BIOMETRICS 58, 209–215
March 2002

Interval Estimation for a Difference Between Intraclass Kappa Statistics

Allan Donner* and Guangyong Zou

Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario N6A 5C1, Canada
**e-mail*: donner@biostats.uwo.ca

SUMMARY. Model-based inference procedures for the kappa statistic have developed rapidly over the last decade. However, no method has yet been developed for constructing a confidence interval about a difference between independent kappa statistics that is valid in samples of small to moderate size. In this article, we propose and evaluate two such methods based on an idea proposed by Newcombe (1998, *Statistics in Medicine*, **17,** 873–890) for constructing a confidence interval for a difference between independent proportions. The methods are shown to provide very satisfactory results in sample sizes as small as 25 subjects per group. Sample size requirements that achieve a prespecified expected width for a confidence interval about a difference of kappa statistic are also presented.

KEY WORDS: Confidence interval; Interobserver agreement; Intraclass correlation; Reliability studies; Sample size.

1. Introduction

There has recently been increased interest in developing model-based inferences for the kappa statistic, particularly as applied to the assessment of interobserver agreement. Most of this research has focused on problems that arise in a single sample of subjects, each assessed on the presence or absence of a binary trait by two observers (e.g., Donner and Eliasziw, 1992; Hale and Fleiss, 1993; Basu, Banerjee, and Sen, 2000; Klar, Lipsitz, and Ibrahim, 2000; Nam, 2000). Extension of this research to single-sample inference problems involving more than two raters and/or multinomial outcomes has also recently appeared (e.g., Altaye, Donner, and Klar, 2001; Bartfay and Donner, 2001).

Although single-sample problems involving the kappa statistic perhaps predominate, the need to compare two or more kappa statistics also frequently arises in practice. For example, it may be of interest to compare measures of interobserver agreement across different study populations or different patient subgroups in a single study. The latter was an objective of a study reported by Ishmail et al. (1987) that focused on interobserver agreement with respect to the presence of the third heart sound (S_3), recognized as an important sign in the evaluation of patients with congestive heart failure. One question of interest in this study was whether the degree of interobserver agreement varied by the gender of the patient, an issue that arose because it may be more difficult to hear S_3 in women compared with men. Other examples of such investigations are given by Barlow, Lai, and Azen (1991), McLellan et al. (1985), and McLaughlin et al. (1987). The comparison of coefficients of interobserver agreement is also the major focus of many measurement studies in educational and psychological research (e.g., Alsawalmeh and Feldt, 1992).

Appropriate testing procedures for this problem have been described by Donner, Eliasziw, and Klar (1996) for the case of independent samples of subjects and Donner et al. (2000) for the case of dependent samples involving the same subjects. Reed (2000) has also addressed this problem, providing an executable Fortran code for testing the homogeneity of kappa statistics in the former case. However, to the best of our knowledge, procedures have not yet been developed for constructing an interval estimate about the difference between two kappa statistics. With the increased emphasis on confidence-interval construction as an alternative to significance testing, this would seem to be an important gap in the literature.

As discussed by Nam (2000), there are two models that have tended to be adopted for constructing inferences for the kappa statistic in the case of two raters and a binary outcome. The first model allows the marginal probabilities of a success to differ between the two raters and leads naturally to Cohen's (1960) kappa. The second model assumes the same marginal probability of success for each rater and leads naturally to the intraclass kappa statistic, identically equal to Scott's (1955) π. Landis and Koch (1977) and Bloch and Kraemer (1989) have presented arguments supporting the use of this model when the main emphasis is directed at the reliability of the measurement process rather than in potential differences among raters. Zwick (1988) also discusses this issue, pointing out that, if marginal differences are small, the value of Cohen's kappa will be close to that of Scott's π and the choice between them will not be important. Furthermore, as shown by Black-

man and Koval (2000), the two estimators are asymptotically equivalent.

In this article, we develop and evaluate a method for constructing a confidence interval for a difference between two intraclass kappa statistics computed from independent samples of subjects. The procedures compared are developed in Section 2, followed in Section 3 by a simulation study evaluating their properties. Section 4 provides guidelines for sample size estimation, while two examples illustrating the procedures are presented in Section 5. The article concludes with overall recommendations for practice and with advice concerning available software.

2. Development of Procedures

Following Donner et al. (1996), we let X_{ijh} denote the rating for the ith subject by the jth rater in sample h ($h = 1, 2$). Letting $\pi_h = \Pr(X_{ijh} = 1)$ be the probability of a success, the probabilities of joint responses within sample h are given by $P_{1h}(\kappa_h) = \Pr(X_{i1h} = 1, X_{i2h} = 1) = \pi_h^2 + \pi_h(1-\pi_h)\kappa_h$, $P_{2h}(\kappa_h) = \Pr(X_{i1h} = 0, X_{i2h} = 1 \text{ or } X_{i1h} = 1,\ X_{i2h} = 0) = 2\pi_h(1-\pi_h)(1-\kappa_h)$, and $P_{3h}(\kappa_h) = \Pr(X_{i1h} = 0, X_{i2h} = 0) = (1-\pi_h)^2 + \pi_h(1-\pi_h)\kappa_h$. This model, previously discussed by several authors, including Mak (1988) and Bloch and Kraemer (1989), has been referred to as the common correlation model because it assumes that the correlation between any pair (X_{i1h}, X_{i2h}) has the same value κ_h.

We now assume that the observed frequencies n_{1h}, n_{2h}, n_{3h} corresponding to $P_{1h}(\kappa_h)$, $P_{2h}(\kappa_h)$, $P_{3h}(\kappa_h)$ follow a multinomial distribution conditional on $N_h = \Sigma_{k=1}^{3} n_{kh}$, the total number of subjects in sample h. Letting $\hat{\pi}_h = (2n_{1h} + n_{2h})/(2N_h)$, Bloch and Kraemer (1989) show that, for a single sample of subjects, the maximum likelihood estimator of κ_h under the common correlation model is given by $\hat{\kappa}_h = 1 - n_{2h}/\{2N_h\hat{\pi}_h(1-\hat{\pi}_h)\}$, with large-sample variance obtained as $\text{var}(\hat{\kappa}_h) = (1-\kappa_h)[(1-\kappa_h)(1-2\kappa_h) + \kappa_h(2-\kappa_h)/\{2\pi_h(1-\pi_h)\}]/N_h$. It is easily verified that $\hat{\kappa}_h$ is simply the standard intraclass correlation coefficient (e.g., Snedecor and Cochran, 1989) as applied to dichotomous outcome data.

We can estimate $\text{var}(\hat{\kappa}_h)$ by substituting $\hat{\pi}_h$ and $\hat{\kappa}_h$ for π_h and κ_h, respectively. A large-sample $100(1-\alpha)\%$ confidence interval for $\kappa_1 - \kappa_2$ is then given by $(\hat{\kappa}_1 - \hat{\kappa}_2) \pm z_{\alpha/2}\{\widehat{\text{var}}(\hat{\kappa}_1) + \widehat{\text{var}}(\hat{\kappa}_2)\}^{1/2}$, where $\widehat{\text{var}}(\hat{\kappa}_h)$ is the sample estimate of $\text{var}(\hat{\kappa}_h)$ and $z_{\alpha/2}$ denotes the $100(1-\alpha/2)\%$ percentile point of the standard normal distribution. We refer to this method as the simple asymptotic (SA) method.

A potential difficulty with the SA method is that the sampling distribution of $\hat{\kappa}_h$ may be far from normal when κ_h is close to one and the nuisance parameter π_h is close to zero or one, a parameter range that is frequently of practical interest (see Bloch and Kraemer (1989) for a detailed discussion). An extreme case of this problem arises at $\hat{\kappa}_h = 1$, where $\text{var}(\hat{\kappa}_h)$ is degenerate. We therefore consider alternatives that are less subject to aberrant results in samples of small to moderate size.

Newcombe (1998) evaluated 11 methods for constructing a confidence interval about the difference between two proportions and recommended a computationally simple hybrid procedure based on the Wilson (1927) score method for a single proportion. He found that this method performs remarkably well under a wide range of conditions. We apply the same idea here to constructing a confidence interval for $\Delta = \kappa_1 - \kappa_2$.

Let (l_h, u_h) denote lower and upper $100(1-\alpha)\%$ limits for κ_h. Then the hybrid confidence interval for Δ is given by (L, U), where

$$L = (\hat{\kappa}_1 - \hat{\kappa}_2) - z_{\alpha/2}\{\text{var}(\hat{\kappa}_1)|_{\kappa_1 = l_1} + \text{var}(\hat{\kappa}_2)|_{\kappa_2 = u_2}\}^{1/2}$$

and

$$U = (\hat{\kappa}_1 - \hat{\kappa}_2) + z_{\alpha/2}\{\text{var}(\hat{\kappa}_1)|_{\kappa_1 = u_1} + \text{var}(\hat{\kappa}_2)|_{\kappa_2 = l_2}\}^{1/2}.$$

Further details concerning the theoretical basis for this procedure are given in the Appendix.

One approach to obtaining the limits (l_h, u_h), $h = 1, 2$, was proposed by Donner and Eliasziw (1992), who used a goodness-of-fit approach to construct a confidence interval for κ_h. This approach, algebraically equivalent to an approach independently proposed by Hale and Fleiss (1993), assumes only that the observed frequencies n_{kh} ($k = 1, 2, 3$) follow a multinomial distribution with corresponding probability P_{kh} conditional on N_h. If estimated probabilities $\hat{P}_{kh}(\kappa_h)$ are obtained by replacing π_h with $\hat{\pi}_h$, it then follows that $\chi_G^2 = \Sigma_{k=1}^{3}\{n_{kh} - N_h P_{kh}(\kappa_h)\}^2 / \{N_h P_{kh}(\kappa_h)\}$ has a limiting chi-square distribution with 1 d.f. One can then obtain the confidence limits (l_h, u_h) for κ_h by finding the two admissible roots to the cubic equation $\chi_G^2 = \chi_{1,1-\alpha}^2$. We refer to this method as the hybrid goodness-of-fit (HGOF) approach.

An alternative method of obtaining (l_h, u_h) is to invert a modified Wald test (Rao and Mukerjee, 1997), which may be constructed by replacing $\text{var}(\hat{\kappa}_h)$ with $\widehat{\text{var}}_p(\hat{\kappa}_h)$, obtained by substituting $\hat{\pi}_h$ for π_h. The confidence limits for κ_h are then given by the two admissible roots of the cubic equation given by $(\hat{\kappa}_h - \kappa_h)^2 / \text{var}_p(\kappa_h) = \chi_{1,1-\alpha}^2$. An approach similar to this was applied by Lee and Tu (1994) to the problem of constructing confidence limits about Cohen's kappa, who refer to it as the profile variance approach. We refer to it here as the hybrid profile variance (HPR) approach. Note that the HGOF and HPR methods differ only in how the confidence limits (l_h, u_h), $h = 1, 2$, are computed.

3. Evaluation

The finite sample properties of the three methods presented in Section 2 are, for the most part, intractable. We therefore compared the performance of the SA, HGOF, and HPR methods using Monte Carlo simulation. Simulation runs having $\hat{\pi}_h(1-\hat{\pi}_h) = 0$ (for which $\hat{\kappa}_h$ is undefined) were discarded until 10,000 runs were obtained. The expected proportion of such runs is given by $\{\pi^2 + \pi(1-\pi)\kappa\}^N + \{(1-\pi)^2 + \pi(1-\pi)\kappa\}^N$, well under 5% for all parameter combinations considered in the simulation. Procedure PROC IML in the statistical software package SAS was used to generate observations from the common correlation model.

The restriction $\pi_1 = \pi_2$ was imposed in the simulation study since many authors (e.g., Thompson and Walter, 1988) caution strongly against the comparison of two or more kappa statistics when the population prevalence for the groups differs. This restriction, although not required by the theoretical development above, is reasonable in practice given the well-known dependence of the kappa statistic on the estimated group prevalence.

For each parameter combination $(N_1, N_2, \pi, \kappa_1, \Delta)$, we computed the empirical coverage level generated by each method for a 95% confidence interval constructed about Δ

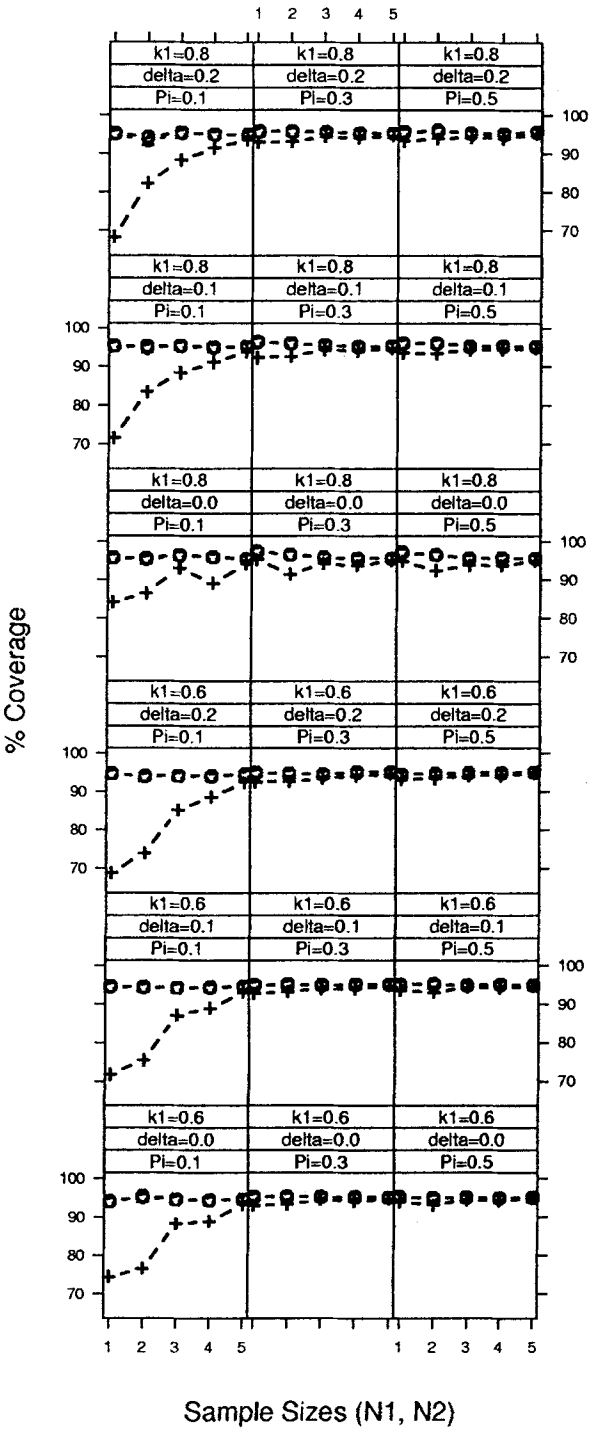


Figure 1. Empirical coverage percentage based on 10,000 runs for the HGOF (O), HPR ($\triangle$), and SA (+) methods as used to construct a 95% two-sided confidence interval for $\Delta = \kappa_1 - \kappa_2$. Sample sizes (N_1, N_2): 1 = (25, 25), 2 = (25, 50), 3 = (50, 50), 4 = (50, 100), 5 = (100, 100).

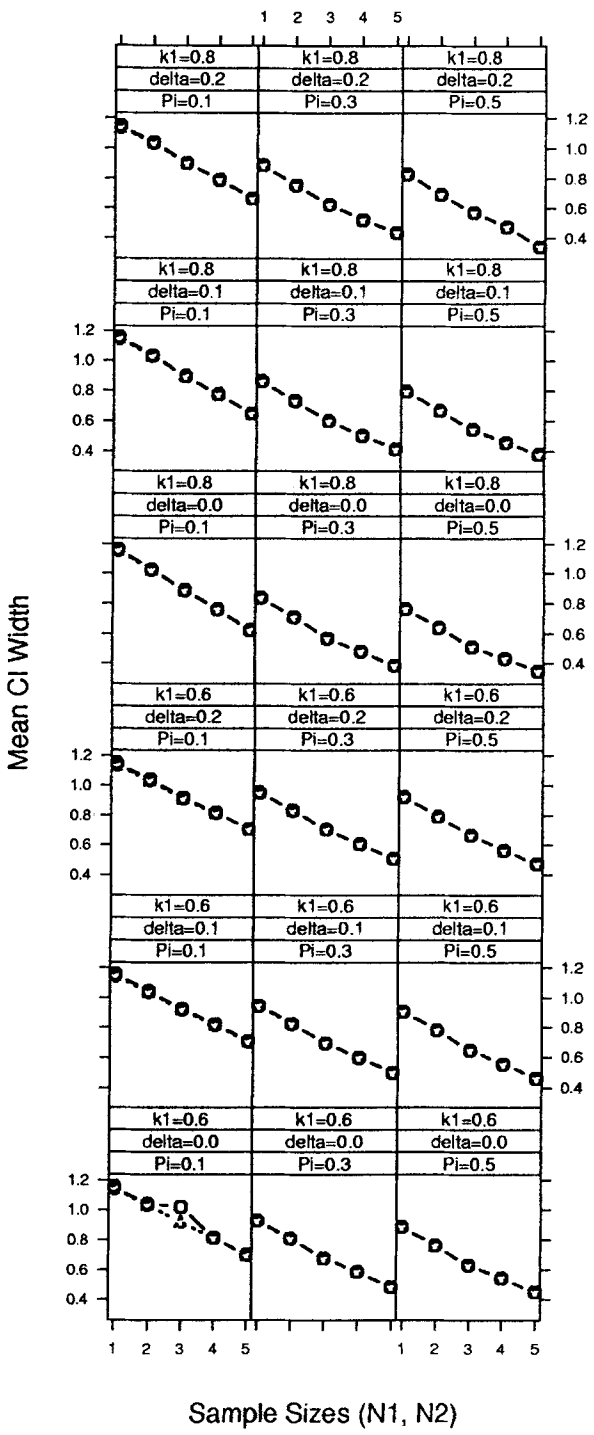


Figure 2. Mean confidence-interval width based on 10,000 runs for the HGOF (O) and HPR ($\triangle$) methods as used to construct a 95% two-sided confidence interval for $\Delta = \kappa_1 - \kappa_2$. Sample sizes (N_1, N_2): 1 = (25, 25), 2 = (25, 50), 3 = (50, 50), 4 = (50, 100), 5 = (100, 100).

$= \kappa_1 - \kappa_2$. These results are shown in Figure 1, with mean confidence-interval widths shown in Figure 2.

The results in Figure 1 are consistent with previous work (e.g., Bloch and Kraemer, 1989; Barlow et al., 1991; Nam, 2000), demonstrating that confidence-interval methods based on the estimated large-sample variance of the kappa statistic give reliable coverage only when sample sizes are large and the prevalence π is not extreme. Otherwise, the actual coverage

Table 1
Number of subjects (N) per group required to ensure a two-sided 95% confidence interval (CI) for Δ ($= \kappa_1 - \kappa_2$) has width W_0 for various values of κ_1, Δ, and π based on the HGOF method

			CI width W_0						CI width W_0		
κ_1	Δ	π	0.20	0.40	0.60	κ_1	Δ	π	0.20	0.40	0.60
0.6	0.0	0.1	1400	342	145	0.8	0.0	0.1	826	220	104
		0.3	590	147	65			0.3	344	93	45
		0.5	493	124	55			0.5	286	77	38
	0.1	0.1	1490	361	151		0.1	0.1	982	252	115
		0.3	636	158	69			0.3	409	107	50
		0.5	534	133	59			0.5	341	89	42
	0.2	0.1	1533	369	152		0.2	0.1	1108	278	123
		0.3	671	166	72			0.3	465	119	54
		0.5	568	141	62			0.5	388	100	46
	0.3	0.1	1520	363	149		0.3	0.1	1196	295	127
		0.3	693	170	74			0.3	511	129	58
		0.5	594	148	65			0.5	429	109	50

level may be substantially less than nominal. The two hybrid methods, on the other hand, maintain coverage levels very close to nominal, even when $N_1 = N_2 = 25$ and $\pi = 0.1$.

Our simulation results also showed that the SA method yields mean confidence-interval widths consistently greater than that of the other two methods, even when its coverage level is below nominal. Because of this observation and the other difficulties associated with this method discussed in Section 2, we report mean widths in Figure 2 only for the two hybrid methods.

The results in this figure clearly show the importance of recruiting a large number of subjects in each comparison group if the investigator's objective is to estimate Δ with a reasonable degree of precision. For example, if $\kappa_1 = 0.8$ and $\kappa_2 = 0.6$, then it is necessary to recruit at least 100 subjects per group to obtain a mean confidence width no more than about 0.50, assuming $\pi = 0.3$. However, a comparison between the HGOF and HPR methods shows that they provide virtually the same mean confidence-interval widths at all parameter values.

4. Sample Size Estimation

The results in Section 2 can also be used to estimate the sample sizes N_1 and N_2 needed to achieve a confidence interval about $\Delta = \kappa_1 - \kappa_2$ having prespecified width. In particular, we may ask how many subjects per sample are needed to construct a $100(1-\alpha)\%$ two-sided confidence interval about Δ having width no greater than a prespecified value W_0. This problem could arise when the objective of a study is to specifically estimate the difference between agreement levels that arise in distinct subgroups of subjects (e.g., Faerstein, Chor, and Lopes, 2001).

For the sake of simplicity, we assume equal sample sizes $N_1 = N_2 = N$ and focus on the hybrid goodness-of-fit (HGOF) approach (results obtained using the hybrid profile variance [HPR] approach were virtually the same). Using equation (2) in the Appendix as applied to obtaining the limits (L, U) given in Section 2, it may be noted that the expression for the interval width $W = U - L$ depends on N, on the number of subjects n_{2h} in sample h having discordant ratings, on the observed values of the parameters π_h, $h = 1, 2$, and on the probability of coverage $1 - \alpha$.

The values $\hat{\pi}_h$ and n_{2h} are unknown prior to the study. However, for the purpose of sample size estimation and assuming a common prevalence π, an approximation to the expected width of the confidence interval may be obtained by replacing $\hat{\pi}_h$ by its anticipated value π, based on information obtained from previous studies or from a pilot investigation. Furthermore, n_{2h} may be replaced by its expected value from the common correlation model, given by $NP_{2h}(\kappa_h) = 2N\pi(1-\pi)(1-\kappa_h)$. These substitutions allow us to estimate the minimum value of N (rounded up to the nearest integer) needed to ensure that $W \leq W_0$ at selected values of κ_1, Δ, π, and α.

The results are shown in Table 1 for $\kappa_1 = 0.6, 0.8$, $\Delta = 0.0, 0.1, 0.2, 0.3$, $\pi = 0.1, 0.3, 0.5$, and $\alpha = 0.05$ (two-sided) for prespecified interval width $W_0 = 0.20, 0.40, 0.60$. These values for κ_1 correspond to what Landis and Koch (1977) have characterized as the upper limits for describing moderate and substantial interobserver agreement, respectively. It is clear from these results that the sample size requirements needed to achieve a confidence-interval width less than 0.40 will often be prohibitive in practice.

As an example, suppose an interobserver agreement study is being planned with the aim of constructing a 95% confidence interval for Δ having width no greater than 0.4 when $\kappa_1 = 0.8$ and $\kappa_2 = 0.6$. Then if $\pi = 0.50$, $N = 100$ subjects are required in each group. At $\pi = 0.10$, however, the required value of N increases to 278, reflecting the substantial sensitivity of these results to the value of the underlying prevalence parameter.

5. Examples

As a first example, we consider data from Landis and Koch (1977), previously analyzed by Barlow (1996). As part of a study on multiple sclerosis reported by Westlund and Kurland (1953), two neurologists classified 149 Winnipeg patients and 69 New Orleans patients on a four-point scale. Using the

same dichotomous classification considered by Barlow (1996), we are interested here in the interobserver agreement with respect to a diagnosis of certain multiple sclerosis versus uncertain. This classification yields $\{n_{11}, n_{21}, n_{31}\} = \{119, 20, 10\}$ and $\{n_{12}, n_{22}, n_{32}\} = \{44, 11, 14\}$ for the Winnipeg and New Orleans samples, respectively, implying $N_1 = 149$ and $N_2 = 69$. We then obtain $\hat{\kappa}_1 = 0.422$ and $\hat{\kappa}_2 = 0.607$, with $\hat{\pi}_1 = 0.866$ and $\hat{\pi}_2 = 0.717$. Applications of both the HGOF and HPR methods yield 95% confidence limits about $\Delta = \kappa_1 - \kappa_2$ given by $(-0.450, 0.122)$, while the SA method yields the slightly narrower limits $(-0.481, 0.112)$. Therefore, there is no indication from this investigation that the level of interobserver agreement varies significantly by geographic region.

As a second example, we consider data presented by Faerstein et al. (2001), who investigated within-person agreement over time with respect to information recorded on the diagnosis and treatment of hypertension. Interviewees involved in this study filled out a health questionnaire that was applied twice within an interval of two weeks.

One aim of this study was to compare the consistency of agreement across this two-week period among different population subgroups, as defined by gender, age, and educational level. Although the focus of this example is on agreement between occasions rather than between observers, the kappa statistic is well suited to make this comparison. We focus here on the comparison of male ($N_1 = 82$) and female ($N_2 = 87$) subjects with respect to level of agreement over time on recorded history of diagnosed hypertension. For this comparison, Faerstein et al. (2001) report the kappa statistics as $\hat{\kappa}_1 = 0.623$ for males and $\hat{\kappa}_2 = 0.876$ for females, with $\hat{\pi}_1 = 0.177$, $\hat{\pi}_2 = 0.167$. The HGOF and HPR methods yield 95% confidence limits given by $(-0.522, 0.022)$ and $(-0.523, 0.023)$, respectively, while the SA method yields $(-0.516, 0.010)$. Therefore, there is some indication that the consistency of agreement over the two-week period is somewhat higher for females, although the difference between $\hat{\kappa}_1$ and $\hat{\kappa}_2$ is not significant at the 5% level.

6. Discussion

Suitable methodology is now available to construct confidence limits about estimates of interobserver agreement in sample sizes typical of those that are conducted in practice. However, as pointed out by Shrout (1998) in the context of psychiatric research, confidence intervals are often not reported by investigators, thus perpetuating the proliferation of underpowered studies. For the comparison of kappa statistics, this problem is even more severe due to the paucity of available methods for constructing confidence intervals. It is hoped that the results presented in this article will help fill this gap.

The HGOF and HPR methods each involve the combination of independently computed single-sample confidence intervals for κ_1 and κ_2, respectively. They differ only in that the goodness-of-fit procedure is used to calculate the single-sample limits for the HGOF method, while an inverted Wald test approach is used to calculate the single-sample limits for the HPR method. Since the two methods provide very similar results in practice, the choice between them is largely a matter of convenience. The single-sample goodness-of-fit procedure, used as an intermediate step in conducting confidence-interval limits using the HGOF method, is also now available in the software package PEPI (Abrahamson and Gohlinger, 1999) and is available at `http://www.usd-inc.com/pepi.html`. One advantage of the goodness-of-fit approach is that it can be readily extended to other inference problems, including hypothesis testing and sample size estimation involving multiple raters (e.g., Altaye et al., 2001).

Acknowledgement

Dr Donner's research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Résumé

Les procédures inférentielles basées sur la modélisation pour la statistique du kappa se sont rapidement développées dans la dernière décennie. Cependant aucune méthode n'a encore été développée pour construire un intervalle de confiance pour la différence de statistiques kappa indépendantes qui soit valide sur des échantillons de taille petite ou moyenne. Dans ce papier nous proposons et évaluons deux méthodes reposant sur une idée suggérée par Newcombe (1998, *Statistics in Medicine* **17,** 873–890) pour construire un intervalle de confiance pour la différence entre deux proportions indépendantes. On montre que les méthodes donnent des résultats très satisfaisants pour des échantillons aussi petits que 25 sujets par groupe. On examine également les tailles d'échantillon requises pour obtenir une largeur moyenne spécifiée pour l'intervalle de confiance d'une différence de statistiques du kappa.

References

Abrahamson, J. and Gohlinger, P. (1999). *PEPI: Computer Programs for Epidemiologists,* Version 3.0. London: Brixton Books.

Alsawalmeh, Y. M. and Feldt, L. S. (1992). Test of the hypothesis that the intraclass reliability coefficient is the same for two measurement procedures. *Applied Psychological Measurement* **16,** 192–205.

Altaye, M., Donner, A., and Klar, N. (2001). Inference procedures for assessing interobserver agreement among multiple raters. *Biometrics* **57,** 584–588.

Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. *Biometrics* **52,** 695–702.

Barlow, W., Lai, M.-Y., and Azen, S. P. (1991). A comparison of methods for calculating a stratified kappa. *Statistics in Medicine* **10,** 1465–1472.

Bartfay, E. and Donner, A. (2001). Statistical inferences for inter-observer agreement studies with nominal outcome data. *The Statistician* **50,** 135–146.

