

## Exact one-sided confidence limits for the difference between two correlated proportions

Chris J. Lloyd\* and Max V. Moldovan

*Melbourne Business School, Carlton, 3053, AUSTRALIA*

### SUMMARY

We construct exact and optimal one-sided upper and lower confidence bounds for the difference between two probabilities based on matched binary pairs using well-established optimality theory of Buehler (1957). Starting with five different approximate lower and upper limits, we adjust them to have coverage probability exactly equal to the desired nominal level and then compare the resulting exact limits by their mean size. Exact limits based on the signed root likelihood ratio statistic are preferred and recommended for practical use. Copyright © 2000 John Wiley & Sons, Ltd.

### 1. INTRODUCTION

Matched binary pairs arise in many statistical contexts. In intervention studies, a population is stratified with respect to some control variable and then one individual per stratum is randomly chosen to be given the treatment and one the control. In cross-over design clinical trials, each subject within a group is randomized to receive a sequence of two treatments thus serving as control for a treatment comparison. In twin studies, one twin is randomly chosen for treatment and the other for control. In repeated measure studies, binary traits are measured before and after some event or treatment of interest on the same individual. In each case, the basic measurement is a pair of possibly dependent binary variables  $Y = (Y_1, Y_2)$  on each stratum, twin pair or individual, taking values in  $\{00, 01, 10, 11\}$ . Outcomes 01 and 10 are called discordant, while outcomes 00 and 11 are called concordant. Let  $X_{jk}$  denote the number of responses ( $jk$ ) out of a total of  $n$  matched binary pairs and let  $T = X_{01} + X_{10}$  be the number of discordant pairs.

To illustrate our ideas, we describe results of a study reported by Kao et al. [1] and also used as a numerical example by Tang, Tang and Chan [2]. The initial objective of the study was to compare the diagnostic accuracy of two alternative procedures for detecting nasopharyngeal carcinomas (NPC). Method 1 is computed tomography (CT) and method 2 is technetium-99m methoxyisobutylisonitrile (Tc-MIBI) single photon emission computed tomography (SPECT).

---

\*Correspondence to: Chris Lloyd, Melbourne Business School, Carlton, 3053, AUSTRALIA

†c.lloyd@mbs.edu

Contract/grant sponsor: Australian Research Council Grant; contract/grant number: 03-1362

The results in Table I are for 25 patients whose true state was negative, each diagnosed by each of the two methods. A correct (i.e. negative) diagnosis is coded as 1 and incorrect as 0. Out of the 25 patients, 22 were correctly diagnosed by both tests and another patient was incorrectly classified by both tests, so the number of concordant observations is 23. Out of the  $t = 2$  discordant observations where the two methods disagreed, Tc-MIBI SPECT was correct on both occasions. Overall, the empirical estimates of the true negative fraction are  $24/25=96\%$  and  $22/25=88\%$  for Tc-MIBI SPECT and CT respectively.

Table I. Results of a comparison of false positive rates of CT and Tc-MIBI SPECT diagnoses of NPC, reported in Kao et al. [1].

	CT (-)	CT (+)	Total
Tc-MIBI SPECT(-)	22( $x_{11}$ )	2( $x_{01}$ )	24
Tc-MIBI SPECT(+)	0( $x_{10}$ )	1( $x_{00}$ )	1
Total	22	3	25( $n$ )

Letting  $p_1$  and  $p_2$  denote the probability of a correct diagnosis (i.e. the true negative fraction) from the two methods, there is interest in the difference  $p_2 - p_1$ , representing here the advantage of using Tc-MIBI SPECT over CT. The main aim of the present paper is to give one-sided confidence limits for the difference between two such correlated proportions, which have exact coverage properties (to the extent that discreteness allows it) and are as tight as possible. One-sided limits set bounds on the superiority or inferiority of one treatment over the other. For instance, to conclude non-inferiority of Tc-MIBI SPECT compared to CT, one calculates a  $100(1 - \alpha)\%$  lower confidence limit for  $p_2 - p_1$  and checks to see if it is greater than  $-\delta$ , a pre-specified non-inferiority margin. This is not the only use of a one-sided confidence limit which is an informative statistic in its own right.

In general, let  $\pi_{jk}$  be the probability that  $Y = (jk)$ , ( $j = 0, 1$ ;  $k = 0, 1$ ) with 1 denoting success and 0 failure. With this notation,  $\phi = \pi_{01} + \pi_{10}$  is the probability of a discordant pair. Let  $p_1 = \pi_{11} + \pi_{10}$  be the marginal success rate for treatment 1, and  $p_2 = \pi_{11} + \pi_{01}$  the success rate for treatment 2. While these probabilities may differ across strata, we are interested in unconditional probabilities, i.e. probabilities averaged across strata. The parameter of primary interest is

$$\theta = p_2 - p_1 = \pi_{01} - \pi_{10},$$

which measures the success rate for treatments 2 minus that for treatment 1.

There is a large and established literature on testing for a treatment effect, beginning with McNemar [3], who suggested the statistic  $(X_{01} - X_{10})/\sqrt{T}$ . This approximately standard normal test statistic depends only on the number of discordant pairs  $T$  and the number  $X_{01}$  of these whose response changes from 0 to 1. For instance, in the illustrative example McNemar's test looks exclusively at the  $t = 2$  individuals for whom the two methods gave conflicting results. Both of these were favorable to Tc-MIBI SPECT and the value of the test statistic is  $(2 - 0)/\sqrt{2} = 1.414$ . The 23 patients for whom the methods agreed play no part in this test, nor does the sample size  $n = 25$ , a fact some have found unintuitive. Tests which do depend on  $n$  have recently been developed that have higher unconditional power for fixed size, see Suissa and Shuster [4], Hsueh, Liu and Chen [5] and Berger and Sidik [6]. There is wide agreement

that tests should not depend on the relative numbers of concordant pairs  $X_{00}$  and  $X_{11}$ , though see Liang and Zeger [7] for a contrary view.

The subject of this paper, however, is not tests of  $\theta = 0$ , but confidence limits for  $\theta$ . Obviously, test statistics suitably generalized to test  $\theta = \theta_0$  can be used as a starting point for generating approximate confidence limits, and conversely confidence limits can be used to generate tests. Our focus is on confidence limits with exact statistical properties and maximum efficiency.

