

Unconditional efficient one-sided confidence limits for the odds ratio based on conditional likelihood

Chris J. Lloyd* and Max V. Moldovan

Melbourne Business School, Carlton, 3053, AUSTRALIA

SUMMARY

We compare various one-sided confidence limits for the odds ratio in a 2×2 table. The first group of limits relies on first order asymptotic approximations and includes limits based on the (signed) likelihood ratio, score and Wald statistics. The second group of limits is based on the conditional tilted hypergeometric distribution, with and without mid-P correction. All these limits have poor unconditional coverage properties and so we apply the general transformation of Buehler (1957) to obtain limits which are unconditionally exact. The performance of these competing exact limits is assessed across a range of sample sizes and parameter values by looking at their mean size. The results indicate that Buehler limits generated from the conditional likelihood have the best performance, with a slight preference for the mid-P version. This confidence limit has not been proposed before and is recommended for general use, especially when the underlying probabilities are not extreme. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Suppose n subjects are randomly assigned to two groups of size n_1 and n_2 and exposed to two alternative treatments resulting in X_i successes in group i . If subjects respond independently then $X_i \sim \text{Bin}(n_i, p_i)$, where p_i is the probability of success in group i . As our main example we use a clinical trial into the effects of tobacco smoke on mice, reported by Essenberg [1]. There were 72 mice randomly allocated to treatment with smoke or control. After one year the 55 surviving mice were checked for the presence of tumour. The results are in Table I.

There are two biological issues of interest here, namely the effect of smoke on mortality and on tumour prevalence given survival. The control group has a mortality rate of 11.1% and tumour prevalence 59.4% while the smoke treatment group has a mortality rate of 36.1% and tumour prevalence 91.3%. The log-odds ratio, given by $\theta = \text{logit}(p_1) - \text{logit}(p_2)$, where $\text{logit}(p_i) = \log(p_i/(1 - p_i))$ is a standard measure of the difference between two rates or proportions. For comparing the treatment group to the control group the log-odds are 1.509

*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

Table I. Results of a clinical trial into the effect of tobacco smoke on tumour prevalence, Essenberg [1].

	Mortality			Tumour Prevalence		
	Mortality	Survival	Total	Tumour	No Tumour	Total
Smoke Treatment	13(x_1)	23(y_1)	36(n_1)	21(x_1)	2(y_1)	23(n_1)
Control	4(x_2)	32(y_2)	36(n_2)	19(x_2)	13(y_2)	32(n_2)
Total	17	55	72(n)	40	15	55(n)

for mortality and 1.972 for tumour prevalence. In our study we are interested in finding exact lower and upper confidence limits for θ .

Due to the inherent symmetry of the log-odds ratio, the upper limit $u(x_1, x_2)$ for (n_1, n_2) corresponds to the lower limit $-u(x_2, x_1)$ for (n_2, n_1) . Clearly, lower and upper limits have absolutely identical properties. Therefore, we will describe our theory for upper limits only while illustrating our ideas with both lower and upper limits. An upper confidence limit $u(X_1, X_2)$ is required to bound the parameter of interest θ from above with probability $1 - \alpha$, as indicated by

$$\Pr(\theta \leq u(X_1, X_2)) \geq 1 - \alpha, \quad \forall p_1, p_2. \quad (1)$$

Subject to this restriction, we want $u(X_1, X_2)$ to be as small as possible. This suggests that we make the left-hand side of (1) as small as possible, ideally equal to $1 - \alpha$ exactly. While this is not possible due to discreteness, it is possible to make the infimum probability equal $1 - \alpha$ exactly. Such limits are called exact, even though it is the bound on coverage, not the coverage itself, which is exact.

There are two general approaches to analyzing 2×2 tables. In the unconditional approach, the joint binomial distribution of X_1 and X_2 is used both to construct the statistical procedure and to evaluate it. This generates tests of $(p_1 = p_2) := (\theta = 0)$ as well as one and two-sided confidence limits for θ , often calibrated against the approximating normal distribution. For instance, the Pearson and deviance tests are in common use as are Wald-type confidence limits for θ . The main weakness of these methods is the fact that, at least for some parameter values (p_1, p_2) , the unconditional coverage of the limits and the size of the tests violate the nominal values, often substantially.

