Biometrics 64, 1270–1275
December 2008
DOI: 10.1111/j.1541-0420.2008.00990.x

Simultaneous Confidence Intervals for Comparing Binomial Parameters

Alan Agresti,[1,*] Matilde Bini,[2] Bruno Bertaccini,[2] and Euijung Ryu[3]

[1]Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.
[2]Department of Statistics, University of Florence, Florence 50134, Italy
[3]Division of Biostatistics, Mayo Clinic, Rochester, Minnesota 55905, U.S.A.
[*]*email:* aa@stat.ufl.edu

Summary. To compare proportions with several independent binomial samples, we recommend a method of constructing simultaneous confidence intervals that uses the studentized range distribution with a score statistic. It applies to a variety of measures, including the difference of proportions, odds ratio, and relative risk. For the odds ratio, a simulation study suggests that the method has coverage probability closer to the nominal value than ad hoc approaches such as the Bonferroni implementation of Wald or "exact" small-sample pairwise intervals. It performs well even for the problematic but practically common case in which the binomial parameters are relatively small. For the difference of proportions, the proposed method has performance comparable to a method proposed by Piegorsch (1991, *Biometrics* **47,** 45–52).

Key words: Bonferroni method; Difference of proportions; Odds ratio; Score test; Studentized range distribution; Tukey multiple comparisons.

1. Introduction

This research arose from a pharmaceutical consultancy collaboration, in which the aim was to construct simultaneous confidence intervals for odds ratios for comparing several groups on a binary response. The originally motivating data are of a similar structure as the data displayed in Table 1, which we will use throughout the article instead, for reasons of confidentiality. This table is based on results shown in an article (Kieburtz, 2001) about a randomized clinical trial to compare four treatments for potentially slowing the functional decline of early Huntington's disease: coenzyme Q10, remacemide hydrochloride, a combination of coenzyme and remacemide, and placebo. Safety measures in the study included tabulation of various adverse events during the 30 months of the study, and Table 1 shows results for nausea. In this article, we construct confidence intervals to make pairwise comparisons of the probabilities of nausea for the four treatments.

For data of this sort, significance tests are often used to summarize the overall evidence against a null hypothesis of no differences among groups or to investigate hypotheses related to the factorial nature of the treatments. But our emphasis here is on making inferences about the sizes of the effects. There is a large literature on multiple comparison methods of interval estimation to compare means of several groups, but there seems to be relatively little literature for multiple comparison of proportions. Most useful seems to be Piegorsch (1991), who proposed simultaneous confidence intervals for pairwise differences between proportions.

Section 2 proposes a method for multiple comparisons using effect measures for binary data with T groups. The method is based on applying the studentized range distribution with a set of approximately standard normally distributed score statistics constructed for the pairs of groups. The $100(1-\alpha)\%$ confidence interval is formed by inverting the test that compares the absolute value of each score statistic to $Q_T(\alpha)/\sqrt{2}$, where $Q_T(\alpha)$ denotes the $100(1-\alpha)$ percentile of the studentized range distribution with an infinite number of degrees of freedom.

Section 3 presents results from a simulation study for simultaneous confidence intervals for odds ratios. The method seems to perform well, considerably better than obvious ad hoc approaches such as applying the Bonferroni method with standard methods for pairwise intervals. Section 4 simulates performance of the method for simultaneous confidence intervals for differences of proportions. The performance is similar to a method Piegorsch (1991) proposed based on a Bayesian approach for the pairwise intervals. The simple approach of using the ordinary Wald interval after adding one outcome of each type to each sample (Agresti and Caffo, 2000) but replacing the normal percentile multiple of the standard error by $Q_T(\alpha)/\sqrt{2}$ also performs well when the true proportions are not very close to 0 or to 1.