The data comprise four counts  $X_{ij}$  with joint multinomial distribution and parameters  $\pi_{jk}$  and  $n$ . This can be expressed in product binomial form as

$$\Pr(x_{11}, x_{01}, x_{10}, x_{00}) = B(t; n, \phi)B(x_{01}; t, \eta)B(x_{11}; n - t, \psi), \quad (1)$$

where  $t = x_{01} + x_{10}$  and  $B(x; n, p)$  is the probability of a binomial with parameters  $(n, p)$  being equal to  $x$ . The probability parameters in this expression are  $\psi = \pi_{11}/(\pi_{00} + \pi_{11})$ ,  $\phi = \pi_{01} + \pi_{10}$ , which is the probability of a discordant pair, and  $\eta = \pi_{01}/\phi$ , which is the probability of responding (01) conditional on a discordant response. Clearly, the last term in (1) has no relevance to inference on  $\theta$  and so denoting  $x_{01} = x$  and noting that  $\eta = 0.5(\theta + \phi)/\phi$  the likelihood becomes

$$L(\theta, \phi; x, t, n) \propto \phi^t (1 - \phi)^{n-t} \eta^x (1 - \eta)^{t-x} \propto (1 - \phi)^{n-t} (\theta + \phi)^x (\theta - \phi)^{t-x}.$$

Two-sided confidence intervals for  $\theta$  have been proposed and investigated by Armitage and Berry [8], Lloyd [9], Newcombe [10], Tango [11], [12] and Agresti and Min [13]. Newcombe compared ten methods, including some based on the exact binomial distribution, but with  $\psi$  replaced by its profile estimate, as well as a mid-p version. Five methods are explicitly described in the next section. All are approximate in their coverage, but we will make use of some of the basic constructions in generating our exact limits. We emphasize that the focus of this paper is *one-sided* confidence limits for  $\theta$ , not intervals. A one-sided upper limit is a statistic  $u(X, T, n)$  such that  $\Pr(\theta \leq u(X, T, n); \theta, \phi)$  is at least  $1 - \alpha$  for all parameter values  $(\theta, \phi)$ . It should be noted that by reversing the definition of success and failure, a lower limit for  $\theta$  is just  $-u(T - X, T, n)$ . Therefore, for the sake of clarity and brevity, we describe our theory and methods only for the problem of constructing upper limits. We also henceforth suppress dependence of a limit on  $n$ .

It is mentioned in the ICH E9 guideline Statistical Principles for Clinical Trials [14] that 'the issue of one-sided or two-sided approaches to inference is controversial and a diversity of views can be found in the statistical literature'. We consider one-sided limits of interest in their own right, as upper or lower statistical limits on an unknown parameter, here a difference in treatment success rates. We have already mentioned that one-sided confidence limits may be used to conclude about superiority and non-inferiority. In so-called bio-equivalence tests, the alternative to be detected is  $|\theta| < \delta$  and can be achieved by two one-sided tests known as the TOST procedure, see Berger and Hsu [15] for a review of this area. It is important to realize that combining upper and lower limits with respective coverage errors  $\alpha_L$  and  $\alpha_U$  gives a confidence interval with coverage error at most  $\alpha_L + \alpha_U$ , but typically such a confidence interval will be highly conservative. In other words, combining exact one-sided limits does not give an exact interval. The conservatism can be reduced by various computationally intensive adjustments, inevitably involving the choice of how much coverage error to place in each tail (see [16], [17], [18]). Shorter confidence intervals can typically be achieved by an unequal allocation of the coverage error, which can even lead to intervals with nearly all the error on one

side, logically undermining the notion of an interval. To put it plainly, there is no satisfactory general optimality theory for constructing exact and short confidence intervals, even after one has decided on a basic test statistic to generate them.

In contrast, the narrower problem of one-sided confidence limits can be addressed quite thoroughly using the theory first presented by Buehler [19] and then explicated and extended by Kabaila and Lloyd [20], [21] and Lloyd and Kabaila [22]. This theory allows any approximate one-sided upper (lower) limit to be adjusted so as to be as small (large) as possible while retaining the essential coverage property. These limits, known in the reliability literature as Buehler bounds, have identical exact coverage properties and so can be compared directly based on their average size, the smallest upper (largest lower) limit to be preferred. The only choice then is the approximate limit to be adjusted into an exact limit. We describe this theory in some detail in Section 3 for upper limits only. The theory for lower limits follows automatically since a lower limit for  $-\theta$  becomes an upper limit for  $\theta$  by a change of sign.

Returning to our example, Table II reports the values of various lower 97.5% confidence limits for  $\theta = p_2 - p_1$ , the point estimate of which was 0.08. The five top figures are all approximate lower limits computed using existing methods that are explained in Section 2. The five lower numbers are Buehler adjusted versions of these approximate limits that have exact coverage properties and also a maximality property. The method of adjustment and the optimality property are described in Section 3. The observed value for exact method 5, namely  $-0.056$ , would support non-inferiority for  $\delta = 0.06$  in contrast to the 'exact' results reported by Tang, Tang and Chan [2]. In this particular example method 5, which is based on the signed likelihood ratio statistic, leads to the largest lower confidence limit. While this is not the case for every data set, the result of our theoretical investigation is to recommend the exact likelihood ratio based method in general.

Table II. Approximate and exact lower confidence limits for data in Kao et al. [1] and Tang, Tang and Chan [2]:  $(x, t, n; \alpha) = (2, 2, 25; 0.025)$ .

Method ( $i$ )	1	2	3	4	5
Approximate ( $s_i$ )	-0.055	-0.224	-0.049	-0.064	-0.021
Exact ( $u_i$ )	-0.208	-0.161	-0.070	-0.082	-0.056

The plan of the paper is as follows. In the next section we describe five approximate upper limits that, in our view, broadly cover the range of available procedures suggested in the literature. In Section 3, we introduce the Buehler algorithm for generating exact confidence limits and identify some of the computational issues that arise with its practical application. In Section 4, we present a comparison of five Buehler upper limits across a range of sample sizes and parameter values. Section 5 gives further numerical examples. Finally, Section 6 summarizes the study and gives recommendations to practitioners.

## 2. APPROXIMATE UPPER LIMITS

In this section we describe five approaches to constructing approximate confidence limits for  $\theta$  that cover the broad range of methods suggested in the literature. Since we are interested in

one-sided upper limits, we specifically present the one-sided versions of these methods. We also show numerically that the coverage properties of these approximate methods are quite poor. It should be born in mind, though, that none of these upper limits is being recommended. Rather they will be used to generate exact upper limits through the theory given in the next section. While there are other candidates we could have added to this list, the exact limits we generate later depend only on how the approximate limits order the sample space, so it is not necessary to include each and every limit that has appeared in the literature.

The first two approximate upper limits to be considered are based on the relationship  $\theta = (2\eta - 1)\phi$  and the fact that the marginal distribution of  $T$  generates inferences on  $\phi$ , while the distribution of  $X$  given  $t$  generates inferences on  $\eta$ . Armitage and Berry [8] suggested constructing a confidence interval for  $\theta$  by combination of a standard binomial confidence interval  $(\eta_L, \eta_U)$ , which give the interval  $((2\eta_L - 1)\phi, (2\eta_U - 1)\phi)$  for known  $\phi$ , and then replacing  $\phi$  by its maximum likelihood (ML) estimate  $\hat{\phi} = t/n$ . The one-sided upper limit generated by this approach is

$$S_1(x, t) = (2\eta_U - 1)\hat{\phi}.$$

This is an ‘estimated’ upper limit in the sense that it is obtained by replacing  $\phi$  by its ML estimate  $\hat{\phi}$  in the exact upper limit  $(2\eta_U - 1)\phi$ . This method takes no account of the variability of  $\hat{\phi}$ . Noticing this defect, Lloyd [9] suggested combining an interval  $(\eta_L, \eta_U)$  for  $\eta$  with coverage error  $\beta$  with an interval  $(\phi_L, \phi_U)$  for  $\phi$  with coverage error  $\gamma$ , which generates a conservative interval for  $\theta$  with coverage error at most  $\alpha := 1 - (1 - \beta)(1 - \gamma)$ . For a one-sided upper limit, this theory leads to the ‘conservative’ upper limit

$$S_2(x, t) = \begin{cases} (2\eta_U - 1)\phi_U & \text{if } \eta_U > 1/2 \\ (2\eta_U - 1)\phi_L & \text{if } \eta_U \leq 1/2 \end{cases}.$$