Basu, S., Banerjee, M., and Sen, A. (2000). Bayesian inference for kappa from single and multiple studies. *Biometrics* **56,** 577–582.

Blackman, N. J.-M. and Koval, J. J. (2000). Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine* **19,** 723–741.

Bloch, D. A. and Kraemer, H. C. (1989). 2×2 Kappa coefficients: Measures of agreement or association. *Biometrics* **45,** 269–287.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20,** 37–46.

Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* **11,** 1511–1519.

Donner, A., Eliasziw, M., and Klar, Neil. (1996). Testing the homogeneity of kappa statistics. *Biometrics* **52,** 176–183.

Donner, A., Shoukri, M. M., Klar, N., and Bartfay, E. (2000). Testing the equality of two dependent kappa statistics. *Statistics in Medicine* **19,** 373–387.

Faerstein, E., Chor, D., and Lopes, C. (2001). Reliability of the information about the history of diagnosis and treatment of hypertension. Differences in regard to sex, age and educational level. The Pró-Saúde Study. *Arquivos Brasileiros de Cardiologia* **76,** 301–304.

Hale, C. A. and Fleiss, J. L. (1993). Interval estimation under two study designs for kappa with binary classifications. *Biometrics* **49,** 523–533.

Ishmail, A. A., Wing, S., Ferguson, J., Hutchinson, T. A., Magder, A., and Flegel, K. M. (1987). Interobserver agreement by auscultation in the presence of a third heart sound in patients with congestive heart failure. *Chest* **91,** 870–873.

Klar, N., Lipsitz, S. R., and Ibrahim, J. G. (2000). An estimating equations approach for modelling kappa. *Biometrical Journal, Journal of Mathematical Methods in Biosciences* **42,** 45–58.

Landis, J. R. and Koch, G. G. (1977). A one-way components of variance model for categorical data. *Biometrics* **33,** 671–679.

Lee, J. J. and Tu, Z. N. (1994). A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes. *Journal of Computational and Graphical Statistics* **3,** 301–321.

Mak, T. K. (1988). Analysing intraclass correlation for dichotomous variables. *Applied Statistics* **37,** 344–352.

McLaughlin, J. K., Diet, M. S., Mehl, E. S., and Blot, W. J. (1987). Reliability of surrogate information on cigarette smoking by type of informant. *American Journal of Epidemiology* **126,** 144–146.

McLellan, A. T., Luborsky, L., Cacciola, J., Griffith, J., Evans, F., Barr, H. L., and O'Brien, C. P. (1985). New data from the addition severity index. *Journal of Nervous and Mental Disease* **173,** 412–423.

Nam, J. (2000). Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics* **56,** 583–585.

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

Rao, C. R. and Mukerjee, R. (1997). Comparison of LR, score, and Wald tests in a non-iid setting. *Journal of Multivariate Analysis* **60,** 99–110.

Reed, J. F. (2000). Homogeneity of kappa statistics in multiple samples. *Computer Methods and Programs in Biomedicine* **63,** 43–46.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* **19,** 321–325.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research* **7,** 301–317.

Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods,* 8th edition. Ames: Iowa State University.

Thompson, W. D. and Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* **41,** 949–958.

Westlund, K. B. and Kurland, L. T. (1953). Studies on multiple sclerosis in Winnipeg, Manitoba, and New Orleans, Louisiana. I. Prevalence, comparison between the patient groups in Winnipeg and New Orleans. *American Journal of Hygiene* **57,** 380–396.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22,** 209–212.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin* **103,** 374–378.

Received June 2001. *Revised November* 2001.
Accepted November 2001.

Appendix

Let $\Delta = \kappa_1 - \kappa_2$ denote the difference between κ_1 and κ_2, estimated by $\hat{\Delta} = \hat{\kappa}_1 - \hat{\kappa}_2$. As N_1 and N_2 become large, it follows from the central limit theorem that

$$\frac{(\hat{\Delta} - \Delta)^2}{\text{var}(\hat{\Delta})} \approx \chi^2_{1,1-\alpha}, \tag{1}$$

which implies that the confidence limits (L, U) may be obtained as

$$\hat{\Delta} \pm \left\{\chi^2_{1,1-\alpha}\text{var}(\hat{\Delta})\right\}^{1/2}. \tag{2}$$

The limits L and U can also be recognized as the minimum and maximum values of Δ that satisfy equation (1). Therefore, we can estimate $\text{var}(\hat{\Delta})$ in (2) by $\widehat{\text{var}}(\hat{\Delta})|_{\Delta=\min\Delta}$ and $\widehat{\text{var}}(\hat{\Delta})|_{\Delta=\max\Delta}$ for L and U, respectively.

Suppose now we compute separate $100(1-\alpha)\%$ confidence limits for κ_1 and κ_2, denoted by (l_1, u_1) and (l_2, u_2). Then the values of $\min\Delta$ and $\max\Delta$ are given by $l_1 - u_2$ and $u_1 - l_2$, respectively, and by equation (2), we have

$$L = \hat{\Delta} - \left\{\chi^2_{1,1-\epsilon}\,\widehat{\text{var}}\,(l_1 - u_2)\right\}^{1/2}$$

and

$$U = \hat{\Delta} + \left\{\chi^2_{1,1-\alpha}\widehat{\text{var}}\,(u_1 - l_2)\right\}^{1/2}.$$

This formulation is analogous to that used by Newcombe (1998) in constructing a confidence interval for the difference between two independent proportions.

Confidence limits for κ_h $(h = 1, 2)$ can be obtained using either the goodness-of-fit approach or the profile variance approach. Using the goodness-of-fit procedure and dropping the subscript h, the limits are obtained by solving the cubic equation given by $A\kappa^3 + B\kappa^2 + C\kappa + D = 0$, where $A =$

$4N\hat{\pi}^2(1-\hat{\pi})^2(\chi^2_{1,1-\alpha}+N)$, $B = 4N\hat{\pi}(1-\hat{\pi})[n_2+\{1-2\hat{\pi}(1-\hat{\pi})\}\chi^2_{1,1-\alpha}]-A$, $C = n_2^2-4N\hat{\pi}(1-\hat{\pi})\{1-4\hat{\pi}(1-\hat{\pi})\}\chi^2_{1,1-\alpha}-A$, and $D = \{n_2^2 - 2N\hat{\pi}(1-\hat{\pi})\}^2 + 4N\hat{\pi}^2(1-\hat{\pi})^2 - A$. Now, let $a_1 = B/A$, $a_2 = C/A$, $a_3 = D/A$, $R = a_1a_2/6 - a_3/2 - a_1^3/27$, and $Q = a_2/3 - a_1^2/9$. Then the confidence limits for κ are given by $l = 2(-Q)^{1/2}\cos\{(\theta + 4\Pi)/3\} - a_1/3$ and $u = 2(-Q)^{1/2}\cos(\theta/3) - a_1/3$, where $\cos\theta = R/(-Q)^{3/2}$ and $\Pi = 3.1415927$.

Similarly, using the single-sample profile variance method, the resulting limits are found by obtaining the solution to the cubic equation $A\kappa^3 + B\kappa^2 + C\kappa + D = 0$, where $A = 1/\{2\hat{\pi}(1-\hat{\pi})\} - 2$, $B = -N/\chi^2_{1,1-\alpha} - 3/\{2\hat{\pi}(1-\hat{\pi})\} + 5$, $C = 2N\hat{\kappa}/\chi^2_{1,1-\alpha} + 1/\{\hat{\pi}(1-\hat{\pi})\} - 4$, and $D = -N\hat{\kappa}^2/\chi^2_{1,1-\alpha} + 1$. Now, if $\hat{\pi} \neq 0.5$, let $a_1 = B/A$, $a_2 = C/A$, $a_3 = D/A$, $R = a_1a_2/6 - a_3/2 - a_1^3/27$, and $Q = a_2/3 - a_1^2/9$. The resulting confidence limits for κ are then given by $l = 2(-Q)^{1/2}\cos\{(\theta+2\Pi)/3\} - a_1/3$ and $u = 2(-Q)^{1/2}\cos\{(\theta + 4\Pi)/3\} - a_1/3$, where $\cos\theta = R/(-Q)^{3/2}$ and $\Pi = 3.1415927$. If $\hat{\pi} = 0.5$, then the limits are given by $(l, u) = \{-C \pm (C^2 - 4BD)^{1/2}\}/(2B)$.

Psychological Methods
2007, Vol. 12, No. 4, 399–413

1082-989X/07/$12.00 DOI: 10.1037/1082-989X.12.4.399

Toward Using Confidence Intervals to Compare Correlations

Guang Yong Zou
University of Western Ontario and Robarts Research Institute

Confidence intervals are widely accepted as a preferred way to present study results. They encompass significance tests and provide an estimate of the magnitude of the effect. However, comparisons of correlations still rely heavily on significance testing. The persistence of this practice is caused primarily by the lack of simple yet accurate procedures that can maintain coverage at the nominal level in a nonlopsided manner. The purpose of this article is to present a general approach to constructing approximate confidence intervals for differences between (a) 2 independent correlations, (b) 2 overlapping correlations, (c) 2 nonoverlapping correlations, and (d) 2 independent R^2s. The distinctive feature of this approach is its acknowledgment of the asymmetry of sampling distributions for single correlations. This approach requires only the availability of confidence limits for the separate correlations and, for correlated correlations, a method for taking into account the dependency between correlations. These closed-form procedures are shown by simulation studies to provide very satisfactory results in small to moderate sample sizes. The proposed approach is illustrated with worked examples.

Keywords: bootstrap, coefficient of determination, confidence interval, hypothesis testing, multiple regression

Supplemental materials: http://dx.doi.org/10.1037/1082-989x.12.4.399.supp

Statistical inference is often conducted through significance testing and confidence interval construction. Although closely related, significance testing focuses on a single priori hypothesis, usually a null value (e.g., $\rho = 0$). In contrast, a confidence interval can avoid such problems by providing a range of plausible parameter values. For a given investigation, the confidence interval reveals both the magnitude and the precision of the estimated effect, whereas the *p* value obtained from significance testing confounds these two aspects of the data. Thus, confidence interval construction in principle is preferred, as discussed by Cohen (1994) in general terms and by Olkin and Finn (1995) in the case of comparing correlations.

Guang Yong Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada, and Robarts Clinical Trials, Robarts Research Institute, London, Ontario, Canada.

This research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. I am grateful to Allan Donner for valuable suggestions that led to substantial improvements in the article.

Correspondence concerning this article should be addressed to Guang Yong Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario N6A 5C1, Canada. E-mail: gzou@robarts.ca

Two distinctive types of correlational analyses are common. The first one involves the simple correlation, which is a measure of linear relationship between two random variables. As a concept, it can be viewed in at least 14 different ways (Rodgers & Nicewander, 1988; Rovine & von Eye, 1997). The second type of correlation analysis usually involves the use of the multiple correlation coefficient to quantify the proportion of criterion variation explained by random predictors (Helland, 1987). This correlation usually appears in the squared form commonly seen in multiple regression models and referred to as R^2 or the coefficient of determination (Cohen, Cohen, West, & Aiken, 2003).

The choice between correlation and the squared correlation as the effect measure in a given investigation may not be as simple as the descriptions given here. A resolution of this issue is not the focus of this article. Readers may consult Ozer (1985) and Steiger and Ward (1987) for different viewpoints. A summary of related effect size measures has recently been provided by Kirk (2007).

The confidence interval for a single correlation ρ is often obtained using Fisher's *r* to *z* transformation because the sampling distribution of *r* is negatively skewed. Specifically, one first forms confidence limits for $z(\rho) = 1/2 \ln[(1 + \rho)/(1 - \rho)]$ and then back-transforms the resultant limits to obtain a confidence interval for ρ. Unfortunately, using the same idea for $\rho_1 - \rho_2$ will fail because the limits

for $z(\rho_1) - z(\rho_2)$ cannot be back-transformed to obtain the interval for $\rho_1 - \rho_2$ (Meng, Rosenthal, & Rubin, 1992; Olkin & Finn, 1995). Therefore, although the confidence interval approach is preferred in general (Olkin & Finn, 1995), the significance testing approach still dominates in the comparison of correlations (Cheung & Chan, 2004; Meng et al., 1992; Olkin & Finn, 1990; Raghunathan, Rosenthal, & Rubin, 1996; Silver, Hittner, & May, 2006; Steiger, 1980).

The problem is even more difficult in the case of squared correlations. It is commonplace for statistical software packages to provide R^2, but rarely with a confidence interval to quantify its precision, primarily because of its complicated sampling distribution (Fisher, 1928). Tabulated exact confidence limits for the population parameter ρ^2 are available (Kramer, 1963; Lee, 1972) but have rarely been used. Simpler approximate confidence interval procedures with good performance have been proposed (Helland, 1987; Lee, 1971), although it is unclear how to use such procedures to obtain confidence intervals for differences between two R^2s. An approach that ignores the asymmetry of the sampling distribution of R^2 has been suggested (Olkin & Finn, 1995) but has poor performance (Algina & Keselman, 1999).

The purpose of this article is to present a general approach to constructing confidence intervals for a difference between correlations. I show how to apply the approach to obtain confidence intervals for (a) differences between two independent simple correlations, (b) differences between overlapping correlations arising from the case in which two correlations involve a common variable, (c) differences between nonoverlapping correlations arising from the case in which two correlated correlations have no common variable involved, and (d) differences between two independent squared multiple correlations. The performance of the proposed approach is compared with that of Olkin and Finn (1995) via a Monte Carlo simulation. I then present worked numerical examples that illustrate the calculations involved in each of the four cases. Finally, I conclude with a brief discussion of the advantages of the approach for practicing researchers.

Confidence Intervals for Differences Between Correlations

Simple Asymptotic Methods

A commonly used approach to setting approximate confidence intervals for a parameter θ is to invoke the central limit theorem, resulting in $100(1 - \alpha)\%$ confidence limits, (L, U), given by

$$(L, U) = \hat{\theta} - z_{\alpha/2}\hat{\sigma}, \hat{\theta} + z_{\alpha/2}\hat{\sigma}, \quad (1)$$

where $\hat{\theta}$ is the sample estimate of θ, $\hat{\sigma}$ is the estimate of its standard deviation, and $z_{\alpha/2}$ is the $100 \cdot \alpha/2$ percentile point of a standard normal distribution. For example, for a 95% two-sided confidence interval, $z_{\alpha/2} = 1.96$. I refer to this approach as the simple asymptotic (SA) method.

Differences Between Two Correlation Coefficients

The construction of a confidence interval for a difference between two correlations may be useful in practice. For instance, a human resources manager may want to use a personnel selection test to select both male and female employees. Before the manager did this, it would be of interest to examine the difference in correlations between job performance and test score, as obtained from data on existing employees. A p value from the null hypothesis testing of $\rho_{male} = \rho_{female}$ would provide far less information than would a confidence interval.

The comparison of correlated correlations has been of interest in practice, as seen by extensive citations of key articles focusing on hypothesis tests of the difference between correlated correlations (Meng et al., 1992; Raghunathan et al., 1996). As a concrete example in psychological research, the examination of whether a variable acts as a suppressor with respect to two other variables can be meaningfully informed by the presentation of the confidence interval for two correlated correlations.

A direct application of the SA method yields a confidence interval for a difference between two correlations, $\rho_1 - \rho_2$, given by (Olkin & Finn, 1995)

$$(L, U) = r_1 - r_2 \mp z_{\alpha/2}\sqrt{\widehat{\text{var}}(r_1) + \widehat{\text{var}}(r_2)}, \quad (2)$$

when the sample estimates r_1 and r_2 are independent, and

$$(L, U) = r_1 - r_2 \mp z_{\alpha/2}\sqrt{\widehat{\text{var}}(r_1) + \widehat{\text{var}}(r_2) - 2\widehat{\text{cov}}(r_1, r_2)}, \quad (3)$$

when r_1 and r_2 are dependent with covariance given by $\widehat{\text{cov}}(r_1, r_2)$.

There are two cases that may be distinguished in comparing dependent correlations. The first case may be referred to as *overlapping*: Two correlations are calculated from the same sample with a common variable involved (Meng et al., 1992). The second case, commonly referred to as *nonoverlapping*, describes a situation in which two correlations are obtained from the same sample without common variables involved (Raghunathan et al., 1996). Although different in interpretation, these two cases are identical from a statistical perspective. The fundamental issue is to take into account the dependency between two correlations.

On the basis of a result from Pearson and Filon (1898, p. 262, Equation xl), the covariance between two nonoverlapping correlations (correlations without a common subscript) may be approximated by

$$\text{cov}(r_{ij}, r_{kl}) = [.5\rho_{ij}\rho_{kl}(\rho_{ik}^2 + \rho_{il}^2 + \rho_{jk}^2 + \rho_{jl}^2) + \rho_{ik}\rho_{jl} + \rho_{il}\rho_{jk} - (\rho_{ij}\rho_{ik}\rho_{il} + \rho_{ij}\rho_{jk}\rho_{jl} + \rho_{ik}\rho_{jk}\rho_{kl} + \rho_{il}\rho_{jl}\rho_{kl})]/n, \quad (4)$$

where n is the sample size. This result has also appeared in Olkin and Siotani (1976). Substituting i for k and j for l in Equation 4 yields the covariance between two overlapping correlations (correlations with a common subscript):

$$\text{cov}(r_{ij}, r_{ik}) = [(\rho_{jk} - .5\rho_{ij}\rho_{ik})(1 - \rho_{ij}^2 - \rho_{ik}^2 - \rho_{jk}^2) + \rho_{jk}^3]/n. \quad (5)$$

Equation 5 reduces further to the variance for a single correlation r_{ij} when k is replaced with j:

$$\text{var}(r_{ij}) = (1 - \rho_{ij}^2)^2/n. \quad (6)$$

Bear in mind that $\rho_{jj} = 1$ because the correlation of a variable with itself is 1. SA variance estimates, denoted by $\widehat{\text{var}}$, and covariance estimates, denoted by $\widehat{\text{cov}}$, may be obtained by replacing each population parameter ρ with corresponding sample value r in Equations 4, 5, and 6. Equations 2 and 3 may then be used to construct confidence intervals. For example, an SA $100(1 - \alpha)\%$ confidence interval for a difference between two independent correlations $\rho_1 - \rho_2$ is given by (L, U), with

$$L = r_1 - r_2 - z_{\alpha/2}\sqrt{\frac{(1 - r_1^2)^2}{n_1} + \frac{(1 - r_2^2)^2}{n_2}} \quad (7)$$

and

$$U = r_1 - r_2 + z_{\alpha/2}\sqrt{\frac{(1 - r_1^2)^2}{n_1} + \frac{(1 - r_2^2)^2}{n_2}}, \quad (8)$$

where n_1 and n_2 are the sample sizes of the two comparing groups.

Differences Between Two Squared Correlation Coefficients

Confidence interval construction for a difference between two independent R^2s, $\rho_1^2 - \rho_2^2$, is useful if one is interested in determining the predictive power of a set of predictors for two independent populations. For example, one might ask how large a difference is present when a battery of entrance tests is used to predict academic performance for male compared with female college students.

The SA approach may also be used to construct confidence intervals for differences between two squared correlations (ρ^2), with variance estimated by the delta (δ) method (Rao, 1973, p. 388). The delta method is a general procedure that uses the Taylor series expansion of a function of one or more random variables to obtain approximations to the mean of the function and to its variance. Suppose $\hat{\theta}$ has mean θ and variance $\text{var}(\hat{\theta})$; then the mean and variance of $g(\hat{\theta})$ are given by $g(\theta)$ and $[g'(\theta)]^2\text{var}(\hat{\theta})$, where $g'(\theta)$ is the derivative of the function g evaluated at $\hat{\theta} = \theta$. Application of the delta method to R^2 yields

$$\text{var}(R^2) = (2\rho)^2\text{var}(r) = 4\rho^2(1 - \rho^2)^2/n,$$

which may be estimated by substituting R^2 for ρ^2. A $100(1 - \alpha)\%$ confidence interval for the difference between two independent R^2s is given by

$$(L, U) = R_1^2 - R_2^2 \mp z_{\alpha/2}\sqrt{4R_1^2(1 - R_1^2)^2/n_1 + 4R_2^2(1 - R_2^2)^2/n_2},$$

where n_1 and n_2 are sample sizes for the two groups.

Deficiency of the SA Approach

The SA confidence intervals in general are simple to apply and thus have become almost universal (DiCiccio & Efron, 1996; Efron & Tibshirani, 1993). However, one needs to be aware that the validity relies crucially on two conditions: (a) The sampling distribution of $\hat{\theta}$ does not change with the value of the underlying parameter θ, and (b) the sampling distribution is close to the standard normal. Deviation from either of the two conditions will invalidate the SA method. In the present context, the SA method could provide confidence intervals containing values outside the range of -1 to 1 for a single correlation and values outside of -2 to 2 for a difference between two correlations. Similarly, the SA method may result in confidence intervals containing values outside the range of 0 to 1 for a squared correlation and values outside the range of -1 to 1 for a difference between two squared correlations. Large sample sizes may improve the performance of the SA method somewhat, but not completely. As Efron (2003, p. 137) pointed out, in the context of a single correlation, the SA method may provide adequate overall coverage with large samples, but it does so in a lopsided fashion. In other words, failure of the confidence interval to capture the true parameter value may be concentrated in one tail. A hidden deficiency of the SA approach is that it may produce results conflicting with that of hypothesis testing, because the latter is usually conducted on the Fisher's z scale (Meng et al., 1992).

It was the above deficiency that motivated the development of bootstrap confidence intervals (DiCiccio & Efron, 1996; Efron, 1979, 1981, 1985, 1987b; Efron & Tibshirani, 1993, chap. 12–14, 22). Increased computational efforts aside, bootstrap confidence intervals are intended to be an improvement over the SA method, although in many cases the question becomes whether the improvement is sufficient to be accurate. The answer to this question relies crucially on whether the assumptions underpinning bootstrap confidence intervals are satisfied. For example, the bias-corrected

method (Efron, 1981) requires the existence of a monotone increasing function g such that $g(\hat{\theta}) - g(\theta)$ has the same normal distribution for all parameter values of θ. See Schenker (1985) for an example of how the bootstrap fails in constructing confidence intervals for a normal variance. The improved bias-corrected bootstrap method, termed bias-corrected and accelerated (*BCa*; Efron, 1987b) relaxes the requirement of g from being both normalizing and variance stabilizing to being only normalizing. This is achieved by calculating an acceleration constant. Therefore, the validity *BCa* depends on the accuracy of the estimated acceleration constant and the existence of the normalizing transformation, although one does not need to know what the transformation is. Currently there are no simple approaches to accurate estimation of the acceleration constant in general (Shao & Tu, 1995, chap. 4). More discussion on bootstrap confidence intervals can be found elsewhere (Carpenter & Bithell, 2000; Young, 1994). Efron (1988) aptly stated

> A good way to think of bootstrap intervals is as a cautious improvement over standard intervals, using large amounts of computation to overcome certain deficiencies of the standard methods, for example its lack of transformation invariance. The bootstrap is not intended to be a substitute for precise parametric results but rather a way to reasonably proceed when such results are unavailable. (p. 295)

In the next section, I apply available results for single correlations (e.g., Fisher's z transformation for correlation and F distribution approximation for the sampling distribution of the squared correlation; Lee, 1971) to construct confidence intervals for differences between correlations. Similar to the bootstrap, these procedures attempt to provide improvement over the SA method. Contrary to most bootstrap methods, which demand intensive computation, the proposed procedures are in closed form and may be calculated using a hand-held calculator.

Modified Asymptotic Methods

The sampling distributions for single r or R^2 are highly skewed, requiring both large sample sizes and middle sized correlations for the SA method to be accurate, that is, to provide adequate coverage in a nonlopsided fashion. On the other hand, Fisher's z transformation for single correlations and F distribution based confidence intervals for ρ^2 have been known to perform very well (Lee, 1971). The accuracy of these procedures originates largely from respecting the asymmetric feature of the sampling distributions.

I now describe a procedure for setting approximate confidence intervals for a difference between two parameters (either two correlations or two squared correlations) that is asymmetry respecting. As I make clear below, this method is an extension of the SA method, which I refer to as the modified asymptotic (MA) method. In what follows, I use uppercase (L, U) to denote confidence limits for differences between correlations and lowercase (l, u), with subscripts representing comparison groups if needed, to represent confidence limits for single correlations.

Confidence limits (l, u) that reflect the asymmetric sampling distribution of $\hat{\theta}$ may be seen as doing so through different variance estimates for l and u, that is,

$$l = \hat{\theta} - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta})_l}$$

and

$$u = \hat{\theta} + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta})_u}.$$

Equivalently, when $\theta = l$,

$$\widehat{\text{var}}(\hat{\theta})_l = (\hat{\theta} - l)^2/z^2_{\alpha/2}, \tag{9}$$

and when $\theta = u$,

$$\widehat{\text{var}}(\hat{\theta})_u = (u - \hat{\theta})^2/z^2_{\alpha/2}. \tag{10}$$

To obtain confidence limits for a difference $\theta_1 - \theta_2$, I exploit the relationship between hypothesis testing and confidence limits, recognizing that the lower (L) and upper (U) confidence limits are the minimum and maximum parameter values that, asymptotically, satisfy

$$\frac{[(\hat{\theta}_1 - \hat{\theta}_2) - L]^2}{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)} = z^2_{\alpha/2}$$

and

$$\frac{[U - (\hat{\theta}_1 - \hat{\theta}_2)]^2}{\text{var}(\hat{\theta}_1) + \text{var}(\hat{\theta}_2)} = z^2_{\alpha/2},$$

respectively.

Suppose now that we have two sample estimates and associated $100(1 - \alpha)\%$ confidence limits obtained from two independent samples $\hat{\theta}_1$ (l_1, u_1) and $\hat{\theta}_2$ (l_2, u_2), which contain the plausible values of θ_1 and θ_2, respectively. Among these plausible values for θ_1 and θ_2, the value closest to the minimum L is $l_1 - u_2$, and the value closest to the maximum U is $u_1 - l_2$. Therefore, it is reasonable to estimate the variance of $\hat{\theta}_1 - \hat{\theta}_2$ when $\theta_1 = l_1$ and $\theta_2 = u_2$ for setting L. With Equations 9 and 10, we have

$$\begin{aligned} L &= \hat{\theta}_1 - \hat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\theta}_1)_{l1} + \widehat{\text{var}}(\hat{\theta}_2)_{u2}} \\ &= \hat{\theta}_1 - \hat{\theta}_2 - z_{\alpha/2}\sqrt{\frac{(\hat{\theta}_1 - l_1)^2}{z^2_\alpha} + \frac{(u_2 - \hat{\theta}_2)^2}{z^2_\alpha}} \\ &= \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2}. \end{aligned} \tag{11}$$

Similar steps result in the upper limit as

$$U = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2}. \quad (12)$$

These confidence limits have been applied to the case of comparing intraclass kappa coefficients, which are indices commonly used in interobserver agreement and reliability studies (Donner & Zou, 2002). Notice that the expressions for L and U may also be extended to include a covariance term when $\hat{\theta}_1$ and $\hat{\theta}_2$ are dependent. Let $\widehat{\text{corr}}\,(\hat{\theta}_1, \hat{\theta}_2)$ be the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$; Equations 11 and 12 may then be extended to

$$L = \hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2 - 2\widehat{\text{corr}}(\hat{\theta}_1, \hat{\theta}_2)(\hat{\theta}_1 - l_1)(u_2 - \hat{\theta}_2)} \quad (13)$$

and

$$U = \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2 - 2\widehat{\text{corr}}(\hat{\theta}_1, \hat{\theta}_2)(u_1 - \theta_1)(\hat{\theta}_2 - l_2)}. \quad (14)$$

I now apply these general results to a difference between two correlations.

Differences Between Two Correlations

For a difference between two correlations, let (l_i, u_i), $i = 1, 2$ be the $(1 - \alpha) \times 100\%$ confidence limits for ρ_i obtained using Fisher's z transformation with data collected on two independent groups. The confidence limits (L, U) for $\rho_1 - \rho_2$ may be obtained, with r_1 and r_2 replacing $\hat{\theta}_1$ and $\hat{\theta}_2$ in Equations 11 and 12, respectively, as

$$\begin{cases} L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2} \\ U = r_1 - r_2 + \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2}. \end{cases} \quad (15)$$

The confidence limits in Equation 15 reduce to those in Equations 7 and 8, if the SA method has been used to obtain the confidence limits for single correlations. This is because

$$l_1, u_1 = r_1 \mp z_{\alpha/2}\sqrt{\frac{(1 - r_1^2)^2}{n_1}},$$

which yields

$$(r_1 - l_1)^2 = (u_1 - r_1)^2 = z_{\alpha/2}^2 \frac{(1 - r_1^2)^2}{n_1},$$

and

$$l_2, u_2 = r_2 \mp z_{\alpha/2}\sqrt{\frac{(1 - r_2^2)^2}{n_2}},$$

which results in

$$(r_2 - l_2)^2 = (u_2 - r_2)^2 = z_{\alpha/2}^2 \frac{(1 - r_2^2)^2}{n_2}.$$

In fact, this relationship between the MA and SA methods holds in general, including in the procedures presented below. This insight also highlights the key feature of the MA approach: It acknowledges the asymmetric nature of the sampling distributions for the single correlations, whereas the SA approach ignores this fact.

