can it be concluded that success rates with nursing care exceed that of doctor-based care -5% ?

Using a noninferiority margin of $\delta_0 = -0.05$ we may calculate several standard approximate test statistics. For instance, the score statistic of Chan (1998) equals 1.676 while the likelihood root statistics equals 1.680. Both these statistics are defined explicitly later. Using their asymptotic standard normal distributions results in respective P-values of 0.0453 and 0.0464. While these two P-values are quite similar in this case, alternative standard methods can give practically different answers in other examples as readers have no doubt experienced themselves. More important, however, is the fact that both these “approximate” P-values are quite inaccurate. Using the joint binomial distribution of the data rather than the asymptotic normal distributions gives rise to so-called exact P-values of 0.0500 for the score statistic and 0.0760 for the likelihood root statistic, see Section 5 for details. At level 5%, we are forced to just reject the null using the score statistic and accept the null using the likelihood root statistic, despite both approximate P-values being clearly less than 0.05. Even for these rather large sample sizes, the standard approximate P-values are quite inaccurate. Moreover their inaccuracy

differs for different statistics in a manner that is difficult to predict.

At least two possible solutions to the inaccurate behavior of standard approximate test statistics have emerged over the past decade. The first alternative is so-called higher order asymptotics that leads to an adjusted likelihood root statistic that is not only closer to normally distributed but also respects conditionality (Brazzale, Davison, and Reid, 2007). The second alternative is the parametric bootstrap, which involves replacing the nuisance parameter by a null estimate. The resulting P-value is also sometimes called a “plug-in” or “estimated” P-value. Both these methods have been shown to have order of error $O(m^{-3/2})$ for continuous models. For discrete models, the order of error is not clear, though it appears to be $O(m^{-1})$, where m is a measure of sample size (DiCiccio and Young, 2008; Lloyd, 2010). In our case, we may take m to be any measure of sample size that increases with $\min(n_1, n_2)$.

The purpose of this article is to describe and evaluate these two alternative methods, for the specific but important case of testing the difference between two independent binomial proportions.

2. First- and Second-Order Test Statistics

Denote the interest parameter by $\delta = p_2 - p_1$, the null value δ_0 and the remaining nuisance parameter by $\lambda = p_1$. Note that λ is restricted to the interval $\Lambda_\delta = [\max(-\delta, 0), \min(1 - \delta, 1)]$. We will use subscript δ to denote estimation under the null hypothesis that $p_2 - p_1$ equals a specified value δ .

There are two first-order statistics in common use for this problem, both with approximate normal distributions when $\min(n_1, n_2)$ is large and (p_1, p_2) is not on the boundary of the unit square. The first is the (signed) likelihood root statistic

$$r(\delta) = \text{sign}(\hat{\delta} - \delta) [2\{\ell(\hat{\delta}, \hat{\lambda}) - \ell(\delta, \hat{\lambda}_\delta)\}]^{1/2}$$

considered by Munk et al. (2005), where $\ell(\delta, \lambda)$ is given in Appendix A. An alternative statistic is the Score statistic, which in this case is

$$t(\delta) = \frac{y_2/n_2 - y_1/n_1 - \delta}{\hat{\sigma}_\delta},$$

where $\sigma^2 = p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$, see Chan (1998).

P-values based on a standard normal approximation to either of these statistics are said to be first-order accurate and suffer errors of $O(m^{-1/2})$ for one-sided P-values. Other first order statistics have been considered (Mehrotra et al., 2003) but the score and likelihood root statistics are preferred among first-order methods.

So-called second-order likelihood inference is based on double saddlepoint approximations to the distribution of the ML estimator conditional on an approximate ancillary component. A readable account is Brazzale et al. (2007). While the general theory is complex it results in a modified likelihood root statistic

$$r^*(\delta) = r(\delta) + r(\delta)^{-1} \log\{q(\delta)/r(\delta)\}, \tag{1}$$

where dependence on y is suppressed and $q(\delta)$ is complicated in general. Letting $w_i = p_i(1 - p_i)$ and $\varphi_i = \text{logit}(p_i)$, the adjustment statistic $q(\delta)$ for the difference of two binomials

simplifies to

$$q(\delta) = \frac{\{\hat{w}_{2\delta}(\hat{\varphi}_2 - \hat{\varphi}_{2\delta}) - \hat{w}_{1\delta}(\hat{\varphi}_1 - \hat{\varphi}_{1\delta})\}}{\sqrt{\hat{w}_{1\delta}/n_1 + \hat{w}_{2\delta}/n_2}} \times \sqrt{\frac{\hat{w}_1 \hat{w}_2}{\hat{w}_{1\delta} \hat{w}_{2\delta}}}. \tag{2}$$

This is derived in Appendix B. This generates a P-value $p^*(Y, \delta_0) = 1 - \Phi(r^*(Y, \delta_0))$. However, the formula for $r^*(\delta)$ diverges under two quite different circumstances. First, saddlepoint approximations are known to fail in the center of the distribution and we see that $r^*(\delta)$ breaks down when $r(\delta) \approx 0$. This is easily corrected by setting $r^*(\delta) = r(\delta)$, say when $|r(\delta)| < 0.001$. These data sets are of little interest anyway.

