Research Report
Statistical Research Unit
Department of Economics
University of Gothenburg
Sweden

Simple conservative confidence intervals for comparing matched proportions

Jonsson, R.

Research Report
2011:1
ISSN 0349-8034

Mailing address:
Statistical Research Unit
P.O. Box 640
SE 405 30 Göteborg
Sweden

Fax
Nat: 031-786 12 74
Int: +46 31 786 12 74

Phone
Nat: 031-786 00 00
Int: +46 31 786 00 00

Home Page:
http://www.statistics.gu.se/

Simple conservative confidence intervals for comparing matched proportions

Robert Jonsson

Department of Economics, University of Goteborg, Box 640, 405 30 Goteborg, Sweden

Summary

Unconditional confidence intervals (CIs) for the difference between marginal proportions in matched pairs data have essentially been based on improvements of Wald's large-sample statistic. The latter are approximate and non-conservative. In some situations it may be of importance that CIs are conservative, e.g. when claiming bio-equivalence in small samples. Existing methods for constructing conservative CIs are computer intensive and are not suitable for sample size determination in planned studies. This paper presents a new simple method by which conservative CIs are readily computed. The method gives CIs that are comparable with earlier conservative methods concerning coverage probabilities and lengths. However, the new method can only be used if the proportions in the discordant cells p and q satisfies $q \le 1 + p - 2\sqrt{p}$, but this is luckily the case in most applications and several examples are given. The new method is compared with previously suggested approximate and exact methods in large-scale simulations.

Key words: Binomial variables, Conservative limits, Pivotal statistic

1 Introduction

Data consisting of matched proportions in a 2 x 2 table arise in many biomedical studies. Typical examples are when measurements are made on the same patients at baseline and after a period of medical intervention, or when the effects of two drugs are compared on the same patients in a medical trial. The hypothesis of equal marginal proportions may be tested by Mc Nemar's test (Mc Nemar, 1947) or some improvement of the latter (Suissa and Shuster, 1991). If the hypothesis is rejected one may want to quantify the magnitude of the difference. Then, focus is on the construction of a confidence interval (CI) for the difference. CIs can be constructed also without first performing a test. E.g. when claiming equivalence between a drug and a reference drug it may be sufficient that the entire (two-sided) CI for the difference falls within predetermined equivalence margins (Lewis, 1999, p. 1921). Wonnacott (1987) gives an interesting discussion on the informative value of classical hypothesis test, p-value and CI and concludes that the CI conveys the most comprehensive information in the one-parameter case.

CIs for the difference can be constructed either by conditioning on the outcomes in the discordant cells, or not. Conditional methods have been shown to be no good, Tango (1998), and therefore only unconditional methods are considered in this paper. CIs can also be classified as approximate and exact. Most approximate methods have been based on asymptotic normality and these are non-conservative in the sense that coverage probabilities can be less than pre-specified nominal levels. However, sometimes exact methods are needed that guarantees that coverage probabilities are within stipulated limits. E.g. in pharmaceutical studies there may be a need to maintain a high safety protection for consumers, or in equivalence studies it may be required that conservative CIs are used. Unfortunately, the few exact mehods that have been proposed are very computer intensive (see Hseueh, Liu and Chen, 2001 and Tang, Tang and Chan, 2005) and hard to use (cf. Section 2.3 below). Since the CI has to be found numerically from each particular sample it is practically impossible to use these methods for sample size determination in a planned study. There seems to be a need for simpler alternative exact methods that are of comparably quality.

In the present paper three approximate methods, two earlier suggested exact methods and one new exact method are compared regarding coverage probability and average lengths of the CIs. The methods are described in Section 2, and in Section 3 the performance of the methods are studied in large-scale simulations. Results in the present paper are also compared with results that have been reported earlier. The paper ends with some concluding remarks.

2 Confidence intervals for the difference between marginal proportions

Consider the following frequencies and theoretical proportions (in parentheses) in the matched 2 x 2 table .

		After		
		Success	Failure	Total
Before	Success	$N_{11}\ (p_{11})$	$N_{10}\ (p)$	$N_{11}+N_{10}\ (p_{1+})$
	Failure	$N_{01}\ (q)$	$N_{00}\ (p_{00})$	$N_{01}+N_{00}\ (p_{0+})$
		$N_{11}+N_{01}\ (p_{+1})$	$N_{10}+N_{00}\ (p_{+0})$	n

Here the notations *p* and *q* are used for simplicity, instead of p_{10} and p_{01}, respectively. The object is to construct a CI for the marginal difference $\delta = p_{1+} - p_{+1} = p - q$. Such intervals can be based on the statistic $D_n / n = (N_{10} - N_{01})/n$, which is unbiased for δ . An expression for the probability function (pf) of D_n is given in Appendix (A1) and from the latter it is seen that: (i) The distribution depends only on the parameters *p* and *q*. (ii) The distribution is only symmetric if $p = q$. (iii) The pf has the property $P(D_n = d : p = a, q = b) =$

$P(D_n = -d : p = b, q = a)$. It is however far from clear how CIs can be constructed from the pd of D_n. Below some approximate CIs of the Wald-type and exact CIs are considered.

2.1 Approximate intervals of the Wald type

All Wald type CIs are based on the fact that the standardized statistic

$$Z_n = \frac{D_n / n - \delta}{\sqrt{V(D_n / n)}}, \text{where} V(D_n / n) = (p + q - \delta^2) / n \tag{1}$$

has a standard normal distribution as $n \to \infty$. However, for small *n* the distribution of Z_n is heavily dependent on *p* and *q*. To demonstrate this, let z_1 be the largest value for which $P(Z_n < z_1) < \alpha / 2$ and let z_2 be the smallest value for which $P(Z_n > z_2) < \alpha / 2$, so

$$P(z_1 \le Z_n \le z_2) \ge 1 - \alpha \tag{2}$$

The percentiles z_1 and z_2 may change substantially even for small variations in *p* and *q* and are far from those of the standard normal distribution. When $\alpha = 0.05$ and *n* = 10 one gets $(z_1 = -2.32, z_2 = 1.66)$ for *p* = 0.05 and *q* = 0.20 and $(z_1 = -1.86, z_2 = 1.86)$ for *p* = 0.05 and *q* = 0.25, the latter values being calculated from the exact distribution of D_n in the Appendix (A1). The statistic in (1) can not be used directly for constructing CIs for δ, but it is the basis for various approaches.

