

# Bootstrap and second order tests of risk difference

BY CHRIS J. LLOYD

*Melbourne Business School, Carlton, 3053, AUSTRALIA*

c.lloyd@mbs.edu

## SUMMARY

Standard approximate tests of the difference of two probabilities have type 1 error that can differ significantly from nominal, even for quite large sample sizes. There are two modern methods of reducing type 1 error. One is to use so-called higher order asymptotics (Reid, 2003) to provide an explicit adjustment to the likelihood ratio statistic. The second is to replace the nuisance parameter in an exact calculation with a null estimate (Young and Lee, 2005), which is a kind of bootstrap. The purpose of this paper is to explain and evaluate these two methods, for testing whether a difference in probabilities  $p_1 - p_0$  exceeds a pre-chosen non-inferiority margin  $-\delta$ . On the basis of an extensive numerical study, we recommend bootstrap P-values as superior both in terms of size accuracy and also power.

*Some Key Words:* nuisance parameters; exact test; r-star; equivalence test; non-inferiority

## 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 non-inferiority 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  $p_1$  is the probability of a positive end-point with the new treatment and  $p_0$  the probability with the old treatment, then the hypotheses under test are

$$\mathcal{H}_0 : p_1 - p_0 \leq \delta \text{ versus } \mathcal{H}_1 : p_1 - p_0 > \delta,$$

for some pre-chosen non-inferiority margin  $\delta$ , often -0.1 or -0.2. The data will typically comprise  $y_0$  responses from  $n_0$  independent individuals with treatment not applied, and  $y_1$  responses from  $n_1$  independent individuals with treatment applied. The same hypotheses can be tested for matched binary data (Liu et al, 2002; Lloyd, 2008) but this design will not be considered here.

Such tests were first considered by Dunnett and Gent (1977) who gave a chi-square goodness of fit statistics 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 Munk et al (2005) who compare the LR test, a score type test of Chan (1998) and tests based on an ordering method of Rohmel and Mansmann (1999), with distributional dependence on the unspecified nuisance parameter handled by maximisation.

The problem with standard approximate test statistics for discrete models is firstly that the sample space is discrete which can cause test performance to depend erratically on sample size. Secondly, distributional approximations tend to break down on the boundary which must be accounted for in any strict frequentist method. There is ample evidence that these effects are serious even for quite large sample sizes, see Berger and Boos (1994) and more recently Lloyd (2008). To put it simply, even for sample sizes in the several hundreds a standard approximate P-value may seriously mis-represent the exact significance of the outcome.

*Example.* A standard example for illustration is from the seminal paper of Dunnett and Gent (1977). They describe the Burlington study reported in Spitzer et al. (1974) comparing patients given doctor based care (regime 0) with those given care administered by nurses only (regime 1). The success rates, appropriately defined, were  $\hat{p}_0 = 148/225 = 65.8\%$  for conventional care and  $\hat{p}_1 = 115/167 = 68.9\%$  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 doctors based care minus 5%?

Using a non-inferiority margin of  $\delta = -0.05$  we may calculate several standard approximate test statistics. For instance, the score and likelihood ratio statistics defined later lead to 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 and readers have no doubt experienced themselves. More importantly however is that fact that both these ‘approximate’ P-values are quite inaccurate. So-called exact P-values based on these are 0.0500 for Chan’s statistic and 0.0760 for the likelihood ratio statistic. So standard approximate P-values are quite inaccurate even for these rather large sample sizes. Moreover their inaccuracy differs for difference statistics in a manner which is difficult to predict.

At least two possible solutions to the erratic behaviour of standard approximate test statistics have emerged over the past decade. The first alternative is so-called higher order asymptotics which leads to an adjusted likelihood ratio statistic which is not only closer to normally distributed but also respects conditionality, see Brazzale, Davison and Reid (2007) for a readable account. The second alternative is the parametric bootstrap, which involves replacing the nuisance parameter by a null estimate. 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, see DiCiccio and Young (2008) and Lloyd (2009).

The purpose of this paper 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 order test statistics

Denote the interest parameter by  $\delta = p_1 - p_0$ , the null value  $\delta_0$  and the remaining nuisance parameter by  $\lambda = p_0$ . We will use subscript  $\delta$  to denote estimation under the null hypothesis that  $p_1 - p_0$  equals a specified value  $\delta$ . Details are in Appendix A.

There are two first order statistics in common use for this problem, both with approximate normal distributions when  $\min(n_0, n_1)$  is large and  $(p_0, p_1)$  is not on the boundary of the unit square. The first is the signed root likelihood statistic

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

considered by Munk et al (2005). An alternative statistic is the standardised estimator

$$t(\delta) = \frac{y_1/n_1 - y_0/n_0 - \delta}{\hat{\sigma}_\delta}$$

of Chan (1998), where  $\sigma^2 = p_0(1 - p_0)/n_0 + p_1(1 - p_1)/n_1$ .

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. Second order methods described below can give one-sided P-values with error  $O(m^{-1})$ .

## 3 Second order test statistics

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, Davison and Reid (2007). While the general theory is complex it results in an modified likelihood root statistic

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

where  $q(\delta)$  is in general complicated. 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 has the simple form

$$q(\delta) = \frac{\{\hat{w}_{1\delta}(\hat{\varphi}_1 - \hat{\varphi}_{1\delta}) - \hat{w}_{0\delta}(\hat{\varphi}_0 - \hat{\varphi}_{0\delta})\}}{\sqrt{n_0\hat{w}_{0\delta}^2 + n_1\hat{w}_{1\delta}^2}} \times \sqrt{\frac{\hat{w}_0\hat{w}_1}{\hat{w}_{0\delta}\hat{w}_{1\delta}}} \quad (2)$$

This is derived in Appendix B. This generates a P-value  $p^*(\delta) = 1 - \Phi(r^*(\delta))$  whose computation does not depend on the sample size. However, the formula for  $r^*(\delta)$  diverges under two quite different circumstances. Firstly, saddlepoint approximations are known to fail in the centre of the distribution and we see that  $r^*(\delta)$  breaks down when  $r(\delta) \approx 0$ . This is easily corrected by setting  $r^*(\delta) = r(\delta)$ . These data sets are of little interest anyway.

Secondly, the adjustment statistic  $q(\delta) \approx 0$  whenever  $\hat{w}_0\hat{w}_1 \approx 0$  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. Three re-definitions of  $r^*(\delta)$  on the boundary will be tried. Again setting it equal to  $r(\delta)$  is a simple alternative. A more complex approach is to add  $\frac{1}{2}$  to all counts, calculate  $r^*(\delta)$  using the formulas above, and then half the resulting P-value, see Pierce and Peters (1992). The third approach is to use the bootstrap statistic which we now define.