Second, the adjustment statistic $q(\delta) \approx 0$ whenever $\hat{w}_1 \hat{w}_2 \approx 0$, which occurs near the boundary of the sample space. It is not widely appreciated that r^* breaks down on the boundary and that this can have quite perverse effects on the performance of the test. This problem must be explicitly handled to define a proper frequentist test. Define the boundary set to be those data points for which $\hat{w}_1 \hat{w}_2 = 0$ and $\hat{\delta} > \delta_0$. In later numerical investigations, we try three re-definitions of $r^*(\delta)$ on this boundary set. The first is to simply set $r^*(\delta_0) = r(\delta_0)$. A more complex approach is to add or subtract $\frac{1}{2}$ to the count (y_1, y_2) to move it off the boundary and then calculate $r^*(\delta_0)$ using the formula above with these corrected data. We then divide the resulting P-value by two, which is motivated by a kind of mid-P argument, see Pierce and Peters (1992). The third approach is to replace $p^*(y, \delta_0)$ by the bootstrap statistic $\hat{P}(y, \delta_0)$ for boundary values y . We now define the bootstrap statistic.

3. Bootstrap P-values

An entirely different approach to obtaining a more accurate test statistic or P-value is to compute the P-value exactly but with the estimated nuisance parameter substituted for the true value. This estimated P-value has the general closed form

$$\hat{P}(y, \delta) = \sum_{y': T(y') \geq T(y)} \Pr(Y = y'; \hat{\theta}_\delta). \tag{3}$$

Treating Y as a random variable, we have the P-value $\hat{P}(Y, \delta_0)$. Even though for this article $\hat{P}(Y, \delta_0)$ is computed directly rather than with simulation, this is a bootstrap P-value in the sense that it is a measure of statistical significance calculated under a distribution estimated from the data. For our application, $y = (y_1, y_2)$ and $\Pr(Y = y')$ is a product of binomial distributions with parameters $\theta = (p_1, p_1 + \delta)$ and we are able to compute $\hat{P}(y, \delta)$ directly by enumeration without resorting to simulation as is typically associated with bootstrap in other contexts. While different test statistics $T(Y)$ generate different bootstrap P-values, the difference is usually slight.

The simplicity of formula (3) suppresses the computational burden as the sample size grows, as we must determine which elements y' of the sample space satisfy $T(y') \geq T(y)$. However, up to sample sizes of several hundred it can be computed quickly. Moreover, when $T(y)$ is monotonic in its arguments the boundary of the tailset in (3) can be found without full enumeration, see Lloyd (2008a) for details. For continuous models, inferences from such P-values incur error of $O(m^{-3/2})$

(Lee and Young, 2005). For discrete models, the error rate is not clear though it appears to be $O(m^{-1})$ (DiCiccio and Young, 2008; Lloyd, 2010). The bootstrap P-value does not require any special modifications near or on the boundary.

Both $p^*(Y, \delta_0)$ and $\hat{P}(Y, \delta_0)$ are claimed to approximately respect any exact or approximate conditionality in the model. Both have error rates better than first-order statistics. What is not clear is how close to exact the implied tests are for discrete models, and how they perform in terms of power. For models where there is an exact conditional inference, $p^*(Y, \delta_0)$ approximates the mid-P tail probability from the conditional distribution. Applying a continuity correction to the data and applying $p^*(Y, \delta_0)$ leads to an approximation to the exact conditional tail probability. We have not investigated this version since exact conditional P-values are known to be unconditionally conservative.

Example cont. For the earlier mentioned Burlington study, the P-value based on Chan's (1998) $t(-0.05)$ was 0.0453 with exact version 0.0500, defined later in equation (5). Using $r(-0.05)$, the approximate P-value was 0.0464 with exact version 0.0760. The statistic r^* here gives a P-value $p^* = 0.0466$ but the exact P-value from this statistic is 0.0760, just as it was for the unadjusted statistic r . The bootstrap P-values are 0.0474 for either Chan's score or likelihood root statistic and the exact versions of these P-values are both 0.0475. For this example then, bootstrap P-values are virtually identical for either statistic and are close to exact.

4. Test Size

All tests to be investigated are expressed in terms of an approximate P-value that generates a nominal size α test by rejecting the null if $P(y) \leq \alpha$. The relative size bias of the test is

$$e(\alpha, \lambda) := \Pr(P(Y, \delta_0) \leq \alpha); \lambda, \lambda + \delta_0 / \alpha - 1, \quad (4)$$

where dependence on (n_1, n_2) is suppressed.

Since $e(\alpha, \lambda)$ depends on λ we will look at two criteria, namely integrated bias $\bar{e}(\alpha) = \int |e(\alpha, \lambda)| d\lambda$ and maximum

liberal bias $e^*(\alpha) = |\sup_{\lambda} e(\alpha, \lambda)|$. The second measures the extent to which the P-value exaggerates the true significance. In particular, noting that the true size of most tests is zero when λ takes extreme values, it does not penalize P-values that have true size smaller than nominal for much of the parameter space. It only penalizes liberality. This is consistent with Bickel and Doksum (1977), who define the size of a test to be the maximum probability of rejection. It is also consistent with theory of Röhmel and Mansmann (1999) and Lloyd (2008a), who show that P-values based on maximizing out nuisance parameters possess strong optimality properties, see equation (5). It is simple to show that if (n_1, n_2) is replaced by (n_2, n_1) then $e(\alpha, \lambda)$ is simply reflected in the interval Λ_{δ} . Consequently, \bar{e} and e^* are unchanged, so we will assume that $n_2 \leq n_1$ without loss of generality.