The most radical way to get rid of the nuicance parameters *p* and *q* in (1) is to replace the variance in the denominator by an estimator. In this way one gets a statistic $\hat{Z}_n$ which only depends on the parameter δ. The inequality in (2) can now be inverted to get a CI for δ. From Slutsky's Theorem (Casella and Berger, 1990, p. 220) it follows that also $\hat{Z}_n$ has a standard normal distribution for large *n*, but the convergence goes slower than for Z_n. The 95 % CIs for δ obtained in this way are (cf. Agresti and Min, 2005, and Tang, Tang and Chan, 2005)

$$D_n / n \pm 1.96\sqrt{\hat{V}_n}\text{, or}$$
$$D_n / n \pm \left(1.96\sqrt{\hat{V}_n} + \frac{1}{n}\right)\text{, where } \hat{V}_n = \left[(N_{10} + N_{01})/n - (D_n/n)^2\right]/n \qquad (3)$$

The first of these CIs is denoted *Wald* and the second *Waldcc* (with correction for continuity). A problem with (3) is that simulations (not presented here) show that the distribution of $\hat{Z}_n$ in small samples is even more dependent on *p* and *q* than the distribution of Z_n, so the use of the percentile 1.96 may be put in question. Anyhow, several studies have shown that *Waldcc* yields a higher degree of conservative CIs in small samples than *Wald* (see e.g. May and Johnson, 1997) and therefore only the former method is considered in the sequel.

Rather than estimating all parameters of the variance in (1) one may only estimate $p+q$, so the variance in the denominator is replaced by $\tilde{V}_n = \left[(N_{10} + N_{01})/n - \delta^2\right]/n$. A CI for δ is obtained from the set of δ-values that satisfies $\left\{\delta : z_1 \leq \tilde{Z}_n \leq z_2\right\}$, where $\tilde{Z}_n = (D_n / n - \delta)/\sqrt{\tilde{V}_n}$. The CI limits for δ are then found as the roots of a quadratic function in δ, see May and Johnson, 1997. Despite the intuitively appealing idea of this method that reduces the number of parameters to be estimated, it was concluded by May and Johnson (1997) that there was no clear choice between this method and *Waldcc* regarding coverage probabilities. One reason for this may be that a large variance in the denominator is likely to increase the length of the CI, and in the Appendix (A2) a proof is given for the rather unexpected result that the variance of $\hat{V}_n$ is smaller than the variance of $\tilde{V}_n$ provided that

$$q < 1 + p - 2\sqrt{p} \qquad (4)$$

This region is depicted in Figure 1 in Section 2.2 in a different context, where it is furthermore demonstrated that in many applications the (p,q)-values are found within this region. The fact that CIs based on $\tilde{V}_n$ are less reliable than those based on $\hat{V}_n$ within the region defined by (4), was furthermore confirmed by simulations (not shown in this paper). Therefore this method is not considered further.

Another way of improving *Wald* was suggested by Agresti and Min (2005).The frequencies N_{10}, N_{01} and *n* were replaced by $N_{10}^* = N_{10} + N/4, N_{01}^* = N_{01} + N/4$ and $n^* = n + N/4$, respectively. The choice *N* = 2 turned out to give the best coverage performance and this was also supported by Bayesian arguments. Let D_n^* and $\hat{V}^*$ denote the statistics based on the new quantities. Then the *Wald+2* CI for δ is

$$D_n^* / n \pm 1.96\sqrt{\hat{V}_n^*} \tag{5}$$

It was noticed above that the percentiles of the *Z*-statistics were far from those given by the standard normal distribution in small samples. One way of improving *Wald* might therefore be to replace 1.96 by percentiles that are closer to the actual percentiles. To this end simulations were performed in order to study how the 2.5 % and 97.5 % percentiles of $\hat{Z}_n$ varied for *n* =10, 25, 50, 100 and *p*, *q* = 0.05,…(0.05)…,0.70, subject to $p + q < 1$. For each value of *n, p, q* a simulation with 50,000 replicates was performed. The distribution of $\hat{Z}_n$ was mostly skew with exception for the case $p = q$ when it was symmetric. It was furthermore found that the variance of $\hat{Z}_n$ was constantly larger than 1 and increased linearly with $p + q$. Table 1 summarizes some characteristics of the percentiles. The absolute percentiles decreased with increasing *n* but were far above 1.96. By fitting a model, 'largest absolute mean percentile' = $z_{ADJ} = a \cdot b^n$, to the four positive means in Table 1 one gets (the coefficient of determination being 98.5 % for the linearized model)

Table 1 Mean, standard deviation (std) and range of the percentiles z_1 and z_2 when *p* and *q* vary between 0.05 and 0.70.

	z_1		z_2	
n	Mean (std)	Range	Mean (std)	Range
10	-2.14 (0.36)	-3.00, -1.50	2.14 (0.34)	1.50, 3.00
25	-2.11 (0.22)	-2.51, -1.63	2.10 (0.25)	1.63, 2.80
50	-2.03 (0.12)	-2.25, -1.83	2.03 (0.11)	1.84, 2.25
100	-1.99 (0.07)	-2.09, -1.87	1.98 (0.06)	1.87, 2.11

$$z_{ADJ} = 2.32 \cdot n^{-1/30}, 10 \le n \le 100 \tag{6}$$

An alternative approximate CI where the percentile has been adjusted, *Waldadj*, is thus obtained by

$$D_n / n \pm z_{ADJ}\sqrt{\hat{V}_n} \tag{7}$$

Several other approximate unconditional methods have been suggested. E.g. Tango (1998) introduced a score method based on a statistic similar to $\tilde{Z}_n$, but where *p* and *q* are replaced by ML estimators subject to the restriction that $p - q = \delta$. The score method was compared with *Wald+2* and it was concluded that the methods are comparable regarding coverage probabilities, Agresti and Min (2005). There are also approximate methods based on the trinomial distribution with estimates inserted for the parameters, Newcombe (1998) and Tang et al. (2005). These require more heavy computations and do not seem to perform substantially better than *Waldcc* and *Wald+2*, although it is hard to draw definite conclusions from the small-scale simulations that have been reported earlier

2.2 Conservative intervals based on a transformation

Introduce new parameters p_1 and p_2 by putting $p = p_1 p_2$ and $q = (1 - p_1)(1 - p_2)$. Then it is shown in the Appendix (A3) that the following holds:

Lemma D_n can be expressed as $A_n + B_n - n$ where A_n and B_n are independent binomial variables such that A_n is $B(n, p_1)$ and B_n is $B(n, p_2)$.