An alternative approach to the analysis of 2×2 tables is to condition on the observed value of $T = X_1 + X_2$. This leads to Fisher's exact test as well as confidence limits for θ based on the tilted hypergeometric distribution, see Thomas [2]. Evaluated conditionally, these procedures have some attractive and logically sound qualities. First, their properties depend only on the parameter of interest since the nuisance parameter is eliminated by conditioning. Second and more important, they possess various optimality characteristics because the conditional distribution belongs to a one-parameter exponential family, see Lehmann [3]. Third, conditional methods are held to lead to inference that is more relevant to the data obtained, rather than averaged over data that might have been obtained, see Lloyd [4], [5]. The main problem for conditional methods is the often excessive discreteness of the conditional distribution which can, however, be partly overcome by using the mid P-value correction of Lancaster [6].

Table II lists lower and upper 95% confidence limits for the log-odds ratio for the illustrative example. The conditional and unconditional limits presented in this table are based on the two most commonly used methods, namely the conditional method of Thomas and the Wald

limits with the usual modification of adding 1/2 to each count. The latter modification is often justified on the basis of reducing asymptotic bias of the estimator. The observed values for these two lower and upper limits are listed in the line labeled “approximate” and it is apparent that the conditional limits are not as tight as the Wald-based limits. This may largely be due to the well-known conservatism of the conditional method of Thomas and the liberalism of the unconditional Wald method. It makes no sense to directly compare the two methods when the extent to which their coverage differs from nominal is unknown and unaccounted for.

Table II. Approximate and exact 95% confidence bounds from data of Essenberg [1], using two standard conditional and unconditional methods. The preferred exact limits are in bold font.

	Mortality: $\hat{\theta} = 1.509$				Tumour Prevalence: $\hat{\theta} = 1.972$			
	<u>Thomas</u>		<u>Wald</u>		<u>Thomas</u>		<u>Wald</u>	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Approximate	0.333	2.803	0.423	2.423	0.471	3.875	0.540	3.028
Exact	0.401	2.677	0.401	3.357	0.616	3.676	0.585	4.173

In the context of hypothesis testing, Fisher’s exact conditional test is known to be excessively conservative having unconditional size much smaller than nominal, which also leads to much lower power as compared to alternative unconditional tests. Fisher’s test has been criticized on this basis, although for testing the null hypothesis of independence it has also been argued that the test should be assessed conditionally, not unconditionally [4], [5]. Due to discreteness, unconditional tests may also have some degree of conservatism, though less than their conditional counterparts. Conservatism can be partly controlled, but not entirely eliminated, by using exact calculations while accounting for the maximum error with respect to nuisance parameters. For any test statistic S , large values of which lead to rejection of the null, one can define the P-value to be the supremum of $\Pr(S \geq s)$ over all parameters consistent with the null. An early pertinent example of this is the test of Boschloo [7], which is identical to the test based on the maximized Fisher’s P-value. Mehrotra, Chan and Berger [8] compared the performance of several common test statistics, after maximization, and found that the conditional likelihood based Boschloo’s test performs as well or better than standard unconditional tests.

Analogously, conditional upper limits are typically conservative having unconditional coverage much larger than nominal for all parameter values. In our paper, we attempt to remove this conservatism while preserving the attractive features of the conditional approach and the essential coverage requirement (1). Specifically, we address the question: “Might one modify conditional methods so as to retain some of the attractive features of logic and efficiency while achieving better unconditional performance?”. In our study, we evaluate all confidence limits for θ unconditionally, regardless of how the bounds are initially derived.

Santner and Snell [9], Chan and Zhang [10] and Agresti and Min [11], amongst others, discussed the properties of ‘exact’ unconditional two-sided confidence limits applied to 2×2 tables. All the suggested confidence intervals are slightly conservative since they generally fail to reach nominal coverage value. More importantly, it is problematic to compare confidence intervals which may have unbalanced probabilities of error on each tail. In other words, it is possible that nearly all of the coverage error will be allocated to one of two tails, which

typically leads to shorter intervals. The uncontrolled balance of coverage error is evidently one of the prime reasons for the lack of theory and general methods for obtaining optimal two-sided limits.

In contrast, clear optimality theory is available for one-sided confidence limits. Buehler [12] gave an algorithm for adjusting a given upper limit so that it is as small as possible subject to both the coverage restriction (1) and ordering of the sample space imposed by the initial limit. Buehler limits have actual coverage equal exactly to nominal for all values of θ corresponding to observable values of the limit, see Kabaila and Lloyd [13]. While it is true that combining two exact one-sided limits gives a two-sided interval which is typically conservative, practitioners might still prefer this approach because the coverage error at each end of the interval is exactly controlled.