Equations 13 and 14 may be applied to overlapping and nonoverlapping correlations, yielding

$$\begin{cases} L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2 - 2\widehat{\text{corr}}(r_1, r_2)(r_1 - l_1)(u_2 - r_2)} \\ U = r_1 - r_2 + \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2 - 2\widehat{\text{corr}}(r_1, r_2)(u_1 - r_1)(r_2 - l_2)}, \end{cases}$$

where the correlation between two correlations can be estimated by

$$\widehat{\text{corr}}(r_1, r_2) = \widehat{\text{cov}}(r_1, r_2)/\sqrt{\widehat{\text{var}}(r_1)\widehat{\text{var}}(r_2)},$$

resulting in difference estimates depending on whether two correlations being compared share a common third variable.

Differences Between Squared Correlation Coefficients

The same idea may be applied to squared multiple correlation coefficients, provided accurate confidence intervals about single R^2 are available. Because of the complexity of the sample distribution for R^2 (Fisher, 1928; Rao, 1973, p. 599), one might consider Fisher's z transformation for R^2. However, Gajjar (1967) has shown that the limiting distribution of Fisher's z transformation of R^2 does not approach normality as sample size increases (see Lee, 1971, for empirical evidence). In addition, Alf and Graf (1999) have pointed out that in this case, "Fisher's z values are severely truncated in the lower tail, resulting in a distribution that is even more positively skewed than is the original distribution of squared multiple correlations" (p. 74). The key here is that one should not confuse the sampling distribution of $\sqrt{R^2}$ with that of r, as the former can only take positive values.

A stand-alone computer program implementing an exact confidence interval for ρ^2 based on Fisher (1928) has been provided by Steiger and Fouladi (1992; available at www.statpower.net). A noncentral F distribution approximate confidence interval (Lee, 1971) can also be obtained using a bisection method implementable with common statistical software (see the Appendix for an outline of the theory and supplemental materials for SAS and SPSS codes). Also note that SAS PROC CANCORR has implemented the approx-

imate confidence interval based on the F distribution (Helland, 1987; Lee, 1971), available with the option SMC (which stands for squared multiple correlations).

Letting (l_i, u_i), $i = 1, 2$ denote $100(1 - \alpha)\%$ lower and upper limits about R^2 obtained using the noncentral F approximation, confidence limits about the difference between two independent R^2s are given (with R_1^2 and R_2^2 replacing $\hat{\theta}_1$ and $\hat{\theta}_2$ in Equations 11 and 12) by

$$\begin{cases} L = R_1^2 - R_2^2 - \sqrt{(R_1^2 - l_1)^2 + (u_2 - R_2^2)^2} \\ U = R_1^2 - R_2^2 + \sqrt{(u_1 - R_1^2)^2 + (R_2^2 - l_2)^2} \end{cases}.$$

Simulation Studies

The theoretical properties of the proposed MA approach are asymptotic. Simulation studies were therefore undertaken to evaluate the performance as compared with that in the SA approach. The simulation evaluation used 10,000 replicates for each parameter combination considered in each of the five cases. All computations were conducted using SAS proc iml, with codes made available in the online supplemental materials to this article for readers interested in exploring additional parameter combinations.

Evaluation Criteria Used

The focus here was on the extent to which the empirical coverage of the confidence interval matched with the nominal 95% level. For this purpose, three criteria commonly seen in psychological literature (Bradley, 1978; Robey & Barcikowski, 1992) were adopted: strict criterion, 94.5%–95.5%; moderate criterion, 93.75%–96.25%; and liberal criterion, 92.5%–97.5%.

All too often the literature has been using overall coverage alone to evaluate confidence interval procedures. However, as Efron (2003, p. 137) pointed out, the worst definition of accuracy in confidence interval evaluation is overall coverage alone. Tail errors are also important. For example, suppose for a sampling distribution for $r_1 - r_2$ arising from 10,000 samples with a true difference of 0, that we would consider, at the 95% confidence level, both the lowest 250 and the highest 250 estimates as too extreme to be the plausible parameter values. On the contrary, it would be awkward if we regard only either the lowest 5% or, alternatively, the highest 5% values as extreme—that is, if the entire error probability was contained in one tail. Therefore, in a given simulation study, if two procedures satisfied the coverage criterion, I considered the procedure having the smaller difference between tail errors as preferable.

Empirical coverage percentage was estimated by the relative frequency out of 10,000 intervals that contained the parameter. Tail errors were estimated by calculating the frequencies of the intervals lying completely to the left of the parameter value (missing from left, ML) and those lying completely to the right of the parameter (missing from the right, MR). Average interval width was also calculated as a secondary criterion in the evaluation.

Differences Between Correlations

Overlapping correlations. For each data set with a sample size of $n = 15$, 50, 100, and 200, an $n \times 3$ matrix $\mathbf{X}$ was first generated as $3n$ independent standard normal variates. The desired correlated data were then obtained as $\mathbf{XU}$, where $\mathbf{U}$ is the is the root matrix (Rao, 1973, p. 36) of the correlation matrix determined by three elements (ρ_{12}, ρ_{13}, ρ_{23}) such that

$$\mathbf{UU}^T = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix},$$

where the superscript T denotes transpose of matrix $\mathbf{U}$. The 95% two-sided confidence intervals for $\rho_{12} - \rho_{13}$ were then constructed using both the SA and the MA methods. Using a $2 \times 2 \times 3$ factorial design, I generated all combinations of $\rho_{23} = .1, .3$; $\rho_{12} = .2, .4$; and $\rho_{13} = .1, .5, .7$.

Simulation results in Table 1 indicate that the SA method does not provide adequate coverage percentage for sample sizes that are not greater than 100. For example, when $n = 100$, $\rho_{23} = .3$, $\rho_{12} = .2$, and $\rho_{13} = .7$, the coverage is only 93.97%, falling short of the strict criterion of 94.5%. This means that the confidence intervals from the SA method are overly narrow. Moreover, as the sample size increases, the coverage percentages tend to reach the nominal level from below, rather than hovering around the nominal level. This result indicates that the asymptotic results of the SA method are reached only with a minimum of 200 cases being necessary for adequate overall coverage in the present study. In contrast, the MA confidence intervals provide adequate coverage according to Bradley's (1978) strict criterion in a nonlopsided fashion, even with a sample size as small as 15.

Nonoverlapping correlations. For each data set with a sample size of $n = 15$, 50, 100, and 200, an $n \times 4$ matrix $\mathbf{X}$ was first generated using $4n$ independent standard normal variates. The desired correlated data were then obtained as $\mathbf{XU}$, where $\mathbf{U}$ is the root matrix (Rao, 1973, p. 36) of the correlation matrix determined by six elements (ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24}, ρ_{34}), that is,

$$\mathbf{UU}^T = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{pmatrix}.$$

The 95% two-sided confidence intervals for $\rho_{12} - \rho_{13}$ and $\rho_{12} - \rho_{34}$ were then constructed using the proposed methods and the SA approach.

Table 1
Performance of Procedures for Constructing Two-Sided 95% Confidence Intervals (CIs) for a Difference Between Two Overlapping Correlations ($\rho_{12} - \rho_{13}$) Based on 10,000 Runs

				Modified asymptotic		Simple asymptotic	
n	ρ_{23}	ρ_{12}	ρ_{13}	Coverage (ML, MR) %	CI width	Coverage (ML, MR) %	CI width
15	.1	.2	.3	95.27 (2.18, 2.55)	1.27	91.40 (4.35, 4.25)	1.21
			.5	94.75 (1.83, 3.42)	1.22	90.77 (4.09, 5.14)	1.14
			.7	94.72 (1.43, 3.85)	1.12	89.70 (3.54, 6.76)	1.04
		.4	.3	94.83 (3.00, 2.17)	1.25	91.35 (4.61, 4.04)	1.17
			.5	95.02 (2.15, 2.83)	1.20	92.48 (3.38, 4.14)	1.10
			.7	94.52 (1.68, 3.80)	1.10	90.66 (3.03, 6.31)	.99
	.3	.2	.3	95.14 (2.11, 2.75)	1.14	91.37 (4.19, 4.44)	1.08
			.5	95.23 (1.15, 3.62)	1.10	91.18 (3.24, 5.58)	1.02
			.7	95.12 (1.04, 3.84)	1.03	89.06 (3.24, 7.70)	0.94
		.4	.3	94.69 (3.17, 2.14)	1.12	91.42 (4.68, 3.90)	1.03
			.5	94.88 (2.09, 3.03)	1.09	92.73 (3.06, 4.21)	0.99
			.7	95.07 (1.23, 3.70)	1.01	90.91 (2.22, 6.87)	0.90
50	.1	.2	.3	95.14 (2.33, 2.53)	0.70	93.79 (3.23, 2.98)	0.69
			.5	94.75 (2.11, 3.14)	0.66	93.58 (2.85, 3.57)	0.65
			.7	95.33 (1.82, 2.85)	0.60	93.83 (2.49, 3.68)	0.59
		.4	.3	95.07 (2.43, 2.50)	0.67	94.15 (2.87, 2.98)	0.66
			.5	94.67 (2.55, 2.78)	0.64	93.83 (2.95, 3.22)	0.62
			.7	94.74 (2.10, 3.16)	0.57	93.71 (2.17, 4.12)	0.55
	.3	.2	.3	95.29 (2.23, 2.48)	0.62	94.35 (2.75, 2.90)	0.62
			.5	95.30 (1.91, 2.79)	0.59	94.31 (2.52, 3.17)	0.58
			.7	95.21 (1.78, 3.01)	0.55	93.62 (2.27, 4.11)	0.54
		.4	.3	94.99 (2.75, 2.26)	0.60	94.12 (3.19, 2.69)	0.59
			.5	95.14 (2.30, 2.56)	0.57	94.62 (2.50, 2.88)	0.55
			.7	94.62 (2.05, 3.33)	0.52	93.71 (2.04, 4.25)	0.50
100	.1	.2	.3	94.89 (2.48, 2.63)	0.50	94.43 (2.82, 2.75)	0.50
			.5	95.02 (2.22, 2.76)	0.47	94.42 (2.67, 2.91)	0.46
			.7	95.29 (2.02, 2.69)	0.42	94.47 (2.38, 3.15)	0.42
		.4	.3	95.06 (2.61, 2.33)	0.48	94.62 (2.78, 2.60)	0.47
			.5	94.52 (2.62, 2.86)	0.44	94.17 (2.78, 3.05)	0.44
			.7	95.24 (1.99, 2.77)	0.40	94.78 (1.99, 3.23)	0.39
	.3	.2	.3	95.21 (2.52, 2.27)	0.44	94.75 (2.87, 2.38)	0.44
			.5	95.15 (1.99, 2.86)	0.42	94.61 (2.36, 3.03)	0.41
			.7	94.71 (2.08, 3.21)	0.38	93.97 (2.26, 3.77)	0.38
		.4	.3	95.32 (2.37, 2.31)	0.43	94.99 (2.53, 2.48)	0.42
			.5	94.92 (2.44, 2.64)	0.40	94.65 (2.53, 2.82)	0.39
			.7	95.30 (1.94, 2.76)	0.36	94.81 (1.87, 3.32)	0.36
200	.1	.2	.3	95.13 (2.20, 2.67)	0.35	94.85 (2.44, 2.71)	0.35
			.5	95.03 (2.35, 2.62)	0.33	94.79 (2.58, 2.63)	0.33
			.5	95.10 (2.32, 2.58)	0.30	94.81 (2.43, 2.76)	0.30
		.4	.3	95.09 (2.50, 2.41)	0.34	94.84 (2.58, 2.58)	0.34
			.5	94.82 (2.60, 2.58)	0.31	94.66 (2.67, 2.67)	0.31
			.7	95.14 (2.17, 2.69)	0.28	94.94 (2.15, 2.91)	0.28
	.3	.2	.3	95.16 (2.36, 2.48)	0.31	94.97 (2.50, 2.53)	0.31
			.5	94.56 (2.32, 3.12)	0.29	94.38 (2.46, 3.16)	0.29
			.7	95.13 (1.92, 2.95)	0.27	94.83 (1.98, 3.19)	0.27
		.4	.3	94.78 (2.67, 2.55)	0.30	94.54 (2.76, 2.70)	0.30
			.5	94.94 (2.47, 2.59)	0.28	94.77 (2.58, 2.65)	0.28
			.7	95.25 (2.24, 2.51)	0.25	95.05 (2.12, 2.83)	0.25

Note. Ideally missing left (ML) and missing right (MR) should be 2.50%. Sample size is *n*.

Table 2 presents typical results for a variety of parameter combinations. The SA method again fails to provide adequate coverage for sample sizes of 100 or fewer. Again, as sample size increased, the coverage percentages for the SA method tended to reach the nominal level through a lopsided fashion. The MA method again performed very well on the basis of Bradley's (1978) strict criterion at all sample sizes considered.

Differences Between Two Independent R^2*s*

For a given parameter ρ^2, I used patterned correlation matrices to generate data to use in the evaluation of procedures for differences between R^2s. Consider a situation in which each of the k predictors has the identical correlation (ρ_y) with the criterion in the population. Similarly, each pair of predictors has a common correlation (ρ_x) in the population. Maxwell (2000) has shown that the population parameter ρ^2 is then given by

$$\rho^2 = \frac{k\rho_y^2}{1 + (k - 1)\rho_x}. \tag{16}$$

Equation 16 can be used to obtain any one of the four parameters ρ^2, ρ_y, ρ_x, or k, when the other three are known. The root matrix (Rao, 1973, p. 36) of the correlation matrix was again used to generate multivariate normal data.

Denote ρ_1^2 and ρ_2^2 as the two population parameter values of squared correlation coefficients and n_1 and n_2 as the corresponding sample sizes for the two comparison groups. I considered 48 parameter combinations ($2n_1 \times 4n_2 \times 3\rho_1^2 \times 2\rho_2^2$): $n_1 = 100, 200$; $n_2 = 50, 100, 200, 500$; $\rho_1^2 = .2, .5, .8$; and $\rho_2^2 = .3, .5$. The results are presented in Table 3. I present only the results for $k = 6$, as $k = 3$ showed similar trends. The SA method resulted in adequate coverage for 9 parameter combinations. Among those outside the range, all fell below 94.56%, even when the sample size was as large as 500. For example, for $\rho_1^2 = \rho_2^2 = 0.5$, the sample size combination of 100 and 500 provided a coverage of only 92.29%. Such poor performance of the SA method is not unexpected and is consistent with previous simulation studies (Algina, 1999; Algina & Keselman, 1999; Algina & Moulder, 2001).

The MA method provided a coverage percentage within the range of 94.5%–95.5%, specified by Bradley's (1978) strict criterion in 18 of 48 parameter combinations. Among those outside this range, all but one case showed coverage within the range of 95.5%–96.00%. Thus, the magnitude of the coverage failures was very small and within the range of Bradley's moderate criterion.

Worked-Out Examples

On the basis of the simulation results that showed that only the proposed procedures may be recommended for practical use, I now illustrate the calculations using examples from the published literature. The first three examples illustrate the value of confidence intervals in highlighting the imprecision of parameter estimates with small sample sizes.

Example 1: Independent Correlations

As older populations increase in industrialized countries, research into the association between diet and diseases in older persons has become a focus of public health researchers. However, meaningful research results rely crucially on valid instruments to quantify relevant nutrients in the diet. Two common instruments for this purpose are interviews by certified nutritionists and self-administered food frequency questionnaires, with the former being more accurate but costly, whereas the latter is more feasible but less accurate. Various questionnaires have been developed and tested for comparative validity, as measured by correlation between the nutrient intake levels estimated from questionnaires and dietary interviews. A high correlation between two instruments may provide rationale for the use of self-administered questionnaires, increasing the feasibility for use in large-scale research.

A study by Morris, Tangney, Bienias, Evans, and Wilson (2003) evaluated the validity of a self-administered food frequency questionnaire, as compared with that of a 24-hr dietary recall interview, in a group of older persons. Among the objectives, it was of interest to determine how big a difference in validity exists between female and male subjects. In particular, it was believed that male subjects may be less patient in completing self-administered questionnaires, thus resulting in lower validity. The magnitude of such a difference is useful to determine whether the self-administered questionnaires should be sent to male subjects. From the data of 145 female subjects, the correlation between self-administered questionnaires and telephone interview was $r_1 = .49$, whereas that for 87 male subjects was $r_2 = .36$. Applying Fisher's z transformation, a 95% confidence interval for the validity correlation of the female group is given by

$$\frac{\exp(2l) - 1}{\exp(2l) + 1}, \quad \frac{\exp(2u) - 1}{\exp(2u) + 1},$$

where l and u are given by

$$\tfrac{1}{2}\ln\frac{1 + 0.49}{1 - 0.49} \mp 1.96\sqrt{\frac{1}{145 - 3}} = .3716, .7005.$$

The resultant 95% confidence interval for female subjects is (.355, .605). Similarly, the confidence interval for male subjects is (.162, .530). Therefore, the 95% confidence interval for the difference between the two correlations is given by

Table 2
Performance of Procedures for Constructing Two-Sided 95% Confidence Intervals (CIs) for a Difference Between Two Nonoverlapping Correlations ($\rho_{12} - \rho_{34}$) Based on 10,000 Runs ($\rho_{14} = \rho_{23} = \rho_{24} = .1$)

				Modified asymptotic		Simple asymptotic	
n	ρ_{13}	ρ_{12}	ρ_{34}	Coverage (ML, MR) %	CI width	Coverage (ML, MR) %	CI width
15	.0	.5	.1	94.78 (3.25, 1.97)	1.26	90.39 (4.77, 4.84)	1.19
			.4	95.44 (2.56, 2.00)	1.20	92.92 (3.96, 3.12)	1.10
			.7	94.64 (2.21, 3.15)	1.03	92.61 (2.70, 4.69)	0.90
		.8	.1	95.14 (2.90, 1.96)	1.08	89.56 (5.71, 4.73)	1.01
			.4	95.07 (3.01, 1.92)	1.00	90.16 (6.91, 2.93)	0.90
			.7	95.23 (2.72, 2.05)	0.79	95.41 (3.27, 1.32)	0.65
	.3	.5	.1	94.70 (3.21, 2.09)	1.25	90.56 (4.47, 4.97)	1.18
			.4	95.09 (2.85, 2.06)	1.19	92.93 (3.84, 3.23)	1.09
			.7	95.18 (1.90, 2.92)	1.02	93.14 (2.53, 4.33)	0.90
		.8	.1	94.98 (3.07, 1.95)	1.07	88.77 (6.42, 4.81)	1.00
			.4	94.94 (3.18, 1.88)	0.99	90.03 (7.01, 2.96)	0.89
			.7	95.09 (2.95, 1.96)	0.77	95.33 (3.49, 1.18)	0.63
50	.0	.5	.1	95.04 (2.67, 2.29)	0.69	93.80 (2.92, 3.28)	0.68
			.4	94.64 (2.91, 2.45)	0.63	94.00 (3.18, 2.82)	0.62
			.7	95.31 (2.29, 2.40)	0.52	94.89 (2.20, 2.91)	0.50
		.8	.1	95.23 (2.77, 2.00)	0.58	93.55 (3.61, 2.84)	0.57
			.4	95.05 (2.80, 2.15)	0.52	93.50 (4.38, 2.12)	0.50
			.7	95.02 (2.96, 2.02)	0.37	95.16 (3.39, 1.45)	0.35
	.3	.5	.1	94.63 (2.97, 2.40)	0.68	93.66 (3.15, 3.19)	0.67
			.4	95.34 (2.58, 2.08)	0.63	94.78 (2.85, 2.37)	0.61
			.7	94.85 (2.25, 2.90)	0.51	94.39 (2.17, 3.44)	0.49
		.8	.1	95.21 (2.62, 2.17)	0.58	93.62 (3.43, 2.95)	0.57
			.4	95.36 (2.65, 1.99)	0.51	93.86 (4.19, 1.95)	0.50
			.7	94.79 (2.61, 2.60)	0.36	95.33 (2.87, 1.80)	0.34
100	.0	.5	.1	95.34 (2.45, 2.21)	0.49	94.85 (2.46, 2.69)	0.48
			.4	95.19 (2.45, 2.36)	0.44	94.83 (2.64, 2.53)	0.44
			.7	95.07 (2.16, 2.77)	0.36	94.80 (2.05, 3.15)	0.35
		.8	.1	95.01 (2.58, 2.41)	0.41	94.29 (2.94, 2.77)	0.41
			.4	94.89 (2.90, 2.21)	0.36	94.32 (3.64, 2.04)	0.36
			.7	95.18 (2.56, 2.26)	0.25	95.26 (2.94, 1.80)	0.24
	.3	.5	.1	95.33 (2.48, 2.19)	0.48	94.75 (2.49, 2.76)	0.48
			.4	94.94 (2.64, 2.42)	0.44	94.47 (2.91, 2.62)	0.43
			.7	94.86 (2.33, 2.81)	0.36	94.55 (2.26, 3.19)	0.35
		.8	.1	95.14 (2.56, 2.30)	0.41	94.31 (3.00, 2.69)	0.41
			.4	94.80 (2.92, 2.28)	0.36	93.96 (3.91, 2.13)	0.35
			.7	95.37 (2.34, 2.29)	0.25	95.64 (2.67, 1.69)	0.24
200	.0	.5	.1	94.95 (2.52, 2.53)	0.34	94.70 (2.48, 2.82)	0.34
			.4	95.11 (2.27, 2.62)	0.31	94.93 (2.34, 2.73)	0.31
			.7	94.96 (2.29, 2.75)	0.25	94.91 (2.10, 2.99)	0.25
		.8	.1	95.11 (2.62, 2.27)	0.29	94.74 (2.77, 2.49)	0.29
			.4	95.32 (2.58, 2.10)	0.25	95.04 (2.98, 1.98)	0.25
			.7	94.11 (3.16, 2.73)	0.17	94.17 (3.50, 2.33)	0.17
	.3	.5	.1	95.29 (2.61, 2.10)	0.34	95.05 (2.51, 2.44)	0.34
			.4	95.41 (2.48, 2.11)	0.31	95.30 (2.57, 2.13)	0.31
			.7	95.04 (2.54, 2.42)	0.25	94.98 (2.33, 2.69)	0.25
		.8	.1	95.55 (2.39, 2.06)	0.29	95.26 (2.53, 2.21)	0.29
			.4	95.11 (2.54, 2.35)	0.25	94.70 (3.08, 2.22)	0.25
			.7	95.13 (2.52, 2.35)	0.17	95.18 (2.82, 2.00)	0.17

Note. Ideally missing left (ML) and missing right (MR) should be 2.50%. Sample size is *n*.

Table 3
Performance of Procedures for Constructing Two-Sided 95% Confidence Intervals (CIs) for a Difference Between Two Independent R^2*s* ($\rho_1^2 - \rho_2^2$) *with Number of Predictors k = 6 Based on 10,000 Runs* ($\rho_x = .1$)

				Modified asymptotic		Simple asymptotic	
n_1	n_2	ρ_1^2	ρ_2^2	Coverage (ML, MR) %	CI width	Coverage (ML, MR) %	CI width
100	50	.2	.3	96.44 (1.61, 1.95)	0.53	93.19 (5.45, 1.36)	0.50
			.5	95.89 (1.56, 2.55)	0.51	93.46 (4.76, 1.78)	0.46
		.5	.3	95.81 (2.41, 1.78)	0.53	91.40 (7.38, 1.22)	0.49
			.5	95.78 (2.06, 2.16)	0.52	92.52 (6.04, 1.44)	0.45
		.8	.3	95.72 (2.12, 2.16)	0.46	88.47 (10.73, .80)	0.43
			.5	95.50 (2.32, 2.18)	0.44	88.60 (10.66, .74)	0.39
	100	.2	.3	95.95 (1.80, 2.25)	0.42	94.36 (2.72, 2.92)	0.41
			.5	96.03 (1.38, 2.59)	0.41	94.54 (2.11, 3.35)	0.39
		.5	.3	96.07 (2.44, 1.49)	0.43	94.68 (3.31, 2.01)	0.40
			.5	95.47 (2.29, 2.24)	0.41	94.39 (2.75, 2.86)	0.38
		.8	.3	95.46 (2.45, 2.09)	0.34	92.53 (5.96, 1.51)	0.33
			.5	95.73 (2.06, 2.21)	0.33	93.85 (4.85, 1.30)	0.30
	200	.2	.3	95.94 (1.74, 2.32)	0.36	94.29 (1.49, 4.22)	0.36
			.5	95.38 (1.80, 2.82)	0.35	93.31 (1.46, 5.23)	0.35
		.5	.3	95.45 (2.62, 1.93)	0.36	94.28 (1.95, 3.77)	0.34
			.5	95.62 (2.25, 2.13)	0.35	94.24 (1.49, 4.27)	0.33
		.8	.3	95.68 (2.44, 1.88)	0.27	94.77 (3.27, 1.96)	0.25
			.5	95.41 (2.55, 2.04)	0.25	94.92 (2.95, 2.13)	0.24
	500	.2	.3	95.46 (2.08, 2.46)	0.32	92.76 (0.99, 6.25)	0.32
			.5	95.48 (2.12, 2.40)	0.31	92.49 (1.00, 6.51)	0.31
		.5	.3	95.00 (2.65, 2.35)	0.32	92.68 (1.28, 6.04)	0.30
			.5	94.95 (2.78, 2.27)	0.31	92.29 (1.28, 6.43)	0.29
		.8	.3	95.19 (2.80, 2.01)	0.20	94.59 (1.79, 3.62)	0.19
			.5	95.37 (2.42, 2.21)	0.20	94.55 (1.37, 4.08)	0.18
200	50	.2	.3	95.81 (1.73, 2.46)	0.48	90.22 (8.79, 0.99)	0.46
			.5	95.88 (1.55, 2.57)	0.46	91.17 (7.82, 1.01)	0.41
		.5	.3	95.86 (2.32, 1.82)	0.48	89.48 (9.72, 0.80)	0.45
			.5	95.26 (2.31, 2.43)	0.46	90.23 (8.84, .93)	0.41
		.8	.3	95.38 (2.34, 2.28)	0.44	87.97 (11.41, .62)	0.42
			.5	95.17 (2.34, 2.49)	0.42	86.57 (12.79, .64)	0.37
	100	.2	.3	95.72 (2.07, 2.21)	0.37	93.88 (4.54, 1.58)	0.36
			.5	95.28 (2.03, 2.69)	0.35	93.92 (4.11, 1.97)	0.33
		.5	.3	95.52 (2.44, 2.04)	0.37	93.45 (5.03, 1.52)	0.35
			.5	95.76 (2.19, 2.05)	0.35	94.03 (4.58, 1.39)	0.33
		.8	.3	95.11 (2.64, 2.25)	0.32	91.80 (7.13, 1.07)	0.31
			.5	95.25 (2.32, 2.43)	0.30	91.96 (6.79, 1.25)	0.28
	200	.2	.3	95.65 (1.91, 2.44)	0.30	94.90 (2.39, 2.71)	0.29
			.5	95.49 (1.99, 2.52)	0.28	94.80 (2.41, 2.79)	0.28
		.5	.3	95.12 (2.73, 2.15)	0.30	94.37 (3.17, 2.46)	0.29
			.5	95.44 (2.08, 2.48)	0.28	94.90 (2.37, 2.73)	0.27
		.8	.3	95.41 (2.32, 2.27)	0.24	93.84 (4.43, 1.73)	0.23
			.5	95.00 (2.32, 2.68)	0.23	93.91 (4.21, 1.88)	0.22
	500	.2	.3	95.49 (1.93, 2.58)	0.24	94.22 (1.61, 4.17)	0.24
			.5	95.38 (2.17, 2.45)	0.24	94.26 (1.61, 4.13)	0.24
		.5	.3	95.69 (2.33, 1.98)	0.24	94.97 (1.73, 3.30)	0.23
			.5	95.51 (2.33, 2.16)	0.24	94.72 (1.64, 3.64)	0.23
		.8	.3	95.85 (2.24, 1.91)	0.17	95.47 (2.31, 2.22)	0.17
			.5	95.10 (2.60, 2.30)	0.16	94.79 (2.56, 2.65)	0.16

Note. Ideally missing left (ML) and missing right (MR) should be 2.50%. Sample sizes for two comparison groups are n_1 and n_2.

$$L = r_1 - r_2 - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2}$$
$$= .49 - .36 - \sqrt{(.49 - .355)^2 + (.530 - .36)^2}$$
$$= -.087$$

and

$$U = r_1 - r_2 + \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2}$$
$$= .49 - .36 + \sqrt{(.605 - .49)^2 + (.36 - .162)^2}$$
$$= .359.$$

This means that with 95% confidence, the difference between validity correlations for male versus female subjects falls between −.087 to .359. Although the difference did not reach statistical significance at the 5% level (because the confidence interval contains 0), the difference could be as high as .36. Thus, efforts may be called for to improve the validity of using self-administered questionnaires in the male group. A null hypothesis significance testing would have missed such information.