I computed the curves $e(\alpha, \lambda)$ exactly at an even grid of 101 values of λ in Λ_{δ} from which I approximated $\bar{e}(\alpha)$ and $e^*(\alpha)$ by the simple average and maximum, respectively. The reader should note that no simulation is involved and the only errors are in approximating the curve by a grid of points. There were five approximate P-values and their bootstrap versions investigated. The five approximate P-values were Chan's $t(\delta_0)$, the likelihood root statistic $r(\delta_0)$, and the modified likelihood root statistic $r^*(\delta_0)$ with three previously mentioned modifications on the boundary, namely bootstrap r_{BT}^* , likelihood root r_{LR}^* , and the continuity correction r_{PP}^* . The following parameter combinations were covered: $\alpha = 0.01, 0.05, 0.10; \delta_0 = -0.1, 0$. Initial sample sizes (4,4), (5,3), and (6,2) were scaled up by the factor $m = 4, 6, \dots, 18, 20$.

Figure 1 gives results for the statistic $t(-0.1)$ for sample sizes $18 \times (5, 3) = (90, 54)$, but are typical of all other sample sizes and statistics, see Lloyd (2008c) for full results. The left panel plots the raw P-values $P(Y; -0.1)$ from the normal approximation to $t(-0.1)$ against the bootstrap versions, each point corresponding to a different possible data set of the 5005 in the sample space. The bootstrap P-values tend to be slightly larger than the raw P-values, correcting for the well-known fact that Wald-type statistics tend to be liberal. There are also some distinct differences in the way the two

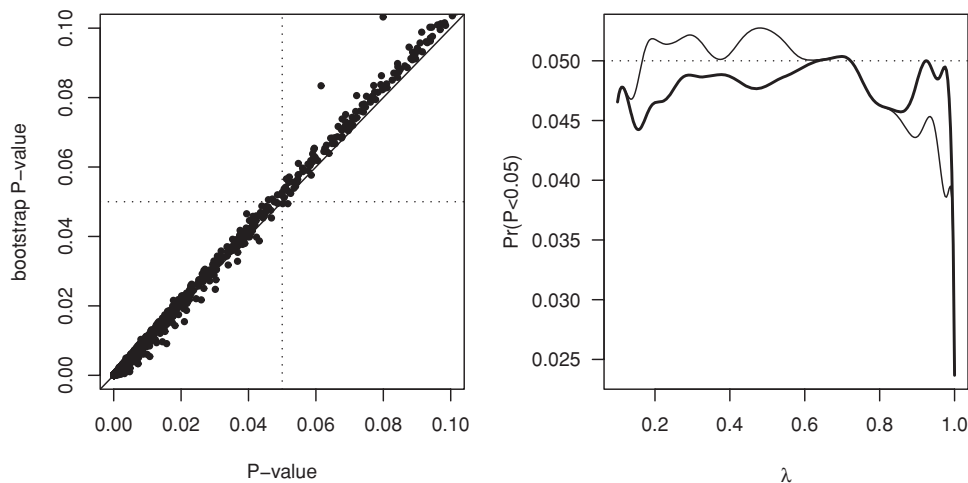


Figure 1. Size bias of P-values from $t(-0.1)$. Left panel. Ordinary P-values based on normal approximation versus their bootstrap versions. Right panel. Tail probability $\Pr(P < \alpha; \lambda)$ versus $\lambda \in \Lambda_{-0.1} = [0.1, 1.0]$.

Table 1

Size bias of five approximate P-values and their bootstrap versions. The table describes absolute relative size bias $e(\alpha, \lambda_1)$ (expressed as a percentage) as defined in (4). Each figure is the average for six different variations of the basic test, namely with $\alpha = 0.01, 0.05, 0.1$, and $\delta_0 = -0.1, 0$. In the upper section, relative size bias is integrated over 100 values of λ (called \bar{e} in the text) and the lower section relative size bias is measured by the maximum over 100 values of λ (called e^* in the text). Smaller values of \bar{e} and e^* are preferred. Sample sizes are $(n_1, n_2) = m \times (5, 3)$.

m	Nonbootstrap P-value					Bootstrap P-value				
	t	r	r* _{BT}	r* _{LR}	r* _{PP}	t	r	r* _{BT}	r* _{LR}	r* _{PP}
4	24.5	28.6	20.0	24.6	24.8	24.5	24.3	23.0	25.1	24.5
6	19.2	24.7	16.0	21.1	24.7	16.4	17.1	16.4	16.8	17.2
8	15.9	16.8	12.8	14.0	18.6	13.9	15.3	14.8	15.0	16.7
10	13.1	12.9	10.1	11.2	14.5	11.9	12.1	11.8	11.4	14.0
12	11.5	12.4	9.7	10.2	14.5	10.8	11.0	10.1	9.7	11.1
14	11.0	10.8	8.7	9.4	12.7	9.2	8.6	8.6	8.4	9.8
16	10.1	9.6	8.4	8.9	12.3	8.3	8.4	8.4	8.2	8.9
18	9.0	8.3	6.8	7.5	10.2	7.1	7.0	7.1	7.1	8.1
20	9.2	7.9	6.9	7.0	10.8	6.8	6.8	6.8	6.8	7.6
4	52.2	114.3	15.8	113.0	77.7	0.2	0.5	0.2	0.3	0.2
6	41.8	143.7	30.7	141.7	164.2	1.6	0.7	3.2	1.2	3.2
8	30.1	87.5	18.5	84.0	235.7	3.5	2.3	3.5	1.2	1.8
10	23.5	122.8	14.2	113.6	165.3	0.6	0.6	0.6	1.9	1.8
12	28.6	93.8	17.5	92.9	207.4	1.8	2.0	0.8	4.1	0.8
14	27.6	70.8	26.8	70.0	185.1	1.4	3.7	2.2	3.7	0.6
16	25.7	64.7	31.1	69.0	130.0	1.9	5.0	0.9	5.0	0.9
18	22.7	83.7	17.9	84.0	151.1	2.4	3.7	3.0	3.8	2.0
20	29.8	61.3	17.8	58.5	173.6	3.8	4.2	2.7	4.2	1.8

P-values order the sample space. The right panel displays the function $\Pr(P < 0.05; \lambda)$ against $\lambda \in \Lambda_{-0.1} = [0.1, 1.0]$. The ordinary line is for the raw P-value and the bold line for the bootstrap P-value. The tail probability is closer to nominal for the bootstrap versions ($\bar{e} = 4.7, e^* = 0.7$) then for the raw P-value ($\bar{e} = 5.9, e^* = 5.4$), especially when assessed by the worst-case measure e^* .