2. Multiple Comparisons with a Binary Response

For T groups, denote the binomial parameters by $\{p_i, i = 1, \ldots, T\}$. For independent samples from the groups, let y_i denote a binomial variate based on n_i observations from group i, and let $\hat{p}_i = y_i/n_i$ denote the sample proportion. The key to the proposed method is to use the studentized range distribution in conjunction with a pairwise test statistic for which the two-sided test using the actual null distribution has P-values that are well approximated by P-values from a large-sample

Table 1
Data example with several groups having a binary response (*taken from Kieburtz,* 2001)

Treatment group	Sample size	Cases with nausea	Proportion
Coenzyme	87	13	0.149
Remacemide	86	27	0.314
Combination	87	22	0.253
Placebo	87	9	0.103

normal distribution. Test results are then inverted to obtain confidence intervals.

For the two-sample case with standard effect measures for binary data, various articles have found that the method of forming a confidence interval by inverting a large-sample score test performs well in terms of having nominal coverage probability approximate well the actual coverage probabilities, even with relatively small sample sizes (e.g., see Miettinen and Nurminen, 1985; Newcombe, 1998; Agresti and Min, 2005). In particular, inverting score tests tends to perform better than inverting likelihood-ratio tests and much better than inverting Wald tests. Mee (1984) and Miettinen and Nurminen (1985) proposed score confidence intervals for the difference of proportions, Cornfield (1956) and Miettinen and Nurminen (1985) proposed score confidence intervals for the odds ratio, and Koopman (1984) and Miettinen and Nurminen (1985) proposed score confidence intervals for the relative risk.

For testing the hypothesis that an effect measure θ of interest applied to groups i and j takes a particular null value $\theta_{ij,0}$, let $\tilde{p}_i$ and $\tilde{p}_j$ denote the maximum likelihood estimates of p_i and p_j under the constraint that the measure equals $\theta_{ij,0}$. For example, $\theta_{ij,0} = \{\tilde{p}_i/(1-\tilde{p}_i)\}/\{\tilde{p}_j/(1-\tilde{p}_j)\}$ for the odds ratio and $\theta_{ij,0} = \tilde{p}_i - \tilde{p}_j$ for the difference of proportions. Let $z_{ij}(\theta_{ij,0})$ denote the score test statistic in the form for which its asymptotic null distribution is standard normal. Let z_a denote the $(1-a)$ quantile of the standard normal distribution. The $100(1-\alpha)\%$ pairwise confidence interval for the measure comparing groups i and j consists of the set of $\theta_{ij,0}$ values for which $|z_{ij}(\theta_{ij,0})| < z_{\alpha/2}$.

The test statistic $z^2_{ij}(\theta_{ij,0})$ is identical to the Pearson chi-squared statistic for testing the hypothesis that the measure equals $\theta_{ij,0}$ (e.g., Bera and Bilias, 2001; Lovison, 2005). That is, $z^2_{ij}(\theta_{ij,0})$ equals the sum over the four cells in rows i and j of the $T \times 2$ table of the Pearson component $\{(\text{observed} - \text{expected})^2/\text{expected}\}$. The test statistic is also algebraically equivalent to

$$z^2_{ij}(\theta_{ij,0}) = \frac{\{n_i(\hat{p}_i - \tilde{p}_i)\}^2}{n_i\tilde{p}_i(1-\tilde{p}_i)} + \frac{\{n_j(\hat{p}_j - \tilde{p}_j)\}^2}{n_j\tilde{p}_j(1-\tilde{p}_j)}.$$

Now consider the set of such statistics $\{z_{ij}(\theta_{ij,0})\}$, for all possible pairwise values $\{\theta_{ij,0}\}$. Evaluated at the actual parameter values $\{\theta_{ij}\}$, for large $\{n_i\}$ the maximum of $|z_{ij}(\theta_{ij})|$ has approximately the $(1/\sqrt{2})$ multiple of the studentized range distribution with infinite degrees of freedom. That studentized range distribution is the distribution of the range between the maximum and minimum of T independent standard normal random variables. Generalizing Hochberg and Tamhane (1987) and Piegorsch (1991), this motivates a multiple comparison approach in which the confidence interval for the measure comparing groups i and j consists of the $\theta_{ij,0}$ values for which

$$|z_{ij}(\theta_{ij,0})| < Q_T(\alpha)/\sqrt{2}.$$

The set of all $T(T-1)/2$ of these intervals has large-sample simultaneous confidence level approximately equal to $(1-\alpha)$. For $T = 2$ this method yields the ordinary score test-based confidence interval for the measure.