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 bootstrap P-value has the general close form

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

For our application,  $y = (y_0, y_1)$  and  $\Pr(Y = y')$  is a product of binomial distributions with parameters  $\theta = (p_0, p_0 - \delta)$ . Notice that a different basic test statistic  $T(Y)$  generates a different bootstrap P-value, though the difference is often slight. The simplicity of the formula suppresses the computational burden as the sample size grows, as we must evaluate each element  $y'$  of the sample space to see if  $T(y') \geq T(y)$ . However, up to sample sizes of several hundred it can be computed quickly. Young and Lee (2005) show that inferences from such P-values incur error of  $O(m^{-3/2})$  for

continuous models. For discrete models the error rate is not clear though it appears to be  $O(m^{-1})$ , see DiCiccio and Young (2008). 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.

*Example ct'd.* For the earlier mentioned Burlington study the P-value based on  $t(-0.05)$  was  $P = 0.0453$  with exact version  $p^* = 0.0500$  while using  $r(-0.05)$  the approximate P-values was 0.0464 with exact version  $P^* = 0.0760$ . The statistic  $r^*$  here gives a P-value 0.0466 but the exact P-value from this statistic is again 0.076. The bootstrap P-values are 0.0474 for either the Chan or LR statistic and the exact versions of these P-values are both 0.0475. For this example then, using the bootstrap P-value gives virtually identical results for both basic statistics and both are close to exact.

## 4 Test size

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

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

where dependence on  $(n_0, n_1)$  and  $\delta_0$  is suppressed. It is simple to show that if  $(n_0, n_1)$  is replaced by  $(n_1, n_0)$  then  $e(\alpha, \lambda)$  is simply reflected in the interval  $\Lambda_\delta$ .

Since  $e(\alpha, \lambda)$  depends on  $\lambda$  we will look at two criteria, namely  $\bar{e}(\alpha) = \int |e(\alpha, \lambda)| d\lambda$  and  $e^*(\alpha) = \sup_\lambda e(\alpha, \lambda)$ . The second measures the extent to which the P-value exaggerates the true significance and ignores parameter values for which  $e(\alpha, \lambda)$  is smaller than  $\alpha$ . This is consistent with Bickel and Doksum (1997), who define the size of a test to be the maximum probability of rejection. It is also consistent with theory of Rohmel and Mansmann (1999) and Lloyd (2008a) who show that P-values based on maximising out nuisance parameters possess strong optimality properties. Both these measures are unchanged when  $(n_0, n_1)$  is replaced by  $(n_1, n_0)$ , so we will assume that  $n_1 \leq n_0$  without loss of generality.

I computed the curves  $e(\alpha, \lambda)$  exactly at an even grid of 101 values of  $\lambda$  in  $\Lambda_\delta$  from which I approximated  $\bar{e}(\alpha)$  and  $e^*(\alpha)$  by the simple average and maximum respectively. There were five basic P-values and their bootstrap versions investigated. The five

basic P-values were Chan's  $t(\delta)$ , the likelihood root statistic  $r(\delta)$ , and the modified likelihood root statistic  $r^*(\delta)$  with three different modifications on the boundary, as mentioned in the previous section. 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  $m = 4, 6, \dots, 18, 20$ .

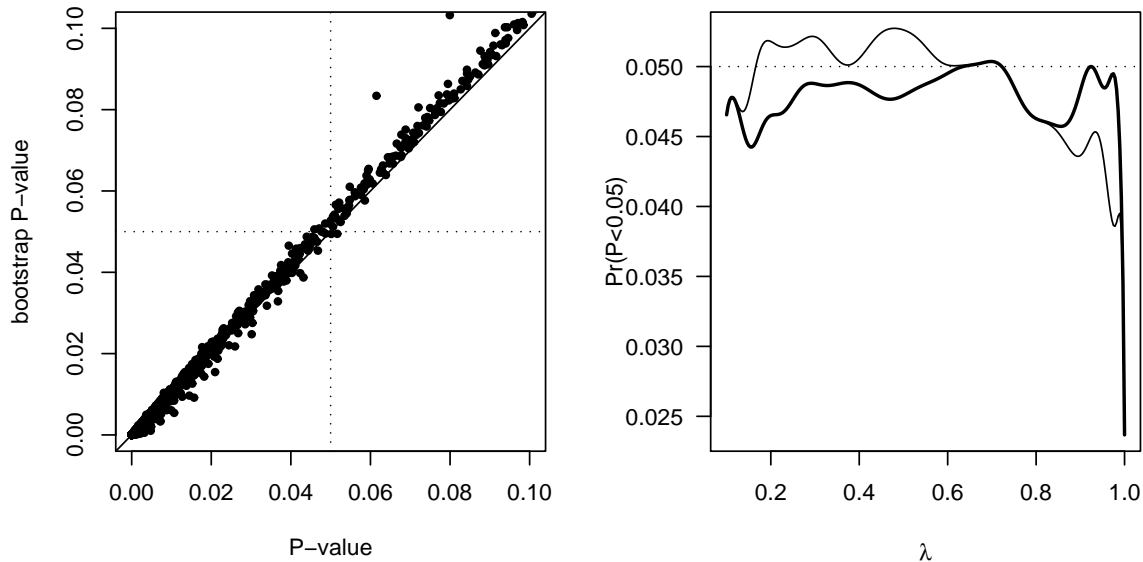


Figure 1: **Size bias of P-values from  $t(-0.1)$ .** *Left panel.* Ordinary P-values based on normal approximation versus their bootstrap versions. *Right panel.* Tail probability  $\Pr(P < \alpha; \lambda)$  versus  $\lambda \in (0.1, 1.0)$ .

Figure 1 gives results for the statistic  $t(-0.1)$  for sample sizes  $18 \times (5, 3) = (90, 54)$ , but are typical of other sample sizes and statistics. 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 know fact that Wald-type statistics tend to be liberal. There are also some distinct differences in the way the two P-values order the sample space. The right panel displays the function  $\Pr(P < 0.05; \lambda)$  against  $\lambda \in [0, 1, 1.0]$ . 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 summarise a rather large amount of numerical evidence, I have taken the average size bias for these six cases and tabulated these in Tables 1-3 for three starting sample

sizes (4, 4), (5, 3) and (6, 2). The upper sections gives mean size bias  $\bar{e}$  and the bottom section maximum size bias  $e^*$ . To aid interpretation, the top left figure 20.4 in Table 1 means that the size of Chan's test for sample size  $(n_0, n_1) = (16, 16)$  deviates from nominal by an average of 24.5% of that nominal value as  $\lambda$  varies. The bottom right figure of 1.8% indicates that the bootstrap test for sample sizes (100, 60) has true size which virtually never exceeds nominal; for the worst value of  $\lambda$  it only exceeds nominal by 1.8% of nominal.