Example 2: Overlapping Correlations

I now consider an example described by Olkin and Finn (1990) in which measurements related to cardiovascular health were collected on a sample of 66 adult Black women. One of the objectives was to determine the predictive value of cardiac measures such as heart rate (pulse) and blood pressure (BP) with fitness as quantified by body mass index (BMI, weight/height^2). The correlations were as follows: BMI and BP, $r_{12} = .396$; BMI and pulse, $r_{13} = .179$; and pulse and BP, $r_{23} = .088$. The question of which of the two cardiac measures has better predictive value may be answered with a confidence interval for $\rho_{12} - \rho_{13}$. Here, BMI is the common variable and thus makes the comparison overlapping.

By the Fisher z transformation, a 95% confidence interval for ρ_{12} is $(l_1, u_1) = (.170, .582)$. A similar approach yields a confidence interval for ρ_{13} of $(l_2, u_2) = (-.066, .404)$. The correlation between r_{12} and r_{13} is given by

$$\widehat{\text{corr}}(r_{12}, r_{13}) = \widehat{\text{cov}}(r_{12}, r_{13}) / \sqrt{\widehat{\text{var}}(r_{12})\widehat{\text{var}}(r_{13})}$$
$$= [(r_{23} - .5r_{12}r_{13})(1 - r_{12}^2 - r_{13}^2 - r_{23}^2) + r_{23}^3] / [(1 - r_{12}^2)(1 - r_{13}^2)] = .0526.$$

Therefore, the confidence limits (L, U) for $\rho_{12} - \rho_{13}$ are given by

$$L = r_{12} - r_{13} - \sqrt{(r_{12} - l_1)^2 + (u_2 - r_{13})^2 - 2\widehat{\text{corr}}(r_{12}, r_{13})(r_{12} - l_1)(u_2 - r_{13})}$$
$$= .396 - .179 - \sqrt{\begin{array}{r}(.396 - .170)^2 + (.404 - .179)^2 - 2(.0526) \\ (.396 - .170)(.404 - .179)\end{array}}$$
$$= -.093$$

and

$$U = r_{12} - r_{13} + \sqrt{\begin{array}{r}(u_1 - r_{12})^2 + (r_{13} - l_2)^2 - 2\widehat{\text{corr}}(r_{12}, r_{13}) \\ (u_1 - r_{12})(r_{13} - l_2)\end{array}}$$
$$= .396 - .179 + \sqrt{\begin{array}{r}(.582 - .396)^2 + [.179 - (-.066)]^2 \\ -2(.0526)(.582 - .396)[.179 - (-.066)]\end{array}}$$
$$= .517.$$

These results indicate that BP may be more predictive, although the difference in correlation did not reach the 5% significance level with $n = 66$.

Example 3: Nonoverlapping Correlations

The same study above (Olkin & Finn, 1990) also collected data on children of those women for the purpose of determining whether the correlation of physiological measures increases with age. Specifically, it may be of interest to estimate the difference between the correlation of BMI and BP for mothers and that for children, that is, $\rho_{12} - \rho_{34}$. The estimated correlations are as follows:

	Mother BP	Child BMI	Child BP
Mother BMI	$r_{12} = .396$	$r_{13} = .208$	$r_{14} = .143$
BP		$r_{23} = .023$	$r_{24} = .423$
Child BMI			$r_{34} = .189$

From Example 2, the confidence interval for ρ_{12} is given by $(l_1, u_1) = (.170, .582)$. Applying Fisher's z transformation to r_{34} yields confidence limits for ρ_{34} as $(-.056, .412)$. The correlation between r_{12} and r_{34} is given by

$$\widehat{\text{corr}}(r_{12}, r_{34}) = \widehat{\text{cov}}(r_{12}, r_{34}) / \sqrt{\widehat{\text{var}}(r_{12})\widehat{\text{var}}(r_{34})}$$
$$= [.5r_{12}r_{34}(r_{13}^2 + r_{14}^2 + r_{23}^2 + r_{24}^2) + r_{13}r_{24} + r_{14}r_{23} - (r_{12}r_{13}r_{14} + r_{12}r_{23}r_{24} + r_{13}r_{23}r_{34} + r_{14}r_{24}r_{34})] / [(1 - r_{12}^2) \times (1 - r_{34}^2)] = .0917.$$

Thus, the confidence interval for $\rho_{12} - \rho_{34}$ is given by

$$L = r_{12} - r_{34} - \sqrt{(r_{12} - l_1)^2 + (u_2 - r_{34})^2 - 2\widehat{\text{corr}}(r_{12}, r_{34})(r_{12} - l_1)(u_2 - r_{34})}$$
$$= .396 - .189$$

$$- \sqrt{(.396 - .170)^2 + (.412 - .189)^2 - 2(.0917)(.396 - .170)(.412 - .189)}$$

$$= -.096$$

and

$$U = r_{12} - r_{34} + \sqrt{(u_1 - r_{12})^2 + (r_{34} - l_2)^2 - 2\widehat{\text{corr}}(r_{12}, r_{34})(u_1 - r_{12})(r_{34} - l_2)},$$

$$= .396 - .189 + \sqrt{(.582 - .396)^2 + [(.189 - (-.056)]^2 - 2(.0917)(.582-.396)[.189-(-.056)]}$$

$$= .500.$$

This result suggests that there is a stronger relationship between BMI and BP in the mothers than in their young children. However, the small sample size leads to a wide confidence interval that includes 0, so the null hypothesis of no difference cannot be rejected.

Example 4: Independent R^2*s*

As an example for independent R^2s, consider a situation in which a battery of four tests has a validity of .52 (R^2 = .522) in a sample of 200 students in School A and a validity of .45 (R^2 = .452) in a sample of 300 students in School B (Alf & Graf, 1999, p. 72). As per the computer codes presented in the Appendix, the 95% confidence intervals are given by (l_1, u_1) = (.1563, .3662) and (l_2, u_2) = (.1175, .2788), respectively. Thus, the confidence limits for the difference are

$$L = R_1^2 - R_2^2 - \sqrt{(R_1^2 - l_1)^2 + (u_2 - R_2^2)^2}$$

$$= .52^2 - .45^2 - \sqrt{(.52^2 - .1563)^2 + (.2788 - .45^2)^2}$$

$$= -.069$$

and

$$U = R_1^2 - R_2^2 + \sqrt{(u_1 - R_1^2)^2 + (R_2^2 - l_2)^2}$$

$$= .52^2 - .45^2 + \sqrt{(.3662 - .52^2)^2 + (.45^2 - .1175)^2}$$

$$= .196.$$

The results suggest that no substantial difference between Schools A and B existed.

Discussion

Confidence interval construction has been advocated for decades as a replacement for significance testing in the reporting of study results (Cohen, 1994; Rozeboom, 1960; Schmidt, 1996; Schmidt & Hunter, 1997; Wilkinson & APA Task Force on Statistical Inference, 1999). The rationale for such action is on the grounds that a p value obtained from the significance testing approach is a mixed product of sample size and a single parameter value, usually 0, regardless of whether that single value is of interest. In contrast, a confidence interval provides a range of plausible parameter values and thus is more informative in reporting research work. As a result, "confidence intervals shift the interpretation from a qualitative judgment about the role of chance as the first (and sometimes only) interpretive goal to a quantitative estimation of the biologic measure of effect" (Rothman, 1986, p. 446).

If so desired, a confidence interval can always answer the same question that a p value answers. Thus, it may be puzzling why the efforts promoting confidence intervals have had only a limited effect on the problem of comparing correlations, as evidenced by the popularity of articles that focus on hypothesis testing (Meng et al., 1992; Raghunathan et al., 1996). I believe that the primary reason for the persistence of significance testing in comparing correlations is the lack of a simple approach for constructing the required confidence intervals.

The previous approach to confidence interval construction for differences between correlations ignores the skewness of the sampling distribution for correlations and thus results in poor performance in terms of overall coverage and tail errors (Algina & Keselman, 1999; Olkin & Finn, 1995). As an alternative, I have presented a general approach to constructing confidence intervals for differences between correlation coefficients. This approach actively takes into account the skewness of the sampling distributions, using results such as Fisher's r to z transformation for correlation and Lee's (1971) approximation for single R^2s. The resulting confidence intervals were shown to perform very well in terms of both overall coverage and tail errors when constructing a confidence interval for (a) a difference between two independent correlations, (b) a difference between two overlapping correlations arising from two correlations sharing a common third variable, (c) a difference between two nonoverlapping correlations arising from two correlations obtained from the same sample but that do not share a third variable, and (d) a difference between two independent R^2s. Because there is not much added complexity in computing, there is little justification for focusing solely on significance testing (Meng et al., 1992; Raghunathan et al., 1996) or for using suboptimal methods (Olkin & Finn, 1995). The method and the worked examples can save practitioners from having to refer to the extensive literature describing suboptimal procedures for each of the four cases discussed above in a piecemeal manner.

One might apply the MA approach, that is, Equations 13 and 14, to obtain a confidence interval for the increase in R^2

($\Delta\rho^2$), a very useful effect size measure that quantifies the gain in R^2 with more predictors adding to a regression model (Cohen et al., 2003). Unfortunately, simulation results (available on request) revealed that this approach failed to provide substantial improvement over the simple asymptotic approach (Algina & Moulder, 2001), unless both the sample size and values of $\Delta\rho^2$ are large. Tentative hypotheses for the suboptimal performance include (a) estimation errors in the correlation between estimates of R^2s from the full and the reduced models, using the multivariate version of the delta methods as previously done (Alf & Graf, 1999; Olkin & Siotani, 1976; Olkin & Finn, 1995), and (b) the MA approach does not inherently take into account the fact that $\Delta\rho^2$ has a constrained parameter space, because by definition $\Delta\rho^2 \geq 0$, so that there is a boundary problem. Future theoretical research is needed to focus on these problems, with a simulation evaluation conducted using an adequate definition for accuracy (Efron, 1987a, 2003).

All the confidence limits presented for single correlations are based on the assumption of a normal distribution, which may not be appropriate in all cases. Thus wherever this assumption becomes unreasonable, alternative confidence limits should be sought. However, the present MA approach would still be applicable as the derivation did not implicitly assume data to be normal. The MA approach derives its validity from that of the validity of the confidence limits for a single correlation.

The MA method presented in this article could contribute to the ongoing statistical reforms, specifically in the aspect of supplanting significance testing with confidence interval construction (Wilkinson & APA Task Force on Statistical Inference, 1999). The method can also be used to avoid the common pitfall of using the overlap of two separate confidence intervals as a criterion for judging the statistical significance of an observed difference (Schenker & Gentleman, 2001). Moreover, the results presented here provide a simple alternative to the rules set for inference by eye (Cumming & Finch, 2005). This is because a confidence interval for a difference is readily available from the confidence limits for single parameters, and the statistical significance is thus known without having to calculate the proportion of overlap of two confidence intervals.

References

Alf, E. F., & Graf, R. G. (1999). Asymptotic confidence limits for the difference between two squared multiple correlations: A simplified approach. *Psychological Methods, 4,* 70–75.

Algina, J. (1999). A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient. *Multivariate Behavioral Research, 34,* 493–504.

Algina, J., & Keselman, H. J. (1999). Comparing squared multiple correlation coefficients. Examination of a confidence interval and a test of significance. *Psychological Methods, 4,* 76–83.

Algina, J., & Moulder, B. C. (2001). Sample sizes for confidence intervals on the increase in the squared multiple correlation coefficient. *Educational and Psychological Measurement, 61,* 633–649.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144–152.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19,* 1141–1164.

Cheung, M. W. L., & Chan, W. (2004). Testing dependent correlation coefficients via structural equation modeling. *Organizational Research Methods, 7,* 206–223.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York: Erlbaum.

Cumming, G., & Finch, S. (2005). Inference by eye—Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals [With discussion]. *Statistical Science, 11,* 189–228.

Donner, A., & Zou, G. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics, 58,* 209–214.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics, 7,* 1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals [With discussion]. *The Canadian Journal of Statistics, 9,* 139–172.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika, 72,* 45–58.

Efron, B. (1987a). Better bootstrap confidence intervals: Rejoinder. *Journal of the American Statistical Association, 82,* 198–200.

Efron, B. (1987b). Better bootstrap confidence intervals [With discussion]. *Journal of the American Statistical Association, 82,* 171–185.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin, 104,* 293–296.

Efron, B. (2003). Second thoughts on the bootstrap. *Statistical Science, 18,* 135–140.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Boca Raton, FL: Chapman & Hall/CRC Press.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London, Series A, 121,* 654–673.

Gajjar, A. V. (1967). Limiting distributions of certain transformations of multiple correlation coefficient. *Metron, 26,* 189–193.

Helland, I. S. (1987). On the interpretation and use of R^2 in regression analysis. *Biometrics, 43,* 61–69.

Kirk, R. E. (2007). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference, 137,* 1634–1646.

Kramer, K. H. (1963). Tables for constructing confidence limits on the multiple correlation coefficient. *Journal of the American Statistical Association, 58,* 1082–1085.

Lee, Y. S. (1971). Some results on the sampling distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society, Series B, 33,* 117–130.

Lee, Y. S. (1972). Tables of upper percentage points of the multiple correlation coefficient. *Biometrika, 59,* 175–189.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5,* 434–458.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175.

Morris, M. C., Tangney, C. C., Bienias, J. L., Evans, D. A., & Wilson, R. S. (2003). Validity and reproducibility of a food frequency questionnaire by cognition in an older biracial sample. *American Journal of Epidemiology, 158,* 1213–1217.

Olkin, I., & Finn, J. D. (1990). Testing correlated correlations. *Psychological Bulletin, 108,* 330–333.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin, 118,* 155–164.

Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ikeda (Ed.), *Essays in probability and statistics* (pp. 235–251). Tokyo, Japan: Shinko Tsusho.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307–315.

Pearson, K., & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London, Series A, 191,* 229–311.

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods, 1,* 178–183.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley & Sons.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45,* 283–288.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician, 42,* 59–66.

Rothman, K. J. (1986). Significance questing [Editorial]. *Annals of Internal Medicine, 105,* 445–447.

Rovine, M. J., & von Eye, A. (1997). A 14th way to look at a correlation coefficient: Correlation as the proportion of matches. *American Statistician, 51,* 42–46.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association, 80,* 360–361.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician, 55,* 182–186.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah NJ: Erlbaum.

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap.* New York: Springer-Verlag.

Silver, N. C., Hittner, J. B., & May, K. (2006). A FORTRAN 77 program for comparing dependent correlations. *Applied Psychological Measurement, 30,* 152–153.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251.

Steiger, J. H., & Fouladi, R. T. (1992). R^2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers, 4,* 581–582.

Steiger, J. H., & Ward, L. M. (1987). Factor analysis and the coefficient of determination. *Psychological Bulletin, 101,* 471–474.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Young, G. A. (1994). Bootstrap: More than a stab in the dark? [With discussion]. *Statistical Science, 9,* 382–415.

Appendix

Sampling Distribution of R^2 Approximated by Noncentral F Distribution

The sampling distribution for R^2 (Fisher, 1928) is computationally challenging even by present standards. However, it has long been recognized that $\tilde{R}^2 = R^2/(1 - R^2)$ is distributed as

$$\frac{(\tilde{\rho}\chi_{n-1} + z)^2 + \chi^2_{k-1}}{\chi^2_{n-k-1}},$$

with $\tilde{\rho}^2 = \rho^2/(1 - \rho^2)$, n denoting sample size, k number of predictors, and z the standard normal variate; χ_f and χ^2_f are chi and chi-square variates on f degrees of freedom. All variates are independent from each other. Note that Fisher (1928) also proved that the asymptotic distribution (as $n \to \infty$) for R^2 is a noncentral chi-square distribution, which describes the sampling distribution of the sum of squared normal distributions each with a nonzero mean. A sum of squared standard normal variates makes up a central chi-square, commonly referred to as a chi-square distribution. Thus, it is apparent that the simple asymptotic procedure should not be used for inference for R^2, let alone for the differences between R^2s.

By matching the first three cumulates of the numerator, Lee (1971) approximated the numerator using a scaled noncentral chi-square distribution $g\chi'^2_v(\lambda)$, where

$$g = [\phi_2 - \sqrt{\phi_2^2 - \phi_1\phi_3}]/\phi_1,$$

$$v = [\phi_2 - 2_{\tilde{\rho}^2}\gamma\sqrt{(n-1)(n-k-1)}]/g^2,$$

and

$$\lambda = \tilde{\rho}^2\gamma\sqrt{(n-1)(n-k-1)/g^2},$$

with

$$\gamma = 1/(1 - \rho^2)$$

and

$$\phi_j = (n-1)(\gamma^{2j} - 1) + k,\ j = 1, 2, 3.$$

By definition, then, $\tilde{R}^2$ may be approximated by a scaled noncentral F distribution. Specifically,

$$\tilde{R}^2 \sim \frac{vg}{n-k-1} F'(v, g; \lambda).$$

Therefore, the confidence limits of ρ^2 may be obtained by iteratively searching for the values that satisfy the equations involving the cumulative noncentral F distribution. For this purpose, we can adopt the bisection method, which works by repeatedly dividing an interval in half and then selecting the subinterval in which the root exists. The method is implemented using SAS (Statistical Analysis Systems) and SPSS (Statistical Package for the Social Sciences). (See the online supplemental materials for more information.)

Received July 6, 2006
Revision received August 20, 2007
Accepted August 22, 2007 ■

STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1693–1702
Published online 24 October 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3095

Construction of confidence limits about effect measures: A general approach

G. Y. Zou[1, 2, *, †] and A. Donner[1, 2]

[1]*Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5C1*
[2]*Robarts Clinical Trials, Robarts Research Institute, London, Ont., Canada N6A 5K8*

SUMMARY

It is widely accepted that confidence interval construction has important advantages over significance testing for the presentation of research results, as now facilitated by readily available software. However, for a number of effect measures, procedures are either not available or not satisfactory in samples of small to moderate size. In this paper, we describe a general approach for estimating a difference between effect measures, which can also be used to obtain confidence limits for a risk ratio and a lognormal mean. Numerical evaluation shows that this closed-form procedure outperforms existing methods, including the bootstrap. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: bootstrap; confidence interval; lognormal; risk ratio; generalized confidence interval

1. INTRODUCTION

Confidence intervals are usually regarded as more informative than significance tests because they provide a range of parameter values that reflect the degree of uncertainty in the estimation procedure. Moreover, given the correspondence between these two approaches, confidence interval estimation encompasses hypothesis testing [1, Chapter 9], and their use in presenting research results is formally recommended in several published guidelines, e.g. the CONSORT statement [2]. Simple procedures and widely available software have also made interval estimation readily accessible to practitioners [3–8].

The purpose of this paper is to first describe a general approach for constructing a confidence interval about a difference between effect measures, which has been previously applied only in

*Correspondence to: G. Y. Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5C1.
†E-mail: gzou@robarts.ca

Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada

Received 6 September 2006
Accepted 4 September 2007

special cases. This approach, which requires only the availability of confidence limits about the two effect measures separately, is then further applied to construct a confidence interval about a risk ratio and a lognormal mean. Its performance is compared with that of more computationally intensive procedures using numerical evaluation studies.

2. CONFIDENCE INTERVAL CONSTRUCTION FOR DIFFERENCES

Let θ_i be the respective parameter of interest for population i, $i=1,2$, with point estimate $\widehat{\theta}_i$. Assuming that $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independently distributed, an approximate two-sided $(1-\alpha)100$ per cent confidence interval (L,U) for $\theta_1-\theta_2$ is traditionally given by

$$(L,U)=\widehat{\theta}_1-\widehat{\theta}_2\mp z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1)+\widehat{\mathrm{var}}(\widehat{\theta}_2)}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, and $\widehat{\mathrm{var}}(\widehat{\theta}_i)$, $i=1,2$, are the respective variance estimators. Unfortunately, this procedure performs well only if sample sizes are large or the sampling distributions of $\widehat{\theta}_i$ are close to normal. The Wald interval for a difference between two proportions is a good example [9].

One reason for the poor performance of the traditional method is that it does not reflect the asymmetry of the underlying sampling distributions [10, p. 190]. In this paper we attempt to improve its performance by obtaining $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ at the neighborhood of the confidence limits L and U separately. We also note that estimating $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ at L and U using an iterative procedure would be equivalent to inverting a test statistic to obtain a confidence interval [11]. Using this approach [1, Section 9.2], we may regard the $(1-\alpha)100$ per cent confidence limits L,U as the minimum and maximum values of $\theta_1-\theta_2$ that satisfy

$$\frac{[(\widehat{\theta}_1-\widehat{\theta}_2)-(\theta_1-\theta_2)]^2}{\mathrm{var}(\widehat{\theta}_1)+\mathrm{var}(\widehat{\theta}_2)}<z_{\alpha/2}^2$$

Let (l_1,u_1) and (l_2,u_2) be the two-sided $(1-\alpha)100$ per cent confidence intervals for θ_1 and θ_2, respectively. Among the plausible parameter values provided by these two sets of limits, l_1-u_2 is near L and u_1-l_2 is near U. Thus, for obtaining L we estimate $\mathrm{var}(\widehat{\theta}_1)$ under $\theta_1=l_1$ and $\mathrm{var}(\widehat{\theta}_2)$ at $\theta_2=u_2$. Similarly, for obtaining U we estimate $\mathrm{var}(\widehat{\theta}_1)$ under $\theta_1=u_1$ and $\mathrm{var}(\widehat{\theta}_2)$ under $\theta_2=l_2$. By again applying the inversion principle, we have

$$\widehat{\mathrm{var}}(\widehat{\theta}_1)=\frac{(\widehat{\theta}_1-l_1)^2}{z_{\alpha/2}^2}$$

under $\theta_1=l_1$, and

$$\widehat{\mathrm{var}}(\widehat{\theta}_1)=\frac{(u_1-\widehat{\theta}_1)^2}{z_{\alpha/2}^2}$$

under $\theta_1=u_1$. Similarly, we have

$$\widehat{\mathrm{var}}(\widehat{\theta}_2)=\frac{(\widehat{\theta}_2-l_2)^2}{z_{\alpha/2}^2}$$

DOI: 10.1002/sim

under $\theta_2 = l_2$, and

$$\widehat{\text{var}}(\widehat{\theta}_2) = \frac{(u_2 - \widehat{\theta}_2)^2}{z_{\alpha/2}^2}$$

under $\theta_2 = u_2$.

Substituting corresponding variance estimators in the expressions for L and U, respectively, we have

$$L = \widehat{\theta}_1 - \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\frac{(\widehat{\theta}_1 - l_1)^2}{z_{\alpha/2}^2} + \frac{(u_2 - \widehat{\theta}_2)^2}{z_{\alpha/2}^2}}$$

$$= \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2} \tag{1}$$

and

$$U = \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2} \tag{2}$$

The advantage of this procedure is that it does not require any specific underlying distributions for $\widehat{\theta}_i$, but only separate confidence limits that have coverage levels close to nominal. It is also trivial to show that it provides the traditional confidence interval if the sampling distributions for $\widehat{\theta}_i$ $(i = 1, 2)$ are symmetric, but is more general in that the symmetry assumption is not required.

We also recognize that this procedure has previously been applied in special cases, including, for example, the construction of a confidence interval for variance components [12], the construction of limits for a difference between two normal means [13] and for a difference between two kappa statistics [14]. It has further been applied to the problem of assessing bioequivalence [15–18] and is closely related to methodology proposed for the analysis of binary data in a variety of contexts [9, 19–23].

It is trivial to show that the proposed procedure satisfies the invariance property in the sense that the confidence interval for $\theta_2 - \theta_1$ is always given by $[-U, -L]$, in contrast to a recent claim [24] which appears to confuse the properties of invariance and symmetry.

3. APPLICATIONS

3.1. Confidence interval about the risk ratio

We now use the approach described above to obtain a confidence interval for the risk ratio by recognizing that a difference on the log scale is equivalent to the log of a ratio. Substitution of $\ln p_i$ $(i = 1, 2)$ for $\widehat{\theta}_i$ and the corresponding confidence limits lp_{li}, lp_{ui} for l_i, u_i in (1) and (2)

DOI: 10.1002/sim

yields a $(1-\alpha)100$ per cent confidence interval for log risk ratio as

$$L = \ln(p_1) - \ln(p_2) - \sqrt{[\ln(p_1) - lp_{l1})]^2 + [lp_{u2} - \ln(p_2)]^2}$$

$$U = \ln(p_1) - \ln(p_2) + \sqrt{[\ln p_{u1} - \ln(p_1)]^2 + [\ln(p_2) - lp_{l2}]^2}$$

Note that this result provides a rebuttal to the assertion that 'it is not possible to obtain a confidence interval for a relative risk by using the confidence limits for the two components absolute risks' [4, p. 789].

There are two methods that can be used for obtaining the confidence limits (lp_{li}, lp_{ui}). The first is the delta method which results in the procedure found in most standard textbooks, e.g. [5, p. 58]. A second approach is to obtain confidence limits for a single proportion using the Wilson method [25, 26] and then use the transformation principle [3, 4] to obtain limits on the log scale. Such limits satisfy $2\ln(p_i) - \ln(1+z^2_{\alpha/2}/n) = lp_{li} + lp_{ui}$, where n is the sample size [25], whereas limits obtained using the delta method satisfy $2\ln(p_i) = lp_{li} + lp_{ui}$. Thus, it is clear that the Wilson approach will provide a narrower interval and, thus, is theoretically preferable.

As an illustration, consider a study reported by Brenner *et al.* [27] where the prevalence of *Helcobacter pylori* infection in preschool children having mothers with a history of duodenal or gastric ulcer is $\frac{6}{22}$. Using the Wilson method a 95 per cent confidence interval about the prevalence is given by (0.132, 0.482), while for children with no parental history of ulcer, the corresponding prevalence is $\frac{112}{842}$ (95 per cent CI 0.112, 0.158). Thus, the 95 per cent confidence interval for the relative risk is given by

$$L = \exp\{\ln\tfrac{6}{22} - \ln\tfrac{112}{841} - \sqrt{(\ln\tfrac{6}{22} - \ln 0.132)^2 + (\ln 0.158 - \ln\tfrac{112}{841})^2}\} = 0.97$$

$$U = \exp\{\ln\tfrac{6}{22} - \ln\tfrac{112}{841} + \sqrt{(\ln 0.482 - \ln\tfrac{6}{22})^2 + (\ln\tfrac{112}{841} - \ln 0.112)^2}\} = 3.71$$

This result is consistent with that obtained from the standard Pearson chi-square test ($P = 0.06$), which, by the duality principle [1, p. 421], can be regarded as a desirable feature. On the other hand, the traditional textbook formula [5, p. 59] provides a 95 per cent confidence interval for the relative risk given by 1.01 to 4.04, inconsistent with the hypothesis testing result [27]. One may alternatively compute an 'exact' confidence interval, although the word exact in the present context does not imply accurate. For example, the exact procedure for a single proportion has been criticized because the method is too conservative [25, 26].

To evaluate the performance of the proposed procedure when using both the delta method and the Wilson method for interval estimation of a single proportion on the log scale, we conducted a numerical evaluation by computing all possible $(n_1+1)(n_2+1)$ outcomes, where n_1 and n_2 are the sizes of two independent samples. The coverage probability for a given interval estimate (L, U) is then easily shown to be given by

$$\sum_{x_1=0}^{n_1}\sum_{x_2=0}^{n_2} 1(\pi_1/\pi_2 \in [L, U]) \prod_{i=1}^{2} \binom{n_i}{x_i} \pi_i^{x_i}(1-\pi_i)^{n_i - x_i}$$

DOI: 10.1002/sim

Table I. Comparative performance of confidence interval procedures for a risk ratio (summary of 2000 parameter combinations).

Method	Mean	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
Delta						
Coverage (per cent)	96.19	95.17	95.49	96.04	96.71	97.66
Left tail * (per cent)	2.33	0.01	1.17	2.42	3.34	4.16
Right tail * (per cent)	1.49	0.00	0.00	1.20	2.64	3.53
Width	21.22	0.60	1.04	4.46	30.10	66.36
Wilson						
Coverage (per cent)	95.41	94.69	94.88	95.15	96.00	96.41
Left tail (per cent)	2.75	1.27	2.23	2.75	3.41	3.98
Right tail (per cent)	1.85	0.00	0.49	2.24	2.81	3.23
Width	15.48	0.58	0.97	4.06	21.78	47.06

*Left tail: the interval lies completely below the parameter; right tail: the interval lies completely above the parameter.

where $1(\pi_1/\pi_2 \in [L, U])$ is 1 if $[L, U]$ contains π_1/π_2, 0 otherwise. To deal with extreme cases, π_i was set to $1/(2n_i)$ if $x_i = 0$, and $1 - 1/(2n_i)$ if $x_i = n_i$, for $i = 1, 2$. Interval width and tail errors were evaluated in a similar manner.