The Buehler exact limits for the illustrative example are listed in the line “exact” of Table II. All these limits have exact 95% coverage and confirm the general tendency of Wald limits to be liberal and conditional limits of Thomas to be conservative. Since alternative Buehler limits have exact coverage, we can directly compare them, preferring limits that are closer to $\hat{\theta}$. For this data it appears that exact limits based on the conditional method of Thomas are preferred, though the Wald lower limit for the mortality log-odds is identical to that based on the conditional method.

The aim of this paper is to compare two groups of Buehler confidence limits for the association parameter θ in a 2×2 table – one group based on the conditional likelihood given t and the second group based on the unconditional likelihood. Bearing in mind the attractive logical properties of the conditional approach, there is nothing intrinsically inefficient in the limit based on the conditional likelihood. Thus, there is reason to expect that a Buehler adjusted version of this limit will perform well unconditionally.

The rest of the paper is structured as follows. In Section 2, standard conditional and unconditional approximate upper limits are described, and the Buehler procedure introduced, which can be used to adjust the standard limits into the exact form. In Section 3, we report the results of a numerical study comparing the average size of the Buehler adjusted exact upper limits generated from the standard methods. An extended illustrative example is given in Section 4. Section 5 summarizes the results of the study.

2. APPROXIMATE AND BUEHLER ADJUSTED EXACT UPPER LIMITS

In this section, we describe a range of methods based on the conditional and unconditional likelihoods that are used to construct upper limits for θ . Each of these upper limits has poor coverage properties which is the motivation for adjusting them to be exact using the Buehler algorithm presented below. We note once again that it is only necessary to study upper limits because if $u(x_1, x_2)$ with (n_1, n_2) is an upper limit for θ then $-u(x_2, x_1)$ with (n_2, n_1) is a lower limit for θ .

Conditional upper limits. The conditional distribution of X_1 given $T = t$ is

$$\Pr(X_1 = x \mid t; \theta) = \kappa^{-1}(\theta) \binom{n_1}{x} \binom{n_2}{t-x} e^{\theta x}, \quad \max(0, t - n_2) \leq x \leq \min(n_1, t),$$

where $\kappa(\theta)$ is a normalizing constant. An upper limit can be obtained by finding the largest value θ_0 for which the one-sided P-value $P(\theta_0) = \Pr(X_1 \leq x \mid t; \theta_0)$ is no smaller than the target

coverage error α . Since the P-value is monotonically decreasing in θ_0 , the solution is unique and is efficiently computed by Newton-Raphson, preferably on the logarithmic scale. Denote this statistic, originally proposed by Thomas [2], by $C(X_1, X_2)$.

This limit is obtained by inverting the exact conditional P-value which is known to be conservative simply because of the limitations of discreteness. The P-value $P(\theta_0)$ is not uniformly distributed but is stochastically larger than uniform with a mean greater than 1/2. Lancaster [6] proposed the mid P-value defined by $P^*(\theta_0) = 0.5 \Pr(X_1 = x|t; \theta_0) + \Pr(X_1 < x|t; \theta_0)$. This leads to an upper limit $C^*(X_1, X_2)$ from solving $P^*(\theta_0) = \alpha$, which, however, as opposed to $C(X_1, X_2)$, is not ‘exact’ and can violate the coverage requirement (1).

We emphasize that all the approximate limits considered in this study, both conditional and unconditional, are used only for imposing a particular ordering on the sample space. Agresti and Coull [14] suggested a continuity correction that can potentially improve this ordering. Therefore, we also investigate the modified versions of $C(X_1, X_2)$ and $C^*(X_1, X_2)$ obtained by adding 1/2 to each count and denoted by $\tilde{C}(X_1, X_2)$ and $\tilde{C}^*(X_1, X_2)$, respectively. The same modification is also applied to all unconditional limits introduced next.