The (5, 3) table has been unpacked into three separate tables for the three values of  $\alpha$ , but still averaged over the two values of  $\delta_0$ . These are presented in Tables 4-6 and display similar patterns, except that relative size bias tends to be larger for  $\alpha = 0.01$  than for  $\alpha = 0.10$ . The results have also been unpacked with respect to  $\delta$ , but averaged over the three values of  $\alpha$ , in Tables 7-8. Examination of these tables suggests very consistent results, as listed below, and that the first table gives a reasonable summary of typical results. The main patterns are as follows.

- (1) Mean size bias decreases with sample size for all P-values. Amongst the basic non-bootstrap P-values,  $r_{BT}^*$  and  $r_{LR}^*$  have somewhat smaller size bias for larger sample sizes. The bootstrap P-values have similar mean size bias to  $r_{BT}^*$ .
- (2) Maximum size bias tells a different story. All the basic P-values have large maximum size bias. The best of them are  $r_{BT}^*$  and we note that Chan's statistic performs much better than the likelihood root statistic. 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 basic P-values it is large for all sample sizes and for the bootstrap it is small for all sample sizes.

On the basis of these results, one could not recommend any of the versions of  $r^*$ , if the aim is to ensure that the true size is close to the nominal size. Any of the bootstrap P-values are much better for this purpose. The simplest to calculate are based on the ordinary likelihood root  $r(\delta)$  or Chan's statistic  $t(\delta)$ .

## 5 Test power

It makes no sense to compare the power of competing tests of differing size, since power can be increased by reducing size. To compare two competing P-values, both must be corrected to have size as close to nominal as possible. There is a very clean theory of exact P-values described in Rohmel and Mansmann (1999) and Lloyd (2008a). The

upshot of this work is that for a given approximate P-value  $P(y, \delta_0)$ , the maximised 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 amongst 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 correct P-value, once the basic test statistic has been decided upon. We therefore assess the power of competing approximate P-values by the power of their exact versions,  $P^*(Y)$ . Specifically the power function

$$\beta(\delta, \lambda) = \Pr(P^*(Y) \leq P^*(y); \lambda, \lambda - \delta),$$

at the non-null parameter values  $\delta_1 = \delta_0 + c/\sqrt{(n_0 n_1)^{1/2}}$  with  $c = 1$ . The non-null values are chosen so that the power is neither close to 0 or 1.

Table 9 summarises power for starting sample size  $(n_0, n_1) = (5, 3)$ . 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 uniformly more powerful than the tests based on the basic statistics. Amongst non-bootstrap 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 power of the bootstrap 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.

The averaging over three sizes does not suppress any anomalous behaviour. For instance, Tables 10-12 give average power for the size 0.01, 0.05 and 0.10 separately, but this time averaged over three sample size configurations. The main differences are that the larger size tests have higher power, but the consistent superiority of the bootstrap tests persists. I also investigated 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 Rohmel 2005) that exact P-values  $P^*(Y, \delta)$  vary rather erratically with the non-inferiority margin  $\delta$ . In principle we would hope that as the non-inferiority margin increases the P-value would decrease smoothly. Unfortunately this is not the case for maximised P-values because the tail set whose exact probability is calculate changes with  $\delta$ .

Figure 2 shows the dependence of various LR absed P-values on  $\delta$  for the specific data set  $(y_0, n_0) = (34, 70)$  and  $(y_1, n_1) = (21, 30)$ . Each plot shows the approximate

P-value (as a line) and the maximised P-value (as points). For the left plot, the approximate P-value is standard normal approximation  $\Psi(-t(-\delta))$  which is a continuous function of  $\delta$ . The maximised P-values are much larger but also vary quite erratically with  $\delta$ . While it is not typical in practice to calculate P-values for a range of non-inferiority margins, the erratic dependence is conceptually worrying. More practically, it means that intervals obtained from inverting such P-values may be discontinuous. The right plot is for the estimated P-value, which is in principle also a discontinuous function of  $\delta$ . However, we have found that usually the discontinuities are so small as to be undetectable as is the case here. The maximised version is more clearly discontinuous, but much less so than for the maximised P-value in the left plot. It is also close to the estimated P-value.

This example is not special and it seems to generally be the case that estimated P-values are close to continuous functions of  $\delta$  (though formally discontinuous) while the maximised versions are more discontinuous but often not practically different to the estimated P-values.

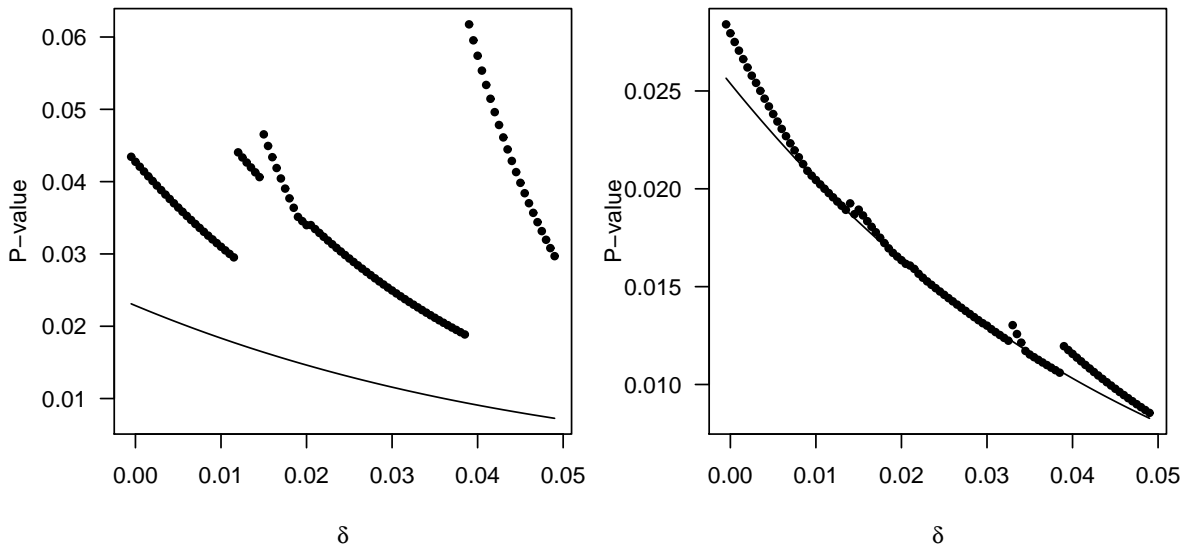


Figure 2: **Dependence of LR based P-values on  $\delta$  for  $(y_0, y_1) = (34, 21)$ .** *Left panel.* Approximate P-value  $P(y; \delta)$  (line) and maximised P-value  $P^*(y; \delta)$ (points). *Right panel.* Estimated P-value  $\hat{P}(y; \delta)$  (line) and estimated then maximised P-value  $\hat{P}^*(y; \delta)$ .points)

## 7 Conclusions

It has been found that the bootstrap adjustment to the basic P-values  $r(\delta)$  and  $t(\delta)$  produce tests whose size is much closer to nominal and exceed nominal by a very small margin. The maximum size bias of standard test statistics does not seem to decrease with sample size, whereas the maximum size bias of the bootstrap based P-values is extremely small for small or large sample sizes. For practical purposes, one can take the bootstrap P-values based tests to be exact, even for the smallest sample sizes considered.