The studentized-range implementation of the score confidence interval has the advantage of generality. It applies with a variety of measures, including odds ratios, relative risks, and differences of proportions, with equal or unequal sample sizes. An alternative approach that applies generally is to implement the score confidence intervals using the Bonferroni inequality. For each of the $q = T(T-1)/2$ pairwise comparisons, one then uses the ordinary score confidence interval but with confidence level equal to $1 - \alpha/q$. Because the score method achieves approximately the nominal level for individual comparisons, one would expect this to be conservative, more so for larger T. One can make this slightly less conservative by using the Šidák inequality (e.g., Hochberg and Tamhane, 1987, p. 366), for which the confidence level for each comparison is $(1-\alpha)^{1/q}$. In practice, this is a very slight difference (e.g., for $\alpha = 0.05$, corresponding to critical values of 2.807 versus 2.800 when $T = 5$ and 3.261 versus 3.254 when $T = 10$). Inspection of percentage points shows that Bonferroni and Šidák intervals are both typically about 2% to 3% wider than the studentized-range type of interval.

In practice, there are cases (e.g., in a regulatory environment) in which it is useful to ensure a lower bound on the simultaneous coverage probability. That is, the goal is to achieve *at least* the nominal coverage probability, using the exact small-sample distributions. Such conservative methods can be used in a multiple comparison setting with the Bonferroni method, to guarantee at least the nominal coverage. For a given overall sample size, one would expect the conservativeness to increase as a function of T, because of the increasing discreteness for each comparison and the conservativeness implicit in the Bonferroni method. Some statisticians prefer the adaptation of exact confidence intervals that invert the test using the mid P-value (which subtracts half the probability of the observed test statistic value from the ordinary P-value and which has null expected value equal to 1/2). In various settings, this method has been found to provide shorter intervals with coverage probability tending to fall nearer the nominal level, although not guaranteed to achieve at least that level. See, for example, Hirji (2005, pp. 50–51, 218–219). Likewise, multiple comparisons can implement this approach using the Bonferroni method.

3. Multiple Comparison of Odds Ratios: Example and Simulations

In practice, especially when the proportions are small, it is often useful to compare groups with the odds ratio or the relative risk. Table 2 shows the results of 95% multiple comparisons of odds ratios for the data in Table 1, using the studentized-range implementation of the score interval. We

Table 2
95% simultaneous confidence intervals for pairwise odds ratios and pairwise difference of proportions for data in Table 1. *Each confidence interval is based on the studentized-range implementation of the pairwise intervals inverting score tests.*

Rows	Odds ratio	Confidence interval	Difference of proportions	Confidence interval
(1, 2)	0.38	(0.15, 1.00)	−0.161	(−0.325, 0.000)
(1, 3)	0.52	(0.20, 1.38)	−0.103	(−0.260, 0.064)
(1, 4)	1.52	(0.48, 4.77)	0.046	(−0.089, 0.184)
(2, 3)	1.35	(0.57, 3.19)	0.061	(−0.114, 0.234)
(2, 4)	3.97	(1.38, 11.31)	0.211	(0.055, 0.364)
(3, 4)	2.93	(1.00, 8.52)	0.149	(0.000, 0.299)

conclude that there is a difference between the remacemide and placebo treatments, with the odds of nausea for the remacemide treatment being between about 1.4 and 11.3 times the odds of nausea with placebo.