To summarize a rather large amount of numerical evidence, Table 1 gives the size bias averaged across the three values of α and two values of δ , for starting sample size (5,3). Full results are listed in Lloyd (2008c). The upper sections give integrated size bias \bar{e} and the bottom section maximizes size bias e^* . To aid interpretation, the top left figure 24.5 means that the size of Chan's test for sample size $(n_1, n_2) = 4 \times (5, 3) = (20, 12)$ deviates from nominal by an average of 24.5% of that nominal value as λ varies. The bottom right figure of 1.8% indicates that the bootstrap test for sample sizes $20 \times (5, 3) = (100, 60)$ has true size that virtually never exceeds nominal; for the worst value of λ it only exceeds nominal by 1.8% of nominal. Examination of this table and those in Lloyd (2008c) suggests very consistent results, as listed below, and also that Table 1 gives a reasonable summary of these results. The main patterns are:

- (1) Integrated size bias decreases with sample size for all P-values. Amongst the nonbootstrap P-values, r^*_{LR} has slightly smaller integrated size bias for larger sample sizes. The bootstrap P-values have similar integrated size bias to r^*_{LR} and slightly better for larger sample sizes.
- (2) Maximum size bias tells a different story. All the non-bootstrap P-values have large maximum size bias. The

best of them is r^*_{BT} (which does employ bootstrap on the boundary). For all the bootstrap P-values, maximum bias is at least an order of magnitude smaller. It is noteworthy that maximum bias is hardly affected by sample size. For the nonbootstrap P-values it is large for all sample sizes and for the bootstrap P-values it is small for all sample sizes.

On the basis of these results, one would recommend bootstrap P-values over any of the versions of r^* , if the aim is to ensure that the true unconditional size is close to the nominal size. Bootstrap P-values have similar or perhaps slightly smaller integrated size bias and much smaller maximum size bias. Moreover, bootstrap of r^* on the problematic boundary is not sufficient to capture these gains. Bootstrap P-values seem to behave very similarly, regardless of the generating statistic, but those based on $t(\delta_0)$ seem slightly better which is fortunate since $t(\delta_0)$ is the simplest of the five to compute. We mention again that the bootstrap P-values for these sample sizes did not require simulation. The term bootstrap is referring to replacing the unknown λ with the estimator $\hat{\lambda}_\delta$ and then computing an exact tail probability.

5. Test Power

It makes no sense to compare the power of competing tests of differing size, since power can be increased by increasing size. Therefore, to compare two competing approximate P-values, both must be corrected to have size as close to nominal as possible without violating nominal. There is a very clean theory of exact P-values described in Röhmel and Mansmann (1999) and Lloyd (2008a). The upshot of this work

Table 2

Powers of exact standard and exact bootstrap tests. Each figure summarizes the average power (expressed as percentage) of an exact test based on the approximate P-value indicated by the column labels. Each figure is the average of three results for $\alpha = 0.01, 0.05, 0.1$ calculated at the local alternative $\delta_1 = \delta_0 + 1/\sqrt{(n_1 n_2)^{1/2}}$. The upper section describes the null $\delta_0 = -0.1$, the lower section $\delta_0 = 0$. Sample sizes are $m \times (5, 3)$.

m	Nonbootstrap P-value					Bootstrap P-value				
	t	r	r* _{BT}	r* _{LR}	r* _{PP}	t	r	r* _{BT}	r* _{LR}	r* _{PP}
4	36.0	35.5	35.5	35.5	35.5	38.7	38.4	38.4	38.4	38.4
6	38.0	34.7	34.7	34.7	34.7	40.4	40.0	40.0	40.0	40.0
8	38.8	33.4	33.4	33.4	33.4	41.0	40.8	40.8	40.8	40.8
10	39.4	32.7	32.7	32.7	32.7	41.9	41.5	41.5	41.5	41.5
12	39.7	32.4	32.4	32.4	32.4	42.3	42.1	42.1	42.1	42.1
14	40.0	32.7	32.7	32.7	32.7	42.5	42.3	42.3	42.3	42.3
16	39.4	32.9	32.9	32.9	32.9	42.8	42.5	42.5	42.5	42.5
18	39.6	32.9	32.9	32.9	32.9	42.8	42.7	42.7	42.7	42.7
20	38.0	32.9	32.9	32.9	32.9	43.0	42.7	42.7	42.7	42.7
4	38.0	37.5	37.5	37.5	37.5	40.8	40.6	40.6	40.6	40.6
6	40.6	37.1	37.1	37.1	37.1	43.1	42.6	42.6	42.6	42.6
8	41.7	36.4	36.4	36.4	36.4	44.0	43.9	43.9	43.9	43.9
10	42.8	35.8	35.8	35.8	35.8	45.3	44.9	44.9	44.9	44.9
12	43.4	35.9	35.9	35.9	35.9	46.0	45.8	45.8	45.8	45.8
14	44.0	36.4	36.4	36.4	36.4	46.6	46.3	46.3	46.3	46.3
16	43.6	37.0	37.0	37.0	37.0	47.1	46.9	46.9	46.9	46.9
18	44.1	37.2	37.2	37.2	37.2	47.4	47.3	47.3	47.3	47.3
20	42.7	37.4	37.4	37.4	37.4	47.8	47.6	47.6	47.6	47.6