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.

The power of the exact tests based on the bootstrap P-values is more powerful than the power of the exact tests based on any of the other basic test statistics, including the second order P-values. Combined with the small size bias of bootstrap P-values, this suggests that bootstrap P-values hardly exceed nominal size and have enhanced power compared to alternative, and are recommended for general use for testing a difference if probabilities.

Computation of the bootstrap P-value required computing the basic test statistic for all possible samples which becomes difficult for sufficient large sample sizes. An R-function is supplied by the author which takes ? seconds for samples sizes (200, 200). For larger sample sizes, a simple simulation algorithm that uses importance sampling is under development.

## References.

- Barndorff-Nielsen, O.E. & Cox, D.R. (1994) *Inference and Asymptotics* London: Chapman and Hall.
- Berger, R.L. and Boos, D.D. (1994). P values maximised over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89**, 1012-1016.
- Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics*. Holden-Day, Oakland.
- Brazzale, A., Davison, A.C. and Reid, N. (2007) *Applied Asymptotics: Case Studies in Small-sample Statistics*. Cambridge University Press, New York.
- Chan, I.S.F. (1998) Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Stat. Med.* **17**, 1403-1413.
- DiCiccio, T.J. and Young, G.A. (2008) Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika* 1–12.
- Dunnnett, C.W. and Gent, M. (1977) Significance testing to establish equivalence between treatments with special reference to data in the form of  $2 \times 2$  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. *Stat. Prob. Letters* **71**, 143–153.
- Lloyd C.J. and Moldovan, M.V. (2008) More powerful exact test of noninferiority from binary matched pairs data. *Statistics in Medicine* **27**, to appear.
- Lloyd, C.J. (2008a) Exact P-values for discrete models obtained by estimation and maximisation. *Austral. and New Zealand J. Statist.* **50**, 329-346.
- Lloyd C.J. (2008b) Exact tests of non-inferiority from independent binomial data based on second order test statistics. MBS working paper.
- Lloyd C.J. (2009a) Estimated P-values in Discrete Models: Asymptotic and non-asymptotic effects. Submitted to *Biometrika*. MBS Working paper.
- Liu J-P, Hsueh H-M, Hsieh E, Chen J.J. (2002) Test for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* **21**, 231–245.
- Lugannani, R. & Rice, S. (1980) Saddlepoint approximation for the distribution of the sum of independent variables. *Adv. Appl. Probab.* **12** 475-490.
- Munk, A., Skipka, G. and Stratmann, B. (2005) testing general hypotheses under binomial sampling; the two sample case – asymptotic theory and exact procedures. *Comp. Statist. data Analysis* **49**, 723-739.

- Pierce D.A. and Peters D., 1992. Practical use of higher order asymptotics for multiparameter exponential families (with discussion) *J. Roy. Statist. Soc B* **54**, 701-737.
- Reid, N. (2003) Asymptotics and the theory of inference. *Ann. Statist.* **31**, 1695-1731.
- Rohmel, J. (2005) Problems with existing procedures to calculate exact unconditional P-values for non-inferiority/superiority and confidence intervals for two binomials and who to resolve them. *Biometrical Journal* **47**, 37-47.
- Rohmel, 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 J.* **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 J. Med* **290**, 251-256.

## Appendix A: First order statistics.

Denote the interest parameter by  $\delta = p_1 - p_0$ , the null value  $\delta_0$  and the remaining nuisance parameter by  $\lambda = p_0$ . The log-likelihood is

$$\ell(\delta, \lambda; y_0, y_1) = \log(\lambda)y_0 + \log(\delta + \lambda)y_1 + (n_0 - y_0) \log(1 - \lambda) + (n_1 - y_1) \log(\delta + \lambda) \quad (6)$$

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

$$\frac{\partial \ell}{\partial \lambda} = \frac{y_0}{\lambda} + \frac{y_1}{\delta + \lambda} - \frac{n_0 - y_0}{1 - \lambda} + \frac{n_1 - y_1}{\delta + \lambda}$$

which re-arranges into a cubic with leading coefficient 1 and remaining coefficients

$$c_2 = \delta(1 + f_0) - 1 - \hat{p}, c_1 = \hat{p} - 2\delta f_0 \hat{p}_0 - \delta + \delta^2 f_0, c_0 = f_0 \hat{p}_0 \delta(1 - \delta)$$

where  $f_0 = n_0/n$ ,  $\hat{p}_0 = y_0/n_0$  and  $\hat{p} = t/n$  is the estimate under  $\delta = 0$ . When  $\delta = 0$ , this reduces to solving a quadratic since  $c_0 = 0$  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  $\Lambda_\delta$ . The cubic based estimator was first noted by Farrington & Manning (1990). Dunnett and Gent (1977) used the simple unbiased estimators

$$\tilde{p}_{0\delta} = \frac{y_0 + y_1}{n_0 + n_1} - \delta \frac{n_1}{n_0 + n_1}, \tilde{p}_{1\delta} = \tilde{p}_{0\delta} + \delta$$

however  $\tilde{p}_{0\delta}$  may violate the interval  $\Lambda_\delta$ .

## 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 (Bardorff-Nielsen & Cox, 1994) and the cumulative tail probability of this distribution is further approximated (Lugannini & 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, Davison and Reid (2007). Suppose that the log-likelihood is of exponential family form

$$\ell(\theta; y) = \varphi(\theta)^T v(y) - c(\varphi(\theta)) \quad (7)$$

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  $d = 2$ . The general formula requires derivation of an adjustment statistic,  $q(\psi)$ , that depends on two quantities. The first is the observed information matrix in the  $(\psi, \lambda)$  parametrisation which we denote  $j_\theta$ . The second is the  $d \times d$  Jacobian matrix  $\varphi_\theta$  of the transform from  $\varphi$  to  $\theta$ . Then the adjustment statistic is

$$q(\psi) = \frac{|\hat{\varphi} - \hat{\varphi}_\psi \quad \varphi_\lambda(\hat{\theta}_\psi)|}{|\varphi_\theta(\hat{\theta})|} \times \frac{|j(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} \quad (8)$$

The notation  $\hat{\varphi} - \hat{\varphi}_\psi$  in the first matrix is shorthand for the column vector giving the deviation of the canonical parameter  $\varphi$  when estimated under the general and null models. When  $\psi$  is a component of the canonical statistic  $\varphi$  then the first term reduces to  $\hat{\psi} - \psi$ . The extent to which  $q(\psi)$  and likelihood root  $r(\psi)$  differ is a measure of the extent to which the first order approximation to the log-likelihood is inadequate. Second order inference is instead based on the modified likelihood root

$$r^*(\psi) = r(\psi) + r(\psi)^{-1} \log\{q(\psi)/r(\psi)\} \quad (9)$$

and its approximate standard normal distribution. The approximate one-sided P-value is  $p^*(\psi) = \Phi(-r^*(\psi))$ . Note that  $r^*(\psi)$  breaks down in the centre of the distribution where  $\hat{\theta}$  and  $\hat{\theta}_\psi$  are close so that  $r(\psi) \approx 0$ . This is of little practical interest however since there is, in this case, no statistical evidence against the value  $\psi$ .