Our third upper limit is based on the ML estimator  $\hat{\theta} = (2X - T)/n$  whose mean is exactly  $\theta$  and variance is  $(\phi - \theta^2)/n$ . A Wald type upper limit is obtained by adding the appropriate number of standard errors to the estimate. Agresti and Min [13] have argued that this statistic should be calculated after adding 1/2 to each of the four counts giving the upper limit

$$S_3(x, t) = \hat{\theta}^* + z_\alpha \sqrt{(\hat{\phi}^* - \hat{\theta}^{*2})/n},$$

where  $\hat{\theta}^* = (2x - t)/(n + 2)$ ,  $\hat{\phi}^* = (t + 1)/(n + 2)$  and  $z_\alpha$  is an upper quintile of the standard normal distribution. Lloyd [9] suggested the alternative correction of adding  $z_\alpha^2/2$  which is very close to 2 for a 95% interval.

The last two approximate limits to be studied both involve the profile ML estimator  $\hat{\phi}_\theta$  of  $\phi$  when  $\theta$  is assumed to be known, given by the larger solution of the quadratic equation

$$\phi^2 - \phi(\hat{\phi} - \hat{\theta}\theta) + \hat{\theta}\theta - (1 - \hat{\phi})\theta^2 = 0.$$

The fourth approximate upper limit  $S_4(x, t)$  is the unique solution greater than  $\hat{\theta}$  of

$$n(\hat{\theta} - \theta)^2 = z_\alpha^2(\hat{\phi}_\theta - \theta^2),$$

proposed by Tango [11] and shown to be equivalent to a score statistic. Newcombe, commenting in Tango [12], notes that these score based intervals give very similar results to the more complex intervals that he recommends in Newcombe [10]. The final upper limit  $S_5(x, t)$  is

based on the signed root likelihood ratio (SRLR) and differs from the Tango limit  $S_4(x, t)$  only in that the difference between  $\hat{\theta}$  and  $\theta$  is measured by the likelihood ratio rather than a quadratic form. This was first suggested by Newcombe [10], but not recommended by him on the basis of its coverage properties. This does not preclude us from using it to generate exact limits. Explicitly,  $S_5(x, t)$  is the unique solution larger than  $\hat{\theta}$  of

$$2(\ell(\hat{\theta}, \hat{\phi}) - \ell(\theta, \hat{\phi}_\theta)) = z_\alpha^2,$$

where the log-likelihood is given by

$$\ell(\theta, \phi) = (n - t) \log(1 - \phi) + x \log(\theta + \phi) + (t - x) \log(\phi - \theta).$$

We briefly summarize the performance of these five approximate methods. The first method  $S_1$  is not even asymptotically correct, as it fails to account for errors in estimation of  $\phi$ , which are of the same order as errors in estimation of  $\theta$ . Of the five methods, only  $S_2$  has guaranteed coverage probability that does not fall below the nominal level. However, in the two-sided case Lloyd [9] found this method to be very conservative, which is also true for the one-sided cases considered in this study. Methods 3, 4 and 5 are all asymptotically equivalent, but can differ considerably for small samples. Indeed,  $S_3$  can even give values greater than 1 and exhibit anomalous behavior at the extremes of the sample space, i.e. when  $x$  is close to  $t$  or  $t$  is close to  $n$ . Newcombe [10] investigated the performance of 10 different approximate two-sided confidence intervals including the Armitage-Berry and the profile likelihood whose one-sided versions correspond respectively to  $S_1$  and  $S_5$  considered in our study. He also investigated various continuity corrected versions of  $S_3$  and  $S_4$  as well as a variant of  $S_1$  based on mid- $p$  intervals for  $\eta$ . However, we restrict our attention to the presented five methods  $S_i, i = 1, \dots, 5$  as being a representative selection of the broad range of techniques available.

The coverage condition for an upper limit can be expressed as the requirement that

$$C_U(\theta) := \inf_{\phi} \Pr(\theta \leq u(X, T); \theta, \phi) \tag{2}$$

does not fall below  $1 - \alpha$  for all values of  $\theta$  and  $\phi$ . The left panel of Figure 1 plots  $C(\theta)$  against  $\theta \in (-1, 1)$  for the SRLR based approximate upper limit  $S_5(X, T)$  when  $n = 10$  and  $\alpha = 0.05$ . This plot is typical, and similar results hold for larger sample sizes, but the function is even more erratic. It is apparent, and already well known, that coverage can fall well below nominal for this and also for the asymptotically equivalent upper limits  $S_3$  and  $S_4$ . All coverage functions tend to look rather similar, having multiple discontinuities and typically increasing in between. The points of discontinuity are values of  $\theta$  equal to observable values of the upper limit. Moving to the right of such a point results in one element of the sample space leaving the confidence set  $\{\theta \leq u(X, T)\}$ . These points are marked in the left panel of Figure 1. Graphical comparison of alternative upper limits can be aided by joining these points of discontinuity by lines and ignoring the rather erratic coverage functions that lie above. The right panel gives the resulting coverage curves, joined by different style lines, for the five approximate upper limits described above with  $n = 10$  and  $\alpha = 0.05$ . The Armitage-Berry upper limit  $S_1$  has quite unacceptable coverage properties, while  $S_2$  is clearly conservative, as was also found by Lloyd [9] for two-sided limits. The asymptotically equivalent upper limits  $S_3, S_4$  and  $S_5$  all violate the coverage requirement to a substantial extent.

An ideal upper confidence limit will be as small as possible, perhaps in some average sense, subject to the coverage condition. Which of the five limits above are smallest? The

simple average value over the whole sample space for these five *approximate* upper limits are respectively 0.346, 0.527, 0.324, 0.337 and 0.328. If these were all valid limits, we would prefer the one with the smallest mean value, here  $S_3$ , but this limit is liberal. Neither is it fair to say that  $S_2$  is the worst having the highest mean value because its coverage is much higher than nominal 95%. The next section presents a method of adjusting each of these approximate limits so that coverage functions equal nominal at each point corresponding to an observable value of a limit. This allows us a direct comparison of the methods since they all have identical coverage properties.

### 3. EXACT UPPER CONFIDENCE LIMITS

The approximate limits described in the previous section had far from ideal coverage properties. One wonders whether one might not adjust them in some way so that the coverage curve was closer to nominal while preserving the essentials of the underlying method. Much of the statistical information in an upper limit is contained in the way it orders the sample space. It answers the question “for which data sets should the upper limit be higher and for which should it be lower”? We would like to find an upper limit which retains this ordering, satisfies the coverage conditions and is as small as possible subject to the first two constraints. Such a limit can be constructed using well-established theory which originally appeared in the reliability literature.