is that for a given approximate P-value $P(y, \delta_0)$, the maximized P-value

$$P^*(y, \delta_0) := \sup_{\lambda} \{\Pr(P(Y, \delta_0) \leq P(y, \delta_0); \lambda, \lambda + \delta_0)\} \quad (5)$$

is as small as possible among P-values that satisfy the size restriction and order the sample space in the same order as $P(Y, \delta_0)$. In short, $P^*(Y, \delta_0)$ has a strong claim to being the exact frequentist P-value, once the approximate test statistic has been decided upon. We therefore assess the power of competing approximate P-values by the power of their exact versions, $P^*(Y, \delta_0)$. Specifically the power function

$$\beta_{\alpha}(\delta, \lambda) = \Pr(P^*(Y, \delta_0) \leq \alpha; \lambda, \lambda + \delta),$$

is evaluated at nonnull parameter values $\delta = \delta_0 + c/\sqrt{(n_1 n_2)^{1/2}}$ with $c = 1$. These nonnull values are chosen so that the power is neither close to 0 or 1.

Table 2 summarizes power for starting sample size $(n_1, n_2) = (5, 3)$. Each figure gives the average of the powers of the exact tests with size 0.01, 0.05, and 0.10. The main message from these figures is that the bootstrap-based test appears to be consistently more powerful than the tests based on the nonbootstrap statistics. Among nonbootstrap tests, Chan’s test is more powerful than the other four, but not as powerful as any of the bootstrap tests. There is evidence that the test based on Chan’s statistic is very slightly more powerful than the other bootstrap tests, but the starker message is that all bootstrap tests are almost equally good.

Again, detailed and exhaustive results are available in Lloyd (2008c). The averaging over three nominal sizes does not suppress any anomalous behavior. The main differences are that the larger size tests have higher power, but the consistent superiority of the bootstrap tests persists. I also inves-

tigated more remote alternatives with $c = 1.5$ and $c = 2$ and again find uniform superiority of the bootstrap test.

6. Varying the Inferiority Margin

It has been noted by Röhmel (2005) that exact P-values $P^*(Y, \delta_0)$ may vary rather erratically with the noninferiority margin δ_0 . In principle, we would hope that, as the noninferiority margin increases, the P-value would decrease smoothly. Unfortunately this is not the case for maximized P-values because the tail set, whose exact probability is calculated, changes with δ_0 . Chan and Zhang (1999) suggested constructing intervals for δ from inverting the exact test which, as noted by Röhmel, leads to problems.

Figure 2 shows the dependence of various likelihood root-based P-values on δ for the specific data set $(y_1, y_2) = (34, 21)$, $(n_1, n_2) = (70, 30)$. Each plot shows the approximate P-value (as a line) and the maximized P-value based on it (as points). For the left plot, the approximate P-value is $1 - \Phi(t(34, 21; -\delta))$ which is a continuous function of δ . The maximized P-value is much larger but also varies quite erratically with δ . While it is not typical in practice to calculate P-values for a range of noninferiority margins, the erratic dependence is conceptually worrying. More practically, it means that intervals obtained from inverting such P-values may be discontinuous, not to mention that computing such regions becomes extremely difficult as the number of discontinuity points increases.

The right plot is for the bootstrap P-value. While the line $\hat{P}((34, 21); \delta)$ looks continuous in δ , it is actually discontinuous, but the sizes of the discontinuities are tiny because the points entering the tail set have very small probability at λ_{δ} . In all other examples checked, the discontinuities were so small as to be visually undetectable. The maximized

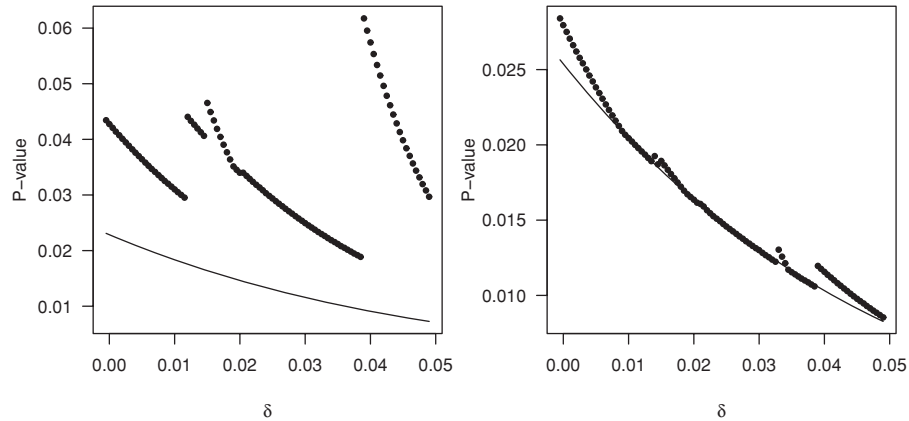


Figure 2. Dependence of likelihood root based P-values on δ for $(y_1, y_2) = (34, 21)$. *Left panel.* Approximate P-value $P(y; \delta)$ (line) and maximized P-value $P^*(y; \delta)$ (points). *Right panel.* Estimated P-value $\hat{P}(y; \delta)$ (line) and estimated then maximized P-value $\hat{P}^*(y; \delta)$ (points).

version is more clearly discontinuous, but much less so than for the maximized P-value in the left plot. It is also quite close to the unmaximized bootstrap P-value.