We now derive an explicit expression for  $q(\psi)$  for the binomial model. For future work, and it turns out also for simplicity, we generalise to testing  $\psi = h(p_1) - h(p_0)$  for

a general link function  $h$  and take the nuisance parameter as  $\lambda = h(p_0)$ . The inverse transformation is

$$p_1 = h^{-1}(\psi + \lambda), p_0 = h^{-1}(\lambda).$$

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

$$\begin{aligned} \ell(\psi, \lambda; y_0, y_1) &= y_0 \varphi\{h^{-1}(\lambda)\} + y_1 \varphi\{h^{-1}(\psi + \lambda)\} \\ &+ n_0 \log(1 - h^{-1}(\lambda)) + n_1 \log(1 - h^{-1}(\psi + \lambda)) \end{aligned} \quad (10)$$

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

$$\begin{aligned} U_\psi(\psi, \lambda) &:= \frac{\partial\ell}{\partial\psi} = \frac{y_1 - n_1 p_1}{p_1(1-p_1)h'(p_1)} \\ U_\lambda(\psi, \lambda) &:= \frac{\partial\ell}{\partial\lambda} = \frac{y_0 - n_1 p_0}{p_0(1-p_0)h'(p_0)} + \frac{y_1 - n_1 p_1}{p_1(1-p_1)h'(p_1)} \end{aligned}$$

and we will denote henceforth  $w(p) = p(1-p)h'(p)$ . The restricted MLE  $\hat{p}_{0\psi}$  satisfies the equation

$$0 = \frac{y_0 - n_1 p_0}{w(p_0)} + \frac{y_1 - n_1 p_1}{w(p_1)} \Leftrightarrow 0 = (y_0 - n_1 p_0)w(p_1) + (y_1 - n_1 p_1)w(p_0)$$

where  $p_1 = h^{-1}(\psi + h(p_0))$ . 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, Skipka and Stratmann (2005) give some theory on restricted ML estimation under this formulation.

Taking the covariance of the score functions gives the information terms

$$j_{\psi\psi} = V_1/w_1^2, j_{\psi\lambda} = V_1/w_1^2, j_{\lambda\lambda} = V_1/w_1^2 + V_0/w_0^2$$

where for  $j = 0, 1$ , we denote  $V_j = n_j p_j (1 - p_j)$  and  $w_j = w(p_j)$ . Hence the non-orthogonality term is

$$\frac{|j(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|} = \frac{\hat{V}_1 \hat{V}_0 / (\hat{w}_1^2 \hat{w}_0^2)}{\hat{V}_{0\psi} / \hat{w}_{0\psi}^2 + \hat{V}_{1\psi} / \hat{w}_{1\psi}^2} \quad (11)$$

It remains to find the first term in (8) which depends on the transformation

$$(\psi, \lambda) \rightarrow (\varphi_1, \varphi_0) = (\beta(\psi + \lambda), \beta(\lambda)).$$

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

$$\varphi_{\theta}(\theta) = \begin{bmatrix} \beta'(\psi + \lambda) & \beta'(\psi + \lambda) \\ 0 & \beta'(\lambda) \end{bmatrix} = \begin{bmatrix} w_1^{-1} & w_1^{-1} \\ 0 & w_0^{-1} \end{bmatrix}$$

with determinant  $1/(w_1 w_0)$  and so the first term in (8) is

$$|\varphi_{\theta}(\hat{\theta})|^{-1} |\hat{\varphi} - \hat{\varphi}_{\psi} \quad \varphi_{\lambda}(\hat{\theta}_{\psi})| = \hat{w}_1 \hat{w}_0 \{ \hat{w}_{0\psi}^{-1} (\hat{\varphi}_1 - \hat{\varphi}_{1\psi}) - \hat{w}_{1\psi}^{-1} (\hat{\varphi}_0 - \hat{\varphi}_{0\psi}) \}$$

Multiplying this by the square root of (11) gives

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

and a slightly cleaner form is

$$q(\psi) = \frac{\{ \hat{w}_{1\psi} (\hat{\varphi}_1 - \hat{\varphi}_{1\psi}) - \hat{w}_{0\psi} (\hat{\varphi}_0 - \hat{\varphi}_{0\psi}) \} \sqrt{\hat{V}_1 \hat{V}_0}}{\sqrt{\hat{V}_{0\psi} \hat{w}_{1\psi} + \hat{V}_{1\psi} \hat{w}_{0\psi}^2}} \quad (12)$$

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

Table 1: **Size bias of 5 P-values and their bootstrap versions.** Each figure summarises the average size bias for  $\alpha = 0.01, 0.05, 0.1$  and  $\delta_0 = -0.1, 0$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (4, 4)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	20.4	29.2	16.4	21.6	24.5	19.5	23.3	19.5	26.9	21.0
6	14.2	18.3	13.1	16.7	20.0	16.8	19.8	17.4	18.7	18.4
8	13.2	17.9	11.5	15.0	17.9	13.6	14.2	13.9	14.2	14.3
10	10.7	14.3	10.0	11.3	15.6	12.0	13.0	12.4	12.6	13.3
12	9.6	11.9	9.8	11.2	14.6	9.8	10.0	9.6	9.7	10.4
14	9.2	11.6	9.6	11.3	13.4	8.5	8.5	8.7	8.5	9.4
16	8.2	9.9	9.1	10.3	14.7	7.3	7.9	7.5	7.6	8.0
18	7.8	9.5	7.6	9.3	10.2	6.6	6.2	6.4	6.2	7.0
20	7.3	8.4	6.9	7.9	10.6	6.9	7.0	7.2	7.0	7.5
4	31.2	102.1	8.5	100.4	120.1	3.2	-0.6	3.2	-4.5	2.9
6	12.8	61.6	16.1	61.6	122.6	0.6	-2.1	0.0	-2.1	2.4
8	20.0	89.2	17.3	86.6	126.6	-1.7	-1.7	-1.3	-1.7	1.6
10	10.7	60.1	18.3	57.3	126.0	-0.8	-2.1	-1.2	-0.8	2.2
12	13.8	53.7	20.7	52.8	114.1	2.1	0.6	0.4	0.8	2.7
14	11.8	60.9	20.6	63.1	103.6	1.0	1.0	1.0	1.0	4.4
16	14.5	51.6	25.0	55.2	158.4	2.5	2.2	1.6	2.2	1.6
18	17.1	65.7	22.0	67.8	85.9	2.4	2.5	1.8	2.5	1.8
20	12.1	47.3	19.3	49.7	126.6	2.0	2.0	1.9	2.0	1.9