Solving for p_1 yields $p_1 = \left(1 + p - q \pm \sqrt{(1 + p - q)^2 - 4p}\right)/2$ and the requirement that p_1 and p_2 are real-valued leads to the condition in (4), but without strict inequality. The admissible (*p*,*q*)-area is shown in Figure 1

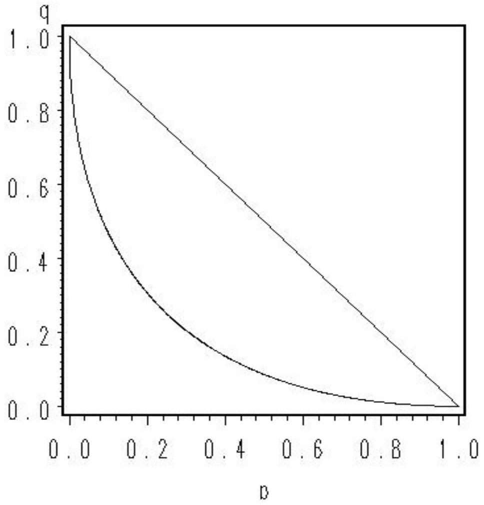

Fig. 1 Plot of the admissible area $q \le \max q = 1 + p - 2\sqrt{p}$ compared with the maximal (*p*,*q*)-area $q \le 1 - p$.

Notice that there is no need to specify an upper limit also for *p*, $p \le 1 + q - 2\sqrt{q}$, since the area defined by the latter is easily seen to be identical with the former admissible area.

In practice one may want to estimate the maximal admissible value of *q*. This can be done by simply insert the *p*-estimate yielding max $\hat{q}_1$, say. Since $\sqrt{\hat{p}}$ is biased for $\sqrt{p}$ it is better to replace $\sqrt{\hat{p}}$ by $\sqrt{\hat{p}} + (1 - \hat{p}) / 8n\sqrt{\hat{p}}$, obtained from a Taylor approximation. The upper limit of *q* estimated in this way is denoted max $\hat{q}_2$. Some examples are quoted in Table 2, where it is seen that all *q*-estimates are well inside the admissible limit.

Table 2 Cell frequencies in six data sets together with *q*-estimates and maximal admissible values. For the meaning of the last two columns see text.

Source	N_{11}	N_{10}	N_{01}	N_{00}	n	$\hat{q}$	max $\hat{q}_1$	max $\hat{q}_2$
Jones and Kenward (1987)	53	8	16	9	86	0.093	0.323	0.318
Ward et al. (2000)	8	3	1	2	14	0.071	0.288	0.258
Kao et al. (2002)	22	2	0	1	25	0	0.514	0.482
Hsueh et al. (2001)	39	5	4	2	50	0.080	0.468	0.453
Karacan (1976)	4	9	3	16	32	0.094	0.221	0.210
Elston and Johnson (1984)	21	17	37	105	180	0.206	0.480	0.476

Since $p - q = p_1 + p_2 - 1$ the initial problem of finding a CI for $p - q$ has been turned into the problem of finding a CI for $p_1 + p_2$ (essentially), based on the distribution $P(A_n + B_n \le x) = \sum_i P(A_n = i) P(B_n \le x - i)$. The latter can in principle be obtained by first generating a sequence of largest lower and smallest upper points (x_L, x_U) such that $P(x_L \le (A_n + B_n) / n \le x_U) \ge 1 - \alpha$ for all possible values of $p_1 + p_2$, thereby creating a confidence contour (cf. Figure 1). The CI produced by the estimate $(A_n + B_n) / n = y$ is then

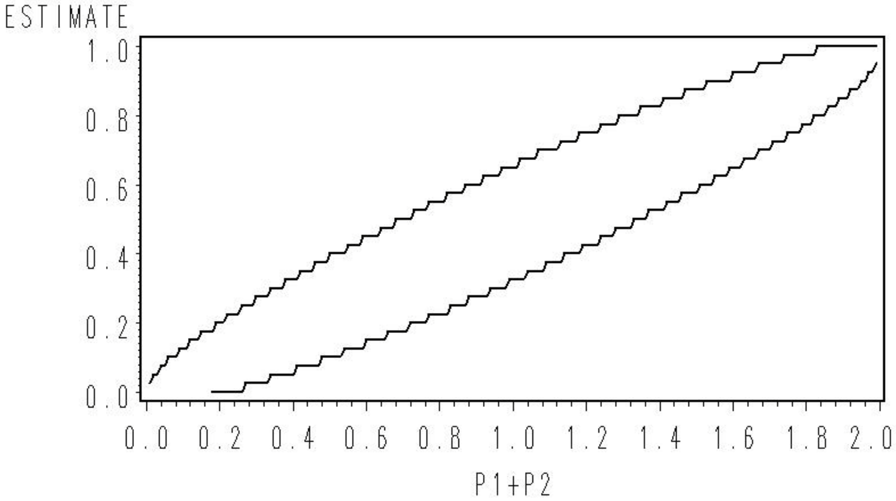

Fig. 2 Confidence contour obtained for $p_1 = p_2$ and $n = 20$

determined by $\left(x_U^{-1}(y), x_L^{-1}(y)\right)$ (cf. Casella and Berger, 1990, p. 420 and Stuart, Ord and Arnold, 1999, p.122). This approach would require a step-wise search over all $p_1 + p_2$ that would not be less computer insive than existing methods, see e.g. Hsueh et al., 2001. However, if $p_1 = p_2 = p_0$ say, then $p - q = 2p_0 - 1$and $A_n + B_n = D_n + n$is distributed $B(2n, p_0)$. Exact conservative CIs for $p_0$can now easily be obtained by using the well-known relation between the binomial and F distributions (Jowett, 1963 and Casella and Berger, 1990, p. 449). It follows that a conservative $100(1-\alpha)\%$ CI for δ is obtained from

$$\left(2\hat{p}_0(L) - 1, 2\hat{p}_0(U) - 1\right), \text{ where}$$

$$\hat{p}_0(L) = \frac{s}{s + (2n - s + 1)F_{1-\frac{\alpha}{2}}\left(2(2n - s + 1), 2s\right)}, \; \hat{p}_0(U) = \frac{(s+1)F_{1-\frac{\alpha}{2}}\left(2(s+1), 2(2n - s)\right)}{2n - s + (s+1)F_{1-\frac{\alpha}{2}}\left(2(s+1), 2(2n - s)\right)}$$

and where s is an observed value of $N_{10} - N_{01} + n.$

(8)