When we set $(n_1, \pi_1) = (22, \frac{6}{22})$ and $(n_2, \pi_2) = (841, \frac{112}{841})$, as in Brenner *et al.* [27], the coverage (left tail error, right tail error) and width for the delta method are 95.53 (0.05, 4.42) and 3.10, while the corresponding results using the Wilson method are given by 95.13 (1.91, 2.95) and 2.71. Thus, the method presented here provides a narrower interval than the traditional method [5, p. 58], which in our evaluation misses the parameter value for 4.42 per cent of time instead of the advertised 2.5 per cent.

We further evaluated the two procedures with n_1 and n_2 both ranging from 10 to 100 at intervals of 10, π_1 ranging from 0.1 to 0.5 at intervals of 0.1, and π_2 varying from 0.05, 0.1, 0.3 and 0.5. The results in Table I based on these 2000 $(10 \times 10 \times 5 \times 4)$ parameter combinations demonstrate clearly the advantage of using Wilson confidence limits for single proportions in constructing a confidence interval for the risk ratio. Thus, it exists in closed form, and has a shorter width, while maintaining coverage close to nominal with equiprobable tail errors. This procedure for risk ratio is also consistent with the recommendation of the Wilson method for a single proportion [5, 25, 26].

3.2. Confidence interval for a lognormal mean

Inferences obtained from positively skewed data are often performed on the log scale under the assumption of an underlying lognormal distribution. This will result in inferences for the geometric means (or the median). However, there are situations where the mean on the original scale is of most interest, as, for example, in the analysis of health costs [28, 29]. Bootstrap methods have been recommended for this purpose [30], although with some limitations that have been recently recognized [31].

Let y_i, $i = 1, 2, \ldots, n$, be observations from a lognormal distribution, implying that $x_i = \ln y_i$ is distributed normally with mean and standard deviation μ and σ, respectively. Then the mean of the lognormal distribution is given by

$$M = \exp[\mu + \sigma^2/2]$$

DOI: 10.1002/sim

This relationship has prompted a remark that 'obtaining the confidence interval for the lognormal estimator is a non-trivial problem since it is a function of two transformed sample estimates [32, p. 422]. However, if we regard μ as θ_1 and $-\sigma^2/2$ as θ_2 in the above procedure, constructing a confidence interval for M becomes straightforward. Since $\bar{x}$ and $-s^2/2$ are independent we can treat $\bar{x}$ and $-s^2/2$ as $\widehat{\theta}_1$ and $\widehat{\theta}_2$ in (1) and (2), respectively. Furthermore, the confidence limits for μ can be obtained from normal distribution theory and those for $-\sigma^2/2$ may be obtained using the chi-square distribution. Substituting these two pairs of limits in (1) and (2) yields a confidence interval for a lognormal mean as

$$L = \exp\left[\bar{x} + \frac{s^2}{2} - \sqrt{z^2_{\alpha/2}\frac{s^2}{n} + \left\{\frac{s^2}{2}\left(1 - \frac{n-1}{\chi^2_{1-\alpha/2,n-1}}\right)\right\}^2}\right]$$

$$U = \exp\left[\bar{x} + \frac{s^2}{2} + \sqrt{z^2_{\alpha/2}\frac{s^2}{n} + \left\{\frac{s^2}{2}\left(\frac{n-1}{\chi^2_{\alpha/2,n-1}} - 1\right)\right\}^2}\right] \tag{3}$$

where $\chi^2_{\alpha/2,\mathrm{df}}$ and $\chi^2_{1-\alpha/2,\mathrm{df}}$ are the $\alpha/2$th and $1-\alpha/2$th percentiles of the chi-square distribution with df degrees of freedom.

As an illustration, consider a study enrolling 26 asthma patients treated with a pressurized metered dose inhaler [31]. Exploratory analysis implied that the observations are lognormally distributed. Analysis of the log-transformed cost data yields $\bar{x} = 5.877$ and $s^2 = 2.158$, and thus by (3) a 95 per cent confidence interval is obtained as (520, 3243), very comparable with the Bayesian parametric interval [31] obtained as (510, 3150).

Since the numerical evaluation approach used in Section 3.1 is not applicable here, we used Monte Carlo simulation in this case to evaluate the performance of the proposed procedure to that of more computationally intensive procedures. The methods evaluated were the traditional t, the jackknife, a method commonly referred to as a generalized confidence interval based on the simulation of pivotal statistics [33–37], and six bootstrap methods [38, Chapter 4], including the Normal, percentile, hybrid, bootstrap-t, bias corrected (BC), and the bias corrected and accelerated (BCa). Each method was used to construct a 95 per cent two-sided confidence interval. We considered sample sizes $n = 20$, 50, 100, 200 and 500, and $\sigma^2 = 0.5$, 1 and 4 with $\mu = -\sigma^2/2$, resulting in a lognormal mean of 1. The simulation was performed using 1000 runs for each of the 15 parameter combinations. For the eight computational intensive methods (six bootstraps, jackknife and generalized interval), we performed 1000 resamples at each run.

Table II presents the observed coverage (per cent), and left and right tail errors (per cent) (defined as missing the parameter from the left or right, respectively), as well as the average interval width. By considering the standards proposed by Burton *et al.* [39], it is seen that only the generalized interval approach and the proposed method (equation (3)) may be regarded as acceptable in terms of coverage. The latter also delivers narrower intervals, with separate tail errors that are closer to 2.5 per cent, advantages that are most obvious at $\sigma^2 = 4$.

DOI: 10.1002/sim

Table II. Comparative performance of 10 procedures for constructing a 95 per cent two-sided confidence interval for a lognormal mean with $\mu=-\sigma^2/2$ based on 1000 runs (computational methods used 1000 resamples for each run).*

		$\sigma^2=0.5$		$\sigma^2=1.0$		$\sigma^2=4.0$	
n	Method	Cover (L, R) per cent†	W	Cover (L, R) per cent	W	Cover (L, R) per cent	W
20	1	88.1 (10.9, 1.0)	0.63	84.3 (15.2, 0.5)	0.93	59.7 (40.3, 0.0)	2.05
	2	88.9 (9.1, 2.0)	0.63	86.5 (12.4, 1.1)	0.95	64.9 (35.0, 0.1)	2.10
	3	89.5 (7.8, 2.7)	0.67	87.3 (10.7, 2.0)	1.05	69.3 (30.2, 0.5)	2.64
	4	87.0 (12.3, 0.7)	0.62	82.7 (17.0, 0.3)	0.91	54.9 (45.1, 0.0)	1.87
	5	88.6 (9.8, 1.6)	0.62	85.2 (13.9, 0.9)	0.91	62.4 (37.5, 0.1)	1.87
	6	93.9 (4.7, 1.4)	0.85	92.1 (7.0, 0.9)	1.60	81.1 (18.8, 0.1)	44.81
	7	89.2 (10.0, 0.8)	0.65	85.0 (14.6, 0.4)	0.96	60.3 (39.7, 0.0)	2.12
	8	90.4 (9.0, 0.6)	0.69	86.6 (13.1, 0.3)	1.03	61.8 (38.2, 0.0)	2.26
	9	95.8 (1.4, 2.8)	0.86	95.2 (1.9, 2.9)	1.64	96.3 (1.5, 2.2)	55.53
	10	95.5 (2.4, 2.1)	0.82	96.1 (2.4, 1.5)	1.56	96.6 (1.9, 1.5)	36.77
50	1	91.4 (7.7, 0.9)	0.42	88.3 (11.1, 0.6)	0.65	67.7 (32.3, 0.0)	1.73
	2	91.1 (6.4, 2.5)	0.42	89.6 (8.6, 1.8)	0.66	72.5 (26.7, 0.8)	1.79
	3	91.4 (5.3, 3.3)	0.44	89.7 (7.2, 3.1)	0.71	76.7 (21.6, 1.7)	2.26
	4	90.4 (8.7, 0.9)	0.42	86.7 (12.9, 0.4)	0.64	63.5 (36.5, 0.0)	1.60
	5	91.6 (6.9, 1.5)	0.42	89.5 (9.5, 1.0)	0.64	70.0 (29.5, 0.5)	1.60
	6	92.9 (4.8, 2.3)	0.48	91.7 (6.3, 2.0)	0.85	83.3 (16.0, 0.7)	10.59
	7	92.1 (7.2, 0.7)	0.43	88.6 (10.9, 0.5)	0.66	68.0 (32.0, 0.0)	1.76
	8	92.3 (7.0, 0.7)	0.44	89.4 (10.1, 0.5)	0.67	68.8 (31.2, 0.0)	1.80
	9	93.9 (3.0, 3.1)	0.45	93.7 (3.5, 2.8)	0.77	94.8 (2.7, 2.5)	4.22
	10	94.4 (3.2, 2.4)	0.46	94.0 (3.6, 2.4)	0.77	94.1 (3.4, 2.5)	3.82
100	1	92.5 (6.4, 1.1)	0.31	89.4 (9.9, 0.7)	0.48	74.0 (26.0, 0.0)	1.44
	2	92.6 (5.4, 2.0)	0.31	90.2 (8.1, 1.7)	0.48	78.0 (21.7, 0.3)	1.48
	3	92.6 (4.6, 2.8)	0.31	91.0 (6.2, 2.8)	0.51	80.7 (17.9, 1.4)	1.80
	4	91.2 (7.8, 1.0)	0.30	88.1 (11.4, 0.5)	0.48	70.5 (29.5, 0.0)	1.35
	5	92.3 (6.0, 1.7)	0.30	89.7 (9.1, 1.2)	0.48	75.7 (24.3, 0.0)	1.35
	6	93.4 (4.1, 2.5)	0.33	92.7 (5.0, 2.3)	0.56	85.5 (14.2, 0.3)	5.40
	7	92.6 (6.2, 1.2)	0.31	89.5 (9.8, 0.7)	0.48	74.3 (25.7, 0.0)	1.44
	8	92.7 (6.1, 1.2)	0.31	89.6 (9.7, 0.7)	0.49	74.6 (25.4, 0.0)	1.46
	9	94.2 (2.4, 3.4)	0.31	94.0 (2.9, 3.1)	0.50	93.6 (2.5, 3.9)	1.90
	10	94.8 (2.9, 2.3)	0.32	94.4 (3.2, 2.4)	0.51	94.7 (3.3, 2.0)	1.84
200	1	93.7 (4.8, 1.5)	0.22	92.0 (6.6, 1.4)	0.35	76.8 (23.2, 0.0)	1.11
	2	92.3 (4.4, 3.3)	0.22	91.7 (5.7, 2.6)	0.35	81.4 (18.0, 0.6)	1.15
	3	92.1 (3.8, 4.1)	0.22	91.9 (4.5, 3.6)	0.36	85.1 (13.4, 1.5)	1.36
	4	92.7 (5.8, 1.5)	0.22	92.3 (7.2, 0.5)	0.34	72.7 (27.3, 0.0)	1.07
	5	93.0 (4.5, 2.5)	0.22	92.0 (6.2, 1.8)	0.34	79.4 (20.4, 0.2)	1.07
	6	92.5 (4.0, 3.5)	0.23	92.4 (4.7, 2.9)	0.38	87.8 (11.6, 0.6)	2.81
	7	93.5 (4.9, 1.6)	0.22	92.2 (6.7, 1.1)	0.35	76.8 (23.2, 0.0)	1.12
	8	93.8 (4.8, 1.4)	0.22	92.5 (6.6, 0.9)	0.35	77.0 (23.0, 0.0)	1.12
	9	91.8 (3.5, 4.7)	0.21	92.3 (3.3, 4.4)	0.34	94.2 (2.3, 3.5)	1.14
	10	93.4 (3.3, 3.3)	0.22	93.7 (3.1, 3.2)	0.35	94.5 (3.0, 2.5)	1.11
500	1	93.5 (4.0, 2.5)	0.14	92.5 (5.6, 1.9)	0.22	79.1 (20.9, 0.0)	0.77
	2	93.7 (3.5, 2.8)	0.14	92.2 (5.0, 2.8)	0.22	83.0 (16.3, 0.7)	0.79
	3	93.1 (3.4, 3.5)	0.14	92.3 (4.0, 3.7)	0.23	85.2 (13.0, 1.8)	0.91
	4	93.2 (4.7, 2.1)	0.14	92.2 (6.2, 1.6)	0.22	76.5 (23.5, 0.0)	0.75

DOI: 10.1002/sim

Table II. *Continued*.

		$\sigma^2=0.5$		$\sigma^2=1.0$		$\sigma^2=4.0$	
n	Method	Cover (L, R) per cent†	W	Cover (L, R) per cent	W	Cover (L, R) per cent	W
	5	93.6 (3.7, 2.7)	0.14	92.9 (5.0, 2.1)	0.22	81.2 (18.5, 0.3)	0.75
	6	93.7 (3.1, 3.2)	0.14	93.5 (3.6, 2.9)	0.23	87.7 (11.5, 0.8)	1.37
	7	93.6 (3.9, 2.5)	0.14	92.7 (5.7, 1.6)	0.22	79.1 (20.9, 0.0)	0.77
	8	93.6 (3.9, 2.5)	0.14	92.7 (5.7, 1.6)	0.22	79.2 (20.8, 0.0)	0.77
	9	91.9 (3.5, 4.6)	0.13	91.6 (4.0, 4.4)	0.20	92.6 (3.4, 4.0)	0.61
	10	94.0 (2.5, 3.5)	0.14	94.0 (2.7, 3.3)	0.22	94.7 (2.8, 2.5)	0.64

*Methods 1–6 are bootstrap methods: normal, BC, BCa, hybrid, percentile and t [38]; Methods 7–10 are Jackknife, traditional t, generalized interval and the proposed, respectively.

†L: The interval lies completely below the parameter; R: the interval lies completely above the parameter.

4. DISCUSSION

We have presented a general approach to confidence interval construction that should prove useful in a wide variety of settings. This has been illustrated by the consideration of two interval estimation problems for which procedures are either currently unavailable or unduly complicated.

The basis of this approach relies on the observation that one can easily obtain a confidence interval for a difference between two effect measures given the availability of a confidence interval method for each effect measure separately. Thus, it provides a relatively simple method of avoiding using the overlap of two separate confidence intervals as a criterion for judging the statistical significance of an observed difference, a procedure which, aside form retaining a hypothesis-testing perspective, is potentially misleading.

The approach presented here may also be extended to the case of two correlated effect measures as obtained, for example, in paired comparisons. Let r be the estimator of the correlation coefficient between $\widehat{\theta}_1$ and $\widehat{\theta}_2$. Then the procedure described may be extended by including covariance terms $r(\widehat{\theta}_1-l_1)(u_2-\widehat{\theta}_2)$ and $r(u_1-\widehat{\theta}_1)(\widehat{\theta}_2-l_2)$ in the expressions given for L and U, respectively. Examples are given in several references [20–23], where the Phi coefficient [40, p. 99] was used to estimate the required correlation parameter.

ACKNOWLEDGEMENTS

The authors' work was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. Casella G, Berger RL. *Statistical Inference* (2nd edn). Duxbury: Pacific Grove, CA, 2002.
2. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Annals of Internal Medicine* 2001; **134**:663–694.
3. Steiger JH, Fouladi RT. Noncentrality interval estimation and the evaluation of statistical models. *What if There Were No Significance Tests*, Harlow LL, Mulaik SA, Steiger JH (eds). Lawrence Erlbaum Association: New Jersey, 1997; 221–257.

DOI: 10.1002/sim

4. Daly LE. Confidence interval made easy: interval estimation using a substitution method. *American Journal of Epidemiology* 1998; **147**:783–790.
5. Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with Confidence* (2nd edn). BMJ Books: Bristol, 2000.
6. Hahn GJ, Meeker WQ. *Statistical Intervals*: *A Guide for Practitioners*. Wiley: New York, 1991.
7. Burdick RK, Graybill FA. *Confidence Intervals on Variance Components*. M. Dekker: New York, 1992.
8. Smithson M. *Confidence Intervals*. Sage Publications: London, 2002.
9. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 1998; **17**:873–890.
10. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical Science* 1996; **11**:189–228.
11. Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1985; **4**(2):213–226.
12. Howe WG. Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variables. *Journal of the American Statistical Association* 1974; **69**:789–794.
13. Wang H, Chow SC. A practical approach for comparing means of two groups without equal variance assumption. *Statistics in Medicine* 2002; **21**:3137–3151.
14. Donner A, Zou G. Interval estimation for a difference between intraclass kappa statistics. *Biometrics* 2002; **58**:209–214.
15. Hyslop T, Hsuan F, Holder DJ. A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine* 2000; **19**(20):2885–2897.
16. Chow SC, Shao J, Wang H. Individual bioequivalence testing under 2×3 designs. *Statistics in Medicine* 2002; **21**(5):629–648.
17. Chow SC, Shao J, Wang H. In vitro bioequivalence testing. *Statistics in Medicine* 2003; **22**(1):55–68.
18. Chervoneva I, Hyslop T, Hauck WW. A multivariate test for population bioequivalence. *Statistics in Medicine* 2007; **26**(6):1208–1223.
19. Fagan T. Exact 95% confidence intervals for differences in binomial proportions. *Computers in Biology and Medicine* 1999; **29**:83–87.
20. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 1998; **17**(22):2635–2650.
21. Newcombe RG. Simultaneous comparison of sensitivity and specificity of two tests in the paired design: a straightforward graphical approach. *Statistics in Medicine* 2001; **20**(6):907–915.
22. Newcombe RG. Estimating the difference between differences: measurement of additive scale interaction for proportions. *Statistics in Medicine* 2001; **20**(19):2885–2893.
23. Newcombe RG. Confidence intervals for the mean of a variable taking the values 0, 1 and 2. *Statistics in Medicine* 2003; **22**(17):2737–2750.
24. Zhou XH, Qin GS. A supplement to: 'A new confidence interval for the difference between two binomial proportions of paired data'. *Journal of Statistical Planning and Inference* 2007; **137**(1):357–358.
25. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872.
26. Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportion. *American Statistician* 1998; **52**(2):119–126.
27. Brenner H, Rothenbacher D, Bode G, Adler G. Parental history of gastric or duodenal ulcer and prevalence of *Helicobacter pylori* infection in preschool children: population based study. *British Medical Journal* 1998; **316**:665.
28. Thompson SG, Barber JA. How should cost data in pragmatic randomized trials be analysed? *British Medical Journal* 2000; **320**:1197–1200.
29. Zhou XH. Inferences about population means of health care costs. *Statistical Methods in Medical Research* 2002; **11**(4):327–339.
30. Barber JA, Thompson SG. Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine* 2000; **19**:3129–3236.
31. O'Hagan A, Stevens JW. Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics* 2003; **12**(1):33–49.
32. Briggs A, Nixon R, Dixon S, Thompson S. Parametric modelling of cost data: some simulation evidence. *Health Economics* 2005; **14**(4):421–428.
33. Krishnamoorthy K, Mathew T. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *Journal of Statistical Planning and Inference* 2003; **115**(1):103–121.

DOI: 10.1002/sim

34. Tian LL. Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. *Statistics in Medicine* 2005; **24**(20):3223–3232.
35. Chen YH, Zhou XH. Interval estimates for the ratio and difference of two lognormal means. *Statistics in Medicine* 2006; **25**(23):4099–4113.
36. Krishnamoorthy K, Mathew T, Ramachandran G. Generalized P-values and confidence intervals: a novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene* 2006; **3**:642–650.
37. Hannig J, Iyer H, Patterson P. Fiducial generalized confidence intervals. *Journal of the American Statistical Association* 2006; **101**:254–269.
38. Shao J, Tu D. *The Jackknife and Bootstrap*. Springer: New York, 1995.
39. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; **25**(24):4279–4292.
40. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions* (3rd edn). Wiley: New York, 2003.

DOI: 10.1002/sim

American Journal of Epidemiology Advance Access published June 25, 2008

American Journal of Epidemiology
Vol. 168, No. 2
DOI: 10.1093/aje/kwn104
Advance Access publication May 28, 2008

Practice of Epidemiology

On the Estimation of Additive Interaction by Use of the Four-by-two Table and Beyond

Guang Yong Zou[1,2]

[1] Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada.
[2] Robarts Clinical Trials, Robarts Research Institute, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada.

Received for publication September 14, 2007; accepted for publication March 27, 2008.

A four-by-two table with its four rows representing the presence and absence of gene and environmental factors has been suggested as the fundamental unit in the assessment of gene-environment interaction. For such a table to be more meaningful from a public health perspective, it is important to estimate additive interaction. A confidence interval procedure proposed by Hosmer and Lemeshow has become widespread. This article first reveals that the Hosmer-Lemeshow procedure makes an assumption that confidence intervals for risk ratios are symmetric and then presents an alternative that uses the conventional asymmetric intervals for risk ratios to set confidence limits for measures of additive interaction. For the four-by-two table, the calculation involved requires no statistical programs but only elementary calculations. Simulation results demonstrate that this new approach can perform almost as well as the bootstrap. Corresponding calculations in more complicated situations can be simplified by use of routine output from multiple regression programs. The approach is illustrated with three examples. A Microsoft Excel spreadsheet and SAS codes for the calculations are available from the author and the *Journal*'s website, respectively.

bootstrap; genotype-environment interaction; logistic regression; proportional hazards models; risk ratio

Abbreviations: AP, attributable proportion due to interaction; CI, confidence interval; OR, odds ratio; RERI, relative excess risk due to interaction; RR, risk ratio; SA, simple asymptotic; SI, synergy index.

In 1976, it was recognized that "[a]s more risk factors become established as probable causes in the elaboration of disease etiology, scientists will turn their attention increasingly to the question of interaction (synergy or antagonism) of the causes" (1, p. 506). Scientists can now study literally thousands of genes and their interactions with environmental factors, thanks to the Human Genome Project.

It has been suggested that at the fundamental core of assessing gene-environment interaction is a four-by-two table; note that the original article refers to the table as a two-by-four table (2). However, conducting proper inferences is the ultimate goal of any research (3, p. 2). Furthermore, on the basis of the sufficient component cause model (4), it is more meaningful to assess interaction on the additive scale (1). This is because information concerning an additive interaction between two factors is more relevant to disease prevention and intervention (5, 6; 7, chapters 6 and 10). For example, if the joint effect of two factors surpasses the sum of their single effects, then reduction of either one would also reduce the risk of the other factor in producing the disease.

There has been little discussion concerning appropriate statistical methods for estimating additive interactions. As

Correspondence to Dr. G. Y. Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada N6A 5C1 (e-mail: gzou@robarts.ca).

a result, a simple asymptotic approach proposed by Hosmer and Lemeshow (8) has proliferated in the literature (9–12), despite its well-documented poor performance (13).

The purpose of this article is to present an alternative approach for constructing accurate confidence intervals (CIs) for measures of additive interaction. The desirable performance of this new approach is the result of incorporating the asymmetric confidence limits for risk ratios (or odds ratios), in contrast to the simple asymptotic approach that forces confidence limits for risk ratios (RRs) to be symmetric. The central idea is to recover the variances needed for measures of interactions from confidence limits for RRs. For the four-by-two table, the calculations involved can be done in a spreadsheet or by a hand-held calculator. Simulation results demonstrate that the new approach is accurate enough to replace the bootstrap. The new approach may also be applied to more complicated situations by using output from standard multiple regression programs. Three worked examples are presented. All calculations were done by use of a Microsoft Excel (Microsoft Corporation, Redmond, Washington) spreadsheet that is available from the author upon request. SAS codes (SAS Institute, Inc., Cary, North Carolina) using routine regression output to obtain confidence limits are supplementary material posted on the *Journal*'s website (http://aje.oupjournals.org/).

THE FOUR-BY-TWO TABLE AND ESTIMATION OF MEASURES OF ADDITIVE INTERACTION

Let G and E denote two risk factors, with their presence and absence reflected by 1 and 0, respectively. In the case of gene-environment interaction, three possible biallelic genotypes may be readily handled. For example, one can assume a dominant mode of gene action so that the genotype *AA* and *Aa* are equivalent and coded as 1, and *aa* coded as 0. Thus, a contingency table may be formed with four rows representing gene and environment combinations 11, 10, 01, and 00 and two columns representing disease status (yes and no) as follows:

G	*E*	Outcomes	
		Yes	No
1	1	*a*	*b*
1	0	*c*	*d*
0	1	*e*	*f*
0	0	*g*	*h*

Depending on the study design, one can estimate either odds ratios (ORs) in a case-control study or RRs in a cohort study, as illustrated in figures 1 and 2, respectively. The three measures of additive interaction devised by Rothman (14, chapter 15), in terms of RR, are relative excess risks due to interaction (RERI),

$$\mathrm{RERI} = \mathrm{RR}_{11} - \mathrm{RR}_{10} - \mathrm{RR}_{01} + 1,$$

attributable proportion due to interaction (AP),

$$\begin{aligned}\mathrm{AP} &= \mathrm{RERI}/\mathrm{RR}_{11} \\ &= 1/\mathrm{RR}_{11} - \mathrm{RR}_{10}/\mathrm{RR}_{11} - \mathrm{RR}_{01}/\mathrm{RR}_{11} + 1,\end{aligned}$$

and the synergy index (SI),

$$\mathrm{SI} = \frac{\mathrm{RR}_{11} - 1}{\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2},$$

which is simpler to investigate analytically on the log scale (1),

$$\ln \mathrm{SI} = \ln(\mathrm{RR}_{11} - 1) - \ln(\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2).$$

Notice that all RRs are estimated with $G = 0$ and $E = 0$ as the reference group. Lack of interaction is reflected by RERI = AP = 0 and SI = 1. It should also be emphasized that one should not rely on the OR in cohort studies, where the RR is readily available (15, 16). Otherwise, the well-known problem of the OR's exaggerating the RR will be even more severe in the current context (17).

CONFIDENCE INTERVALS FOR MEASURES OF ADDITIVE INTERACTION

Because the sampling distributions for single RRs are positively skewed, introductory texts in epidemiology thus suggest that inferences be conducted on the log scale. Since log-transformation cannot be applied to RERI (it could be negative), Hosmer and Lemeshow (8) suggested a simple asymptotic (SA) approach by which the 95 percent confidence limits may be obtained by subtracting from and adding to the point estimate a quantity of 1.96 times the standard error.

Both RERI and AP may be parameterized as

$$\theta_1 - \theta_2 - \theta_3 + 1,$$

with $\theta_1 = \mathrm{RR}_{11}$, $\theta_2 = \mathrm{RR}_{10}$, and $\theta_3 = \mathrm{RR}_{01}$ for RERI and $\theta_1 = 1/\mathrm{RR}_{11}$, $\theta_2 = \mathrm{RR}_{10}/\mathrm{RR}_{11}$, and $\theta_3 = \mathrm{RR}_{01}/\mathrm{RR}_{11}$ for AP. Therefore, the problem reduces to constructing CIs for $\theta_1 - \theta_2 - \theta_3 + 1$ using estimates $\hat{\theta}_i$ ($i = 1, 2, 3$) and the corresponding confidence limits, which are shown in figures 1–3 for case-control and cohort studies.