Table 2: **Size bias of 5 P-values and their bootstrap versions.** Each figure summarises the average size bias for  $\alpha = 0.01, 0.05, 0.1$  and  $\delta_0 = -0.1, 0$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (5, 3)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
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

Table 3: **Size bias of 5 P-values and their bootstrap versions.** Each figure summarises the average size bias for  $\alpha = 0.01, 0.05, 0.1$  and  $\delta_0 = -0.1, 0$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (6, 2)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	32.2	39.9	39.9	39.9	39.9	31.2	29.8	29.8	29.8	29.8
6	28.8	28.4	28.4	28.4	28.4	21.9	21.8	21.8	21.8	21.8
8	23.3	23.6	23.6	23.6	23.6	16.3	16.5	16.5	16.5	16.5
10	23.3	19.8	19.8	19.8	19.8	13.8	14.1	14.1	14.1	14.1
12	20.3	17.2	17.2	17.2	17.2	11.5	11.5	11.5	11.5	11.5
14	19.5	16.4	16.4	16.4	16.4	11.1	11.4	11.4	11.4	11.4
16	17.6	12.4	12.4	12.4	12.4	9.3	9.2	9.2	9.2	9.2
18	15.8	11.3	11.3	11.3	11.3	8.7	8.6	8.6	8.6	8.6
20	15.6	10.4	10.4	10.4	10.4	8.2	8.2	8.2	8.2	8.2
4	65.1	158.7	158.7	158.7	158.7	-2.0	-0.8	-0.8	-0.8	-0.8
6	73.9	130.6	130.6	130.6	130.6	1.1	1.1	1.1	1.1	1.1
8	80.8	138.9	138.9	138.9	138.9	2.8	3.1	3.1	3.1	3.1
10	73.7	129.3	129.3	129.3	129.3	3.4	3.2	3.2	3.2	3.2
12	70.8	120.3	120.3	120.3	120.3	2.2	1.6	1.6	1.6	1.6
14	71.8	158.3	158.3	158.3	158.3	2.2	1.4	1.4	1.4	1.4
16	78.4	112.3	112.3	112.3	112.3	2.9	2.5	2.5	2.5	2.5
18	69.8	102.6	102.6	102.6	102.6	2.3	2.2	2.2	2.2	2.2
20	69.8	78.2	78.2	78.2	78.2	4.0	2.1	2.1	2.1	2.1

Table 4: **Size bias of 5 P-values and their bootstrap versions.**  $\alpha = 0.01$  Each figure summarises the average size bias for  $\delta_0 = -0.1, 0$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (5, 3)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	38.6	39.0	28.0	32.0	31.2	35.2	35.2	35.2	37.8	35.2
6	32.2	38.8	22.7	30.5	29.6	20.9	24.6	20.8	24.6	20.8
8	27.6	25.2	16.1	19.8	25.0	17.3	20.3	20.2	21.2	22.2
10	22.4	17.8	13.2	16.3	16.3	15.5	15.5	15.5	15.5	18.2
12	19.2	17.0	13.4	14.2	18.1	14.8	14.2	13.8	13.3	13.8
14	18.4	14.2	11.6	11.4	16.4	12.2	10.6	10.6	10.6	12.8
16	16.9	12.9	9.9	10.2	14.8	11.3	11.5	11.6	11.0	12.0
18	15.4	12.4	9.4	10.0	12.8	9.2	9.1	9.1	9.1	10.1
20	16.3	12.6	9.4	8.7	15.0	8.6	8.6	8.6	8.6	9.6
4	67.3	93.6	11.1	89.9	11.1	-4.2	-4.2	-4.2	-4.2	-4.2
6	35.2	198.4	37.0	194.5	173.4	0.5	-3.2	-0.3	-3.2	-0.3
8	35.8	122.8	20.7	114.7	362.0	9.2	5.8	3.0	-1.0	-1.0
10	17.2	202.3	18.8	202.3	122.8	0.2	0.2	0.2	0.2	-2.4
12	19.0	105.1	23.0	105.1	221.6	-0.1	4.3	0.1	4.9	0.1
14	18.3	72.1	38.3	69.8	308.9	1.6	5.1	5.1	5.1	0.3
16	15.9	75.3	35.3	70.9	154.4	4.0	6.5	-0.1	6.5	-0.1
18	18.0	115.6	18.2	112.4	189.8	3.8	1.6	2.0	2.0	2.0
20	20.2	69.4	22.3	61.0	244.9	5.4	2.8	2.8	2.8	0.2

Table 5: **Size bias of 5 P-values and their bootstrap versions.**  $\alpha = 0.05$  Each figure summarises the average size bias for  $\delta_0 = -0.1, 0$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (5, 3)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	20.4	25.6	17.2	22.8	19.7	21.2	20.4	19.2	20.4	19.2
6	14.6	19.5	10.9	15.4	19.8	14.7	12.2	15.7	12.2	17.2
8	11.6	14.6	14.0	12.4	15.8	15.1	16.3	15.1	14.4	16.3
10	10.2	12.4	10.2	8.8	15.2	11.2	11.1	11.1	9.6	12.6
12	8.4	12.5	9.2	8.9	13.4	10.0	10.8	10.3	9.1	12.2
14	8.6	10.0	8.6	9.3	12.8	9.4	9.4	9.2	8.8	9.8
16	7.6	9.6	8.8	9.2	12.2	8.1	7.8	8.1	7.8	8.8
18	7.0	7.2	5.7	6.3	9.9	6.9	6.6	7.4	7.2	8.4
20	6.5	5.8	6.3	6.2	9.4	6.8	7.0	6.8	6.6	7.7
4	34.6	118.2	19.9	118.2	72.9	2.2	2.7	2.2	2.7	2.2
6	27.5	149.9	25.2	147.7	138.8	3.2	5.7	3.4	5.7	3.4
8	12.1	82.4	11.5	80.0	176.8	0.6	0.4	0.6	3.8	0.4
10	10.4	83.4	9.6	55.8	204.6	0.2	0.0	0.0	4.1	0.0
12	14.0	115.6	15.2	113.0	226.2	3.9	0.4	0.4	5.4	0.4
14	22.6	70.5	21.3	70.4	149.8	1.3	4.8	0.1	4.8	0.1
16	8.8	63.0	31.4	77.2	132.2	-0.2	5.6	-0.2	5.6	-0.2
18	8.6	83.6	13.4	83.6	154.4	1.1	5.9	0.7	5.9	0.6
20	13.3	54.2	11.9	54.2	161.8	3.2	5.9	1.4	5.9	1.4