In (8) $F_{1-\frac{\alpha}{2}}(f_1, f_2)$ denotes the $100(1-\alpha/2)\%$ percentile of the *F*-distribution with f_1 and f_2 degrees of freedom. The expression in (8) does not cover the cases $s = 0$ and $s = 2n$. In the former case the lower end-point is put equal to 0 and in the latter case the upper end-point is put equal to 1 (Casella and Berger, 1990, p. 449). This method for constructing CIs is denoted *Trans*.

The CI in (8) is based on the assumption that $p_1 = p_2 = p_0$, but which are the properties of the CIs when $p_1 \neq p_2$? To study this, 95 % CIs for δ were simulated with $p_1, p_2 =$ 0.1,…(0.1)…0.9, n = 10 and with 50,000 replicates for each value of p_1, p_2 and n. The results are summarized in Table 3a (coverage probabilities) and Table 3b (average lengths). Both tables are symmetric around $p_1 = p_2$ with few exceptions due to random deviations since 50,000 replicates are not sufficient to reach stability in all three figures after the decimal point (cf. Section 3.1). All coverage probabilities are above the stipulated level 95 %. However, when $|p_1 - p_2| = |h|$ is large the coverage probabilities tend to be very large, indicating over-conservativeness. Also the lengths tend to increase as $|h|$ increases. CIs determined by (8) will thus perform well provided that $|h|$ is not too large, but it is hard to determine how likely this is without making further distributional assumptions. Assume e.g. a uniform distribution of (p, q) over the admissible region in Figure 1. Then it is easily shown that in the present example h has a probability function given by $p(h) = (9 + 10h)/81, h = -0.8,...,-0.1, 0$ and $p(h) = (9 - 10h)/81, h = 0, 0.1,..., 0.8$. From the latter it is seen that large values of $|h|$ are less probable.

Table 3a Coverage probabilities (%) obtained by using (8) for various p_1 and p_2 .

p_1	$p_2 =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1		98.9	97.9	98.2	97.3	98.8	98.8	99.2	99.6	99.8
0.2		98.0	97.8	96.4	97.9	97.6	97.8	98.3	99.0	99.6
0.3		98.2	96.2	97.6	97.0	96.5	96.9	97.4	98.3	99.1
0.4		97.3	97.9	96.9	96.3	96.2	96.2	96.8	97.8	98.8
0.5		98.8	97.7	96.8	96.0	95.7	96.0	96.7	97.5	98.8
0.6		98.9	97.8	96.9	96.3	96.0	96.3	97.0	97.9	97.3
0.7		99.2	98.3	97.5	96.9	96.7	96.9	97.6	96.2	98.3
0.8		99.5	98.8	98.3	97.8	97.7	97.9	96.3	97.7	98.0
0.9		99.9	99.5	99.1	98.8	98.8	97.4	98.1	97.9	98.9

Table 3b Average lengths obtained by using (8) for various p_1 and p_2.

p_1	$p_2 =$ 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.584	0.668	0.736	0.788	0.830	0.862	0.884	0.898	0.905
0.2	0.669	0.736	0.786	0.827	0.853	0.880	0.893	0.890	0.898
0.3	0.736	0.786	0.827	0.855	0.877	0.890	0.895	0.893	0.884
0.4	0.789	0.827	0.858	0.876	0.888	0.893	0.890	0.879	0.861
0.5	0.830	0.858	0.880	0.888	0.892	0.888	0.877	0.858	0.830
0.6	0.862	0.880	0.893	0.893	0.888	0.876	0.856	0.826	0.789
0.7	0.884	0.893	0.890	0.890	0.877	0.855	0.826	0.786	0.736
0.8	0.898	0.899	0.880	0.880	0.857	0.827	0.785	0.735	0.669
0.9	0.905	0.848	0.862	0.862	0.830	0.789	0.735	0.669	0.583

(8) can only be used if p and q falls within the admissible area, otherwise there is no guarantee that the CI is conservative. To illustrate this consider the case $p = 0.50$ and $q = 0.10$. Here q is larger than the upper limit $\max q = 1 + 0.50 - 2\sqrt{0.50} = 0.0858$, but close to it. Simulation with 50,000 replicates and $n = 100$ yielded a 95 % coverage probability of just 94.8. When the (p, q)-values are far from the admissible area the coverage probability can be much smaller.

Since the CI in (8) is given in closed form it is readily calculated. It can also be used to determine the sample size needed to obtain a CI of desired length, either from a pilot study or from reasonable assumptions about the magnitude of s (cf. Altman, 1990, p. 160), the choice $s = n$ yielding the widest interval. As an example consider the data from Ward et al. in Table 2 where n = 14. The observed difference between the marginal proportions is $\hat{\delta} = (3-1)/14$ = 0.093 and by using (8) with s = 3-1+14 a 95 % CI for δ is (- 0.256, 0.511), indicating non-significant difference from zero at the 5 % level. Assume that the data is the outcome of a pilot study used for determining the sample size in a final study. Two extreme cases are $s = n$ (corresponding to $\hat{\delta} = 0$) and $s = 1.5n$ (corresponding to $\hat{\delta} = 0.5$). By using (8) one may study how the length of 95 % and 90 % CIs depend on n. This is demonstrated in Figure 3 from which several conclusions can be drawn, e.g. that a sample size of about 50 is likely to yield a 95 % CI that is half of the one obtained with n = 14.