Any approximate limit  $S(X, T)$  can be transformed into a new upper limit  $u_S(X, T)$  that has the following three properties. First,  $u_S$  satisfies the coverage requirement (2). Second,  $u_S$  respects the ordering of  $S$ , i.e. if  $S(x_1, t_1) \geq S(x_2, t_2)$  then  $u_S(x_1, t_1) \geq u_S(x_2, t_2)$ . Third,  $u_S$  is as small as possible subject to the first two properties. A formula for constructing  $u_S(x, t)$  from  $S$  was given by Buehler [19]. It is the largest solution for  $\theta$  of

$$\sup_{\phi} \Pr(S(X, T) \leq S(x, t); \theta, \phi) > \alpha. \quad (3)$$

The three properties of so-called Buehler bounds listed above were first proven by Jobe and David [23] and extended by Lloyd and Kabaila [22]. Under mild regularity conditions, one finds that the coverage function  $C(\theta)$  of a Buehler bound exactly touches the nominal level  $1 - \alpha$  when  $\theta$  equals an observable value of the upper limit. Thus, for Buehler bounds the plots given in the right panel of Figure 1 are a single horizontal line at height  $1 - \alpha$ , meaning that the coverage functions satisfy the coverage restriction exactly. From the computational side, the key issues are firstly finding the tail set  $R(x, t) := \{S(X, T) \leq S(x, t)\}$ , which in our application is made much simpler by using the known monotonicity properties of the chosen statistics  $S(X, T)$ , and secondly finding the supremum of what can be a poorly behaved objective function such that it corresponds to maximum  $\theta$  subject to (3). Further details of the computation of  $u_S(x, t)$  are given in the appendix.

The exact upper limit of Buehler given by (3) bears a more than superficial resemblance to a construction given by various authors in various contexts (see [2], [5] and [24] for matched binary pairs, and [25], [26] and [27] for alternative problems involving  $2 \times 2$  tables), which can be expressed as the largest solution for  $\theta$  of

$$\sup_{\phi} \Pr(S(Y, \theta) \leq S(y, \theta); \theta, \phi) > \alpha. \quad (4)$$

The key difference here is that  $S(Y, \theta_0)$  is a test statistic for testing  $\theta = \theta_0$  and depends on  $\theta_0$ . The quantity on the left-hand side of (4) is a generalized P-value and the method can be thought of as an inversion of the test. An alternative version of this approach is

$$\sup_{\phi} \Pr(|S(Y, \theta)| \leq |S(y, \theta)|; \theta, \phi) > 2\alpha, \quad (5)$$

which is an inversion of the two-sided test, see [2]. Despite the apparent similarities, the ideas behind (3) and (4) are quite different in conception. The Buehler algorithm takes an approximate upper limit and adjusts it to have certain ideal properties, namely minimality subject to coverage and ordering. It is a mapping from one random variable  $S$  to a new random variable  $U_S$  with better statistical properties. Test inversion generates the limit directly from the test statistic, which is the more classical approach. However, there is no theory at present to demonstrate that these limits possess optimality properties, while our numerical investigation indicated that at least some of them, in fact, do not. Perhaps just as important, test inversion is numerically much more difficult than Buehler adjustment, precisely because in test inversion elements of the sample space enter and exit the set  $\{S(Y, \theta) \leq S(y, \theta)\}$  over the computational procedure. This results in the function having jump and drop discontinuities, which greatly complicates the computation of maximum  $\theta$  in (4). Some details and the beginnings of more general theory on test inversion limits can be found in Kabaila [28]. Due to unclear optimality properties and excessive computational intensity we will not look at test inversion based limits in this paper. At the same time, the theory for Buehler limits is well-established implying a strong optimality property of the limits that are relatively easy to compute. There is a clear need to clarify the relation between test inversion and Buehler limits, but this will need to be done in a more general and theoretical paper.

For the specific problem of a difference between correlated binomial proportions, we will look at a range of exact upper limits based on the Buehler method. Specifically, in the next section we will be comparing the five Buehler bounds  $U_i$  generated respectively from the five alternative approximate upper limits  $S_i$ ,  $i = 1, \dots, 5$ . Our objective is to find exact upper limits that are smallest in mean. It should be noted that none of considered exact Buehler limits for a difference in correlated proportions have been published or even studied previously.

#### 4. NUMERICAL RESULTS

We calculated 95% Buehler limits based on the suggested five approximate upper limits for all possible data sets, with  $n = 10, 15, 20, 25, 50, 100$ . It should be noted that for  $S_2$  one must decide how much of the total coverage error  $\alpha$  to spend on the intervals for  $\phi$  and  $\eta$ . We found numerically that spending around one third (i.e.  $\gamma \simeq \alpha/3$ ) on  $\phi$  and two thirds (i.e.  $\beta \simeq 2\alpha/3$ ) on  $\eta$  ultimately leads to exact upper limits that have the best statistical properties. For the sake of brevity, we will only be reporting results for this specific version of  $S_2$ , which in any case turns out to be uncompetitive.

Recall that each Buehler upper limit is as small as possible subject to the coverage restriction and we would prefer the Buehler limit that is the smallest. We will measure the size of a limit  $u_S$  by  $\sqrt{n}(u_S(X, T) - \hat{\theta})$ . Of course, the  $\sqrt{n}$  term is not essential, but it leads to comparable values across different sample sizes. Table III gives the simple average of this measure over the entire sample space for the five alternative Buehler limits and six sample sizes. On the

Table III. Average values of  $\sqrt{n}(u_S(X, T) - \hat{\theta})$  for five Buehler limits

	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 50$	$n = 100$
$S_1$	1.650	1.644	1.625	1.608	1.569	1.568
$S_2$	1.367	1.384	1.380	1.384	1.388	1.389
$S_3$	1.247	1.239	1.234	1.239	1.208	1.193
$S_4$	1.226	1.206	1.205	1.197	1.179	1.167
$S_5$	1.234	1.212	1.203	1.197	1.178	1.166

basis of these figures the score and SRLR based upper limits  $S_4$  and  $S_5$  appear to lead to the most desirable (i.e. smallest) exact upper limits  $U_4$  and  $U_5$  for all considered sample sizes. The reducing pattern from left to right is largely due to the distribution becoming less discrete with larger sample sizes.

The simple average gives equal weight to each element of the sample space. To better reveal the differences in the methods we use the mean value of  $\sqrt{n}(u_S(X, T) - \hat{\theta})$  giving model based weights to different points of the sample space. Since  $\hat{\theta}$  is unbiased for  $\theta$  this becomes

$$m_n(\theta, \phi) = \sqrt{n} (E(u_S(X, T); \theta, \phi) - \theta),$$

where  $E(\cdot; \theta, \phi)$  is mean with respect to the distribution of  $(X, T)$ . The distribution of  $(X, T)$  can be expressed as the product of the binomial distribution for  $T$  and the conditional binomial distribution of  $X$  given  $T$ , see equation (1). Clearly, the results depend on the parameter values  $(\theta, \phi)$  which vary over the triangular region  $\{\theta \in (-1, 1), \phi \in (0, 1); |\theta| \leq \phi\}$ . We calculated  $m_n(\theta, \phi)$  at 10100 points uniformly spread over the entire parameter space. Table 2 lists the proportions of the parameter space where each of the five Buehler limits is ranked smallest or second smallest.