We used a simulation study to evaluate the performance of this method, considering essentially the same cases as Piegorsch (1991) did in evaluating a method he proposed. He considered three sets of proportion values: $p_1 = 0.02$, 0.05, and 0.10, with the remaining $T - 1$ parameters equally spaced between p_1 and $p_T = 5\ p_1$. We used $T = 2$, 3, 5, 8 in order to study behavior ranging from the baseline of a single comparison ($T = 2$) to a relatively large number of groups. As in Piegorsch's analysis, we considered cases in which the sample sizes were equal, with values 25, 50, and 100. We also considered a mixed case with sample sizes 25 for half the groups and 50 for the other half, with 25 for the extra group when T is an odd number. Table 3 shows the estimated probability (based on 10,000 simulations) that at least one of the $T(T-1)/2$ intervals failed to contain the true parameter value. The proposed method seems to work well, perhaps better than one might expect in cases with relatively small sample sizes and small proportions. This method tended to be a bit conservative, overall, so we do not show results here for the Bonferroni implementation of the score interval, because that is necessarily more conservative.

Table 3 also shows results of using Bonferroni methods with the pairwise exact conditional confidence interval due to Cornfield (1956) based on inverting two one-sided tests using the noncentral hypergeometric distribution. The ordinary conservatism for two groups tends to worsen with more groups, even though the overall sample size increases in these analyses. When the mid-P adaptation of the exact approach was implemented with the Bonferroni method, the error probabilities (not shown in Table 3) fell between those for the "exact" intervals and for the score intervals. Hence, this method also tended to be quite conservative. For example, when $p_1 = 0.05$ with $n = 50$, its estimated error probabilities for $T = (2, 3, 5, 8)$ were (0.035, 0.027, 0.023, 0.023), compared to (0.015,

Table 3
Estimated error probabilities (nominal level 0.05*) for simultaneous confidence intervals for odds ratios for T groups. Methods compared are multiple range implementation of score intervals, Bonferroni method with Wald intervals, and Bonferroni method with "exact" small-sample intervals. True proportions are equally spaced between* $p_1 = 0.02$, 0.05, 0.10 *and* $p_T = 5\ p_1$.

		$p_1 = 0.02$			$p_1 = 0.05$			$p_1 = 0.10$		
T	n_i	Score	Exact	Wald	Score	Exact	Wald	Score	Exact	Wald
2	25	0.057	0.009	0.008	0.042	0.015	0.026	0.036	0.018	0.024
	50	0.034	0.009	0.028	0.042	0.015	0.029	0.053	0.028	0.034
	100	0.039	0.013	0.024	0.045	0.024	0.033	0.049	0.034	0.044
	Mixed	0.054	0.009	0.031	0.042	0.012	0.029	0.042	0.018	0.029
3	25	0.046	0.003	0.002	0.058	0.011	0.011	0.047	0.013	0.027
	50	0.056	0.008	0.010	0.039	0.014	0.025	0.053	0.024	0.031
	100	0.047	0.012	0.021	0.049	0.025	0.030	0.048	0.028	0.037
	Mixed	0.059	0.005	0.007	0.050	0.013	0.021	0.044	0.015	0.028
5	25	0.034	0.001	0.000	0.044	0.008	0.005	0.046	0.014	0.020
	50	0.044	0.005	0.002	0.043	0.015	0.014	0.045	0.018	0.025
	100	0.042	0.012	0.011	0.044	0.019	0.028	0.049	0.028	0.032
	Mixed	0.049	0.003	0.002	0.051	0.009	0.011	0.047	0.015	0.022
8	25	0.029	0.001	0.000	0.046	0.007	0.002	0.044	0.012	0.015
	50	0.038	0.004	0.001	0.047	0.013	0.010	0.046	0.016	0.025
	100	0.040	0.009	0.004	0.044	0.017	0.020	0.046	0.022	0.026
	Mixed	0.058	0.002	0.002	0.054	0.008	0.014	0.046	0.015	0.018

0.014, 0.015, 0.013) for the "exact" method and (0.042, 0.039, 0.043, 0.047) for the studentized-range implementation of the score interval.