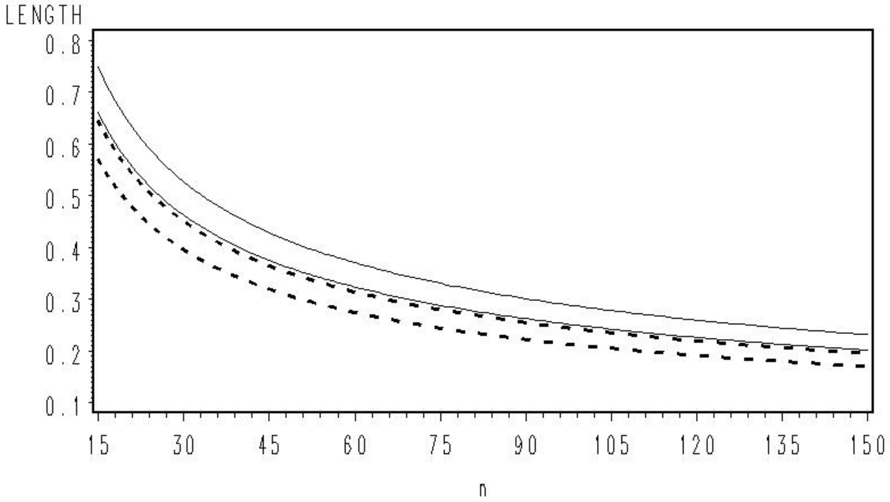

Fig. 3 Length of 95 % CIs (filled lines) and 90 % CIs (dotted lines) plotted aginst sample size (n). The upper of each of the two lines is obtained with s = n and the lower with s = 1.5n.

2.3 Some previously suggested exact conservative intervals

Hsueh et al. (2001) suggested a method that is based on inverting two one-sided score test statistics. By using the correspondence between hypothesis testing and CI, CIs can be constructed that are conservative. However, the CI limits have to be found by numerical calculations of trinomial tail probabilities and this makes the procedure extremely computer intensive. It was reported that a computer time of about 200 minutes was needed in order to construct a CI from a particular sample with $n = 50$. A similar conservative procedure was suggested by Tang et al. (2005), the latter being based on inverting one two-sided score test statistic. These two methods, denoted *EUM_1* and *EUM_2*, respectively, were compared in the latter article and it was found that the coverage probabilities of *EUM_1* never were smaller than those of *EUM_2,* and that *EUM_1* yielded wider CIs. However, by using these methods it would not be feasible to determine the sample size in planned studies.

3 Simulation results

In this section the three approximate methods *Waldcc*, *Wald+2*, *Waldadj* and the conservative method *Trans* are compared in large-scale simulations. The results are then compared with those obtained in previous studies. The method *Trans* is also compared with the two conservative methods mentioned above. Since it is important that the simulation results are reliable a first section is devoted to design considerations.

3.1 Design of the simulation study

Two properties of the CIs were evaluated, coverage probability and average length. It was the aim to compare results from the present study with previous ones, but this turned out to be troublesome for several reasons. The outcome of a simulation is determined by choice of sample size *n*, choice of parameters *p* and *q*, and choice of the number of replicates in each simulation. Choice of *n* was the least problem since many of the previous studies use *n* = 10, 25 and 100. Choice of *p* and *q* was more cumbersome since the latter are seldom reported, but only the value of $\delta = p - q$. The number of replicates used in each simulation has in earlier studies varied between 100 and 10,000, but the latter numbers were found to be too small, especially for estimating coverage probabilities. The reliability when estimating coverage probabilities and lengths are illustrated in Table 4 with 1000 and 50,000 replicates. From the table it is concluded that even 50,000 replicates are not sufficient for reaching three stable figures after the decimal point.

Table 4 Variation limits of estimates computed from 10 simulations each with 1000 and 50,000 replicates and with n = 10, $p = 0.25 = q$

	Coverage probability of 95 % CI		Average length	
Method	1000 replicates	50,000 replicates	1000 replicates	50,000 replicates
Waldcc	96.5 - 97.8	97.2 - 97.3	1.010 - 1.026	1.015 - 1.017
Wald+2	92.5 - 97.8	93.2 - 93.4	0.759 - 0.766	0.763 - 0.764
Waldadj	92.5 - 94.4	93.2 - 93.4	0.932 - 0.952	0.942 - 0.944
Trans	95.0 - 96.7	95.6 - 95.8	0.890 - 0.893	0.892

Based on these considerations it was decided to use n = 10, 25, 50, 100 and δ =0, 0.2, 0.4, 0.6, 0.8 with various p and q inside the admissible region. (For δ = 0.6 and 0.8 just one combination of p and q was used since the admissible region is very narrow.) When comparing *Trans* with the exact methods *EUM_1* and *EUM_2* the case δ =0.3 was furthermore considered. Only positive values of δ were used since negative values yielded the same coverage probabilities and lengths. Simulations, each with 50,000 replicates, were performed sequentially in steps until three stable figures after the decimal point of the sequential averages was reached and it turned out that 2-4 steps were needed. All simulations were based on random number functions in SAS version 9.1. A computer program is available from the author on request.

3.2 Results

The performance of the four methods is summarized in Table 5, from which the following conclusions are drawn.

Waldcc: For $n \geq 25$ all coverage probabilities were above 95 %. CIs were generally wider than those obtained by the other approximate methods and occasionally even wider than those obtained by the conservative method *Trans*. The method seems thus to be reliable but yields wide CIs.

Wald+2: Rather unexpectedly, this method was more reliable for the smallest sample size n = 10, in which case 5 out of 8 coverage probabilities in the table were above 95 %. For larger n the reliability was lower, even if the coverage probabilities were just below 95 %. On the other hand this method produced the shortest CIs among the compared methods, with few exceptions.

Waldadj: The method was extremily poor for n = 10, where only 1 out of 8 coverage probabilities were acceptable. For larger n it was slightly more reliable than *Wald+2* but the latter yielded shorterCIs. A conclusion is that very little is gained by trying to adjust the percentiles in small samples by means of an adjustment to the mean percentiles.

Trans: The method works well as far as $p_1$does not deviates too much from p_2. E.g when $p = 0.25$ and $q = 0.25$ (corresponding to $p_1 = p_2 = 0.50$) the lengths are in most cases shorter than those of the approximate cases. On the other hand, when $p = 0.05$ and $q = 0.05$ (corresponding to $p_1 = 0.05$ and $p_2 = 0.95$) the method yields coverage probabilities of 100 % or just below and the price that has to be paid for this are wide CIs.