Unconditional upper limits. The unconditional likelihood is

$$\ell(\theta, \psi) = x_1\theta + (x_1 + x_2)\psi - n_1 \log(1 + e^{\theta+\psi}) - n_2 \log(1 + e^\psi),$$

where $\theta = \text{logit}(p_1) - \text{logit}(p_2)$ is the parameter of interest and we take $\psi = \text{logit}(p_2)$ to be the nuisance parameter. This likelihood generates several standard approximate unconditional upper limits, each unaffected by the choice of nuisance parameter. The most straightforward is based on the maximum likelihood (ML) estimator of θ and has the form

$$W(x_1, x_2) = \hat{\theta} + z_\alpha \sqrt{\frac{1}{n_1 \hat{p}_1 (1 - \hat{p}_1)} + \frac{1}{n_2 \hat{p}_2 (1 - \hat{p}_2)}},$$

where $\hat{p}_i = x_i/n_i$ are the ML estimates of p_i . We denote the standard and modified versions of the Wald limit by $W(X_1, X_2)$ and $\tilde{W}(X_1, X_2)$, respectively.

Let $\hat{\psi}_\theta$ be the ML estimate of ψ for fixed θ , which can be found by solving a quadratic, see Lloyd [15] p. 130. It is well known that $2(\ell(\hat{\theta}, \hat{\psi}) - \ell(\theta, \hat{\psi}_\theta))$ has approximate χ_1^2 distribution. If z_α is the upper α quantile of the standard normal distribution, then the solution for $\theta > \hat{\theta}$ of

$$\text{sign}(\theta - \hat{\theta}) \sqrt{2(\ell(\hat{\theta}, \hat{\psi}) - \ell(\theta, \hat{\psi}_\theta))} = z_\alpha$$

is a sign root likelihood ratio (SRLR) based upper limit for θ . Denote the standard and modified versions of this limit by $L(X_1, X_2)$ and $\tilde{L}(X_1, X_2)$, respectively. Finally, Rao’s score statistic is

$$(x_1 - n_1 \hat{p}_{1\theta}) \sqrt{\frac{1}{n_1 \hat{p}_{1\theta} (1 - \hat{p}_{1\theta})} + \frac{1}{n_2 \hat{p}_{2\theta} (1 - \hat{p}_{2\theta})}},$$

where $\hat{p}_{j\theta}$ are the ML estimates of p_j for known θ . An approximate upper limit follows from finding the value of $\theta > \hat{\theta}$ for which this equals z_α , as described by Agresti and Min [16]. We denote this limit by $R(X_1, X_2)$ and its modified version by $\tilde{R}(X_1, X_2)$.

Exact upper limits. Buehler [12] suggested a general procedure for obtaining an exact upper limit for a parameter of interest θ in the presence of a nuisance parameter ψ . Starting with

an initial approximate confidence limit S , called the *designated* statistic, we compute the new confidence limit u_S by finding the largest solution for θ of

$$\sup_{\psi} \Pr(S \leq s; \theta, \psi) > \alpha.$$

The values of u_S depends solely on the tail set $\{S \leq s\}$ rather than the actual values of S . Thus, as noted earlier, the exact limit depends on S only through the ordering it imposes on the sample space. Again, we point out that the choice of nuisance parameterization has no effect on this definition.

The transformed statistic u_S satisfies the coverage requirement and is as small as possible amongst non-decreasing functions of S , proven by Jobe and David [17]. With a properly designed computational algorithm, Buehler limits can be reliably and quickly computed for sample sizes up to several hundred. Lloyd and Moldovan [18], [19] give a detailed description of the computational issues behind the Buehler algorithm and also explain the key differences between this approach and the test inversion approach. The latter is not considered in this study due to its excessive computational intensity and unresolved optimality properties.

3. NUMERICAL STUDY

We have numerically compared Buehler exact upper limits $u_S(X_1, X_2)$ based on the five introduced basic generating statistics with and without the “add-1/2” modification, ten in all. The comparison is based on computing all values of each Buehler bound for sample sizes 5, 10, 20, 50 and 100. We base our analysis on the enumeration of the entire sample space, not on simulation. It is sufficient to consider only cases where $n_1 \geq n_2$ because (n_1, n_2) and (n_2, n_1) lead to identical results. We report the results for the two best conditional and three best unconditional methods. Specifically, the results for the following five approximate limits are reported: Conditional hypergeometric $C(X_1, X_2)$, conditional hypergeometric with mid-P correction $C^*(X_1, X_2)$, modified Wald $\tilde{W}(X_1, X_2)$, score $R(X_1, X_2)$ and modified SRLR $\tilde{L}(X_1, X_2)$.