Because we are not aware of any literature on multiple comparisons for odds ratios, for another comparison we considered an ad hoc method we thought practitioners are likely to employ. In practice, the Wald interval is commonly used, and an obvious multiple comparison approach is to implement it using the Bonferroni method. The Wald confidence interval for the log odds ratio is known to be somewhat conservative (Agresti, 1999). With sample odds ratio $\hat{\theta}_{ij}$ for groups i and j, the Wald confidence interval exponentiates

$$\log(\hat{\theta}_{ij}) \pm z_{\alpha/2} \times \sqrt{(n_i\hat{p}_i)^{-1} + \{n_i(1-\hat{p}_i)\}^{-1} + (n_j\hat{p}_j)^{-1} + \{n_j(1-\hat{p}_j)\}^{-1}}.$$

Table 3 also shows results for the Bonferroni implementation of the Wald interval. As expected, it is quite conservative. The order of conservatism is the same as the table shows for the Bonferroni method applied with the "exact" interval, but that approach has the advantage of necessarily achieving at least the nominal confidence level. We found similar behavior with adjustments of the Wald interval that have been recommended, such as implementing it after smoothing the table by adding 0.50 to each cell count or adding $2n_{i+}n_{+j}/n^2$ to cell count n_{ij} in group i making response j (Agresti, 1999).

4. Multiple Comparison of Difference of Proportions

Next, we apply the proposed method of using the studentized range together with the score statistic to the difference of proportions. For $T = 2$ groups, the confidence interval is then equivalent to one that Mee (1984) proposed. When none of the restricted maximum likelihood estimates falls at the boundary, the score interval corresponds to inverting the test having test statistic

$$z_{ij}(\theta_{ij,0}) = \frac{(\hat{p}_i - \hat{p}_j) - \theta_{ij,0}}{\sqrt{\{\tilde{p}_i(1-\tilde{p}_i)/n_i\} + \{\tilde{p}_j(1-\tilde{p}_j)/n_j\}}}.$$

Newcombe (1998) and Agresti and Min (2005) found that this method performs well. Table 2 also shows the results of 95% multiple comparisons for the difference of proportions for Table 1, using the studentized-range implementation of the score confidence interval. Substantive results are the same as with the odds ratio. There may be a considerable difference between remacemide and placebo, with the probability of nausea being between about 0.06 and 0.36 higher for remacemide.

Piegorsch (1991) considered methods for simultaneous confidence intervals for the difference of proportions. He first considered (1) the Bonferroni approach applied with the Wald confidence interval, and (2) a method implemented in Hochberg and Tamhane (1987, p. 275) using the Wald interval together with the studentized range distribution. These performed poorly, having error rates substantially higher than the nominal level when the sample sizes are small. This is no surprise, because the Wald method behaves poorly when $T = 2$, especially when the parameters are not near 0.50. In particular, the midpoint for each interval is the difference between the sample proportions. Piegorsch showed that better performance is obtained with a reformulated pairwise interval, motivated partly by a Bayesian approach of Beal (1987), for which the midpoint of the interval shrinks the sample differences of proportions toward 0. When implemented for multiple comparisons using the studentized range distribution, with equal sample sizes, $n_i = n$, the interval for $p_i - p_j$ is

$$(1+d^2)^{-1}\left[(\hat{p}_i - \hat{p}_j) \pm d\{(2-\tilde{\theta}_{ij})\tilde{\theta}_{ij}\ (1+d^2) - (\hat{p}_i - \hat{p}_j)^2\}^{1/2}\right],$$

with

$$\tilde{\theta}_{ij} = \{n(\hat{p}_i + \hat{p}_j) + 1\}/(n+1)$$

and

$$d = Q_T(\alpha)/2\sqrt{n}.$$

For $T = 2$, the center of the 95% confidence interval is approximately $\{n/(n+2)\}(\hat{p}_1 - \hat{p}_2)$. McCann and Tebbs (2007) extended this method to pooled data in which a group is classified as "positive" if at least one subject in the group is positive.

Table 4 shows simulation results comparing the studentized-range implementation of the score interval with this method of Piegorsch's. Overall, results are comparable. Neither method performs well for the smallest sample size (25 per group) when the true proportions are very small. This is not surprising, as in these cases the number of "successes" would tend to be near 0 for each group.