The Appendix details a general approach to construction of the CI for linear combination of parameters. Since the basic idea is to recover variance estimates needed for setting confidence limits for functions of parameters, the method may be referred to as "MOVER," indicating the method of variance estimates recovery. By Appendix equations A7 and A8, a $(1 - \alpha)100$ percent CI (L, U) for $1 + \theta_1 - \theta_2 - \theta_3$ is given by

$$\begin{aligned}L = 1 + \hat{\theta}_1 - \hat{\theta}_2 - \hat{\theta}_3 - [&(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2 + (u_3 - \hat{\theta}_3)^2 \\ &- 2r_{12}(\hat{\theta}_1 - l_1)(u_2 - \hat{\theta}_2) - 2r_{13}(\hat{\theta}_1 - l_1)(u_3 - \hat{\theta}_3) \\ &+ 2r_{23}(u_2 - \hat{\theta}_2)(u_3 - \hat{\theta}_3)]^{1/2} \qquad (1)\end{aligned}$$

and

G	E	Outcome: Yes	Outcome: No	OR	var(lnOR)
1	1	a	b	$\mathrm{OR}_{11} = \frac{ah}{bg}$	$\frac{1}{a}+\frac{1}{b}+\frac{1}{h}+\frac{1}{g}$
1	0	c	d	$\mathrm{OR}_{10} = \frac{ch}{dg}$	$\frac{1}{c}+\frac{1}{d}+\frac{1}{g}+\frac{1}{h}$
0	1	e	f	$\mathrm{OR}_{01} = \frac{eh}{fg}$	$\frac{1}{e}+\frac{1}{f}+\frac{1}{g}+\frac{1}{h}$
0	0	g	h	1	

Measures of additive interaction

$$\mathrm{RERI} = \mathrm{OR}_{11} - \mathrm{OR}_{10} - \mathrm{OR}_{01} + 1 = \widehat{\theta}_1 - \widehat{\theta}_2 - \widehat{\theta}_3 + 1$$

$$r_{12} = \frac{1/g+1/h}{\sqrt{\mathrm{var}(\ln \mathrm{OR}_{11})\mathrm{var}(\ln \mathrm{OR}_{10})}} \qquad (l_1, u_1) = \mathrm{OR}_{11} \exp[\mp 1.96\sqrt{\mathrm{var}(\ln \mathrm{OR}_{11})}]$$

$$r_{13} = \frac{1/g+1/h}{\sqrt{\mathrm{var}(\ln \mathrm{OR}_{11})\mathrm{var}(\ln \mathrm{OR}_{01})}} \qquad (l_2, u_2) = \mathrm{OR}_{10} \exp[\mp 1.96\sqrt{\mathrm{var}(\ln \mathrm{OR}_{10})}]$$

$$r_{23} = \frac{1/g+1/h}{\sqrt{\mathrm{var}(\ln \mathrm{OR}_{10})\mathrm{var}(\ln \mathrm{OR}_{01})}} \qquad (l_3, u_3) = \mathrm{OR}_{01} \exp[\mp 1.96\sqrt{\mathrm{var}(\ln \mathrm{OR}_{01})}]$$

$$\mathrm{AP} = 1/\mathrm{OR}_{11} - \mathrm{OR}_{10}/\mathrm{OR}_{11} - \mathrm{OR}_{01}/\mathrm{OR}_{11} + 1 = \widehat{\theta}_1 - \widehat{\theta}_2 - \widehat{\theta}_3 + 1$$

$$r_{12} = \frac{1/a+1/b}{\sqrt{\mathrm{var}(\ln \mathrm{OR}_{11})(1/a+1/b+1/c+1/d)}} \qquad (l_1, u_1) = \frac{1}{\mathrm{OR}_{11}} \exp[\pm 1.96\sqrt{\mathrm{var}(\ln \mathrm{OR}_{11})}]$$

$$r_{13} = \frac{1/a+1/b}{\sqrt{\mathrm{var}(\ln \mathrm{OR}_{11})(1/a+1/b+1/e+1/f)}} \qquad (l_2, u_2) = \frac{\mathrm{OR}_{10}}{\mathrm{OR}_{11}} \exp(\mp 1.96\sqrt{\tfrac{1}{a}+\tfrac{1}{b}+\tfrac{1}{c}+\tfrac{1}{d}})$$

$$r_{23} = \frac{1/a+1/b}{\sqrt{(1/a+1/b+1/c+1/d)(1/a+1/b+1/e+1/f)}} \qquad (l_3, u_3) = \frac{\mathrm{OR}_{01}}{\mathrm{OR}_{11}} \exp(\mp 1.96\sqrt{\tfrac{1}{a}+\tfrac{1}{b}+\tfrac{1}{e}+\tfrac{1}{f}})$$

$$\ln(\mathrm{SI}) = \ln(\mathrm{OR}_{11} - 1) - \ln(\mathrm{OR}_{10} + \mathrm{OR}_{01} - 2) = \widehat{\theta}_1 - \widehat{\theta}_2$$

$$\mathrm{var}(\mathrm{OR}_{11} - 1) = \mathrm{OR}_{11}^2 \mathrm{var}(\ln \mathrm{OR}_{11}) \Rightarrow \mathrm{var}[\ln(\mathrm{OR}_{11} - 1)] = \mathrm{OR}_{11}^2 \mathrm{var}(\ln \mathrm{OR}_{11})/(\mathrm{OR}_{11} - 1)^2$$

$$\mathrm{var}(\mathrm{OR}_{10} + \mathrm{OR}_{01} - 2) = \mathrm{OR}_{10}^2 \mathrm{var}(\ln \mathrm{OR}_{10}) + \mathrm{OR}_{01}^2 \mathrm{var}(\ln \mathrm{OR}_{01}) + 2\mathrm{OR}_{10}\mathrm{OR}_{01}\ (1/g + 1/h)$$

$$\mathrm{var}[\ln(\mathrm{OR}_{10} + \mathrm{OR}_{01} - 2)] = \mathrm{var}(\mathrm{OR}_{10} + \mathrm{OR}_{01} - 2)/(\mathrm{OR}_{10} + \mathrm{OR}_{01} - 2)^2$$

$$r = (\mathrm{OR}_{10} + \mathrm{OR}_{01})\ (1/g + 1/h)\ /\sqrt{\mathrm{var}(\ln \mathrm{OR}_{11})\mathrm{var}(\mathrm{OR}_{10} + \mathrm{OR}_{01} - 2)}$$

FIGURE 1. The four-by-two table for a case-control study assessing gene (*G*)-environment (*E*) interaction. OR, odds ratio; var(lnOR), variance of the data in the natural log of the odds ratio; RERI, relative excess risks due to interaction; AP, attributable proportion due to interaction; exp, exponent; SI, synergy index.

$$U = 1 + \hat{\theta}_1 - \hat{\theta}_2 - \hat{\theta}_3 + [(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2 + (\hat{\theta}_3 - l_3)^2 - 2r_{12}(u_1 - \hat{\theta}_1)(\hat{\theta}_2 - l_2) - 2r_{13}(u_1 - \hat{\theta}_1)(\hat{\theta}_3 - l_3) + 2r_{23}(\hat{\theta}_2 - l_2)(\hat{\theta}_3 - l_3)]^{1/2}. \quad (2)$$

The estimated correlation r_{ij}, $i = 1, 2$, $j = 2, 3$ may be obtained as

$$r_{ij} = \frac{\widehat{\mathrm{cov}}(\hat{\theta}_i, \hat{\theta}_j)}{\sqrt{\widehat{\mathrm{var}}(\hat{\theta}_i)\widehat{\mathrm{var}}(\hat{\theta}_j)}}. \quad (3)$$

For tabulated case-control data (figure 1), Rothman (1, p. 507) showed that

$$\mathrm{cov}(\mathrm{OR}_{01}, \mathrm{OR}_{10}) = \mathrm{OR}_{01} \times \mathrm{OR}_{10}(1/g + 1/h),$$

reflecting that OR_{10} and OR_{01} have g and h in common. An application of the delta method yields

$$\mathrm{var}(\mathrm{OR}_{10}) = \mathrm{OR}_{10}^2 \times \mathrm{var}(\ln \mathrm{OR}_{10})$$

and similarly found for $\mathrm{var}(\mathrm{OR}_{01})$. By the definition of correlation,

$$\mathrm{corr}(\mathrm{OR}_{10}, \mathrm{OR}_{01}) = \frac{\mathrm{cov}(\mathrm{OR}_{01}, \mathrm{OR}_{10})}{\sqrt{\mathrm{var}(\mathrm{OR}_{10}) \times \mathrm{var}(\mathrm{OR}_{01})}} = \frac{1/g + 1/h}{\sqrt{\mathrm{var}(\ln \mathrm{OR}_{10})\mathrm{var}(\ln \mathrm{OR}_{01})}}.$$

Other correlations shown in figure 1 may be obtained in the same fashion. Correlation estimates for a cohort study may be obtained by replacing $1/h$ with $-1/n_{00}$ and ORs with

G	E	Outcome: Yes	Outcome: No	Total	RR	var(lnRR)
1	1	a	b	n_{11}	$\mathrm{RR}_{11} = \frac{an_{00}}{gn_{11}}$	$\frac{1}{a} - \frac{1}{n_{11}} + \frac{1}{g} - \frac{1}{n_{00}}$
1	0	c	d	n_{10}	$\mathrm{RR}_{10} = \frac{cn_{00}}{gn_{10}}$	$\frac{1}{c} - \frac{1}{n_{10}} + \frac{1}{g} - \frac{1}{n_{00}}$
0	1	e	f	n_{01}	$\mathrm{RR}_{01} = \frac{en_{00}}{gn_{01}}$	$\frac{1}{e} - \frac{1}{n_{01}} + \frac{1}{g} - \frac{1}{n_{00}}$
0	0	g	h	n_{00}	1	

Measures of additive interaction

$$\mathrm{RERI} = \mathrm{RR}_{11} - \mathrm{RR}_{10} - \mathrm{RR}_{01} + 1 = \widehat{\theta}_1 - \widehat{\theta}_2 - \widehat{\theta}_3 + 1$$

$$r_{12} = \frac{1/g - 1/n_{00}}{\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})\mathrm{var}(\ln \mathrm{RR}_{10})}} \qquad (l_1, u_1) = \mathrm{RR}_{11} \exp[\mp 1.96\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})}]$$

$$r_{13} = \frac{1/g - 1/n_{00}}{\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})\mathrm{var}(\ln \mathrm{RR}_{01})}} \qquad (l_2, u_2) = \mathrm{RR}_{10} \exp[\mp 1.96\sqrt{\mathrm{var}(\ln \mathrm{RR}_{10})}]$$

$$r_{23} = \frac{1/g - 1/n_{00}}{\sqrt{\mathrm{var}(\ln \mathrm{RR}_{10})\mathrm{var}(\ln \mathrm{RR}_{01})}} \qquad (l_3, u_3) = \mathrm{RR}_{01} \exp[\mp 1.96\sqrt{\mathrm{var}(\ln \mathrm{RR}_{01})}]$$

$$\mathrm{AP} = 1/\mathrm{RR}_{11} - \mathrm{RR}_{10}/\mathrm{RR}_{11} - \mathrm{RR}_{01}/\mathrm{RR}_{11} + 1 = \widehat{\theta}_1 - \widehat{\theta}_2 - \widehat{\theta}_3 + 1$$

$$r_{12} = \frac{1/a - 1/n_{11}}{\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})(1/a - 1/n_{11} + 1/c - 1/n_{10})}} \qquad (l_1, u_1) = (1/\mathrm{RR}_{11}) \exp[\mp 1.96\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})}]$$

$$r_{13} = \frac{1/a - 1/n_{11}}{\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})(1/a - 1/n_{11} + 1/e - 1/n_{01})}} \qquad (l_2, u_2) = (\mathrm{RR}_{10}/\mathrm{RR}_{11}) \exp(\mp 1.96\sqrt{\tfrac{1}{a} - \tfrac{1}{n_{11}} + \tfrac{1}{c} - \tfrac{1}{n_{10}}})$$

$$r_{23} = \frac{1/a - 1/n_{11}}{\sqrt{(1/a - 1/n_{11} + 1/c - 1/n_{10})(1/a - 1/n_{11} + 1/e - 1/n_{01})}} \qquad (l_3, u_3) = (\mathrm{RR}_{01}/\mathrm{RR}_{11}) \exp(\mp 1.96\sqrt{\tfrac{1}{a} - \tfrac{1}{n_{11}} + \tfrac{1}{e} - \tfrac{1}{n_{01}}})$$

$$\ln(\mathrm{SI}) = \ln(\mathrm{RR}_{11} - 1) - \ln(\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2) = \widehat{\theta}_1 - \widehat{\theta}_2$$

$$\mathrm{var}(\mathrm{RR}_{11} - 1) = \mathrm{RR}_{11}^2 \mathrm{var}(\ln \mathrm{RR}_{11}) \Rightarrow \mathrm{var}[\ln(\mathrm{RR}_{11} - 1)] = \mathrm{RR}_{11}^2 \mathrm{var}(\ln \mathrm{RR}_{11})/(\mathrm{RR}_{11} - 1)^2$$

$$\mathrm{var}(\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2) = \mathrm{RR}_{10}^2 \mathrm{var}(\ln \mathrm{RR}_{10}) + \mathrm{RR}_{01}^2 \mathrm{var}(\ln \mathrm{RR}_{01}) + 2\mathrm{RR}_{10}\mathrm{RR}_{01}(1/g - 1/n_{00})$$

$$\mathrm{var}[\ln(\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2)] = \mathrm{var}(\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2)/(\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2)^2$$

$$r = (\mathrm{RR}_{10} + \mathrm{RR}_{01})\,(1/g - 1/n_{00})\,/\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})\mathrm{var}(\mathrm{RR}_{10} + \mathrm{RR}_{01} - 2)}$$

FIGURE 2. The four-by-two table for a cohort study assessing gene (*G*)-environment(*E*) interaction. RR, risk ratio; var(lnRR), variance of the data in the natural log of the risk ratio; RERI, relative excess risks due to interaction; exp, exponent; AP, attributable proportion due to interaction; SI, synergy index.

RRs, resulting in the expressions in figure 2. In the case of the multiplicative regression model, applying the delta method and definition of correlation, for $i \neq j$,

$$\begin{aligned}\mathrm{corr}(e^{b_i}, e^{b_j}) &= \frac{\mathrm{cov}(e^{b_i}, e^{b_j})}{\sqrt{\mathrm{var}(e^{b_i})\mathrm{var}(e^{b_j})}} \\ &= \frac{e^{b_i}e^{b_j}\mathrm{cov}(b_i, b_j)}{\sqrt{(e^{b_i})^2\mathrm{var}(b_i)(e^{b_j})^2\mathrm{var}(b_j)}} \\ &= \frac{\mathrm{cov}(b_i, b_j)}{\sqrt{\mathrm{var}(b_i)\mathrm{var}(b_j)}}.\end{aligned}$$

Applying the same idea to other correlations results in the expressions in figure 3.

It can be shown with equations 1 and 2 that the SA method by Hosmer and Lemeshow (8) is a consequence of assuming symmetric confidence limits for RRs. To see this for RERI, one needs to replace $\hat{\theta}_1 - l_1$ and $u_1 - \hat{\theta}_1$ by $z_{\alpha/2}\mathrm{RR}_{11}\sqrt{\mathrm{var}(\ln \mathrm{RR}_{11})}$, $\hat{\theta}_2 - l_2$ and $u_2 - \hat{\theta}_2$ by $z_{\alpha/2}\mathrm{RR}_{10}\sqrt{\mathrm{var}(\ln \mathrm{RR}_{10})}$, and $\hat{\theta}_3 - l_3$ and $u_3 - \hat{\theta}_3$ by $z_{\alpha/2}\mathrm{RR}_{01}\sqrt{\mathrm{var}(\ln \mathrm{RR}_{01})}$. Similar exercises will result in the SA CI for AP. This brings out the failing point of the SA approach, that it has implicitly assumed that confidence limits for the RR are given by $\mathrm{RR} \mp z_{\alpha/2}\mathrm{RR}\sqrt{\mathrm{var}(\ln \mathrm{RR})}$. Failing to see this point may have resulted in the proliferation of the SA method (9–12).

Now, since the derivation of the MOVER method (equations 1 and 2) did not assume symmetric confidence limits for θ_i, one can use sensible confidence limits for RRs, such as $\mathrm{RR}\exp(\mp z_{\alpha/2}\sqrt{\mathrm{var}(\ln \mathrm{RR})})$, in the construction of confidence interval for RERI and AP.

	Estimate	Variance-covariance			
G	b_1	v_1	v_{12}	v_{13}	$\text{var}(b_1+b_2+b_3)=v_1+v_2+v_3+2(v_{12}+v_{13}+v_{23})$
E	b_2		v_2	v_{23}	$\text{var}(b_1+b_3)=v_1+v_3+2v_{13}$
$G\times E$	b_3			v_3	$\text{var}(b_2+b_3)=v_2+v_3+2v_{23}$

$$\text{RERI}=e^{b_1+b_2+b_3}-e^{b_1}-e^{b_2}+1=\widehat{\theta}_1-\widehat{\theta}_2-\widehat{\theta}_3+1$$
$$r_{12}=(v_1+v_{12}+v_{13})/\sqrt{v_1\text{var}(b_1+b_2+b_3)}$$
$$r_{13}=(v_{12}+v_2+v_{23})/\sqrt{v_2\text{var}(b_1+b_2+b_3)}$$
$$r_{23}=v_{12}/\sqrt{v_1v_2}$$

$$\text{AP}=1/e^{b_1+b_2+b_3}-1/e^{b_2+b_3}-1/e^{b_1+b_3}+1=\widehat{\theta}_1-\widehat{\theta}_2-\widehat{\theta}_3+1$$
$$r_{12}=(v_{12}+v_{13}+v_2+2v_{23}+v_3)/\sqrt{\text{var}(b_2+b_3)\text{var}(b_1+b_2+b_3)}$$
$$r_{13}=(v_1+v_{12}+2v_{13}+v_{23}+v_3)/\sqrt{\text{var}(b_1+b_3)\text{var}(b_1+b_2+b_3)}$$
$$r_{23}=(v_{12}+v_{23}+v_{13}+v_3)/\sqrt{\text{var}(b_2+b_3)\text{var}(b_1+b_3)}$$

$$\ln(\text{SI})=\ln(e^{b_1+b_2+b_3}-1)-\ln(e^{b_1}+e^{b_2}-2)=\widehat{\theta}_1-\widehat{\theta}_2$$
$$\text{var}[\ln(e^{b_1+b_2+b_3}-1)]=\left[\frac{e^{b_1+b_2+b_3}}{e^{b_1+b_2+b_3}-1}\right]^2\text{var}(b_1+b_2+b_3)$$
$$\text{var}(e^{b_1}+e^{b_2}-2)=e^{2b_1}v_1+e^{2b_2}v_2+2e^{b_1+b_2}v_{12}$$
$$\text{var}[\ln(e^{b_1}+e^{b_2}-2)]=\frac{\text{var}(e^{b_1}+e^{b_2}-2)}{(e^{b_1}+e^{b_2}-2)^2}$$
$$r=[e^{b_1}(v_1+v_{12}+v_{13})+e^{b_2}(v_{12}+v_2+v_{23})]/\sqrt{\text{var}(b_1+b_2+b_3)\text{var}(e^{b_1}+e^{b_2}-2)}$$

FIGURE 3. Confidence interval construction for measures of additive interaction using output from multiplicative regression programs. *G*, gene; var, variance; *E*, environment; RERI, relative excess risks due to interaction; AP, attributable proportion due to interaction; SI, synergy index.

Furthermore, denoting $\theta_1=\ln(\text{RR}_{11}-1)$ and $\theta_2=-\ln(\text{RR}_{10}+\text{RR}_{01}-2)$, Appendix equations A5 and A6 may be applied to ln SI that, in turn, can be used to obtain confidence limits for SI. With the expressions in figures 1–3, the results for SI will be identical to those obtained by use of the methods proposed by Rothman (1).

SIMULATION STUDY

Despite the justification provided in the Appendix, the proposed procedure for measures of interaction is based on asymptotic theory. Simulation studies were therefore undertaken to evaluate its performance.

For AP, a method based on ln(1 − AP) (18) was also included. The studies were performed in the context of a case-control design, with the understanding that the statistical theory is identical regardless of whether the OR, RR, or hazard ratio is selected as the effect measure.

The first study used 20 OR combinations ($2\text{RR}_{10}\times 2\text{RR}_{01}\times 5\text{RR}_{11}$) and a sample size of 250 in each case and control group as in the study reported by Assmann et al. (13). Compared with the MOVER approach, the approaches were the SA approach (8) and the bias-corrected and accelerated (BCa) bootstrap approach (3, pp. 184–188). For each parameter combination, 1,000 replicates were performed. The number of resamples for the bootstrap was also set to 1,000. The proportions of control subjects exposed to *G* alone, *E* alone, and both *G* and *E* were 0.1, 0.2, and 0.1, respectively. The exposure probability distribution for the case subjects was then calculated by use of the specific values of RR_{11}, RR_{10}, and RR_{01}. Data for the cases and controls were generated separately from multinomial distributions. Cells with 0 count were added by 0.5 so that ORs could be calculated.

An additional simulation study without the bootstrap was conducted to see whether the SA approach could perform reasonably well in sample sizes of 1,000 in each of the case and control groups.

A third simulation study was performed to assess the performance of the MOVER method compared with the SA method in situations with small exposure probabilities. With the other parameters set as in the first simulation, the

TABLE 1. Coverage properties of the 95% two-sided confidence intervals for relative excess risks due to interaction and attributable proportion due to interaction based on 1,000 runs*

OR_{10}†	OR_{01}	OR_{11}	RERI†	$\widehat{RERI}$	SA†		BCa		MOVER†	
					Rate	95% CI†	Rate	95% CI	Rate	95% CI
4.0	5.0	20.000	12.000	12.86	94.4	5.6, 0.0	96.2	2.1, 2.7	94.9	3.1, 2.0
		12.000	4.000	4.27	96.2	3.8, 0.0	95.7	2.4, 1.9	96.1	2.3, 1.6
		8.000	0.000	−0.01	99.0	0.0, 0.1	95.6	2.3, 2.1	95.6	2.0, 2.4
		6.000	−2.000	−2.09	98.5	0.8, 0.7	95.9	2.2, 1.9	95.5	2.2, 2.3
		4.000	−4.000	−4.20	96.7	0.4, 2.9	95.8	2.2, 2.0	95.5	1.8, 2.7
2.0	5.0	15.000	9.000	9.55	95.0	5.0, 0.0	96.4	1.9, 1.7	95.4	2.9, 1.7
		9.000	3.000	3.15	97.1	2.9, 0.0	95.9	2.7, 1.4	96.1	2.8, 1.1
		6.000	0.000	−0.03	98.9	0.9, 0.2	95.5	3.0, 1.5	95.9	2.7, 1.4
		4.500	−1.500	−1.61	98.4	0.4, 1.2	95.3	2.7, 2.0	95.5	2.4, 2.1
		3.000	−3.000	−3.20	97.3	0.1, 2.6	96.1	2.3, 1.6	96.5	1.5, 2.0
4.0	2.5	13.750	8.250	8.81	94.4	5.6, 0.0	95.9	1.9, 2.2	95.6	2.6, 1.8
		8.250	2.750	2.95	96.7	3.3, 0.0	95.7	2.3, 2.0	95.6	2.6, 1.8
		5.500	0.000	0.04	98.2	1.6, 0.2	95.4	2.1, 2.5	95.6	2.1, 2.3
		4.125	−1.375	−1.41	97.4	1.5, 1.1	95.9	2.0, 2.1	95.4	2.6, 2.0
		2.750	−2.750	−2.87	97.0	0.5, 2.5	96.4	1.8, 1.8	96.1	2.0, 1.9
2.0	2.5	8.750	5.250	5.52	94.4	5.6, 0.0	96.1	2.3, 1.6	95.5	3.1, 1.4
		5.250	1.750	1.83	97.0	2.9, 0.1	95.9	2.8, 1.3	95.9	3.0, 1.1
		3.500	0.000	−0.01	98.3	1.2, 0.5	95.3	2.8, 1.9	95.4	2.8, 1.8
		2.625	−0.875	−0.94	98.1	0.7, 1.2	95.6	2.3, 2.1	95.4	2.5, 2.1
		1.750	−1.750	−1.85	96.2	0.6, 3.2	95.4	2.6, 2.0	95.3	2.4, 2.3

			AP†	$\widehat{AP}$	SA		BCa		ln(1 − AP)†		MOVER	
					Rate	95% CI	Rate	95% CI	Rate	95% CI	Rate	95% CI
4.0	5.0	20.000	0.60	0.58	95.7	0.3, 4.0	96.0	2.0, 2.0	96.7	2.3, 1.0	96.1	3.0, 0.9
		12.000	0.33	0.30	95.0	0.1, 4.9	96.1	1.7, 2.2	96.2	2.1, 1.7	95.4	3.4, 1.2
		8.000	0.00	−0.05	94.8	0.0, 5.2	95.9	2.0, 2.1	95.8	2.6, 1.6	95.6	3.4, 1.0
		6.000	−0.33	−0.41	95.0	0.0, 5.0	95.5	2.6, 1.9	95.7	2.7, 1.6	95.2	3.7, 1.1
		4.000	−1.00	−1.12	94.5	0.1, 5.4	95.5	2.6, 1.9	95.3	3.0, 1.7	94.9	3.8, 1.3
2.0	5.0	15.000	0.60	0.58	93.9	0.1, 6.0	96.8	2.1, 1.1	96.7	2.2, 1.1	96.6	2.5, 0.9
		9.000	0.33	0.30	95.5	0.0, 5.0	96.5	2.1, 1.4	96.1	2.4, 1.5	96.2	2.9, 0.9
		6.000	0.00	−0.06	94.5	0.0, 5.5	96.0	2.8, 1.2	96.1	3.1, 0.8	95.1	4.2, 0.7
		4.500	−0.33	−0.42	94.3	0.0, 5.7	95.8	2.5, 1.7	95.6	3.0, 1.4	94.8	4.3, 0.9
		3.000	−1.00	−1.15	94.8	0.0, 5.2	96.0	2.5, 1.5	95.4	3.0, 1.6	95.1	3.9, 1.0
4.0	2.5	13.750	0.60	0.58	94.6	0.2, 5.2	96.5	1.6, 1.9	96.0	2.4, 1.6	95.8	2.9, 1.3
		8.250	0.33	0.31	95.0	0.1, 4.9	96.3	1.8, 1.9	95.5	2.6, 1.9	95.4	3.0, 1.6
		5.500	0.00	−0.05	94.5	0.2, 5.3	95.8	2.0, 2.2	95.8	2.6, 1.6	95.0	3.7, 1.3
		4.125	−0.33	−0.41	95.0	0.0, 5.0	95.6	2.4, 2.0	95.4	3.1, 1.5	95.0	3.8, 1.2
		2.750	−1.00	−1.13	95.2	0.0, 4.8	95.7	2.5, 1.8	95.4	3.2, 1.4	94.8	3.9, 1.3
2.0	2.5	8.750	0.60	0.58	94.9	0.3, 4.8	96.2	2.2, 1.5	96.1	2.5, 1.4	95.9	3.2, 0.9
		5.250	0.33	0.30	95.0	0.0, 5.0	96.8	2.1, 1.1	96.3	2.7, 1.0	96.0	3.2, 0.8
		3.500	0.00	−0.06	94.4	0.0, 5.6	96.0	2.5, 1.5	95.4	3.4, 1.2	95.1	3.8, 1.1
		2.625	−0.33	−0.42	94.2	0.0, 5.8	95.7	2.7, 1.6	95.2	3.6, 1.2	95.0	4.2, 0.8
		1.750	−1.00	−1.16	94.1	0.1, 5.8	95.9	2.3, 1.8	95.2	3.0, 1.8	94.0	4.9, 1.1

* Each entry is the coverage rate (left miscoverage, right miscoverage) based on 250 cases and 250 controls. The bias-corrected and accelerated (BCa) bootstrap approach was based on 1,000 resamples. The proportions of controls exposed to 10, 01, and 11 were 0.1, 0.2, and 0.1, respectively.

† OR, odds ratio; RERI, relative excess risks due to interaction; SA, simple asymptotic; MOVER, method of variance estimates recovery; CI, confidence interval; AP, attributable proportion due to interaction; ln(1 − AP), natural log of (1 − AP).