Table IV. Proportions of the parameter space for each method ranked 1 or 2. The methods ranked according to the values of binomially weighted averages. The smaller value leads to the higher rank.

	rank	n=10	n=15	n=20	n=25	n=50	n=100
Estimated ( $U_1$ )	1	1%	0%	1%	2%	2%	2%
	2	0%	1%	1%	5%	1%	1%
Lloyd ( $U_2$ )	1	0%	1%	0%	1%	0%	1%
	2	1%	1%	3%	1%	0%	0%
Wald ( $U_3$ )	1	12%	5%	6%	4%	9%	5%
	2	16%	16%	14%	11%	13%	14%
Score ( $U_4$ )	1	76%	75%	49%	55%	46%	48%
	2	13%	14%	33%	24%	33%	32%
SRLR ( $U_5$ )	1	11%	19%	43%	38%	42%	44%
	2	70%	68%	50%	59%	52%	53%

Over the vast majority of the parameter space, either  $U_3$ ,  $U_4$  or  $U_5$  are preferred. Figure 2 shows plots of  $m_n(\theta, \phi)$  versus  $n$  for four representative parameter values  $(\theta, \phi)$  and indicates that  $U_3$  is typically worse than  $U_4$  and  $U_5$ , which have similar performance. We have omitted  $U_1$  and  $U_2$  from these plots as they are uncompetitive.

The regions of the parameter space where each of  $U_3$ ,  $U_4$  or  $U_5$  are smallest are displayed as an image plot in the Figure 3, for the case  $n = 25$  with  $\alpha = 0.05$ . Very similar plots are obtained for other sample sizes and values of  $\alpha$ , but for this sample size the choice between the three asymptotically equivalent methods was least clear. The general pattern is that  $U_4$  (brightest) or  $U_5$  (second brightest) are best over the vast majority of the parameter space, while  $U_3$  (darkest) performs well only in some boundary areas. Although  $U_4$  and  $U_5$  show somewhat similar performance, it is useful to compare the relative sizes of these limits. The contour lines describe the percentage difference in mean size between  $U_4$  and  $U_5$ , positive values indicating that  $U_5$  is superior. In regions where  $U_4$  performs better the difference is slight – of the order of 1%. At the same time, well-designed matched-pairs studies are likely to result in a large proportion of concordant pairs, so  $\phi$  is typically small, say less than 0.25. In this practically interesting area of the parameter space where  $\phi$  is small,  $U_5$  is superior by the order of 6-8%. This feature becomes even more pronounced with larger sample size  $n$ .

## 5. FURTHER NUMERICAL EXAMPLES

For further numerical illustration we present two examples. For both of these examples, the original studies were primarily interested in a test, whereas we are interested in one-sided confidence limits. For illustrational purposes, we report both lower and upper limits. To conclude non-inferiority or superiority, the lower confidence limit must lie respectively in  $(-\delta, \infty)$  or  $(0, \infty)$ , with  $-\delta$  again referring to a pre-specified non-inferiority margin. It is also possible to combine lower and upper exact limits into intervals in order to conclude about the equivalence, in which case both limits must lie in  $(-\delta, \delta)$ . However, we do not recommend doing so because such intervals will not be exact, but normally be highly conservative.

*Cross-over trial example.* The first example is a cross-over trial described in Jones and Kenward [29] and used as a numerical example by Agresti and Min [13]. The initial study compares the low- and high-dose analgesics for relief of primary dysmenorrhoea given to 86 patients. The treatment was successful at both high and low doses for 53 patients and was unsuccessful at both high and low doses for 9 patients. Of the  $t = 24$  patients with a different response at high or low dose,  $x_{01} = 16$  were favorable to the high dose. In our notation the data are  $(x, t, n) = (16, 24, 86)$ . The ML estimate of  $\theta$  is  $(16 - 8)/86 = 0.093$ .

*Comparison of two binary diagnostic tests.* Our second example is a clinical trial that involved repeated binary measures on a sample of patients of known disease status. The data are originally from Liu et al. (unpublished manuscript) and were also used by Liu et al. [30] and Hsueh, Liu and Chen [5] to illustrate tests on non-inferiority and bio-equivalence. The purpose of the original trial was to compare two diagnostic procedures in which patients with liver lesions are first given a non-invasive procedure based on magnetic resonance imaging (MRI) (method 2) and then an invasive procedure based on computer tomography arterio portography (CTAP) (method 1). From the  $n = 50$  patients, there were  $t = 9$  discordant results and of these  $x_{01} = 5$  patients were correctly diagnosed with the non-invasive MRI method. The data for this example are  $(x, t, n) = (5, 9, 50)$ , and the ML estimate of  $\theta$  is  $(5 - 4)/50 = 0.02$ .

Table V. Approximate and exact lower and upper confidence limits for data in Jones and Kenward [29] and Agresti and Min [13]:  $(x, t, n; \alpha) = (16, 24, 86; \alpha)$  for  $\alpha = 0.01, 0.05$ .

$i$	1		2		3		4		5	
$s_i(\alpha = 0.05)$	-0.012	0.180	-0.031	0.265	-0.001	0.183	-0.001	0.188	0.000	0.187
$u_i(\alpha = 0.05)$	-0.054	0.256	-0.003	0.235	-0.007	0.190	-0.004	0.191	-0.009	0.191
$s_i(\alpha = 0.01)$	-0.050	0.205	-0.089	0.321	-0.039	0.221	-0.042	0.229	-0.040	0.227
$u_i(\alpha = 0.01)$	-0.129	0.316	-0.075	0.301	-0.053	0.237	-0.053	0.235	-0.052	0.235

Table VI. Approximate and exact lower and upper confidence limits for data in Liu et al. [30] and Hsueh, Liu and Chen [5]:  $(x, t, n; \alpha) = (5, 9, 50; \alpha)$  for  $\alpha = 0.01, 0.05$ .

$i$	1		2		3		4		5	
$s_i(\alpha = 0.05)$	-0.090	0.119	-0.177	0.228	-0.081	0.119	-0.086	0.127	-0.082	0.123
$u_i(\alpha = 0.05)$	-0.174	0.199	-0.148	0.177	-0.092	0.131	-0.092	0.131	-0.092	0.131
$s_i(\alpha = 0.01)$	-0.118	0.142	-0.253	0.298	-0.122	0.161	-0.138	0.180	-0.129	0.171
$u_i(\alpha = 0.01)$	-0.242	0.292	-0.225	0.267	-0.158	0.187	-0.147	0.186	-0.147	0.187

Tables V and Table VI give the values of both approximate and exact lower and upper 95% and 99% confidence limits, the lower limit quoted first. The approximate limits  $S_i$  cannot be directly compared due to poor coverage properties. Consistently with characteristics reported in Section 2,  $S_2$  are conservative and  $S_1, S_3, S_4$  and  $S_5$  are liberal as can be judged from the corresponding limits  $U_i, i = 1, \dots, 5$  that all have coverage equal exactly to nominal. The exact limits  $U_i$  broadly confirm the results of our numerical study reported in the previous section, although the recommended SRLR based method  $U_5$  does not always lead to smallest upper (largest lower) confidence limit values.