Buehler upper limits have exact coverage properties and are as small as possible subject to an ordering constraint. Amongst competing upper limits we would prefer the smallest, perhaps in some average sense. It is convenient to take the parameter of interest as $\delta = \text{expit}(\theta)$, where $\text{expit}(\theta) = (1 + e^{-\theta})^{-1}$ is the inverse of the logit transform. This parameter is bounded within $[0, 1]$. We judge the smallness of an observed upper limit $u_S(x_1, x_2)$ by

$$v_S(x_1, x_2) = \text{expit}(u_S(x_1, x_2)) - \hat{\delta},$$

where $\hat{\delta}$ is the ML estimate of δ . When $(n_1 - x_1)x_2 = 0$, both the ML estimate $\hat{\delta}$ and the corresponding upper limit equal 1. The performance of each method is measured by the average value \bar{m}_s of v_S across the sample space, excluding those uninteresting outcomes for which $\hat{\delta} = 1$. As an additional indicator of performance, the mean value surface $m_S(p_1, p_2) = E(v_S(X_1, X_2) | \hat{\delta} < 1; p_1, p_2)$ is used to identify which parameter values are associated with better or worse performance. To relate these two measures, one can think of the global measure \bar{m}_S as a summary of the surface $m_S(p_1, p_2)$.

Figure 1 plots the mean values \bar{m}_S for all five 95% Buehler limits. The two methods C and C^* based on the conditional likelihood have very similar performance and uniformly dominate

the limits based on the unconditional likelihood, the best of the latter being the modified SRLR method \tilde{L} . The dominance of the conditional methods is modest when $n_1 = n_2$, but becomes substantially more pronounced when $n_1 \neq n_2$. We have obtained very similar results for 99% upper limits.

The global measure \tilde{m}_S suppresses possibly important details in the mean surface $m_S(p_1, p_2)$. We evaluated $m_S(p_1, p_2)$ on a 100 by 100 uniform grid of values for $p_1, p_2 \in (0, 1)$. We first compare the best conditional and unconditional methods identified by the global efficiency measure \tilde{m}_S , namely Buehler limits based on the mid-P method C^* and on the modified SRLR method \tilde{L} . Table III reports the percentage of these parameter values where u_{C^*} is smaller in mean than $u_{\tilde{L}}$. For every case considered, u_{C^*} dominates $u_{\tilde{L}}$ over more than 50% of the parameter space and this dominance quickly becomes clearer for larger sample sizes, especially with $n_1 \neq n_2$. Figure 2 displays the surface $m_{\tilde{L}}(p_1, p_2) - m_{C^*}(p_1, p_2)$ for two cases, namely $(n_1, n_2) = (10, 10)$ which is one of few cases where the dominance of mid-P is less pronounced and $(n_1, n_2) = (100, 20)$ which is more typical. Even for the former case, there are no parameter values for which u_{C^*} is substantially worse in mean than $u_{\tilde{L}}$.

Table III. Relative efficiency of u_{C^*} against $u_{\tilde{L}}$. Percentage of parameter space where $m_{C^*}(p_1, p_2)$ is smaller than $m_{\tilde{L}}(p_1, p_2)$.

	$n_1 = 5$	$n_1 = 10$	$n_1 = 20$	$n_1 = 50$	$n_1 = 100$
$n_2 = 5$	55%	88%	91%	96%	97%
$n_2 = 10$	-	76%	89%	94%	96%
$n_2 = 20$	-	-	86%	92%	94%
$n_2 = 50$	-	-	-	90%	92%
$n_2 = 100$	-	-	-	-	88%

What may cause the difference in performance between C^* and \tilde{L} observed in different areas of the parameter space? In the left panel of Figure 2, the dark contour represents the efficiency border between u_{C^*} and $u_{\tilde{L}}$. Apparently, $u_{\tilde{L}}$ can have higher efficiency in the top right and bottom left corners, which correspond to both probabilities (p_1, p_2) being extreme. It is precisely under these conditions that $\text{Var}(T) = n_1 p_1 (1 - p_1) + n_2 p_2 (1 - p_2)$ becomes small. Lloyd [4], [5] has argued that conditional methods give a more correct assessment of statistical significance when the variance of the conditioning variable, in this case T , is high. In the top left and bottom right corners, one probability is close to 1 while the other is close to zero. This implies a very high probability for the data sets that lead to the highest or lowest possible values of both C^* and \tilde{L} . In this case, the Buehler limits calculated from these two designated statistics will be identical since the sets $\{C^*(X_1, X_2) \leq C^*(x_1, x_2)\}$ and $\{\tilde{L}(X_1, X_2) \leq \tilde{L}(x_1, x_2)\}$ will be identical. Thus, the relative efficiency approaches equality towards these corners.