This example is not special and it seems to generally be the case that bootstrap P-values are close to continuous functions of δ (though formally discontinuous) while their maximized versions are more discontinuous but often not practically different to the bootstrap P-values. One way to avoid these discontinuity problems is to base confidence limits not on test inversion but on adjusting an approximate limit to be exact, see Buehler (1957). Such limits possess a strong optimality property but do not always agree with the exact test based on the same test statistics.

7. Conclusions

In this article, I have looked at a range of modern methods for calculating more accurate test statistics for risk difference and investigated their performance in terms of size accuracy, power, and dependence on the null parameter. It has been found that the bootstrap adjustment to the P-values based on $r(\delta)$ and $t(\delta)$ produces tests whose size is closer to nominal than other approaches and exceeds nominal by a very small margin.

It is not well appreciated that the maximum size bias of standard test statistics does not seem to decrease with sample size, and can be very large even for sample sizes in the several hundreds. In contrast, the maximum size bias of the bootstrap based P-values is extremely small for small or large sample sizes. Notwithstanding the asymptotic properties of bootstrap P-values in Lee and Young (2005), the excellent performance seems to have little to do with asymptotics. For practical purposes, one can take the bootstrap P-value based tests to be exact, even for the smallest sample sizes considered here.

Second-order based P-values display somewhat smaller size bias than standard P-values, particularly with the bootstrap adjustment on the boundary. However, their maximum size bias is still an order of magnitude larger than bootstrap

P-values. To be fair, r^* is constructed to approximate certain conditional tail probabilities and for discrete problems, conditional performance need not translate into unconditional performance. So supporters may argue that they are being held to a false, i.e., unconditional, standard. On the other hand, DiCiccio and Young (2008) show that for continuous models parametric bootstrap P-values do respect conditionality approximately in a similar manner to r^* .

I have argued that the only sensible way to compare the powers of approximate tests is to compare the powers of the exact test they generate. The powers of the exact tests based on the bootstrap P-values consistently exceed the power of the exact tests based on any of the nonbootstrap P-values, including the second-order P-values. So not only do bootstrap P-values hardly exceed nominal size, they have enhanced power compared to alternatives. They are recommended for general use for testing risk difference. Similarly favorable results for bootstrap P-values have been demonstrated for other models (Lloyd, 2008b, 2010; Lloyd and Moldovan, 2008).

Computation of the bootstrap P-value requires computing the approximate test statistic for all possible samples, which becomes difficult for sufficient large sample sizes. General computational issues are considered in Lloyd (2008a). An R function is supplied by the author, which takes only a few seconds for samples sizes (200,200). For larger sample sizes, a simple simulation algorithm that uses importance sampling is under development.

REFERENCES

- Barndorff, O. and Cox, D. R. (1994). *Inferences and Asymptotics*. London: Chapman and Hall.
- Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*. Oakland, California: Holden-Day.
- Brazzale, A., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-sample Statistics*. New York: Cambridge University Press.
- Buehler, R. J. (1957). Confidence intervals for the product of two binomial probabilities. *Journal of the American Statistical Association* **52**, 482–493.

Chan, I. S. F. (1998). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine* **17**, 1403–1413.

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

Davison, A. C., Fraser, D. A. S., and Reid, N. (2006). Improved likelihood inference for discrete data. *Journal of the Royal Statistical Society, Series* **68**, 495–508.

DiCiccio, T. J. and Young, G. A. (2008). Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika* **95**, 1–12.

Dunnett, C. W. and Gent, M. (1977). Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables. *Biometrics* **33**, 593–602.

Farrington, C. P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine* **9**, 1447–1454.

Lee, S. M. S. and Young, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Statistics and Probability Letters* **71**, 143–153.

Liu, J.-P., Hsueh, H.-M., Hsieh, E., and Chen, J. J. (2002). Test for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* **21**, 231–245.

Lloyd, C. J. (2008a). Exact P-values for discrete models obtained by estimation and maximisation. *Australian and New Zealand Journal of Statistics* **50**, 329–346.

Lloyd, C. J. (2008b). Exact tests based on pre-estimation and second order pivots: non-inferiority trials. To appear in *Journal of Statistical Computation and Simulation* **15**.

Lloyd, C. J. (2008c). A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics* **64**, 716–723.

Lloyd, C. J. (2008d). Bootstrap and second order tests of risk difference. MBS Working Paper.

Lloyd, C. J. (2010). Estimated P-values in discrete models: Asymptotic and non-asymptotic effects. To appear in *Journal of Statistical Computation and Simulation*.

Lloyd, C. J. and Moldovan, M. V. (2008). More powerful exact test of noninferiority from binary matched pairs data. *Statistics in Medicine* **27**, 3540–3549.

Lugannani, R. and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent variables. *Advances in Applied Probability* **12**, 475–490.

Mehrotra, D. V., Chan, I. S. F., and Berger, R. L. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* **59**, 441–450.

Munk, A., Skipka, G., and Stratmann, B. (2005). Testing general hypotheses under binomial sampling; The two sample case— asymptotic theory and exact procedures. *Computational Statistics and Data Analysis* **49**, 723–739.

Pierce, D. A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion). *Journal of the Royal Statistical Society, Series B* **54**, 701–737.

Reid, N. (2003). Asymptotics and the theory of inference. *Annals of Statistics* **31**, 1695–1731.