Good performance would also occur with studentized-range implementation of other approaches that provide such shrinkage and have good coverage performance on a pairwise comparison basis. The overall performance should reflect whether the pairwise coverage probabilities tend to be either conservative or liberal. For example, one such possibility for each pairwise interval is the Wald confidence interval constructed after adding one observation of each type to each sample (Agresti and Caffo, 2000). Let $\tilde{p}_i = (y_i + 1)/(n_i + 2)$. The confidence interval for $p_i - p_j$ is

$$(\tilde{p}_i - \tilde{p}_j) \pm \{Q_T(\alpha)/\sqrt{2}\}se$$

where

$$se = \sqrt{\frac{\tilde{p}_i(1-\tilde{p}_i)}{n_i+2} + \frac{\tilde{p}_j(1-\tilde{p}_j)}{n_j+2}}.$$

With equal sample sizes, the center of the interval is $\{n/(n+2)\}(\hat{p}_i - \hat{p}_j)$. Agresti and Caffo (2000) noted that this method has good pairwise performance but tends to be conservative when the p_i are near the boundary. Simulations with this method, summarized also in Table 4, indicated that it performs well unless p_i are near 0. Another pairwise interval that provides shrinkage and performs quite well on a pairwise basis uses the Bayesian approach (Agresti and Min, 2005). The posterior distribution for $p_i - p_j$ is induced by independent binomial samples having relatively uninformative beta priors.

5. Comments and Summary

We have recommended using the studentized-range method by inverting the score test rather than the likelihood-ratio test. This is because, for two groups with the measures considered, the score test has been observed to perform better. However, our simulations did evaluate the performance

Table 4
Estimated error probabilities (nominal level 0.05*) for Piegorsch (P) method and for studentized-range implementation of score intervals and Agresti–Caffo (A–C) intervals for the difference between proportions. True proportions are equally spaced between* $p_1 = 0.02, 0.05, 0.10$ *and* $p_T = 5\ p_1$.

		$p_1 = 0.02$			$p_1 = 0.05$			$p_1 = 0.10$		
T	n_i	Score	P	A–C	Score	P	A–C	Score	P	A–C
2	25	0.028	0.079	0.006	0.067	0.055	0.044	0.056	0.049	0.042
	50	0.060	0.040	0.035	0.051	0.055	0.046	0.050	0.051	0.054
	100	0.047	0.045	0.046	0.048	0.049	0.049	0.049	0.050	0.049
	Mixed	0.029	0.018	0.023	0.039	0.016	0.042	0.054	0.021	0.046
3	25	0.010	0.012	0.003	0.042	0.048	0.031	0.058	0.055	0.047
	50	0.031	0.053	0.023	0.055	0.049	0.042	0.052	0.048	0.045
	100	0.052	0.047	0.034	0.053	0.050	0.041	0.054	0.050	0.052
	Mixed	0.035	0.015	0.005	0.046	0.036	0.029	0.050	0.040	0.043
5	25	0.003	0.002	0.002	0.034	0.043	0.026	0.051	0.042	0.050
	50	0.019	0.027	0.011	0.049	0.042	0.034	0.052	0.050	0.047
	100	0.042	0.039	0.028	0.053	0.045	0.041	0.053	0.046	0.045
	Mixed	0.025	0.008	0.003	0.041	0.043	0.028	0.051	0.040	0.046
8	25	0.001	0.000	0.000	0.023	0.035	0.026	0.045	0.045	0.049
	50	0.010	0.022	0.005	0.043	0.039	0.035	0.052	0.046	0.050
	100	0.037	0.029	0.020	0.054	0.047	0.041	0.056	0.049	0.050
	Mixed	0.028	0.011	0.003	0.044	0.033	0.028	0.050	0.040	0.051

of the studentized-range method together with inverting the likelihood-ratio test in standard normal form. The estimated coverage probabilities tended to be farther from the nominal value than obtained with the score test, typically erring in the liberal direction. For example, using the likelihood-ratio test with the difference of proportions, for the case $p_1 = 0.05$ with $n = 50$ the estimated error probabilities for $T = (2, 3, 5, 8)$ were (0.056, 0.062, 0.077, 0.085), compared to (0.051, 0.055, 0.049, 0.043) for the score test.