Table 5 Coverage probabilities of the 95 % CIs and average lengths of four methods when *p, q* and *n* are varying. Bold italic figures signals that coverage probability is below 95 %.

				Coverage probability				Average length			
p	*q*	*δ*	Method	*n* =10	*n*=25	*n*=50	*n*=100	*n*=10	*n*=25	*n*=50	*n*=100
0.05	0.05	0	*Waldcc*	100.0	99.5	99.1	97.9	0.493	0.307	0.209	0.142
			Wald+2	99.8	99.5	97.5	95.9	0.441	0.262	0.180	0.126
			Waldadj	***65.0***	***90.7***	***94.8***	95.8	***0.338***	***0.245***	***0.177***	0.124
			Trans	100.0	100.0	100.0	100.0	0.908	0.578	0.406	0.285
0.25	0.25	0	*Waldcc*	97.2	96.1	96.3	96.0	1.018	0.620	0.427	0.295
			Wald+2	***93.4***	***94.3***	***94.5***	***94.6***	***0.765***	***0.521***	***0.380***	***0.273***
			Waldadj	***93.4***	95.4	95.3	95.0	***0.945***	0.586	0.406	0.281
			Trans	95.8	96.5	96.4	95.9	0.892	0.573	0.404	0.283
0.21	0.01	0.2	*Waldcc*	***90.4***	97.0	97.0	96.3	***0.668***	0.401	0.271	0.185
			Wald+2	***90.4***	***91.4***	***94.1***	95.0	***0.530***	***0.335***	***0.236***	0.167
			Waldadj	***89.7***	***91.4***	***94.5***	95.0	***0.541***	***0.349***	***0.243***	0.168
			Trans	99.9	99.9	99.9	99.9	0.889	0.556	0.398	0.280
0.36	0.16	0.2	*Waldcc*	96.1	96.4	96.4	96.1	0.995	0.609	0.419	0.290
			Wald+2	***94.3***	***94.8***	***94.7***	***94.8***	***0.755***	***0.514***	***0.373***	***0.268***
			Waldadj	***93.7***	95.4	95.5	95.0	***0.921***	0.572	0.398	0.275
			Trans	96.4	97.1	95.9	95.7	0.876	0.563	0.396	0.279
0.41	0.01	0.4	*Waldcc*	***94.5***	96.0	96.9	96.4	***0.786***	0.470	0.319	0.219
			Wald+2	95.6	***93.2***	***94.7***	***94.3***	0.621	***0.398***	***0.282***	***0.200***
			Waldadj	***94.5***	95.8	95.9	***94.6***	***0.682***	0.424	0.293	***0.203***
			Trans	99.7	99.3	99.1	98.9	0.833	0.531	0.374	0.262
0.49	0.09	0.4	*Waldcc*	95.0	96.3	96.0	96.0	0.927	0.574	0.394	0.272
			Wald+2	95.5	95.3	95.1	***94.8***	0.732	0.488	0.352	***0.251***
			Waldadj	95.0	***94.5***	95.0	95.0	0.852	***0.536***	0.372	0.258
			Trans	97.4	97.1	96.2	95.5	0.826	0.529	0.373	0.262
0.64	0.04	0.6	*Waldcc*	***92.7***	95.0	96.7	96.1	***0.781***	0.509	0.349	0.240
			Wald+2	96.0	***94.5***	96.2	***94.9***	0.886	***0.442***	0.313	***0.222***
			Waldadj	92.7	***94.4***	95.3	***94.5***	0.710	***0.486***	0.324	***0.224***
			Trans	97.9	96.6	96.8	96.0	0.734	0.467	0.328	0.230
0.81	0.01	0.8	*Waldcc*	***87.7***	96.3	97.1	96.1	***0.524***	0.377	0.270	0.184
			Wald+2	97.4	95.9	95.8	95.7	0.601	0.367	0.249	0.171
			Waldadj	***87.4***	***88.8***	***93.6***	***93.8***	***0.455***	***0.337***	***0.241***	***0.168***
			Trans	98.9	97.0	95.5	96.7	0.584	0.361	0.251	0.175

As expected, the lengths of the CIs decreased with increasing n for all methods. A similar, but less apparent pattern, is seen for the coverage probabilities, which tend to approach 95 % as n increases. From the table it is evident that, when different methods are to be compared, it is not enough to just study the performance of the methods for various $\delta = p - q$ without taking account of both p and q.

3.3 Comparison with previous studies

A large number of comparative studies have been published on the issue, especially on the performance of approximate methods. It is beyond the scope of this article to review all of these, so below just a few are reviewed that seems to be relevant for this study. The findings in these are then contrasted with the results in the preceding section. First approximate methods are considered and then exact conservative methods.

May and Johnson (1997) compared *Waldcc*, *Wald* and the method based on the statistic $\tilde{Z}_n$ (cf. Section 2.1). Here n = 50, 75, 100, 500 and the discordant cell proportions p and q were chosen such that p+q ranged from 0.055 to 0.1040. The number of replicates was 10,000. It was concluded that *Waldcc* performed better than *Wald*, but there was no clear choice between *Waldcc* and the method based on $\tilde{Z}_n$ regarding coverage probabilities. However the CIs obtained with *Waldcc* were wider.

Tango (1998) compared *Wald* with a proposed score method. In this study n =30, 50, 80, p = 0, 0.05, 0.10, 0.20 and two values of q were chosen such that $p - q$ (Δ in their notation) = 0 and 0.1. Each simulation was performed with just 1000 replicates (in contrast to 10,000 replicates that was used to study the power of the corresponding tests.). From the table on p. 902 in the latter paper it is evident that the score method is more reliable than *Wald*. A curious pattern, that is not commented in the paper, is that the score method seems to be less reliable for the largest sample size n = 80.

Agresti and Min (2005) compared *Wald*, *Wald+2* and the score method of Tango. The used n = 25 and varied the marginal proportions p_{1+} and p_{+1} (not p and q) such that δ = 0 and 0.1. The number of replicates in the simulations was not reported. The conclusion was that *Wald+2* performed better than *Wald* and that the coverage probabilities obtained with *Wald+2* were comparable with those obtained with the scoring method of Tango, but yielding wider CIs.