TABLE 2. Coverage properties of the 95% two-sided confidence intervals for relative excess risks due to interaction and attributable proportion due to interaction based on 1,000 runs*

OR_{10}†	OR_{01}	OR_{11}	RERI†	$\widehat{RERI}$	SA†		MOVER†	
					Rate	95% CI†	Rate	95% CI
4.0	5.0	20.000	12.000	12.18	94.8	4.7, 0.5	95.8	2.6, 1.6
		12.000	4.000	4.09	95.6	3.5, 0.9	95.5	2.4, 2.1
		8.000	0.000	0.05	96.1	2.5, 1.4	95.1	2.5, 2.4
		6.000	−2.000	−1.97	95.7	1.9, 2.4	95.0	2.2, 2.8
		4.000	−4.000	−4.00	95.4	0.9, 3.7	95.4	1.4, 3.2
2.0	5.0	15.000	9.000	9.12	95.1	4.4, 0.5	95.5	2.7, 1.8
		9.000	3.000	3.06	96.3	2.6, 1.1	95.6	2.2, 2.2
		6.000	0.000	0.03	97.1	1.5, 1.4	96.4	1.6, 2.0
		4.500	−1.500	−1.49	96.5	1.2, 2.3	95.8	1.8, 2.4
		3.000	−3.000	−3.01	95.2	0.8, 4.0	95.6	1.5, 2.9
4.0	2.5	13.750	8.250	8.38	95.4	4.2, 0.4	96.1	2.7, 1.2
		8.250	2.750	2.82	95.3	3.6, 1.1	94.6	3.1, 2.3
		5.500	0.000	0.04	96.6	2.3, 1.1	95.6	2.3, 2.1
		4.125	−1.375	−1.36	95.3	2.3, 2.4	94.6	2.6, 2.8
		2.750	−2.750	−2.74	95.8	1.1, 3.1	95.5	1.7, 2.8
2.0	2.5	8.750	5.250	5.32	95.1	4.4, 0.5	96.0	2.6, 1.4
		5.250	1.750	1.79	95.5	3.0, 1.5	95.6	2.6, 1.8
		3.500	0.000	0.02	96.4	2.3, 1.3	95.6	2.3, 2.1
		2.625	−0.875	−0.86	96.1	1.2, 2.7	95.8	1.6, 2.6
		1.750	−1.750	−1.76	95.8	1.0, 3.2	95.8	1.6, 2.6

			AP†	$\widehat{AP}$	SA		ln(1 − AP)†		MOVER	
					Rate	95% CI	Rate	95% CI	Rate	95% CI
4.0	5.0	20.000	0.60	0.60	94.8	1.2, 4.0	94.9	3.0, 2.1	94.8	3.2, 2.0
		12.000	0.33	0.33	95.0	1.1, 3.9	95.1	2.3, 2.6	95.1	2.8, 2.1
		8.000	0.00	−0.01	94.6	1.2, 4.2	95.3	2.7, 2.0	95.4	3.0, 1.6
		6.000	−0.33	−0.34	94.9	1.0, 4.1	95.4	2.8, 1.8	95.2	3.1, 1.7
		4.000	−1.00	−1.01	95.3	0.8, 3.9	95.9	2.1, 2.0	95.8	2.3, 1.9
2.0	5.0	15.000	0.60	0.60	95.2	0.9, 3.9	95.1	2.4, 2.5	95.3	2.6, 2.1
		9.000	0.33	0.33	94.1	0.8, 5.1	95.5	2.0, 2.5	95.3	2.4, 2.3
		6.000	0.00	−0.01	94.9	0.9, 4.2	96.2	1.9, 1.9	96.2	2.3, 1.5
		4.500	−0.33	−0.34	95.3	1.1, 3.6	95.9	2.0, 2.1	96.0	2.1, 1.9
		3.000	−1.00	−1.02	95.1	0.6, 4.3	95.6	2.1, 2.3	95.5	2.5, 2.0
4.0	2.5	13.750	0.60	0.60	95.3	1.1, 3.6	95.6	2.7, 1.7	95.6	2.8, 1.6
		8.250	0.33	0.33	94.6	1.2, 4.2	94.5	2.9, 2.6	94.7	3.0, 2.3
		5.500	0.00	−0.01	95.1	1.2, 3.7	95.6	2.4, 2.0	95.5	2.7, 1.8
		4.125	−0.33	−0.34	95.2	1.0, 3.8	94.7	3.0, 2.3	94.9	3.0, 2.1
		2.750	−1.00	−1.01	94.7	0.7, 4.6	95.6	2.1, 2.3	95.7	2.5, 1.8
2.0	2.5	8.750	0.60	0.60	94.3	0.7, 5.0	95.0	2.1, 2.9	95.2	2.3, 2.5
		5.250	0.33	0.33	94.3	1.0, 4.7	95.4	2.3, 2.3	95.4	2.5, 2.1
		3.500	0.00	−0.01	95.5	0.7, 3.8	95.8	2.5, 1.7	95.5	3.0, 1.5
		2.625	−0.33	−0.34	95.4	0.7, 3.9	95.4	2.5, 2.1	95.7	2.9, 1.4
		1.750	−1.00	−1.02	95.9	0.6, 3.5	96.2	2.1, 1.7	95.5	3.0, 1.5

* Each entry is the coverage rate (left miscoverage, right miscoverage) based on 1,000 cases and 1,000 controls. The bias-corrected and accelerated (BCa) bootstrap approach was based on 1,000 resamples. The proportions of controls exposed to 10, 01, and 11 were 0.1, 0.2, and 0.1, respectively.

† OR, odds ratio; RERI, relative excess risks due to interaction; SA, simple asymptotic; MOVER, method of variance estimates recovery; CI, confidence interval; AP, attributable proportion due to interaction; ln(1 − AP), natural log of (1 − AP).

TABLE 3. Coverage properties of the 95% two-sided confidence intervals for relative excess risks due to interaction and attributable proportion due to interaction based on 1,000 runs*

OR_{10}†	OR_{01}	OR_{11}	RERI†	$\widehat{RERI}$	SA†		MOVER†	
					Rate	95% CI†	Rate	95% CI
4.0	5.0	20.000	12.000	13.7	94.5	5.5, 0	94.1	3.5, 2.4
		12.000	4.000	4.35	97.8	2.2, 0	94.2	3.9, 1.9
		8.000	0.000	0.12	100	0, 0	95.0	3.0, 2.0
		6.000	−2.000	−2.35	99.6	0.1, 0.3	94.9	3.1, 2.0
		4.000	−4.000	−4.42	98.2	0, 1.8	95.7	2.9, 1.4
2.0	5.0	15.000	9.000	10.02	94.9	5.1, 0	94.5	4.1, 1.4
		9.000	3.000	3.06	98.2	1.8, 0	94.3	4.2, 1.5
		6.000	0.000	−0.07	99.6	0.4, 0	96.0	2.5, 1.5
		4.500	−1.500	−1.75	99.4	0.1, 0.5	95.5	2.2, 2.3
		3.000	−3.000	−3.28	98.3	0, 1.7	95.6	2.5, 1.9
4.0	2.5	13.750	8.250	8.94	94.9	5.1, 0	94.1	4.3, 1.6
		8.250	2.750	3.03	98.0	2.0, 0	95.4	3.3, 1.3
		5.500	0.000	−0.04	99.4	0.5, 0.1	93.5	4.3, 2.2
		4.125	−1.375	−1.75	99.7	0, 0.3	94.5	3.8, 1.7
		2.750	−2.750	−3.11	98.7	0.1, 1.2	95.6	2.6,1.8
2.0	2.5	8.750	5.250	5.32	95.2	4.8, 0	94.6	4.2, 1.2
		5.250	1.750	1.85	99.0	1.0, 0	96.0	2.5, 1.5
		3.500	0.000	−0.19	99.4	0.5, 0.1	94.5	4.0, 1.5
		2.625	−0.875	−1.00	98.9	0.5, 0.6	94.4	3.3, 2.3
		1.750	−1.750	−2.10	98.0	0.1, 1.9	94.6	4.3, 1.1

			AP†	$\widehat{AP}$	SA		ln(1 − AP)†		MOVER	
					Rate	95% CI	Rate	95% CI	Rate	95% CI
4.0	5.0	20.000	0.60	0.55	92.9	0, 7.1	94.3	3.9, 1.8	92.9	6.0, 1.1
		12.000	0.33	0.25	93.3	0, 6.7	95.0	3.1, 1.9	93.2	5.4, 1.4
		8.000	0.00	−0.10	92.6	0, 7.4	95.6	2.4, 2.0	95.0	4.0, 1.0
		6.000	−0.33	−0.51	93.2	0, 6.8	95.0	3.2, 1.8	93.9	4.8, 1.3
		4.000	−1.00	−1.26	92.2	0, 7.8	96.3	2.4,1.3	93.9	5.4, 0.7
2.0	5.0	15.000	0.60	0.56	91.8	0, 8.2	95.9	2.4, 1.7	95.1	3.8, 1.1
		9.000	0.33	0.23	91.8	0, 8.2	95.2	3.0, 1.8	93.5	5.3, 1.2
		6.000	0.00	−0.12	93.1	0, 6.9	96.3	2.2, 1.5	94.9	3.9, 1.2
		4.500	−0.33	−0.51	91.7	0, 8.3	95.9	2.0, 2.1	94.5	3.9, 1.6
		3.000	−1.00	−1.26	92.2	0, 7.8	96.3	2.2, 1.5	94.6	4.2, 1.2
4.0	2.5	13.750	0.60	0.54	93.6	0, 6.4	95.4	3.0, 1.6	93.5	5.5, 1.0
		8.250	0.33	0.26	92.6	0, 7.4	95.6	2.4, 2.0	94.5	4.2, 1.3
		5.500	0.00	−0.13	92.2	0, 7.8	94.3	3.7, 2.0	92.6	5.8, 1.6
		4.125	−0.33	−0.56	93.2	0, 6.8	94.9	3.3, 1.8	92.7	6.1, 1.2
		2.750	−1.00	−1.34	93.3	0, 6.7	95.2	3.0, 1.8	93.0	5.8, 1.2
2.0	2.5	8.750	0.60	0.52	93.6	0, 6.4	94.7	3.8, 1.5	92.3	7.1, 0.6
		5.250	0.33	0.24	93.1	0, 6.9	96.2	2.3, 1.5	94.2	5.1, 0.7
		3.500	0.00	−0.18	93.1	0, 6.9	94.9	3.6, 1.5	92.0	7.0, 1.0
		2.625	−0.33	−0.53	92.3	0, 7.7	95.2	2.4, 2.4	92.2	6.1, 1.7
		1.750	−1.00	−1.44	93.6	0, 6.4	95.8	3.5, 0.7	90.4	9.4, 0.2

* Each entry is the coverage rate (left miscoverage, right miscoverage) based on 250 cases and 250 controls. The proportions of controls exposed to 10, 01, and 11 were all set to 0.05.

† OR, odds ratio; RERI, relative excess risks due to interaction; SA, simple asymptotic; MOVER, method of variance estimates recovery; CI, confidence interval; AP, attributable proportion due to interaction; ln(1 − AP), natural log of (1 − AP).

probabilities of controls exposed to G alone, E alone, and both G and E were 0.05, 0.05, and 0.05, respectively.

Because the main focus was on the extent to which the empirical coverage of the CI matched with the nominal 95 percent level, the first criterion was whether or not the coverage rate was within the range of 93.6–96.4 percent. The difference between the two miscoverage rates was the second criterion, where smaller differences were preferred. The reason to set balanced miscoverage errors as the second criterion is that a CI should contain possible parameter values that are not too large and not too small. An advertised 95 percent CI is supposed to miss about 2.5 percent from each side.

The coverage rate for each method was calculated as the proportion of the 1,000 CIs constructed that contained the values of the additive interaction. The left miscoverage rate was obtained by calculating the proportion of the upper limits that were less than the parameter value, while the right miscoverage rate was obtained as the proportion of the lower limits larger than the parameter value.

The results in table 1 show that, for RERI, the SA approach missed the target coverage range of 93.6–96.4 percent in 14 of 20 cases. The poor performance is more pronounced when the miscoverage rates are considered. In contrast, the MOVER approach provided coverage rates that are all in the range, with only a single one with 96.5 percent. The overall performance is very comparable to that of the bootstrap.

For AP, the SA approach actually provides overall coverage rates that are within the range, but in a lop-sided manner. In particular, the high right miscoverage rates indicate that this approach tends to provide lower confidence limits that are too high. A possible consequence is false positive results. Table 1 also demonstrates that the MOVER approach and the ln(1 − AP) approach provided slightly better coverage results, but the miscoverage rates are not balanced as is the case for the bootstrap approach. Nonetheless, these miscoverage rates seem to be reasonable from a practical perspective. Table 2 shows that increasing sample sizes to 2,000 subjects can have only a limited effect in improving the performance of the SA approach, especially when the miscoverage rates are considered.

As predicted by the theoretical results above, further simulation results with small exposure probabilities (table 3) demonstrate that, for RERI, the MOVER approach performed satisfactorily, while the SA deteriorated. Interestingly, the ln(1 − AP) approach performed very well. These results also demonstrate that there exists room for improvement in the case of AP when the exposure probabilities are small. Since the MOVER approach draws its validity for the confidence limits for ORs, future research may focus on adopting better CIs for OR (19) or for RR (20).

EXAMPLES

Example 1: negative confidence limits for ORs used by the SA method to obtain those for RERI

This example concerns smoking and alcohol use in relation to oral cancer among male veterans (8; 14, chapter 15). The data are presented in a four-by-two table in figure 4. An application of the naive SA method results in a 95 percent CI of −1.83, 9.31 for RERI (8). As discussed above, this CI is a consequence of applying symmetric intervals for ORs in equations 1 and 2. In other words, the SA method has implicitly used a symmetric interval for OR given by $\mathrm{OR} \mp z_{\alpha/2}\mathrm{OR}\sqrt{\mathrm{var}(\ln \mathrm{OR})}$. For the data in figure 4, such an approach provides 95 percent CIs for OR_{11}, OR_{10}, and OR_{01} as −2.07, 20.12; −1.86, 8.52; and −1.40, 7.32, respectively. This hidden feature of the SA method has escaped notice for the past 16 years, although it is well known that a better interval for OR is given by $\mathrm{OR}\exp(\mp z_{\alpha/2}\sqrt{\mathrm{var}(\ln \mathrm{OR})})$. For the data at hand, the CIs for the above three ORs are 2.64, 30.9; 0.70, 15.84; and 0.68, 12.89, respectively. It is interesting to note that these limits were presented in the article by Hosmer and Lemeshow (8, p. 454), but the SA method has no ability to use them. Applying these asymmetric intervals to equations 1 and 2 yields a 95 percent CI of −11.41, 21.84 for RERI.

		Oral cancer	
Alcohol	Smoke	Case	Control
1	1	$a = 225$	$b = 166$
1	0	$c = 6$	$d = 12$
0	1	$e = 8$	$f = 18$
0	0	$g = 3$	$h = 20$

	Estimate	Symmetric CI (for naive method)	Asymmetric CI (for proposed method)
OR_{11}	9.027270	-2.071834, 20.126374	2.6399015, 30.869183
OR_{10}	3.3300589	-1.862949, 8.5230672	0.7001614, 15.838194
OR_{01}	2.9600318	-1.395426, 7.3154895	0.6796192, 12.892202
r_{12}	0.7674552		
r_{13}	0.8133598		
r_{23}	0.6412791		

		SA	MOVER
RERI	3.74	-1.83, 9.31	-11.41, 21.84

FIGURE 4. Calculation of 95% confidence intervals (CIs) for relative excess risks due to interaction (RERI) by use of data from example 1. OR, odds ratio; SA, simple asymptotic; MOVER, method of variance estimates recovery.

Example 2: falsely claimed interaction resulting from the SA method

Consider a data set arising from a case-cohort design in the Atherosclerosis Risk in Communities (ARIC) Study (21, 22), where it is of interest to determine the interaction between a susceptibility genotype, glutathione *S*-transferase M1 polymorphism (*GSTM1*), and smoking on the risk of incident coronary heart disease. A total of 458 incident cases of coronary heart disease occurred in the population of 14,239 eligible participants during the period from 1989

Example 2. *GSTM1* and smoking on coronary heart disease (12), with b_i's from a Cox model for case-cohort design (23).

	Estimate	Variance-covariance	
GSTM1	$b_1 = 0.0543$		
Smoking	$b_2 = 0.2826$		
$G \times S$	$b_3 = 0.5869$		
GSTM1	0.08169	0.03124	-0.08129
Smoking		0.06543	-0.05929
$G \times S$			0.13954

	For RERI			For AP		
	Estimate	delta CI	Convention CI	Estimate	delta CI	Convention CI
θ_1	2.52	1.23, 3.81	1.51, 4.20	0.40	0.19, 0.60	0.24, 0.66
θ_2	1.06	0.46, 1.65	0.60, 1.85	0.42	0.18, 0.66	0.24, 0.75
θ_3	1.33	0.66, 1.99	0.80, 2.19	0.53	0.28, 0.78	0.33, 0.85
r_{12}	0.4246			0.4742		
r_{13}	0.5605			0.4846		
r_{23}	0.4273			0.4243		
		SA	MOVER		SA	MOVER
Measure	1.14	0.05, 2.22	-0.013, 2.505	45.1%	10.8, 79.4%	-2.3, 72.9%

Example 3. Age and BMI on hypertension (17).

		Hypertension	
Age	BMI	Yes	No
1	1	278	743
1	0	100	581
0	1	153	1,232
0	0	79	1,731

	For RERI			For AP		
	Estimate	delta CI	Convention CI	Estimate	delta CI	Convention CI
θ_1	6.24	4.75, 7.72	4.92, 7.91	0.16	0.12, 0.20	0.13, 0.20
θ_2	3.36	2.43, 4.31	2.54, 4.46	0.54	0.43, 0.65	0.44, 0.66
θ_3	2.53	1.87, 3.20	1.95, 3.29	0.41	0.33, 0.48	0.34, 0.49
r_{12}	0.6945			0.2044		
r_{13}	0.7454			0.2351		
r_{23}	0.6297			0.2702		
		SA	MOVER		SA	MOVER
Measure	1.34	0.39, 2.30	0.31, 2.37	21.5%	7.2, 35.9%	5.6, 34.6%

FIGURE 5. Calculation of 95% confidence intervals (CIs) for measures of additive interaction by use of data from examples 2 and 3. *GSTM1*, glutathione *S*-transferase M1 polymorphism; $G \times S$, product term for *GSTM1* and smoking; RERI, relative excess risks due to interaction; AP, attributable proportion due to interaction; SA, simple asymptotic; MOVER, method of variance estimates recovery; BMI, body mass index.

to the end of 1993. A cohort of 986 participants including 36 incident cases were selected from the eligible population. Excluding 118 subjects with missing *GSTM1* data, the final sample of 1,290 with the outcome variable "time to coronary heart disease diagnosis" was analyzed by Cox proportional hazards regression, taking into account the feature of the case-cohort design by using a weighting scheme (23). Specifically, the weights in the denominator of the pseudolikelihood are one for cases that arise outside the subcohort and the inverse of the sampling fraction for subcohort controls. In addition, the subcohort cases are weighted by the inverse of the sampling fraction before failure and by one at failure. Valid variances can then be estimated using the sandwich error approach (23).

After adjustment for 10 covariates, it was reported (12) that the estimated coefficients for the *GSTM1* susceptibility genotype (yes/no), ever smoking (yes/no), and their product term are given, respectively, by $\hat{\beta}_1 = 0.0543$, $\hat{\beta}_2 = 0.2826$, and $\hat{\beta}_3 = 0.5869$. Application of the MOVER approach with the use of figure 5 results in a 95 percent CI for RERI of -0.013, 2.505. On the basis of the simulation results presented above, one should doubt that "we found a statistically significant additive interaction between susceptibility genotype and ever smoking for the risk of incident CHD [coronary heart disease]" (12, p. 232), which was based on the CI of 0.052, 2.222 that was derived using the SA method. As regards to attributable fraction due to interaction, the MOVER approach also provided a 95 percent CI of -0.023, 0.729, which is very comparable to that from the $\ln(1 - \text{AP})$ transformation method: -0.025, 0.706, but very different from the one provided by the SA method: 0.108, 0.794. (Note that reference 12 contains errors in the expression for $\widehat{\text{AP}}$). Again, there is no sufficient evidence to suggest additive interaction as claimed (12, 22).

Example 3: exaggerated interaction using ORs in a cohort study

This data set arose from a cohort study in which it was of interest to investigate the effect of age and body mass index (weight (kg)/height $(\text{cm})^2$) on diastolic blood pressure (17). To form a four-by-two table, we coded age $\geq$40 years as 1 and age $<$40 years as 0, while body mass index $\geq$25 was coded as 1 and body mass index $<$25 as 0. The outcome, diastolic blood pressure $\geq$90 mmHg, was classified as hypertension and coded as 1, and $<$90 mmHg was coded as 0. The four-by-two table and associated calculation are given in figure 5. Although the RERI = 1.3 in terms of RRs, the data were analyzed by logistic regression with the percentile bootstrap, resulting in a RERI of 2.7 (95 percent CI: 1.3, 4.4) (17).

As the measures of interaction are defined in terms of RRs, it is much more appropriate to discuss additive interaction in terms of RRs when it is possible, using either regression programs (15, 16) or the formulas presented here (figure 5). With figure 5, the estimated RERI is 1.34 (95 percent CI: 0.31, 2.37), and AP is 21.5 percent (95 percent CI: 5.6, 34.6). When estimating measures of interaction in terms of ORs, the new approach would result in RERI = 2.71 (95 percent CI: 1.25, 4.45) and AP = 33.0 percent (95 percent CI: 16.1, 46.0). Although the direction of the interaction would be unchanged, the magnitude would be exaggerated if ORs were used as the effect measure (17). The intuitive explanation is that the first term in RERI is a product of three RRs, and thus a slight exaggeration of each will result in a large overestimation of RERI.

Concluding remarks

This article has proposed a simple approach to construction of confidence intervals for measures of additive interaction. This approach works because it acknowledges the fact that confidence limits for risk ratios are asymmetric. The article has also demonstrated that one can appropriately analyze the four-by-two table without having to use a statistical program. In the case of multivariable models, there is no need to recode the risk variables prior to using a regression program (8–11).

Furthermore, this article has shown that the RR should always be the first choice of effect measure for single risk factors because, as shown in the third example (17), the exaggeration of the OR can be more pronounced in assessing additive interaction. Regression models resulting in RR should be adopted if covariate adjustment is desired (15, 16, 24).

Although additive regression models are available for assessing additive interaction (25), multiplicative models are still more accessible and commonly used by epidemiologists (26), even when assessment of additive interaction is desired (14, chapter 15). This may be in part because additive models require specialized software to fit, and in part because it is straightforward to estimate measures of additive interaction using routinely available output from multiplicative software (14, chapter 15). The results in this article help to remove the obstacle of CI construction for measures of additive interaction and thus face a real challenge of gene-environment interaction, that is, conducting appropriate inferences for disease prevention (5–7).

ACKNOWLEDGMENTS

Guang Yong Zou is a recipient of the Early Researcher Award, Ontario Ministry of Research and Innovation, Canada. This work was also partially supported by an Individual Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

The author gratefully acknowledges Julia Taleban for comments on drafts of the manuscript and help on the Excel spreadsheet.

Conflict of interest: none declared.

REFERENCES

1. Rothman KJ. The estimation of synergy or antagonism. Am J Epidemiol 1976;103:506–11.
2. Botto LD, Khoury MJ. Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. Am J Epidemiol 2001;153:1016–20.
3. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York, NY: Chapman & Hall/CRC, 1993.
4. Rothman KJ. Causes. Am J Epidemiol 1976;104:587–92.
5. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. Am J Epidemiol 1980;112:467–70.
6. Darroch J. Biologic synergy and parallelism. Am J Epidemiol 1997;145:661–8.
7. Szklo M, Nieto FJ. Epidemiology: beyond the basics. 2nd ed. Sudbury, MA: Jones and Bartlett Publishers, Inc, 2006.
8. Hosmer DW, Lemeshow S. Confidence interval estimation of interaction. Epidemiology 1992;3:452–6.
9. Lundberg M, Fredlund P, Hallqvist J, et al. A SAS program calculating three measures of interaction with confidence intervals. (Letter). Epidemiology 1996;7:655–6.

10. Andersson T, Alfredsson L, Kallberg H, et al. Calculating measures of biological interaction. Eur J Epidemiol 2005; 20:575–9.
11. Kallberg H, Ahlbom A, Alfredsson L. Calculating measures of biological interaction using R. Eur J Epidemiol 2006;21: 571–3.
12. Li R, Chambless L. Test for additive interaction in proportional hazards models. Ann Epidemiol 2007;17:227–36.
13. Assmann SF, Hosmer DW, Lemeshow S, et al. Confidence intervals for measures of interaction. Epidemiology 1996;7: 286–90.
14. Rothman KJ. Modern epidemiology. Boston, MA: Little, Brown & Co, 1986.
15. Zou GY. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004;159: 702–6.
16. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. Am J Epidemiol 2005;162: 199–200.
17. Knol MJ, van der Tweel I, Grobbee DE, et al. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. Int J Epidemiol 2007;36: 1111–18.
18. Walker AM. Proportion of disease attributable to the combined effect of two factors. Int J Epidemiol 1981;10:81–5.
19. Gart JJ, Thomas DG. The performance of three approximate confidence limit methods for the odds ratio. Am J Epidemiol 1982;115:453–70.
20. Zou GY, Donner A. Construction of confidence limits about effect measures: a general approach. Stat Med 2008;27: 1693–702.
21. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study. Am J Epidemiol 1989;129: 687–702.
22. Li R, Boerwinkle E, Olshan AF, et al. Glutathione *S*-transferase genotype as a susceptibility factor in smoking-related coronary heart disease. Atherosclerosis 2000;149:451–62.
23. Barlow WE. Robust variance estimation for the case-cohort design. Biometrics 1994;50:1064–72.
24. Wacholder S. Binomial regression in GLIM—estimating risk ratios and risk differences. Am J Epidemiol 1986;123: 174–84.
25. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med 1983;2:243–51.
26. Levy PS, Stolte K. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. Stat Methods Med Res 2000;9:41–55.
27. Zou GY. Toward using confidence intervals to compare correlations. Psychol Methods 2007;12:399–413.
28. Kolmogorov AN. Introductory real analysis. Silverman RA, trans-ed. New York, NY: Dover Publications, 1975.

APPENDIX

Construction of Confidence Interval for Linear Functions of Parameters

Recall that the asymmetric confidence limits for RRs are readily available either by hand calculation or from regression programs. The strategy here is to use these limits to recover the variance estimates, without destroying the asymmetric feature of sampling distribution for RRs, in setting confidence intervals for a linear function of several RRs. The underlying principle has been discussed in the case of constructing CIs for differences between two parameters in general (20) and applied to correlations in particular (27). A summary is presented here followed by a generalized framework for a linear combination of parameters.

To begin, consider construction of a $(1 - \alpha)100$ percent two-sided confidence interval for $\theta_1 + \theta_2$, where the two estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ are for the moment assumed to be independently distributed. The lower limit may be given by

$$L = (\hat{\theta}_1 + \hat{\theta}_2) - z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\theta}_1) + \mathrm{var}(\hat{\theta}_2)}, \qquad \text{(A1)}$$

and the upper limit by

$$U = (\hat{\theta}_1 + \hat{\theta}_2) + z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\theta}_1) + \mathrm{var}(\hat{\theta}_2)}, \qquad \text{(A2)}$$

where $z_{\alpha/2}$ denotes the upper $\alpha/2$ quantile from the standard normal distribution.

Equations A1 and A2 contain unknown terms $\mathrm{var}(\hat{\theta}_i)$ ($i = 1, 2$), which may be estimated by two approaches: one assumes that $\mathrm{var}(\hat{\theta}_i)$ is independent of θ_i, while the other makes no such assumption. Confidence limits from the former are symmetric, and those from the latter are asymmetric and usually perform better (3, p. 180). The focus here is to derive asymmetric confidence intervals since the variance for the estimated RR is a function of RR itself.

By the duality between hypothesis testing and confidence interval construction, a $(1 - \alpha)100$ percent two-sided CI should contain all such parameter values that cannot be rejected by a test at the α level (3, p. 157). In other words, L is the minimum value of $\theta_1 + \theta_2$, satisfying

$$\frac{[(\hat{\theta}_1 + \hat{\theta}_2) - (\theta_1 + \theta_2)]^2}{\mathrm{var}(\hat{\theta}_1 + \hat{\theta}_2)} = z^2_{\alpha/2}.$$

To reflect the fact that variance for $\hat{\theta}_1 + \hat{\theta}_2$ in general depends on the true parameter value $\theta_1 + \theta_2$, the variance estimate for obtaining L should be estimated in the neighborhood of L, or $\min(\theta_1) + \min(\theta_2)$. Among the plausible values provided by the two pairs of confidence limits (l_1, u_1 and l_2, u_2), $\theta_1 + \theta_2 = l_1 + l_2$ is close to L. This implies that the variance can be estimated at $\theta_1 = l_1$ and $\theta_2 = l_2$.