Röhmel, J. (2005). Problems with existing procedures to calculate exact unconditional P-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. *Biometrical Journal* **47**, 37–47.

Röhmel, J. and Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal* **41**, 149–170.

Spitzer, W. O., Sackett, D. L., Sibley, J. C., Roberts, R. S., Gent, M., Kergin, D. J., Hackett, B. C., and Olynich, A. (1974). The

Burlington randomised trial of the nurse practitioner. *New England Journal of Medicine* **290**, 251–256.

Received February 2009. Revised July 2009.
Accepted September 2009.

APPENDIX A

First-Order Statistics

Denote the interest parameter by $\delta = p_2 - p_1$, the null value δ_0 , and the remaining nuisance parameter by $\lambda = p_1$. The log-likelihood is

$$\ell(\delta, \lambda; y_1, y_2) = y_1 \logit(\lambda) + y_2 \logit(\delta + \lambda) + n_1 \log(1 - \lambda) + n_2 \log(1 - \delta - \lambda)$$

where λ is restricted to the interval $\Lambda_\delta = [\max(-\delta, 0), \min(1 - \delta, 1)]$. The maximum likelihood (ML) estimator $\hat{\lambda}_\delta$ of λ for fixed δ satisfies the equation

$$\frac{\partial \ell}{\partial \lambda} = \frac{y_1}{\lambda} + \frac{y_2}{\delta + \lambda} - \frac{n_1 - y_1}{1 - \lambda} + \frac{n_2 - y_2}{1 - \delta - \lambda},$$

which rearranges into a cubic with leading coefficient 1 and remaining coefficients

$$c_2 = \delta(1 + f_1) - 1 - \hat{p}, c_1 = \hat{p} - 2\delta f_1 \hat{p}_1 - \delta + \delta^2 f_1, c_1 = f_1 \hat{p}_2 \delta(1 - \delta),$$

where $f_1 = n_1/n$, $\hat{p}_1 = y_1/n_1$, and $\hat{p} = t/n$ is the estimate under $\delta = 0$. When $\delta = 0$, this reduces to solving a quadratic since $c_1 = 0$ whose solutions are \hat{p} and 1, while 0 is also a solution of the cubic. Cubics can be solved quickly and accurately using the **R**-function *polyroot*. Except for boundary cases there is a unique real root within the interval Λ_δ . The cubic-based estimator was first noted by Farrington and Manning (1990). Dunnett and Gent (1977) used the simple unbiased estimators

$$\tilde{p}_{1\delta} = \frac{y_1 + y_2}{n_1 + n_2} - \delta \frac{n_2}{n_1 + n_2}, \tilde{p}_{2\delta} = \tilde{p}_{1\delta} + \delta$$

however $\tilde{p}_{1\delta}$ may violate the interval Λ_δ .

APPENDIX B

Second-Order Statistic

So-called second-order likelihood inference is based on a local decomposition of the model into a sufficient and ancillary component. An approximate formula for the conditional distribution is available (Barndorff-Nielsen and Cox, 1994) and the cumulative tail probability of this distribution is further approximated (Lugannani and Rice, 1980), see Reid (2003) for an extensive review. 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 Brazzale et al. (2007). Suppose that the log-likelihood is of exponential family form

$$\ell(\theta; y) = \varphi(\theta)^T v(y) - c(\varphi(\theta)),$$

where $\varphi(\theta)$ is the canonical parameter of dimension d and $v(y)$ is the sufficient statistic. For our application, the canonical

parameters are the log-odds parameters and $d = 2$. Suppose we are interested in inference on the parameter $\delta(\theta)$. For simplicity we reparametrize the model in terms of $\theta = (\delta, \lambda)$ where λ denotes a nuisance parameter.

The general formula for second-order inference requires derivation of an adjustment statistic, $q(\delta)$, that depends on two quantities. The first is the observed information matrix in the (δ, λ) parameterization, denoted $j(\theta)$. The second is the $d \times d$ Jacobian matrix φ'_θ of the transform from φ to θ . Then the adjustment statistic is

$$q(\delta) = \frac{|\hat{\varphi} - \hat{\varphi}_\delta \quad \varphi'_\lambda(\hat{\theta}_\delta)|}{|\varphi'_\theta(\hat{\theta})|} \times \frac{|j(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\delta)|}. \tag{A.1}$$

By $\hat{\varphi}_\delta$ we mean the ML estimate of the parameter vector φ when δ is fixed.

The first term depends only on the functional relationship of δ to the canonical parameter φ . The notation $\hat{\varphi} - \hat{\varphi}_\delta$ in the first matrix is shorthand for the column vector giving the deviation of the canonical parameter φ when estimated under the general and null models. By φ'_λ we mean the submatrix of φ'_θ with the first column removed. When δ is a component of the canonical statistic φ , then the first term reduces to $\hat{\delta} - \delta$. The second term is often called the nonorthogonality term since it equals unity when the interest parameter δ and nuisance parameter λ are orthogonal. Note that $j_{\lambda\lambda}$ refers to the submatrix of $j(\theta)$ obtained by removing the first row and column. Second-order inference is based on the modified likelihood root defined earlier in (1).