To summarize the results of our numerical study, we found that the Buehler limit u_{C^*} shows the best performance across the vast majority of the parameter space for all considered samples sizes (n_1, n_2) . This method is based on the conditional likelihood with mid-P correction and we recommend it for practical use.

4. AN EXTENDED ILLUSTRATIVE EXAMPLE

In the introduction we gave partial results for the example of Essenberg [1]. In this section we extend this example reporting values for all five considered approximate and Buehler adjusted exact confidence limits. Table IV shows 95% and 99% approximate and exact lower and upper limits for the log-odds of mortality and Table V for the log-odds of tumour prevalence given survival, among the mice under experiment.

Table IV. Approximate and exact 95% and 99% confidence bounds on log-odds of mortality from data of Essenberg [1] using five methods. The data are $(x_1, x_2) = (13, 4)$ and $(n_1, n_2) = (36, 36)$. The tighter exact limits are in bold font.

	Thomas		Mid-P		modif. Wald		Score		modif. SRLR	
$\alpha = 0.01$	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
s	-0.044	3.348	0.066	3.167	0.009	2.837	0.096	2.897	0.086	2.990
u_s	0.071	3.204	0.077	3.128	0.052	4.467	0.077	4.467	0.077	3.629
$\alpha = 0.05$	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
s	0.333	2.803	0.460	2.618	0.423	2.423	0.490	2.520	0.466	2.492
u_s	0.401	2.677	0.476	2.619	0.401	3.357	0.479	2.855	0.476	2.855

Table V. Approximate and exact 95% and 99% confidence bounds on log-odds of tumour prevalence from data of Essenberg [1] using five methods. The data are $(x_1, x_2) = (21, 19)$ and $(n_1, n_2) = (23, 32)$. The tighter exact limits are in bold font.

	Thomas		Mid-P		modif. Wald		Score		modif. SRLR	
$\alpha = 0.01$	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
s	0.042	4.764	0.179	4.419	0.025	3.543	0.197	3.690	0.198	3.900
u_s	0.162	4.508	0.205	4.430	0.068	5.801	0.205	5.583	0.205	5.333
$\alpha = 0.05$	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
s	0.471	3.875	0.639	3.532	0.540	3.028	0.676	3.248	0.636	3.187
u_s	0.616	3.676	0.645	3.499	0.585	4.173	0.618	3.955	0.645	4.173

The tables confirm the well-known liberal tendencies of the approximate Wald, score and SRLR methods and the conservative tendency of the conditional method of Thomas. It is also apparent that the mid-P correction leads to the noticeable reduction of conservatism, though sometimes leading to liberal limits. The Buehler adjusted exact limits are broadly consistent with the results of our numerical study. The method based on the conditional likelihood with mid-P correction dominates the other methods as indicated by the closeness of confidence limit values to the corresponding ML estimates. We remind the reader that this method will not show the best performance for each and every data set. For example, for mortality, the exact score method gives the largest lower limit for $\alpha = 0.05$.

5. CONCLUSION

In the presented study, we have numerically compared ten Buehler upper confidence limits for the log-odds ratio in a 2×2 table. Four of these are generated from a designated statistics based on the conditional likelihood and the other six on the unconditional likelihood. We used average size to assess their performance across a range of sample sizes and parameter values. We found that Buehler limits based on the conditional limits have highest efficiency across the vast majority of the parameter space for all considered (n_1, n_2) , especially when the samples sizes are unequal. We suggest that this result is related to the attractive logical properties of the conditional method underlying C and C^* , see [4], which, after application of the Buehler procedure, manifest as smaller average value unconditionally.

To our knowledge, the upper limits u_C or u_{C^*} have not been proposed before. These limits retain the attractive features of conditional methods while remaining fully efficient unconditionally. Based on the results of our study, we recommend the use of u_{C^*} to practitioners. This method is based on the conditional likelihood with mid-P correction and almost uniformly dominates other methods, except when (p_1, p_2) are both close to either 0 or 1. In this case no method clearly dominates because there is little available information about θ using any method. Buehler lower confidence limits have efficiency properties absolutely identical to the upper limits examined in this paper and we provide software for the computation of both (www.mbs.edu/home/lloyd/homepage/research/bcu.html).