## 6. CONCLUSION

Confidence limits are one of the key tools in statistical analysis. It is important to ensure that conclusions based on confidence limits and actions that may follow are neither excessively conservative nor too liberal. We have studied the problem of constructing upper and lower confidence limits for the difference between two proportions based on matched-pairs design. In particular, we have applied the general optimality theory of Buehler to five alternative approximate upper confidence limits, which broadly represent the main methods proposed in the literature. After the Buehler adjustment all five upper limits become exact in their coverage and can then be directly compared according to their average size, smaller size to be preferred. Our numerical analysis over the whole parameter space has shown that only three of five considered methods, Wald type limit  $U_3$ , score type limit  $U_4$  and likelihood ratio type

limit  $U_5$ , are competitive. More detailed analysis has revealed that  $U_4$  and  $U_5$  take the smallest mean values across the vast majority of the parameter space. We also found that the SRLR based method  $U_5$  performs substantially better than  $U_4$  in the areas of the parameter space that are the most plausible under typical matched-pairs studies. Therefore, we conclude that the SRLR based method  $U_5$  has the best performance out of five considered methods and we recommend it to practitioners. Lower confidence limits have absolutely identical optimality properties to the upper limits considered in this study and we provide the software tools for computation of both (URL supplied by publisher). Of course, the superiority of our method is not uniform and for particular data sets the limits  $U_3$  or  $U_4$  may be tighter.

There are alternative parameters for describing the difference between two probabilities in matched binary pairs design. We have not looked at confidence limits for the ratio  $p_2/p_1$  in this study where there has been recent progress on construction of exact intervals by Chan et al. [31] and approximate intervals by Bonett and Price [32]. Another common parameter is the odds ratio, see Agresti and Min [13]. Both of these problems involve two nuisance parameters and so the application of the Buehler procedure, although still possible, becomes computationally challenging. Special methods must be developed and used for the search over multidimensional parameter space and this is a question for further research.

#### REFERENCES

1. Kao CH, Shiau YC, Shen YY, Yen RF. Detection of recurrent or persistent nasopharyngeal carcinomas after radiotherapy with technetium-99m methoxyisobutylisonitrile single photon emission computed tomography and computed tomography: comparison with 18-fluoro-2-deoxyglucose positron emission tomography. *textitCancer* 2002; **94**:1981–1986.
2. Tang M-L, Tang N-S, Chan ISF. Confidence interval construction for proportion difference in small-sample paired design. *Statistics in Medicine* 2005; **24**:3565–3579.
3. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **17**:153–157.
4. Suissa S, Shuster JJ. The  $2 \times 2$  matched pairs trial: exact unconditional design and analysis. *Biometrics* 1991; **47**:361–372.
5. Hsueh H-M, Liu J-P, Chen JJ. Unconditional exact tests for equivalence or noninferiority for paired binary endpoints. *Biometrics* 2001; **57**:478–483.
6. Berger RL, Sidik K. Exact unconditional tests for a  $2 \times 2$  matched-pairs design. *Statistical Methods in Medical Research* 2003; **12**:91–108.
7. Liang KY, Zeger SL. On the use of concordant pairs in matched case-control studies. *Biometrics* 1988; **44**:1145–1156.
8. Armitage P, Berry G. *Statistical methods in medical research*. London: Blackwell Scientific Publications, 1987.
9. Lloyd CJ. Confidence intervals from the difference between two correlated proportions. *Journal of the American Statistical Association* 1990; **85**:1154–1158.
10. Newcombe RG. Improved confidence intervals for the difference between Binomial proportions based on paired data. *Statistics in Medicine* 1998; **17**:2635–2650.
11. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* 1998; **17**:891–908.
12. Tango T. Letter to the editor. *Statistics in Medicine* 1999; **18**:3511–3513.
13. Agresti A, Min Y. Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine* 2005; **24**:729–740.
14. Statistical principles for clinical trials (with an introductory note by JA Lewis). *Statistics in Medicine* 1999; **18**:1903–1942.
15. Berger RL, Hsu JC. Bio-equivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996; **11**:283–319.
16. Blyth CR, Still HA. Binomial confidence intervals. *Journal of the American Statistical Association* 1983; **78**:108–116.

17. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics* 2000; **28**:783–798.
18. Reiczigel J. Confidence intervals for the binomial parameter: some new considerations. *Statistics in Medicine* 2003; **22**:611–621.
19. Buehler RJ. Confidence intervals for the product of two binomial parameters. *Journal of the American Statistical Association* 1957; **52**:482–493.
20. Kabaila P, Lloyd CJ. Tight upper confidence limits from discrete data. *Australian Journal of Statistics* 1997; **39**:193–204.
21. Kabaila P, Lloyd CJ. The Importance of the Designated Statistic on Buehler Upper Limits on a System Failure Probability. *Technometrics* 2002; **44**:390–395.
22. Lloyd CJ, Kabaila P. On the optimality and limitations of Buehler bounds. *Australian Journal of Statistics* 2003; **45**:167–174.
23. Jobe JM, David HT. Buehler confidence bounds for reliability-maintainability measure. *Technometrics* 1992; **34**:214–222.
24. Sidik K. Exact unconditional tests for testing non-inferiority in matched-pairs design. *Statistics in Medicine* 2003; **22**:265–278.
25. Chan ISF, Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* 1999; **55**:1202–1209.
26. Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001; **57**:963–971.
27. Tang N-S, Tang M-L. Exact unconditional inference for risk ratio in a correlated  $2 \times 2$  table with structural zero. *Biometrics* 2002; **58**:972–980.
28. Kabaila P. Computation of exact confidence intervals from discrete data using studentized test statistics. *Statistics and Computing* 2005; **15**:71–78.
29. Jones B, Kenward MG. Modeling binary data from a three-period cross-over trial. *Statistics in Medicine* 1987; **6**:555–564.
30. Liu J-P, Hsueh H-M, Hsieh E, Chen JJ. Test for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* 2002; **21**:231–245.
31. Chan ISF, Tang NS, Tang NL, Chan PS. Statistical analysis of non-inferiority with a rate ratio in small sample matched-pair designs. *Biometrics* 2003; **59**:1170–1177.
32. Bonnet DG, Price RM. Confidence intervals for a ratio of binomial proportions based on paired data. *Statistics in Medicine* 2005; : In press.
33. Kabaila P. Computation of exact confidence limits from discrete data. *Computational Statistics* 2005; **20**:401–414.

#### Appendix. Computational implementation of Buehler bounds.

Buehler’s algorithm generates upper confidence limits for  $\theta$  starting with a statistic  $S(Y)$  called the designated statistic. This statistic should be chosen to be itself an approximate upper limit, see Kabaila and Lloyd [21]. The exact upper limit  $u_S(y)$  for a given observed value  $y$  of the data  $Y$  is given as the largest solution for  $\theta$  (more formally, the supremum of  $\theta$ ) of the set defined by the inequality

$$\sup_{\phi} \Pr(S(Y) \leq S(y); \theta, \phi) > \alpha. \quad (6)$$

The key quantities that determine the computational difficulty of this algorithm are the cardinality  $N$  of the sample space, and the dimension  $k$  of the nuisance parameter  $\phi$ .