In summary, our proposed method using the studentized range distribution with a score statistic is applicable for all the standard measures for comparing binomial parameters. For the odds ratio, it seems to perform better than the obvious ad hoc approaches one might use, such as the Bonferroni method applied with a standard confidence interval. For the difference of proportions, the method seems to give results comparable to a method proposed by Piegorsch (1991). Thus, the proposed method seems to be a useful general-purpose way to obtain simultaneous confidence intervals comparing several binomial parameters.

In future research, it would be useful to develop simultaneous confidence intervals in other contexts. An important case is when the groups to be compared are ordered (such as doses of a drug) and it is sensible to assume monotonicity of the proportions. Then, more efficient approaches undoubtedly exist, especially with small sample sizes. Various closed-testing procedures have been proposed that are more efficient than the Bonferroni method for multiple significance testing. These apply when the hypotheses tested are closed under intersection, and the test statistics are ordered from largest to smallest, applying less stringent significance levels to the second, third, and so on (e.g., Holm, 1979). Such methods might be useful if they could be applied to interval estimation. Some such methods are especially useful for ordered groups (e.g., Marcus, Peritz, and Gabriel, 1976; Rom, Costello, and Connell, 1994). There is also the challenge of obtaining realistic coverage probabilities when the confidence intervals formed are themselves suggested by a large number of preliminary tests, such as in the extension by Benjamini and Yekutieli (2005) of the false discovery rate from multiple testing to selected multiple interval estimation. Finally, our method applies with independent samples, and there is scope for developing simultaneous confidence intervals for dependent samples.

A program using `R` for implementing the proposed studentized-range-score method with the odds ratio and the difference of proportions is available from the authors.

Acknowledgements

The research of AA was partially supported by funds from the University of Florence, Italy. The authors appreciate helpful comments from Walter Piegorsch, Anna Gottard, and two referees.

References

Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics* **55,** 597–602.

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *The American Statistician* **54,** 280–288.

Agresti, A. and Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 contingency tables. *Biometrics* **61,** 515–523.

Beal, S. L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics* **43,** 941–950.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate—adjusted multiple confidence intervals for selected

parameters. *Journal of the American Statistical Association* **100,** 71–81.

Bera, A. K. and Bilias, Y. (2001). Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: An essay on historical developments and some new results. *Journal of Statistical Planning and Inference* **97,** 9–44.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman (ed.), **4,** 135–148.

Hirji, K. (2005). *Exact Analysis of Discrete Data*. Boca Raton, Florida: Chapman & Hall.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: John Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6,** 65–70.

Kieburtz, K. and Huntington Study Group (2001). A randomized, placebo-controlled trial of coenzyme Q10 and remacemide in Huntington's disease. *Neurology* **57,** 397–404.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics* **40,** 513–517.

Lovison, G. (2005). On Rao score and Pearson X^2 statistics in generalized linear models. *Statistical Papers* **46,** 555–574.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63,** 655–660.

McCann, M. H. and Tebbs, J. M. (2007). Pairwise comparisons for proportions estimated by pooled testing. *Journal of Statistical Planning and Inference* **137,** 1278–1290.

Mee, R. W. (1984). Confidence bounds for the difference between two probabilities (letter). *Biometrics* **40,** 1175–1176.

Miettinen, O. and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine* **4,** 213–226.

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine* **17,** 873–890.

Piegorsch, W. W. (1991). Multiple comparisons for analyzing dichotomous response. *Biometrics* **47,** 45–52.

Rom, D. M., Costello, R. J., and Connell, L. T. (1994). On closed test procedures for dose-response analysis. *Statistics in Medicine* **13,** 1583–1596.

Received July 2007. *Revised November* 2007.
Accepted November 2007.