Table 6: **Size bias of 5 P-values and their bootstrap versions.**  $\alpha = 0.1$  Each figure summarises the average size bias for  $\delta_0 = -0.1, 0$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (5, 3)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	14.6	21.4	14.8	19.0	23.6	17.0	17.2	14.5	17.0	19.0
6	10.9	15.8	14.3	17.3	24.8	13.6	14.6	12.8	13.6	13.6
8	8.4	10.6	8.4	9.8	15.0	9.4	9.4	9.0	9.4	11.6
10	6.8	8.4	6.8	8.4	12.0	8.9	9.6	8.9	8.9	11.3
12	7.0	7.8	6.6	7.5	12.1	7.7	8.2	6.2	6.8	7.4
14	6.1	8.2	5.9	7.5	9.0	5.9	5.9	5.9	5.9	6.8
16	5.8	6.2	6.4	7.2	9.8	5.6	6.0	5.6	5.6	6.0
18	4.8	5.3	5.4	6.2	8.1	5.3	5.4	4.8	5.1	5.6
20	4.6	5.4	5.2	6.1	8.2	4.9	4.8	5.2	5.2	5.6
4	54.6	131.0	16.2	131.0	149.2	2.6	3.1	2.6	2.6	2.6
6	62.8	82.8	29.8	82.8	180.6	1.0	-0.4	6.6	1.0	6.6
8	42.4	57.2	23.2	57.2	168.4	0.8	0.8	6.8	0.8	6.2
10	42.9	82.6	14.0	82.6	168.5	1.5	1.5	1.5	1.5	7.8
12	52.7	60.8	14.2	60.8	174.6	1.4	1.4	2.1	2.1	2.1
14	42.1	69.7	20.8	69.7	96.7	1.4	1.4	1.4	1.4	1.4
16	52.4	55.9	26.6	59.0	103.4	2.0	2.8	2.8	2.8	2.8
18	41.4	52.0	22.1	56.0	109.2	2.2	3.5	6.4	3.5	3.5
20	56.0	60.3	19.2	60.3	114.3	2.7	3.8	3.8	3.8	3.8

Table 7: **Size bias of 5 P-values and their bootstrap versions.**  $\delta = -0.1$  Each figure summarises the average size bias for  $\alpha = 0.01, 0.05, 0.1$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (5, 3)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	26.3	38.2	38.2	38.2	38.2	25.5	25.4	25.4	25.4	25.4
6	22.3	24.8	24.8	24.8	24.8	16.7	17.4	17.4	17.4	17.4
8	17.7	19.0	19.0	19.0	19.0	9.8	11.4	11.4	11.4	11.4
10	17.8	17.0	17.0	17.0	17.0	9.8	10.3	10.3	10.3	10.3
12	14.6	14.0	14.0	14.0	14.0	8.3	8.7	8.7	8.7	8.7
14	15.4	12.9	12.9	12.9	12.9	7.8	8.2	8.2	8.2	8.2
16	13.8	8.7	8.7	8.7	8.7	6.6	6.4	6.4	6.4	6.4
18	11.0	7.3	7.3	7.3	7.3	5.5	5.6	5.6	5.6	5.6
20	11.4	6.7	6.7	6.7	6.7	5.5	6.0	6.0	6.0	6.0
4	27.5	192.1	192.1	192.1	192.1	0.9	-1.7	-1.7	-1.7	-1.7
6	38.0	153.8	153.8	153.8	153.8	2.2	2.2	2.2	2.2	2.2
8	49.2	161.3	161.3	161.3	161.3	4.9	4.1	4.1	4.1	4.1
10	31.5	141.7	141.7	141.7	141.7	5.2	5.2	5.2	5.2	5.2
12	26.2	119.9	119.9	119.9	119.9	1.0	0.3	0.3	0.3	0.3
14	28.7	191.9	191.9	191.9	191.9	-0.3	-0.4	-0.4	-0.4	-0.4
16	40.4	115.3	115.3	115.3	115.3	0.3	0.5	0.5	0.5	0.5
18	21.8	97.8	97.8	97.8	97.8	-0.1	-0.3	-0.3	-0.3	-0.3
20	19.1	54.2	54.2	54.2	54.2	2.0	0.4	0.4	0.4	0.4

Table 8: **Size bias of 5 P-values and their bootstrap versions.**  $\delta = 0$  Each figure summarises the average size bias for  $\alpha = 0.01, 0.05, 0.1$  (expressed as percentage). The upper section describes average bias  $\bar{e}$  and the lower section maximum bias  $e^*$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (5, 3)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	38.1	41.6	41.6	41.6	41.6	36.8	34.2	34.2	34.2	34.2
6	35.2	31.9	31.9	31.9	31.9	27.1	26.2	26.2	26.2	26.2
8	29.0	28.2	28.2	28.2	28.2	22.9	21.5	21.5	21.5	21.5
10	28.8	22.6	22.6	22.6	22.6	17.7	17.9	17.9	17.9	17.9
12	26.1	20.3	20.3	20.3	20.3	14.7	14.3	14.3	14.3	14.3
14	23.6	19.9	19.9	19.9	19.9	14.3	14.5	14.5	14.5	14.5
16	21.4	16.0	16.0	16.0	16.0	12.0	12.0	12.0	12.0	12.0
18	20.6	15.3	15.3	15.3	15.3	11.9	11.6	11.6	11.6	11.6
20	19.7	14.1	14.1	14.1	14.1	10.8	10.3	10.3	10.3	10.3
4	102.8	125.3	125.3	125.3	125.3	-4.9	0.1	0.1	0.1	0.1
6	109.8	107.3	107.3	107.3	107.3	0.0	0.0	0.0	0.0	0.0
8	112.5	116.4	116.4	116.4	116.4	0.7	2.1	2.1	2.1	2.1
10	116.0	116.8	116.8	116.8	116.8	1.6	1.1	1.1	1.1	1.1
12	115.5	120.7	120.7	120.7	120.7	3.5	3.0	3.0	3.0	3.0
14	114.9	124.7	124.7	124.7	124.7	4.6	3.1	3.1	3.1	3.1
16	116.4	109.2	109.2	109.2	109.2	5.5	4.6	4.6	4.6	4.6
18	117.8	107.4	107.4	107.4	107.4	4.7	4.8	4.8	4.8	4.8
20	120.4	102.2	102.2	102.2	102.2	6.0	3.8	3.8	3.8	3.8

Table 9: **Powers of exact tests.** Each figure summarises the average power (expressed as percentage) for  $\alpha = 0.01, 0.05, 0.1$  at the local alternative  $\delta_1 = \delta_0 + 1/\sqrt{(n_0 n_1)^{1/2}}$ . The upper section describes the null  $\delta_0 = -0.1$ , the lower section  $\delta_0 = 0$ . The multiplier  $m$  is applied to starting sample size  $(n_0, n_1) = (5, 3)$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
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