Tang et al. (2005) investigated the performance of five approximate methods, including *Waldcc* and the scoring method of Tango. Sample sizes were chosen as n = 7, 10, 15, 20, 25, 30 and δ = 0, 0.3, 0.6, 0.95, without reporting what values of p and q that where used. In the table on p. 3574 it is seen that *Waldcc* yielded coverage probabilities below 95 % in 8 out of 24 cases, whereas the same figures for the score method was 6 out of 24 cases. The CIs produced by *Waldcc* were however wider.

Results from several other simulations have been reported but they are hard to compare with those above. E.g. Newcombe (1998) studied coverage probabilities of ten unconditional methods. 100 triplets of three functions of (p_{11}, p, q) were chosen from uniformly distributed random number, using n =10, 11,...,100. Coverage probabilities were then calculated from the 100 x 91 =9100 outcomes. Here it is hard to draw conclusions about the effect of *n, p* and *q* upon coverage probability and length.

Previous studies seem to confirm that *Waldcc* and *Wald+2* have about the same reliability. This is not in accordance with the results in Table 5 which clearly shows that *Waldcc* is more reliable than *Wald+2* for $n \geq 25$. Earlier studies have also demonstrated that *Waldcc* yields wider CIs than *Wald+2* and this agrees with the results in Table 5.

Now, consider the conservative methods *EUM_1*, *EUM_2* and the new method *Trans*. The performance of these methods are summarized in Table 6. Since no values of *p* and *q* were reported in Tang et al. (2005) but only of δ, it is hard to draw any definite conclusions about the merits of the methods. However, from the table it is seen that the method *Trans* is comparable with the other methods. The coverage probabilities obtained by *Trans* are found between those obtained by the other methods, with exception for the case $p = q = 0.05$ which yields over-conservativeness (cf . Section 3.2). Also, the lengths of the CIs obtained by *Trans* are not generally larger than those obtained by the other methods.

Table 6 Comparison between three exact methods for constructing CIs. Figures for *EUM_1* and *EUM-2* are quoted from tables IV and VI in Tang et al. (2005). Figures for *Trans* are taken from Table 5. When $\delta = 0$ two figures are shown, one with $p = q = 0.25$ marked with *, and one with $p = q = 0.05$ marked with **. Figures for Trans when $\delta = 0.3$ are obtained with $p = 0{,}42, q = 0.12$.

		Coverage probability				Average length			
n	δ	*EUM_1*	*EUM_2*	*Trans*		*EUM_1*	*EUM_2*	*Trans*	
10	0	97.838	97.838	95,8*	100.0**	0.911	0.823	0.892*	0.908**
10	0.3	98.367	96.460	96.9		0.932	0.865	0.855	
10	0.6	98.897	96.577	97.8		0.863	0.831	0.734	
25	0	96.769	95.936	96.5*	100.0**	0.557	0.537	0.573*	0.578**
25	0.3	96.922	95.541	96.4		0.589	0.562	0.549	
25	0.6	97.118	96.391	96.6		0.536	0.512	0.467	

4 Conclusions and suggestions for further studies

In a first round three approximate methods and one conservative method to construct CIs for the difference between marginal proportions were compared. The approximate methods were based on improvements of Wald’s large-sample statistic. Of these, Wald’s method with continuity correction (*Walcc*) was found to be more reliable than methods that either adjust the percentiles (*Waldadj*) or the standard error (*Wald+2*), but *Waldcc* yielded wider CIs. *Waldadj* was based on an adjustment of the percentiles to the actual mean percentiles, but other types of adjustments may be taken into consideration, e.g. adjustment to the actual median percentiles. Also combinations of the methods may be worth considering, so that the high reliability of *Waldcc* is maintained while the length of the CI is reduced. One argument for using CIs of the Wald type being simplicity and the possibility to determine sample sizes in planned studies. However, CIs produced by the new method *Trans* has the same properties but having the advantage of being conservative provided that *p* and *q* are in the admissible region. *Trans* was shown to have coverage probabilities and lengths that were comparable with those obtained by the much more labouring exact methods *EUM_1* and *EUM_2*. However, the comparisons were made for just a few *p*,*q*-values and a more extensive study is required to reach any definite conclusions.

Appendix

(A1) The probability function (pf) of $\mathbf{D_n}$

Let $p(x,y) = \dfrac{n!}{x!y!(n-x-y)!} p^x q^y r^{n-x-y}$, where $r = 1-p-q$, be the pf of (N_{10}, N_{01}). Then, for $j = 0,1,...,n$, $P(D_n = n-j)$ is obtained by summing $p(x,y)$ over $\{(x,y): x-y = n-j\}$. In this way one gets the expressions

$$P(D_n = n-j) = \sum_{x=j/2}^{j} C(n,x,j) p^{n-x} q^{j-x} r^{2x-j} \ (j \text{ even}), = \sum_{x=(j+1)/2}^{j} C(n,x,j) p^{n-x} q^{j-x} r^{2x-j} \ (j \text{ odd}),$$

where $C(n,x,j) = \binom{n}{x}\binom{x}{j-x}$. Similarly,

$$P(D_n = j-n) = \sum_{x=j/2}^{j} C(n,x,j) p^{j-x} q^{n-x} r^{2x-j} \ (j \text{ even}), = \sum_{x=(j+1)/2}^{j} C(n,x,j) p^{j-x} q^{n-x} r^{2x-j} \ (j \text{ odd})$$

(A2) $V(\hat{p}+\hat{q}-\hat{\delta}^2) < V(\hat{p}+\hat{q}-\delta^2)$ for $q < 1 + p - 2\sqrt{p}$.