The first step in the algorithm is to determine the set  $R(y) := \{S(Y) \leq S(y)\}$ . At least in general, this would require checking whether or not each sample point  $\tilde{y}$  satisfies the condition  $S(\tilde{y}) \leq S(y)$ , and so would require  $O(N)$  computations. For even moderately complex problems this will be impossible. Consider, for instance, the cardinality of the sample space for a logistic regression modeling, say,  $n=50$  binary responses in terms of a continuous covariate. The sample space will then contain  $N = 2^{50}$  elements, which is clearly computationally infeasible to enumerate. In our application,  $Y = (X, T)$  and the cardinality is  $N = (n + 1)(n + 2)/2$ . Computation can be reduced by using known logical properties of the statistic  $S(y)$ . In our

case,  $S(x, t)$  is non-decreasing in  $x$  for fixed  $t$  and so the tail set  $R(x, t)$  can be expressed as a disjoint union of sets  $R_t := \{x \leq x_t\}$ . This reduces the number of computations for set  $R(y)$  to  $O(n \log n)$ . It also allows efficient computation of the tail probability via

$$\Pr(R(y)) = \sum_{t=0}^n \Pr(R_t),$$

with  $\Pr(R_t)$  referring to easily and efficiently computable cumulative binomial probability of  $R_t := \{x \leq x_t\}$  for fixed  $t$ .

The second part of the computation is finding the supremum of  $\Pr(R(y); \theta, \phi)$  with respect to  $\phi$ , keeping in mind that this optimization is carried out multiple times in the search for the largest  $\theta$  that satisfies (6). In general, this function does not enjoy any special properties that can reduce or guarantee computation of the optimum. Indeed, it typically has multiple maxima and is rarely convex. Maximizing such a function will always be problematic, especially so if the dimension  $k$  of the nuisance parameter  $\phi$  is large.

In problems like the present one with  $k = 1$ , we recommend one of three computational approaches. Firstly, an ordinary grid search can be applied with moderate number of points  $50 \leq M \leq 200$  in the nuisance parameter space. We found this approach with  $M = 100$  to be especially efficient in our case of difference between two correlated binomial probabilities mostly due to dependence of the range of  $\phi$  on  $\theta$ :  $\phi \geq |\theta|$ . Secondly, the combination of a grid search and numerical optimization can be applied. In this case, one should select a sufficient number of grid points  $M$  in order to reliably localize the interval in the nuisance parameter space that contains the global maximum. Then the numerical optimization routine should be applied within this interval in order to arrive to the supremum probability solution. Finally, an alternative approach can be applied that follows from a result of Lloyd and Kabaila [22], who show that under mild regularity conditions  $u_S(y) = \sup_{\phi} u_S(y; \phi)$ , where  $u_S(y; \phi)$  solves (6) for known  $\phi$ . This approach sometimes can have advantages since, in contrast to  $\Pr(R(y); \theta, \phi)$ ,  $u_S(y; \phi)$  can often be a well-behaved function of  $\phi$ . Having selected the technique for finding the probability supremum, the search for maximum  $\theta$  can be implemented by replacing the inequality (6) by its corresponding equality and applying one of numerical equation solving algorithms (e.g. bisection). This can be done since the function (6) is continuously non-increasing in  $\theta$  for any reasonable choice of designated statistics, including all five statistics considered in the preceding study (see Kabaila [33] for a proof of this property).

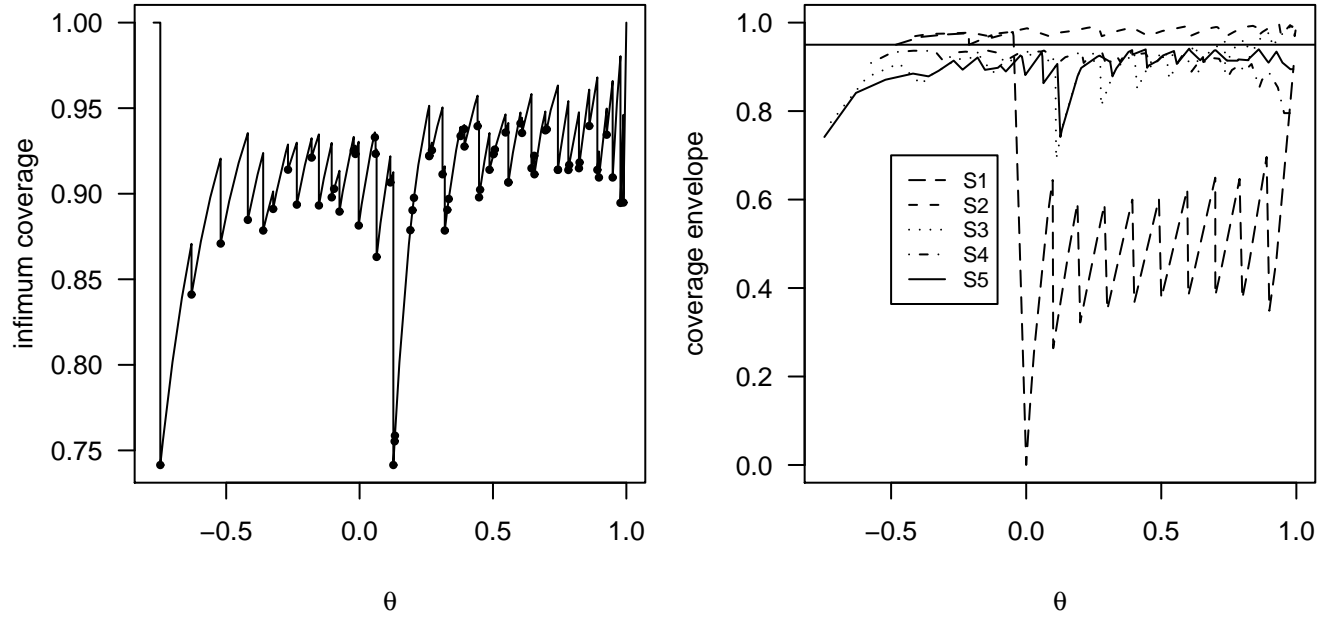


Figure 1. Coverage functions of five upper limits for  $\theta$ ,  $n=10$ ,  $\alpha = 0.05$ . *Left.* LR upper limit, discontinuities marked by points. *Right.* Lower coverage envelopes against  $\theta$ .