Table 10: **Powers of size  $\alpha = 0.01$  exact tests.** Each figure summarises the average power (expressed as percentage) over three samples sizes configurations (4, 4), (5, 3) and (6, 2) at the local alternative  $\delta_1 = \delta_0 + 1/\sqrt{(n_0 n_1)^{1/2}}$ . The upper section describes the null  $\delta_0 = -0.1$ , the lower section  $\delta_0 = 0$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	16.1	18.5	18.5	18.5	18.5	19.4	20.1	20.1	20.1	20.1
6	18.9	18.4	18.4	18.4	18.4	22.6	22.2	22.2	22.2	22.2
8	21.3	19.5	19.5	19.5	19.5	23.7	23.3	23.3	23.3	23.3
10	22.9	16.1	16.1	16.1	16.1	24.6	24.1	24.1	24.1	24.1
12	23.1	21.4	21.4	21.4	21.4	25.3	25.4	25.4	25.4	25.4
14	24.1	21.4	21.4	21.4	21.4	26.6	26.5	26.5	26.5	26.5
16	23.6	22.4	22.4	22.4	22.4	27.3	27.3	27.3	27.3	27.3
18	24.6	22.0	22.0	22.0	22.0	27.5	27.5	27.5	27.5	27.5
20	25.0	19.4	19.4	19.4	19.4	28.0	28.2	28.2	28.2	28.2
4	15.2	17.9	17.9	17.9	17.9	19.1	19.1	19.1	19.1	19.1
6	19.4	20.1	20.1	20.1	20.1	21.8	21.5	21.5	21.5	21.5
8	20.3	20.6	20.6	20.6	20.6	22.7	22.6	22.6	22.6	22.6
10	20.9	20.5	20.5	20.5	20.5	24.0	23.9	23.9	23.9	23.9
12	22.1	21.1	21.1	21.1	21.1	25.2	25.2	25.2	25.2	25.2
14	22.9	20.2	20.2	20.2	20.2	25.3	24.9	24.9	24.9	24.9
16	22.8	21.1	21.1	21.1	21.1	25.9	25.7	25.7	25.7	25.7
18	23.5	21.1	21.1	21.1	21.1	26.4	26.1	26.1	26.1	26.1
20	23.9	21.5	21.5	21.5	21.5	26.5	26.5	26.5	26.5	26.5

Table 11: **Powers of size  $\alpha = 0.05$  exact tests.** Each figure summarises the average power (expressed as percentage) over three samples sizes configurations (4, 4), (5, 3) and (6, 2) at the local alternative  $\delta_1 = \delta_0 + 1/\sqrt{(n_0 n_1)^{1/2}}$ . The upper section describes the null  $\delta_0 = -0.1$ , the lower section  $\delta_0 = 0$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	42.9	36.1	36.1	36.1	36.1	44.7	42.6	42.6	42.6	42.6
6	43.4	33.2	33.2	33.2	33.2	45.7	45.7	45.7	45.7	45.7
8	45.6	38.0	38.0	38.0	38.0	47.9	47.9	47.9	47.9	47.9
10	46.5	38.2	38.2	38.2	38.2	49.2	48.9	48.9	48.9	48.9
12	47.0	38.7	38.7	38.7	38.7	49.8	49.8	49.8	49.8	49.8
14	47.3	37.1	37.1	37.1	37.1	50.4	50.3	50.3	50.3	50.3
16	47.8	46.1	46.1	46.1	46.1	50.6	50.5	50.5	50.5	50.5
18	48.2	44.0	44.0	44.0	44.0	51.2	51.1	51.1	51.1	51.1
20	48.7	45.0	45.0	45.0	45.0	51.4	51.1	51.1	51.1	51.1
4	40.5	38.8	38.8	38.8	38.8	43.3	43.3	43.3	43.3	43.3
6	43.0	40.0	40.0	40.0	40.0	45.5	45.7	45.7	45.7	45.7
8	44.1	41.4	41.4	41.4	41.4	47.4	47.6	47.6	47.6	47.6
10	46.0	42.4	42.4	42.4	42.4	47.9	48.0	48.0	48.0	48.0
12	46.6	41.8	41.8	41.8	41.8	48.7	48.7	48.7	48.7	48.7
14	46.0	41.3	41.3	41.3	41.3	49.7	49.5	49.5	49.5	49.5
16	46.3	42.0	42.0	42.0	42.0	49.9	49.8	49.8	49.8	49.8
18	46.4	42.1	42.1	42.1	42.1	50.1	49.9	49.9	49.9	49.9
20	46.8	42.5	42.5	42.5	42.5	50.9	50.5	50.5	50.5	50.5

Table 12: **Powers of size  $\alpha = 0.10$  exact tests.** Each figure summarises the average power (expressed as percentage) over three samples sizes configurations (4, 4), (5, 3) and (6, 2) at the local alternative  $\delta_1 = \delta_0 + 1/\sqrt{(n_0 n_1)^{1/2}}$ . The upper section describes the null  $\delta_0 = -0.1$ , the lower section  $\delta_0 = 0$ .

$m$	Basic P-value					Bootstrap P-value				
	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$	$t$	$r$	$r_{BT}^*$	$r_{LR}^*$	$r_{1/2}^*$
4	58.8	49.4	49.4	49.4	49.4	58.8	58.6	58.6	58.6	58.6
6	59.3	52.2	52.2	52.2	52.2	60.4	60.2	60.2	60.2	60.2
8	61.5	50.6	50.6	50.6	50.6	61.7	61.6	61.6	61.6	61.6
10	60.4	50.5	50.5	50.5	50.5	62.6	62.4	62.4	62.4	62.4
12	62.5	55.9	55.9	55.9	55.9	62.9	62.8	62.8	62.8	62.8
14	62.8	57.0	57.0	57.0	57.0	63.0	62.9	62.9	62.9	62.9
16	62.3	54.4	54.4	54.4	54.4	63.4	63.3	63.3	63.3	63.3
18	62.1	54.9	54.9	54.9	54.9	63.7	63.6	63.6	63.6	63.6
20	61.8	57.8	57.8	57.8	57.8	63.8	63.6	63.6	63.6	63.6
4	52.5	50.3	50.3	50.3	50.3	59.1	58.1	58.1	58.1	58.1
6	54.5	51.1	51.1	51.1	51.1	58.7	58.5	58.5	58.5	58.5
8	55.4	49.3	49.3	49.3	49.3	59.8	59.5	59.5	59.5	59.5
10	55.4	49.6	49.6	49.6	49.6	61.1	61.0	61.0	61.0	61.0
12	56.4	49.3	49.3	49.3	49.3	61.8	61.4	61.4	61.4	61.4
14	56.6	50.2	50.2	50.2	50.2	62.0	62.0	62.0	62.0	62.0
16	56.9	50.4	50.4	50.4	50.4	62.5	62.5	62.5	62.5	62.5
18	57.1	50.6	50.6	50.6	50.6	63.0	63.0	63.0	63.0	63.0
20	55.8	50.8	50.8	50.8	50.8	63.3	63.2	63.2	63.2	63.2