We now derive an explicit expression for $q(\delta)$ for the binomial model. For future work, and it turns out also for simplicity, we generalize to testing $\delta = h(p_2) - h(p_1)$ for a general link function h and take the nuisance parameter as $\lambda = h(p_1)$. The inverse transformation is

$$p_2 = h^{-1}(\delta + \lambda), p_1 = h^{-1}(\lambda).$$

Let φ denote the logit transform that is also the canonical link. With abuse of notation, we later denote the logit parameters $\text{logit}(p_j)$ by φ_j . The log-likelihood is

$$\ell(\delta, \lambda; y_1, y_2) = y_1\varphi\{h^{-1}(\lambda)\} + y_2\varphi\{h^{-1}(\delta + \lambda)\} + n_1 \log(1 - h^{-1}(\lambda)) + n_2 \log(1 - h^{-1}(\delta + \lambda)).$$

Since $\partial\varphi/\partial p = 1/(p(1-p))$, the score functions are given by

$$U_\delta(\delta, \lambda) := \frac{\partial\ell}{\partial\delta} = \frac{y_2 - n_2 p_2}{p_2(1-p_2)h'(p_2)}$$

$$U_\lambda(\delta, \lambda) := \frac{\partial\ell}{\partial\lambda} = \frac{y_1 - n_2 p_1}{p_1(1-p_1)h'(p_1)} + \frac{y_2 - n_2 p_2}{p_2(1-p_2)h'(p_2)}$$

and we will denote henceforth $w(p) = p(1-p)h'(p)$. The restricted MLE $\hat{p}_{1\delta}$ satisfies the equation

$$0 = \frac{y_1 - n_2 p_1}{w(p_1)} + \frac{y_2 - n_2 p_2}{w(p_2)} \Leftrightarrow 0 = (y_1 - n_2 p_1)w(p_2) + (y_2 - n_2 p_2)w(p_1),$$

where $p_2 = h^{-1}(\delta + h(p_1))$. Solution requires numerical methods in general, but in our case the equation reduces to a quadratic. When h is the identity link it reduces to a cubic. Munk et al. (2005) give some theory on restricted ML estimation under this formulation.

Taking the covariance of the score functions gives the information terms

$$j_{\delta\delta} = V_2/w_2^2, j_{\delta\lambda} = V_2/w_2^2, j_{\lambda\lambda} = V_2/w_2^2 + V_1/w_1^2,$$

where $V_j = n_j p_j(1-p_j)$ and $w_j = w(p_j)$. Hence the nonorthogonality term is

$$\frac{|j(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\delta)|} = \frac{\hat{V}_2 \hat{V}_1 / (\hat{w}_2^2 \hat{w}_1^2)}{\hat{V}_{1\delta} / \hat{w}_{1\delta}^2 + \hat{V}_{2\delta} / \hat{w}_{2\delta}^2}. \tag{A.2}$$

It remains to find the first term in (A.1), which depends on the transformation

$$(\delta, \lambda) \rightarrow (\varphi_2, \varphi_1) = (\beta(\delta + \lambda), \beta(\lambda)),$$

where $\beta(v) = \varphi(h^{-1}(v))$. The derivative of this function is

$$\beta'(v) = \frac{\varphi'\{h^{-1}(v)\}}{h'\{h^{-1}(v)\}} = \frac{1}{p(1-p)h'(p)} = \frac{1}{w(p)},$$

where $p = h^{-1}(v)$. Therefore, the Jacobian matrix is

$$\varphi_\theta(\theta) = \begin{bmatrix} \beta'(\delta + \lambda) & \beta'(\delta + \lambda) \\ 0 & \beta'(\lambda) \end{bmatrix} = \begin{bmatrix} w_2^{-1} & w_2^{-1} \\ 0 & w_1^{-1} \end{bmatrix}$$

with determinant $1/(w_2 w_1)$ and so the first term in (A.1) is

$$|\varphi_\theta(\hat{\theta})|^{-1} |\hat{\varphi} - \hat{\varphi}_\delta \quad \varphi_\lambda(\hat{\theta}_\delta)| = \hat{w}_2 \hat{w}_1 \{ \hat{w}_{1\delta}^{-1} (\hat{\varphi}_2 - \hat{\varphi}_{2\delta}) - \hat{w}_{2\delta}^{-1} (\hat{\varphi}_1 - \hat{\varphi}_{1\delta}) \}.$$

Multiplying this by the square root of (A.2) gives

$$q(\delta) = \frac{\{ \hat{w}_{1\delta}^{-1} (\hat{\varphi}_2 - \hat{\varphi}_{2\delta}) - \hat{w}_{2\delta}^{-1} (\hat{\varphi}_1 - \hat{\varphi}_{1\delta}) \} \sqrt{\hat{V}_2 \hat{V}_1}}{\sqrt{\hat{V}_{1\delta} / \hat{w}_{1\delta}^2 + \hat{V}_{2\delta} / \hat{w}_{2\delta}^2}}.$$

Rearranging and denoting $v_j = p_j(1-p_j)$ and $h'_j = h'(p_j)$, a slightly cleaner form is

$$q(\delta) = \frac{\{ \hat{v}_{2\delta} h'_{2\delta} (\hat{\varphi}_2 - \hat{\varphi}_{2\delta}) - \hat{v}_{1\delta} h'_{1\delta} (\hat{\varphi}_1 - \hat{\varphi}_{1\delta}) \}}{\sqrt{\hat{v}_{2\delta} \hat{h}_{2\delta}^2 / n_2 + \hat{v}_{1\delta} \hat{h}_{1\delta}^2 / n_1}} \sqrt{\frac{\hat{v}_2 \hat{v}_1}{\hat{v}_{2\delta} \hat{v}_{1\delta}}}.$$

For inference on the rate difference, $h(p) = p$ and so $w(p) = p(1-p)$. The formula is parameterization invariant though Davison, Fraser, and Reid (2006) seem to only claim invariance to linear transformations.