REFERENCES

1. Essenberg JM. Cigarette smoke and the incidence of primary neoplasm of the lung in albino mice, *Science* **116**, 561-562. textitScience 1952; **116**:561–562.
2. Thomas DG. Algorithm AS 36: Exact confidence limits for the odds ratio in a 2×2 table. *Applied Statistics* 1971; **22**:105–110.
3. Lehmann EL. *Testing statistical hypothesis*. New York: Wiley, 1959.
4. Lloyd CJ. Some issues arising from the analysis of 2×2 contingency tables. *Australian Journal of Statistics* 1988; **30**:35–46.
5. Lloyd CJ. Effective conditioning. *Australian Journal of Statistics* 1992; **34**:241–260.
6. Lancaster HO. Significance tests in discrete distributions. *Journal of the American Statistical Association* 1961; **56**:223–234.
7. Boschloo RD. Raised conditional level of significance for the 2×2 table when testing the equalities of probabilities. *Statistica Neerlandica* 1970; **24**:1–35.
8. Mehrotra DV, Chan ISF, Berger RL. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* 2003; **59**:441–450.
9. Santner TJ, Snell MK. Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. *Journal of the American Statistical Association* 1980; **75**:386–394.
10. Chan ISF, Zhang Z. Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* 1999; **55**:1202–1209.
11. Agresti A, Min Y. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* 2001; **57**:963–971.
12. Buehler RJ. Confidence intervals for the product of two binomial parameters. *Journal of the American Statistical Association* 1957; **52**:482–493.
13. Kabaila P, Lloyd CJ. Tight upper confidence limits from discrete data. *Australian Journal of Statistics* 1997; **39**:193–204.
14. Agresti A, Coull BA. Approximate better than ‘exact’ for interval estimation of binomial proportions. *The American Statistician* 1998; **52**:119–126.
15. Lloyd CJ. *Statistical Analysis of Categorical Data*. New York: Wiley, 1999.
16. Agresti A, Min Y. Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics* 2002; **3**:379–386.

17. Jobe JM, David HT. Buehler confidence bounds for reliability-maintainability measure. *Technometrics* 1992; **34**:214–222.
18. Lloyd CJ, Moldovan MV. Exact one-sided confidence limits for the difference between two correlated proportions. *Statistics in Medicine* 2007; in press.
19. Lloyd CJ, Moldovan MV. Exact one-sided confidence bounds for the risk ratio in 2×2 tables with structural zero. *Biometrical Journal* 2007; in press.