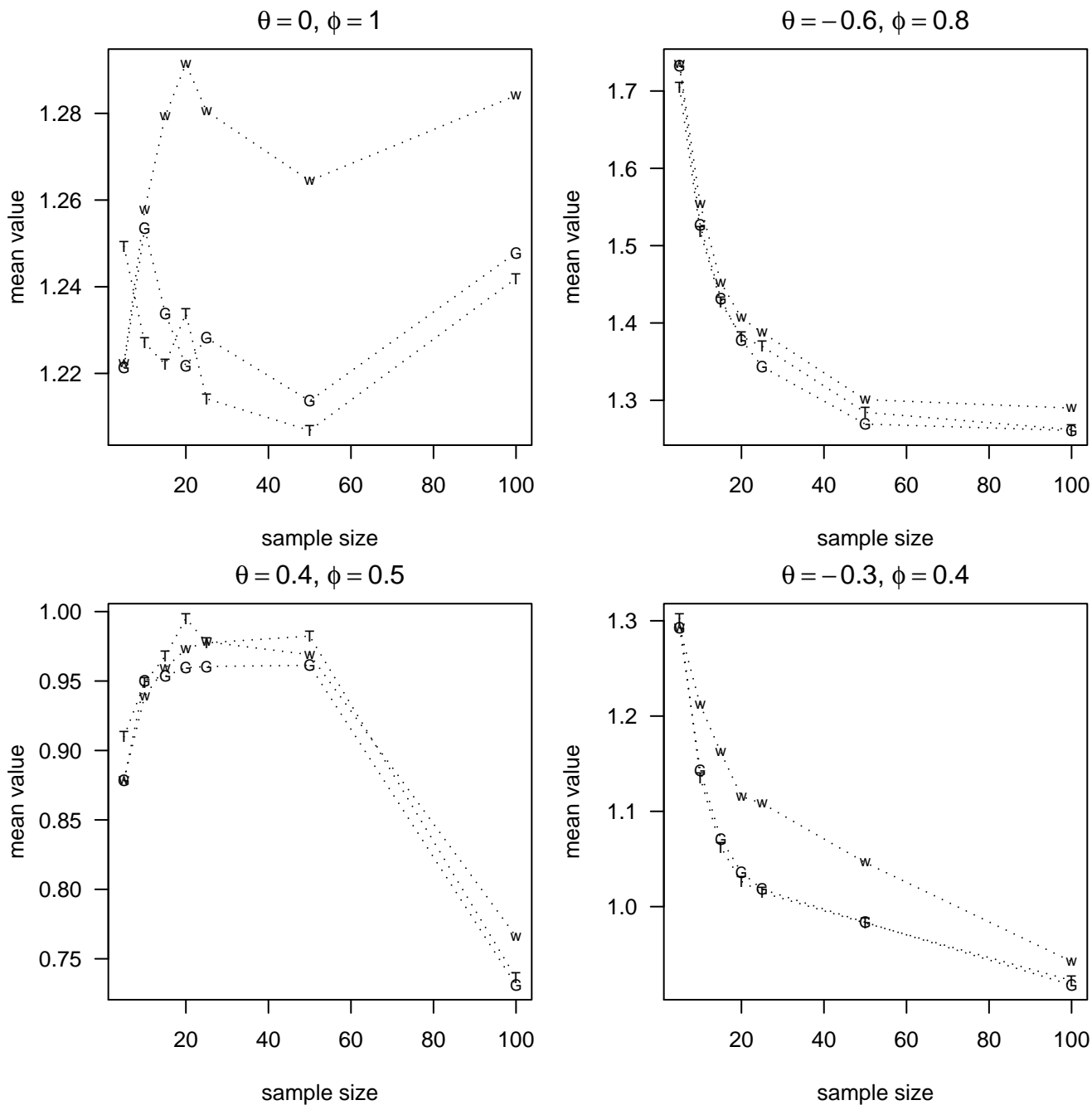


Figure 2. Mean values  $m_n(\theta, \phi)$  for  $U_3, U_4, U_5$  versus sample size  $n$  for some arbitrary  $\theta$  and  $\phi$ .

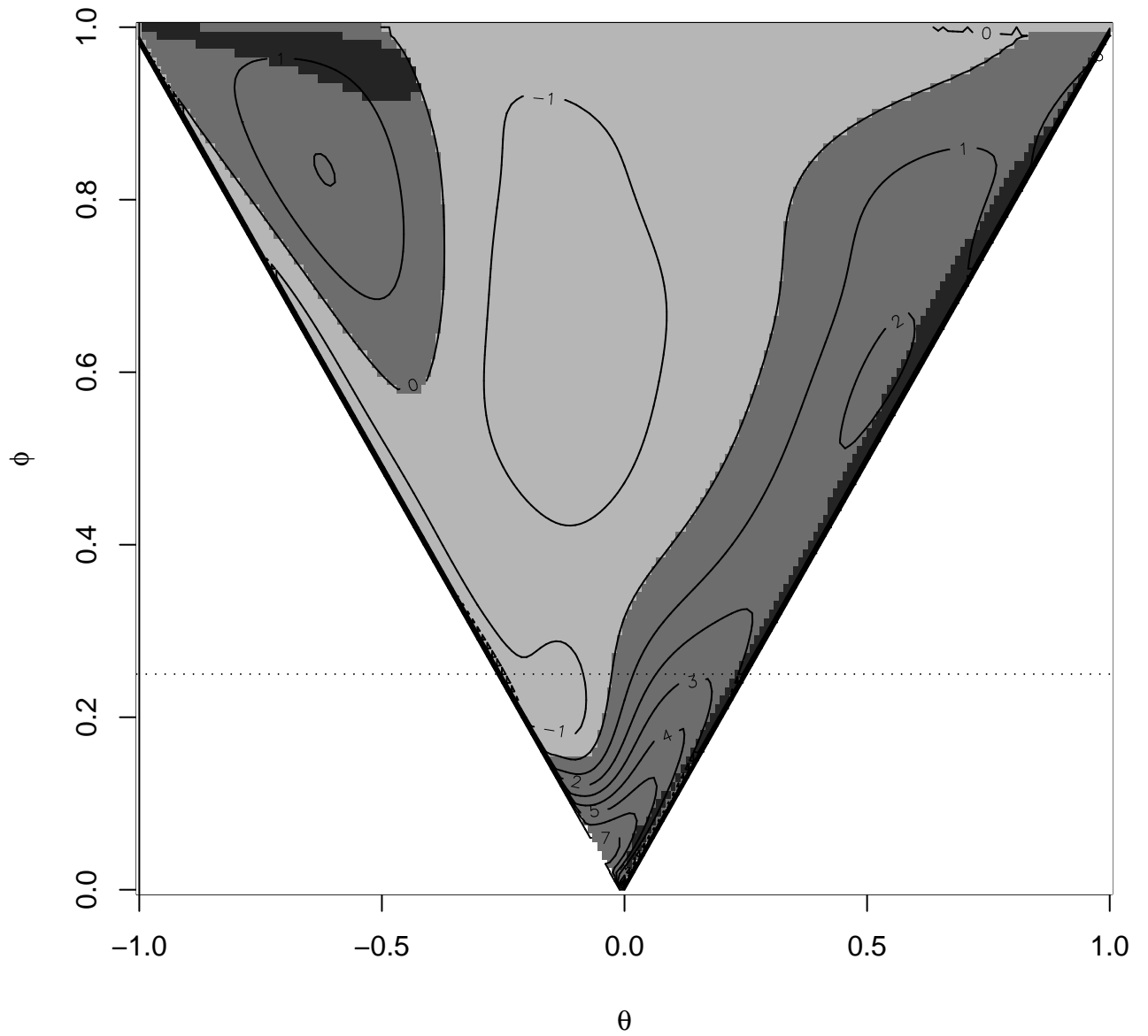


Figure 3. The areas of the parameter space for which either  $U_3$ ,  $U_4$  or  $U_5$  have the minimal  $m_n(\theta, \phi)$ . Contours describe the % difference between  $U_4$  and  $U_5$ .