Proof $V(\hat{p}+\hat{q}-\delta^2) = \dfrac{p(1-p)}{n} + \dfrac{q(1-q)}{n} - 2\dfrac{pq}{n} = \dfrac{1}{n}(p+q)[1-(p+q)]$. The variance on the left hand side is more complicated, so a Taylor approximation is made. Put $g = p+q-(p-q)^2$ with $D_p g = 1-2(p-q)$ and $D_q g = 1+2(p-q)$. Then $V(\hat{p}+\hat{q}-\hat{\delta}^2) \approx$

$$(D_p g)^2 V(\hat{p}) + (D_q g)^2 V(\hat{q}) + 2(D_p g)(D_q g) Cov(\hat{p},\hat{q}) = \frac{1}{n}\begin{Bmatrix} (p+q)[1+8(p-q)^2] - (p+q)^2 \\ -4(p-q)^2 - 4(p-q)^4 \end{Bmatrix}.$$

Notice that for $p = q$ the two variances are equal. For $p \neq q$ it is easily seen that the variance on the left hand side is smaller if $2(p+q) < 1+(p-q)^2$, i.e. if $q < 1 + p - 2\sqrt{p}$.

(A3) Proof of the Lemma in Section 2.2.2

D_n can be expressed as $\sum_{i=1}^{n} Z_i$, where $P(Z_i = 1) = p, P(Z_i = -1) = q, P(Z_i = 0) = 1 - p - q$ and where the Z_i s are independent. Z_i has the probability generating function (pgf) $G_{Z_i}(s) =$ $sp + 1 - p - q + s / q$. Putting $p = p_1 p_2$ and $q = (1 - p_1)(1 - p_2)$ yields $G_{Z_i}(s) = (sp_1 + 1 - p_1) \cdot$ $(sp_2 + 1 - p_2) \cdot s^{-1}$, which is the pgf of $Y_{1i} + Y_{2i} - 1$ where Y_{1i} and Y_{2i} are independent Bernoulli variables. Therefore, $G_{D_n}(s) = (sp_1 + 1 - p_1)^n (sp_2 + 1 - p_2)^n s^{-n} = G_{A_n + B_n - n}(s)$ and the Lemma follows (cf. Feller (1968), Chapter X1).

References

Agresti, A. and Min, Y. (2005). Simple improved confidence intervals for companng matched proportions. *Statistics in Medicine* 24, 729-740.

Altman, D. G. (1991). *Practical Statisticsfor Medical Research,* Chapman and Hall, London.

Casella, G. and Berger, R. L. (1990). *Statistical Inference,* Duxburry Press, Belmont, California.

Elstone, R. C. and· Johnson, W. D. (1994). *Essentials of Biostatistics,* 2nd edn, F. A. Davis and Company, Philadelphia.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications,* Vol. 1, Third ed., Wiley, New York.

Hsueh, HM., Liu, JP., and Chen, J. J. (2001). Unconditional exact tests for equivalence or noninferiority for paired binary endpoints. *Biometrics* 57, 478-483.

Jones, B. and Kenward, M. G. (1987). Modelling binary data from a three-period cross-over trial, *Statistics in Medicine* 6, 555-564.

Jowett, G. H. (1963). The relationship between the binomial and F distributions. *The Statistician* 13, 55-57.

Kao, CH., Shiau, ye., Shen, YY., and Ven, RF. (2002). Detection of recurrent or persistent nasopharyngeal carcinomas after radiotherapy with technetium-99m methoxyiso-butylisonitrile single photon emission computed tomography and computed tomography: comparison with 18-fluoro-2-deoxyglucose positron emission tomography. *Cancer* 94, 1981-1986.

Karacan, L, Fernandez, S. A., and Coggins, W. S. (1976). Sleep electrocephalographic-electrooculographic characteristics of chronic marijuana users: part 1, *New York Academy of Science* 282,348-374.

Lewis, J. A. (1999). Statistical principles for clinical trials (ICH E9): An introductory note on an international guidline. *Statistics in Medicine* 18, 1903-1942.

May, W. L. and Johnson, W. D. (1997). Confidence intervals for differences in correlated binary proportions. *Statistics in Medicine* 16, 2127-2136.

Mc Nemar, Q. (1947). Note on the sampling error of the differences between correlated proportions ofpercentages. *Psychometrica* 12, 153-157.

Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 17, 2635-2650.

Stuart, A., Ord, K., and Arnold, S. (1999). *Kendall* 's *Advanced Theory ofStatistics, Vol 2A Classical Inference and the Linear Model,* 6th edn, Arnol<;l, London.

Suissa, S. and Shuster, J. J. (1991). The 2 x 2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics* 47, 361-372.

Tang, ML., Tang, NS., and Chan, S. F. (2005). Confidence interval construction for proportion difference in small-sample paired studies. *Statistics in Medicine* 24,3565-3579.

Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sarnple design. *Statistics in Medicine* **17,** 891-908.

Ward, S., Donovan, HS., Owen, B., Grosen, E., and Serlin, R. (2000). An individualized intervention to overcome patient-related baITiers to pain management in women with gynaecologic cancers. *Research in Nursing* & *Health* 23, 393-405.

Wonnacott, T. (1987). Confidence intervals or hypothesis tests? *Journal of Applied Statistics* 14, 185-201.

Research Report

2007:14	Pettersson, K.	Unimodal regression in the two-parameter exponential family with constant or known dispersion parameter.
2007:15	Pettersson, K.	On curve estimation under order restrictions.
2008:1	Frisén, M.	Introduction to financial surveillance.
2008:2	Jonsson, R.	When does Heckman's two-step procedure for censored data work and when does it not?
2008:3	Andersson, E.	Hotelling´s T2 Method in Multivariate On-Line Surveillance. On the Delay of an Alarm.
2008:4	Schiöler, L. & Frisén, M.	On statistical surveillance of the performance of fund managers.
2008:5	Schiöler, L.	Explorative analysis of spatial patterns of influenza incidences in Sweden 1999—2008.
2008:6	Schiöler, L.	Aspects of Surveillance of Outbreaks.
2008:7	Andersson, E & Frisén, M.	Statistiska varningssystem för hälsorisker
2009:1	Frisén, M., Andersson, E. & Schiöler, L.	Evaluation of Multivariate Surveillance
2009:2	Frisén, M., Andersson, E. & Schiöler, L.	Sufficient Reduction in Multivariate Surveillance
2010:1	Schiöler, L	Modelling the spatial patterns of influenza incidence in Sweden
2010:2	Schiöler, L. & Frisén, M.	Multivariate outbreak detection
2010:3	Jonsson, R.	Relative Efficiency of a Quantile Method for Estimating Parameters in Censored Two-Parameter Weibull Distributions
2010:4	Jonsson, R.	A CUSUM procedure for detection of outbreaks in Poisson distributed medical health events