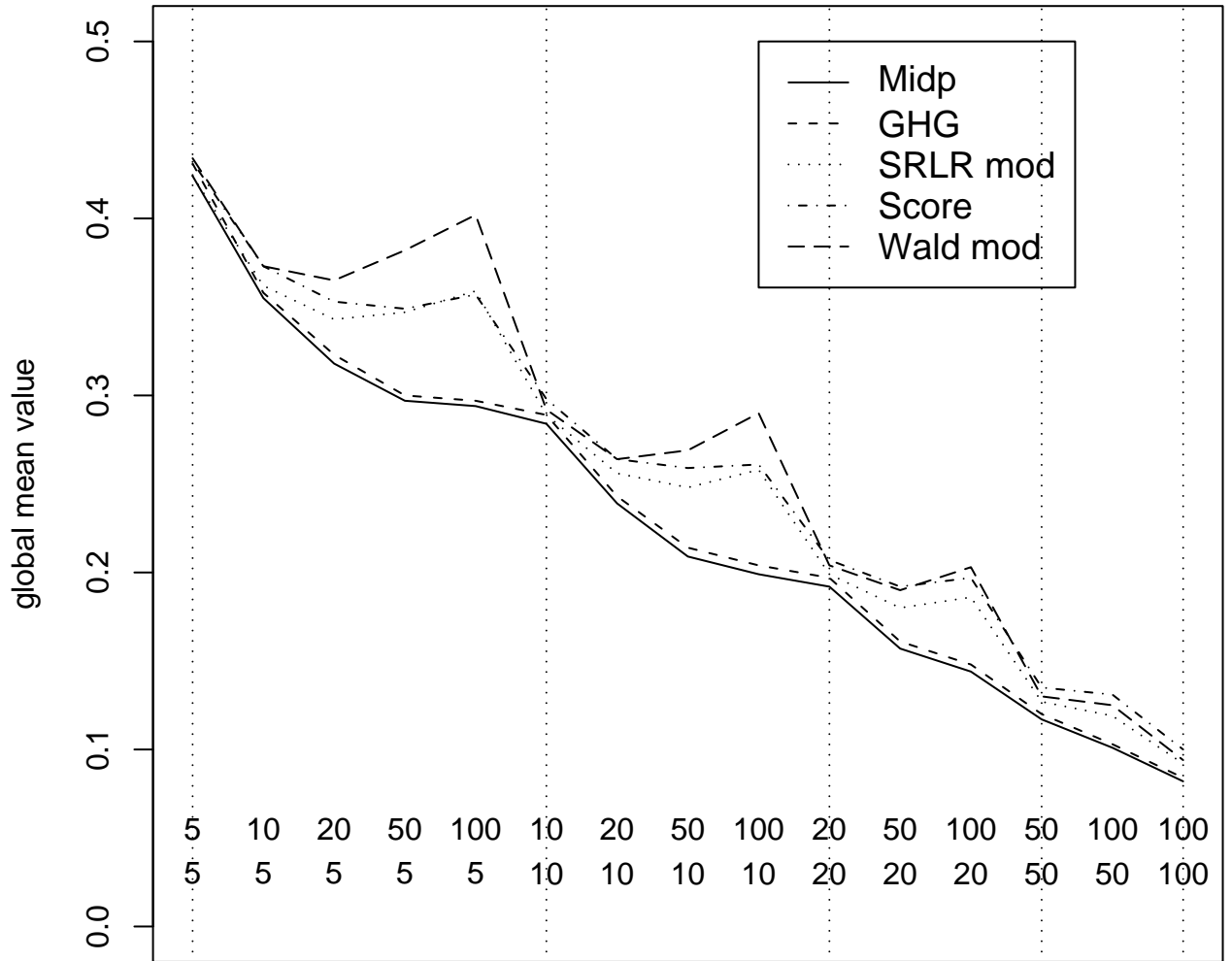


Figure 1. Global mean values \bar{m}_S of five 95% Buehler upper limits for 15 combinations of sample size n_1 (top line of values) and n_2 (bottom line of values).

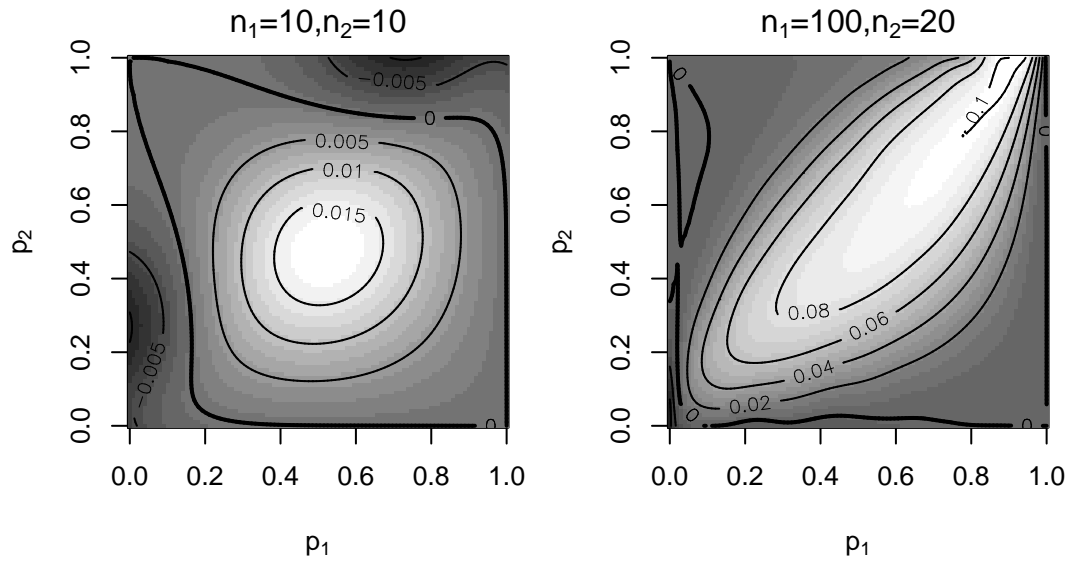


Figure 2. Contour plots of $m_{\bar{L}}(p_1, p_2) - m_{C^*}(p_1, p_2)$ for $(n_1, n_2) = (10, 10)$ and $(n_1, n_2) = (100, 20)$, for $\alpha = 0.05$.