Again, by the duality of hypothesis testing and confidence interval construction,

$$l_i = \hat{\theta}_i - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\theta}_i)}, \quad i = 1, 2,$$

which recovers the variance estimates as

$$\widehat{\mathrm{var}}(\hat{\theta}_i) = \frac{(\hat{\theta}_i - l_i)^2}{z^2_{\alpha/2}}.$$

Plugging these estimates back into equation A1 yields

$$L = (\hat{\theta}_1 + \hat{\theta}_2) - \sqrt{(\hat{\theta}_1 - l_1)^2 + (\hat{\theta}_2 - l_2)^2}. \qquad \text{(A3)}$$

Analogous arguments result in the upper limit

$$U = (\hat{\theta}_1 + \hat{\theta}_2) + \sqrt{(u_1 - \hat{\theta}_1)^2 + (u_2 - \hat{\theta}_2)^2}. \qquad \text{(A4)}$$

Equations A3 and A4 may be extended to $\theta_1 - \theta_2 = \theta_1 + (-\theta_2)$ by recognizing that the confidence limits for $-\theta_i$ are given by $(-u_i, -l_i)$. These equations may also be extended to incorporate dependency between $\hat{\theta}_1$ and $\hat{\theta}_2$. Let $r = \widehat{\text{corr}}(\hat{\theta}_1, \hat{\theta}_2)$, and then a confidence interval for $\theta_1 + \theta_2$ is given by

$$L = (\hat{\theta}_1 + \hat{\theta}_2) - \sqrt{(\hat{\theta}_1 - l_1)^2 + (\hat{\theta}_2 - l_2)^2 + 2r(\hat{\theta}_1 - l_1)(\hat{\theta}_2 - l_2)} \quad \text{(A5)}$$

and

$$U = (\hat{\theta}_1 + \hat{\theta}_2) + \sqrt{(u_1 - \hat{\theta}_1)^2 + (u_2 - \hat{\theta}_2)^2 + 2r(u_1 - \hat{\theta}_1)(u_2 - \hat{\theta}_2)}. \quad \text{(A6)}$$

By mathematical induction (28, pp. 28–29), it can be shown that a $(1 - \alpha)100$ percent confidence interval (L, U) for $\sum_i c_i\theta_i$, where c_i's are constants, is given by

$$L = \sum_i c_i\hat{\theta}_i - \sqrt{\sum_i \left[c_i\hat{\theta}_i - \min(c_i l_i, c_i u_i)\right]^2 + 2\text{cov}_L}, \quad \text{(A7)}$$

where

$$\text{cov}_L = \sum_{i<j} \text{sgn}(c_i c_j) r_{ij} \left[c_i\hat{\theta}_i - \min(c_i l_i, c_i u_i)\right] \times \left[c_j\hat{\theta}_j - \min(c_j l_j, c_j u_j)\right]$$

and

$$U = \sum_i c_i\hat{\theta}_i + \sqrt{\sum_i \left[c_i\hat{\theta}_i - \max(c_i l_i, c_i u_i)\right]^2 + 2\text{cov}_U}, \quad \text{(A8)}$$

where

$$\text{cov}_U = \sum_{i<j} \text{sgn}(c_i c_j) r_{ij} \left[c_i\hat{\theta}_i - \max(c_i l_i, c_i u_i)\right] \times \left[c_j\hat{\theta}_j - \max(c_j l_j, c_j u_j)\right]$$

and

$$\text{sgn}(c_i c_j) = \begin{cases} +1, & \text{if } c_i \times c_j > 0 \\ -1, & \text{if } c_i \times c_j < 0. \end{cases}$$

ENVIRONMETRICS
Environmetrics 2009; **20**: 172–180
Published online 24 April 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.919

Simple confidence intervals for lognormal means and their differences with environmental applications

G. Y. Zou[1,2,*,†], Cindy Yan Huo[3] and Julia Taleban[1]

[1]*Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada*
[2]*Robarts Clinical Trials, Robarts Research Institute, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada*
[3]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*

SUMMARY

The lognormal distribution has frequently been applied to approximate environmental data, with inference focusing on arithmetic means. Confidence interval estimation involving lognormal means in small to moderate sample sizes has received much attention over the years without a simple procedure in sight. We therefore propose a closed-form procedure for constructing confidence intervals for a lognormal mean and a difference between two lognormal means. The advantage of our procedure is that it only requires confidence limits for a normal mean and variance. The results of a numerical study show that our method performs as well as the generalized confidence interval (GCI) approach, which relies completely on computer simulation. Two real datasets are used to illustrate the methodology.

KEY WORDS: generalized confidence interval; log-normal; coverage; bootstrap

1. INTRODUCTION

It has become a tradition to fit the lognormal distribution to empirical data in environmental sciences (e.g., El-Shaarawi and Viveros, 1997; El-Shaarawi and Lin, 2007), due largely to the multiplicative central limit theorem (Limpert *et al*., 2001) in the sense that multiplication of a large number of random variables will result in a composite variable which can be approximated by the lognormal distribution.

A simple approach to analyzing lognormal data would be to log-transfer the data prior to employing standard statistical methods. The resultant inference would then be in terms of the median, which is less than the mean, and thus may provide substantial underestimates if the mean is the parameter of interest.

Inference in terms of lognormal means has received widespread attention in the literature, with two volumes devoted to the topic (Aitchison and Brown, 1957; Crow and Shimizu, 1988). Statistical methods for inference involving lognormal means have also appeared frequently in this journal, ranging from

*Correspondence to: G. Y. Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada N6A 5C1.
†E-mail: gzou@robarts.ca

Received 28 August 2007
Accepted 2 March 2008

computationally intensive methods such as the Gibbs sampler and bootstraps (Wild *et al*., 1996) to a *t*-distribution-based method (El-Shaarawi and Lin, 2007). It seems evident that the results for single lognormal means are not entirely satisfactory. Furthermore, there has not been much discussion on methods of comparing two lognormal means.

The purpose of this paper is to present a closed-form confidence interval procedure for a single lognormal mean and a difference between two lognormal means. We show that this closed-form procedure, requiring only confidence limits for a normal mean and variance, performs at least as well as the generalized confidence interval (GCI) approach which relies entirely on computer simulation.

The rest of the paper is structured as follows. Section 2 presents the new procedure, after summarizing the GCI (Krishnamoorthy and Mathew, 2003) and the modified Cox method (Armstrong, 1992; El-Shaarawi and Lin, 2007). In Section 3, we perform simulation studies to compare the performance of our method with previous ones. Two real datasets in an environmental context are used to illustrate the methods in Section 4. The paper closes with a discussion.

2. METHODS

Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed (iid) as lognormal with parameters μ and σ^2. This is to say that the log-transformed variables $X_1 = \ln Y_1, X_2 = \ln Y_2, \ldots, X_n = \ln Y_n$ are iid normal, denoted here as $N(\mu, \sigma^2)$. It is well known that the lognormal mean is $M = E(Y) = \exp(\mu + \sigma^2/2)$, estimated by

$$\widehat{M} = \exp\left(\bar{x} + s^2/2\right)$$

where $\bar{x}$ and s^2 are the sample mean and variance obtained using the log-transformed observations. Note that $\bar{x}$ and s^2 are independent of each other.

2.1. *Confidence interval for a single lognormal mean*

2.1.1. Existing methods. Land (1971) proposed an exact confidence interval by inverting the uniformly most powerful unbiased test. The procedure is computationally tedious and requires extensive tables. Thus, Land (1972) searched for simple approximate approaches and ended up with the one suggested by DR Cox in a personal communication showing promising performance. This method uses the property that $\bar{x}$ and s^2 are independent, with respective variances given by s^2/n and $s^4/[2(n-1)]$. Thus, as n becomes large, the $100(1-\alpha)\%$ confidence limits for $\mu + \sigma^2/2$ are given by

$$\left[\bar{x} + s^2/2\right] \pm z_{1-\alpha/2}\sqrt{s^2/n + s^4/[2(n-1)]}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. These limits can then be exponentiated to obtain a confidence interval for $\exp(\mu + \sigma^2/2)$.

As pointed by Land (1972), this method is not entirely satisfactory, particularly in the case of small n or large σ^2. To improve the performance in small samples, Armstrong (1992) and El-Shaarawi and Lin (2007) suggested replacing $z_{1-\alpha/2}$ with critical values from the *t*-distribution. This approach ignores the fact that the sampling distribution for s^2, which is distributed as chi-squared, is right-skewed.

DOI: 10.1002/env

Recently, a computer simulation-based method termed GCI appeared to perform very well. Krishnamoorthy and Mathew (2003) provide an algorithm as follows:

1. Obtain $\bar{x}$ and s^2 from log-transformed data.
2. Compute

$$T = \exp\left(\bar{x} - \frac{Z}{U/\sqrt{n-1}} \cdot \frac{s}{\sqrt{n}} + \frac{s^2}{2U^2/(n-1)}\right)$$

   where Z and U^2 are random numbers generated independently from the standard normal and chi-square distribution with $n-1$ degrees of freedom, respectively.
3. Repeat step 2 a large number (m) of times.
4. Sort the values of T. The $m(\alpha/2)^{\text{th}}$ and $m(1-\alpha/2)^{\text{th}}$ values are the $100(1-\alpha)\%$ confidence limits for $\exp(\mu + \sigma^2/2)$.

2.1.2. The proposed method. Before presenting our method for a single lognormal mean, we propose a general approach to setting confidence limits for a sum of two parameters, $\theta_1 + \theta_2$.

The conventional $100(1-\alpha)\%$ two-sided limits are

$$\widehat{\theta}_1 + \widehat{\theta}_2 - z_{1-\alpha/2}\sqrt{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)}$$

and

$$\widehat{\theta}_1 + \widehat{\theta}_2 + z_{1-\alpha/2}\sqrt{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)}$$

assuming $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent of each other. Besides the application of the central limit theorem, these limits are immediate results of assuming $\widehat{\theta}_i$ ($i = 1, 2$) and $\text{var}(\widehat{\theta}_i)$ are statistically independent of each other. Except for a normal mean $\bar{x}$, this is unlikely to hold in general.

Our idea is to exploit the dependence between $\widehat{\theta}_i$ and $\text{var}(\widehat{\theta}_i)$ in confidence interval construction. Specifically, we strive to estimate the variance of $\widehat{\theta}_1 + \widehat{\theta}_2$ in the vicinity of the limits (L, U) for $\theta_1 + \theta_2$.

By the duality between hypothesis testing and confidence interval construction, we recognize L as the minimum and U as the maximum value of $\theta_1 + \theta_2$ such that

$$\frac{\left[\widehat{\theta}_1 + \widehat{\theta}_2 - (\theta_1 + \theta_2)\right]^2}{\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)} \approx z^2_{1-\alpha/2} \tag{1}$$

Thus, we should estimate the variances for $\widehat{\theta}_1$ and $\widehat{\theta}_2$ in the vicinity of $\min(\theta_1 + \theta_2)$ for L and that of $\max(\theta_1 + \theta_2)$ for U.

Now suppose the confidence limits for θ_i are readily obtained as (l_i, u_i), for $i = 1, 2$. Among the plausible values provided by the two pairs of confidence limits (l_1, u_1) and (l_2, u_2), the plausible minimum is $l_1 + l_2$ and the plausible maximum is $u_1 + u_2$. This implies that to obtain L, we need to estimate $\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)$ under the condition $\theta_1 = l_1$ and $\theta_2 = l_2$. Similarly, to obtain U, we need to estimate $\text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2)$ under the condition $\theta_1 = u_1$ and $\theta_2 = u_2$.

DOI: 10.1002/env

Again by the duality between hypothesis testing and confidence interval construction, l_i is $\min(\theta_i)$ satisfying

$$\frac{\left(\widehat{\theta}_i - l_i\right)^2}{\mathrm{var}(\widehat{\theta}_i)} \approx z^2_{1-\alpha/2}$$

which results in the estimated variance $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ under the condition $\theta_i = l_i$ of

$$\widehat{\mathrm{var}}_l(\widehat{\theta}_i) \approx \frac{\left(\widehat{\theta}_i - l_i\right)^2}{z^2_{1-\alpha/2}}$$

Similarly, the estimated variance $\widehat{\mathrm{var}}(\widehat{\theta}_i)$ under the condition $\theta_i = u_i$ is

$$\widehat{\mathrm{var}}_u(\widehat{\theta}_i) \approx \frac{\left(u_i - \widehat{\theta}_i\right)^2}{z^2_{1-\alpha/2}}$$

Substituting these variance estimates back into Equation (1) yields the confidence limits (L, U) for $\theta_1 + \theta_2$ as

$$\begin{cases} L = \widehat{\theta}_1 + \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (\widehat{\theta}_2 - l_2)^2} \\ U = \widehat{\theta}_1 + \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (u_2 - \widehat{\theta}_2)^2} \end{cases} \tag{2}$$

These limits can now be applied in the current context, where $\theta_1 = \mu$ and $\theta_2 = \sigma^2/2$, with respective confidence intervals given by $(l_1, u_1) = (\bar{x} - z_{1-\alpha/2}\sqrt{s^2/n}, \bar{x} + z_{1-\alpha/2}\sqrt{s^2/n})$ and $(l_2, u_2) = \left[\frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{2\chi^2_{\alpha/2,n-1}}\right]$. Exponentiation of these limits yields a confidence interval for the lognormal mean. Specifically, the limits (*LL*, *UL*) are given by

$$LL = \widehat{M} \exp\left[-\left(\frac{z^2_{1-\alpha/2}s^2}{n} + \left(\frac{s^2}{2} - \frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}}\right)^2\right)^{1/2}\right] \tag{3}$$

and

$$UL = \widehat{M} \exp\left[\left(\frac{z^2_{1-\alpha/2}s^2}{n} + \left(\frac{(n-1)s^2}{2\chi^2_{\alpha/2,n-1}} - \frac{s^2}{2}\right)^2\right)^{1/2}\right] \tag{4}$$

DOI: 10.1002/env

The Cox method can be obtained the same way by replacing the confidence interval for $\sigma^2/2$ with

$$(l_2, u_2) = \left(s^2 \left[\frac{1}{2} - z_{1-\alpha/2} \sqrt{\frac{1}{2(n-1)}} \right], s^2 \left[\frac{1}{2} + z_{1-\alpha/2} \sqrt{\frac{1}{2(n-1)}} \right] \right)$$

This is equivalent to treating the confidence interval of σ^2 as symmetric, indicating that for $n < 8$ a 95% confidence interval contains negative variance values. Replacing $z_{1-\alpha/2}$ with the t-value will not reduce the problem since the $t_{1-\alpha/2,n-1}$ is larger than that of a Normal distribution.

2.2. *Confidence intervals for a difference between two lognormal means*

Denoting a difference between two lognormal means as

$$\Delta = \exp(\mu_1 + \sigma_1^2/2) - \exp(\mu_2 + \sigma_2^2/2)$$

the correspondent estimator is

$$\widehat{\Delta} = \exp(\bar{x}_1 + s_1^2/2) - \exp(\bar{x}_2 + s_2^2/2)$$

with $(\bar{x}_1, s_1^2)$ and $(\bar{x}_2, s_2^2)$ computed from the log-transformed observations from two independent samples.

2.2.1. Generalized confidence interval approach. Krishnamoorthy and Mathew (2003) proposed the following algorithm for obtaining a $100(1-\alpha)\%$ confidence interval for Δ:

1. Compute $(\bar{x}_1, s_1^2)$ and $(\bar{x}_2, s_2^2)$.
2. Compute

$$T_\Delta = \exp\left(\bar{x}_1 - \frac{Z_1}{U_1/\sqrt{n_1 - 1}} \cdot \frac{s_1}{\sqrt{n_1}} + \frac{s_1^2}{2U_1^2/(n_1)} \right) - \exp\left(\bar{x}_2 - \frac{Z_2}{U_2/\sqrt{n_2 - 1}} \cdot \frac{s_2}{\sqrt{n_2}} + \frac{s_2^2}{2U_2^2/(n_2)} \right)$$

   where Z_i and U_i^2 are random numbers generated independently from the standard normal and chi-squared distribution with $n_i - 1$ degrees of freedom from two independent samples ($i = 1, 2$);
3. Repeat step 2 a large number of, say m, times.
4. Sort the T_Δ values from step 3. The confidence limits are given by the $m(\alpha/2)^{\text{th}}$ and $m(1-\alpha/2)^{th}$ T_Δ values.

2.2.2. The proposed method. Our alternative is first to obtain confidence limits for $M_1 = \exp(\mu_1 + \sigma_1^2/2)$ and $M_2 = \exp(\mu_2 + \sigma_2^2/2)$ using Equations (3) and (4), then to treat M_1 as θ_1 and $-M_2$ as θ_2 in the application of Equation (2). Note here that the limits for M_2, obtained using Equations (3) and (4), must be multiplied by -1 and then switched positions before plugging into Equation (2).

DOI: 10.1002/env

Straightforward algebra yields the $100(1-\alpha)\%$ confidence interval (L_Δ, U_Δ) for the difference between two lognormal means as

$$L_\Delta = \widehat{M}_1 - \widehat{M}_2 - \sqrt{(M_1 - LL_1)^2 + (UL_2 - M_2)^2}$$

and

$$U_\Delta = \widehat{M}_1 - \widehat{M}_2 + \sqrt{(UL_1 - M_1)^2 + (M_2 - LL_2)^2}$$

where

$$LL_i = \widehat{M_i} \exp\left[-\left(\frac{z_{1-\alpha/2}^2 s_i^2}{n_i} + \left(\frac{s_i^2}{2} - \frac{(n_i-1)s_i^2}{2\chi_{1-\alpha/2,n_i-1}^2}\right)^2\right)^{1/2}\right]$$

and

$$UL_i = \widehat{M_i} \exp\left[\left(\frac{z_{1-\alpha/2}^2 s_i^2}{n_i} + \left(\frac{(n_i-1)s_i^2}{2\chi_{\alpha/2,n_i-1}^2} - \frac{s_i^2}{2}\right)^2\right)^{1/2}\right]$$

for $i = 1, 2$.

3. SIMULATION

The confidence interval procedures described above are all asymptotic, meaning that their performance such as average percentage and tail errors may depend on sample size and parameter values. Before making any recommendations, we must evaluate their performance in finite sample sizes. For this purpose, we use Monto Carlo simulations to compare the procedures for the 95% confidence interval in terms of the percentage of times the interval contains the parameter value (coverage%). For a given parameter value, we assess the performance of a procedure using the percentage of times the confidence interval lies completely below or above the parameter value, termed left and right tail errors, respectively. We used 10 000 replicates for each parameter combination, with 10 000 resamples for the GCI approach. Using two standard errors of the nominal coverage rate as the criterion, we regarded coverage as within $(.95 \pm 2\sqrt{0.95 \times 0.05/10\,000})$, or (94.6–95.4) as adequate.

The second criterion is the balance between left and right tail errors (Jennings, 1987; Efron, 2003). We used confidence width as the third criterion to distinguish procedures satisfying the first and second criteria equally. Without loss of generality (Land, 1972, p. 147), we set $\mu = -\sigma^2/2$ in the simulation study.

For a single lognormal mean, we considered $n = 10$, 15, 25, and 50; $\sigma^2 = 0.1$, 0.5, 1.0, 1.5, and 2.0. The performance of the modified Cox method, our proposed method, and the generalized confidence interval are shown in Table 1. These results indicate that all three methods have acceptable coverage percentages. As expected, the modified Cox method has unbalanced tail errors, while the other two methods deliver reasonably balanced tail errors, with the proposed method showing consistently narrower average width.

DOI: 10.1002/env

Table 1. Comparative performance of three procedures for constructing a 95% two-sided confidence interval for a lognormal mean with $\mu = -\sigma^2/2$ based on 10 000 runs

		$n = 10$		$n = 15$		$n = 25$		$n = 50$	
σ^2	Method	Cover (ML , MR)	W	Cover (ML , MR)	W	Cover (ML , MR)	W	Cover (ML , MR)	W
0.1	MCox	95.23 (3.31, 1.46)	0.46	95.33 (3.03, 1.64)	0.36	94.90 (3.15, 1.95)	0.27	95.08 (2.88, 2.04)	0.18
	Proposed	93.27 (3.87, 2.86)	0.44	93.85 (3.53, 2.62)	0.34	94.13 (3.32, 2.55)	0.26	94.55 (2.96, 2.49)	0.18
	GCI	95.10 (2.20, 2.70)	0.50	95.00 (2.27, 2.73)	0.37	94.88 (2.38, 2.74)	0.27	94.92 (2.33, 2.75)	0.19
0.5	MCox	94.88 (4.48, 0.64)	1.29	94.93 (4.29, 0.78)	0.94	94.84 (3.87, 1.29)	0.68	94.44 (3.97, 1.59)	0.46
	Proposed	94.50 (3.24, 2.26)	1.67	94.54 (3.30, 2.16)	1.05	94.79 (2.97, 2.24)	0.71	94.54 (3.20, 2.26)	0.47
	GCI	94.84 (1.94, 3.22)	1.90	94.76 (2.33, 2.91)	1.14	95.03 (1.95, 3.02)	0.75	94.57 (2.62, 2.81)	0.48
1.0	MCox	93.99 (5.82, 0.19)	2.53	94.44 (5.17, 0.39)	1.67	94.69 (4.70, 0.61)	1.14	94.89 (3.89, 1.22)	0.73
	Proposed	94.68 (3.39, 1.93)	5.36	94.89 (3.12, 1.99)	2.28	95.02 (3.18, 1.80)	1.30	94.87 (2.80, 2.33)	0.77
	GCI	94.49 (2.40, 3.11)	6.12	94.42 (2.23, 3.35)	2.49	94.86 (2.42, 2.72)	1.37	94.77 (2.29, 2.94)	0.79
1.5	MCox	93.76 (6.19, 0.05)	4.84	94.24 (5.58, 0.18)	2.60	94.07 (5.40, 0.53)	1.63	95.00 (4.05, 0.95)	1.01
	Proposed	95.37 (2.94, 1.69)	24.34	95.28 (2.89, 1.83)	4.48	94.74 (3.08, 2.18)	2.06	94.99 (2.64, 2.37)	1.11
	GCI	94.89 (2.06, 3.05)	27.52	95.18 (2.04, 2.78)	4.91	94.53 (2.34, 3.13)	2.17	95.04 (2.11, 2.85)	1.14
2.0	MCox	93.32 (6.65, 0.03)	10.63	93.71 (6.16, 0.13)	4.14	94.58 (5.02, 0.40)	2.27	94.72 (4.50, 0.78)	1.31
	Proposed	95.15 (2.98, 1.87)	497.08	94.83 (3.09, 2.08)	9.84	95.24 (2.73, 2.03)	3.19	94.82 (2.70, 2.48)	1.49
	GCI	94.71 (2.15, 3.14)	899.26	94.38 (2.41, 3.21)	10.80	95.07 (2.14, 2.79)	3.38	94.64 (2.41, 2.95)	1.53

MCox, the modified Cox method; GCI, generalized confidence interval; ML, the confidence interval lies completely below the parameter; MR, the confidence interval lies completely above the parameter; W, average interval width.

For a difference between two lognormal means, we considered $n_1 = 10, 15, 20, 25$, and 50; $n_2 = 10, 20, 25$, and 50; $\sigma_1^2 = 0.1, 0.5, 1.0, 1.5, 2.0$; $\sigma_2^2 = 0.5, 1.5$, and 2.0. The performance of the generalized confidence interval method and the proposed method with modified Cox method for single means for these 300 parameter combinations are presented using summary statistics (Table 2). These results clearly show that the Modified Cox method provides severely unbalanced tails with coverage percentage ranging from 93.17 to 98.11%. Our proposed method is very competitive with the computer simulation-based GCI, both having coverage rates outside the range of 94.6 to 95.4% when $n \leq 15$.

Table 2. Comparative performance of three procedures for constructing a 95% two-sided confidence interval for a difference between two lognormal means with $\mu_i = -\sigma_i^2/2$, $i = 1, 2$ (summary of 300 parameter combinations with 10 000 runs for each combination)

Method		Mean	Min	10th pctl	25th pctl	50th pctl	75th pctl	90th pctl	Max
MCox	Cover	95.57	93.17	94.55	95.09	95.63	96.13	96.52	98.11
	ML	1.86	0.04	0.25	0.66	1.50	2.85	3.98	6.19
	MR	2.57	0.07	0.49	1.09	2.42	3.86	4.90	6.76
	Width	2.36	0.49	1.05	1.46	2.11	2.96	3.90	8.25
Proposed	Cover	95.32	94.26	94.92	95.13	95.32	95.52	95.75	96.32
	ML	2.28	1.74	1.97	2.12	2.27	2.42	2.59	3.17
	MR	2.40	1.69	2.06	2.19	2.36	2.59	2.84	3.42
	Width	4.11	0.49	1.13	1.73	2.92	5.32	9.54	29.67
GCI	Cover	95.25	94.29	94.86	95.03	95.23	95.48	95.71	96.18
	ML	2.40	1.76	2.04	2.18	2.35	2.59	2.83	3.40
	MR	2.34	1.82	2.06	2.16	2.32	2.51	2.68	3.25
	Width	4.47	0.51	1.18	1.83	3.07	5.91	10.81	33.10

MCox, the modified Cox method; GCI, generalized confidence interval; ML, the confidence interval lies completely below the parameter; MR, the confidence interval lies completely above the parameter.

DOI: 10.1002/env

4. ILLUSTRATIVE EXAMPLES

As an example of a simple lognormal mean, we consider air lead levels (μ g/m^3) of $n = 15$ sites at the Alma American Labs, Fairplay, Colorado on 23 February 1989 (Krishnamoorthy *et al*., 2006): 200, 120, 15, 7, 8, 6, 48, 61, 380, 80, 29, 1000, 350, 1400, 110. The lognormal distribution was found to fit the data well. Log-transformation of the data yields $\bar{x} = 4.333$ and $s = 1.739$. Therefore, we have the 95% confidence limits for $\theta_1 = \mu$ given by $[4.333 - 1.96 \times 1.739/\sqrt{15}, 4.333 + 1.96 \times 1.739/\sqrt{15}]$, i.e., (3.452584, 5.213141) and that for $\theta_2 = \sigma^2/2$ given by:

$$\frac{1}{2}\left[\frac{(15-1)\times 1.739^2}{\chi^2_{0.975,14}}, \frac{(15-1)\times 1.739^2}{\chi^2_{0.025,14}}\right]$$

that is, (0.8108892, 3.762765). Substituting these limits into Equations (3) and (4) yields the 95% two-sided confidence interval for $\exp(\mu + \sigma^2/2)$ as (112, 3873), comparable with the GCI of (122, 4280) based on 100 000 simulations.

As an example for a difference between two lognormal means. We consider a dataset from the Data and story Library (http://lib.stat.cmu.edu/DASL). In April–May 1993, an oil refinery near San Francisco submitted $n = 31$ daily CO emission measurements from its stacks to the Bay Area Air Quality Management District for establishing a baseline. It was of interest to see whether the refinery had over-measured CO emission, as compared to nine measurements taken by the Management District person between September 1990 to March 1993. The data are given as:

Refinery ($n_1 = 31$): 45, 30, 38, 42, 63, 43, 102, 86, 99, 63, 58, 34, 37, 55, 58, 153, 75, 58, 36, 59, 43, 102, 52, 30, 21, 40, 141, 85, 161, 86, 71.

District management ($n_2 = 9$): 12.5, 20, 4, 20, 25, 170, 15, 20, 15.

Recognizing the temporal dependence among the measurements, we nevertheless treat them as independent for illustration purposes. The lognormal distribution fits both dataset well (Krishnamoorthy and Mathew, 2003), with $\bar{x}_1 = 4.074252$, $s_1^2 = 0.252081$, $\bar{x}_2 = 2.963333$, and $s_2^2 = 0.949618$. Using our approach, the estimated mean and 95% confidence interval of the refinery data are given by 66.70583 (55.57714, 81.69155) and that of the district Management data are given by 31.12906 (15.66019, 128.6178). Application of our procedure yields the difference and 95% confidence interval of 35.58 (−62.55, 57.11). Again, comparable with those from the GCI (−79.15, 57.47) based on 100 000 simulations.

5. DISCUSSION

We have presented a simple approach to confidence interval estimation concerning lognormal means. The resultant procedures for a single lognormal mean and a difference between two lognormal means are in closed-form, requiring only methods found in introductory textbooks. The performance of our procedure has been shown to do at least as well as the GCI approach, which relies on computer simulation. Moreover, although exact in theory, even with the same dataset the latter approach may result in different answers from different analysts or the same analyst performing analyses at different times.

We note that the method we described here can be readily applied to lognormal regression models (Bradu and Mundlak, 1970; El-Shaarawi and Viveros, 1997; El-Shaarawi and Lin, 2007). Exten-

DOI: 10.1002/env

sions and applications of this method in other contexts can be found elsewhere (Zou, 2007; Zou and Donner, 2008).

We did not consider bootstrap methods for lognormal data, as it has been revealed that such methods fail even for a normal variance (Schenker, 1985). It is then inevitable for bootstrap to fail for the lognormal mean because it is a function of the normal mean and variance. We refer to Zhou and Dinh (2005) for simulation results showing that bootstrap methods fail terribly in the case of lognormal data. Interestingly, many papers have appeared by merely implementing a bootstrap method, as if it is the gold standard. This practice is a result of overlooking the fact that bootstrap is also asymptotically reliable and requires evaluation on a case-by-case basis (DiCiccio and Efron, 1996).

ACKNOWLEDGEMENTS

Dr Zou is a recipient of an Early Researcher Award from Ontario Ministry of Research and Innovation, Canada. His research is also supported partially by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Aitchison J, Brown JAC. 1957. *The Lognormal Distribution*. Cambridge University Press: Cambrige.

Armstrong BG. 1992. Confidence intervals for arithmetic means of lognormally distributed exposures. *American Industrial Hygiene Association Journal* **53**: 481–485.

Bradu D, Mundlak T. 1970. Estimation in lognormal linear models. *Journal of the American Statistical Association* **65**: 198–211.

Crow EL, Shimizu K. 1988. *Lognormal Distributions: Theory and Applications*. Dekker: New York.

DiCiccio TJ, Efron B. 1996. Bootstrap confidence intervals. *Statistical Science* **11**: 189–228.

Efron B. 2003. Second thoughts on the bootstrap. *Statistical Science* **18**: 135–140.

El-Shaarawi AH, Lin J. 2007. Interval estimation for log-normal mean with applications to water quality. *Environmetrics* **18**: 1–10.

El-Shaarawi AH, Viveros R. 1997. Inference about the mean in log-regression with environmental applications. *Environmetrics* **8**: 569–582.

Jennings DE. 1987. How do we judge confidence-interval adequacy? *The American Statistician* **41**: 335–337.

Krishnamoorthy K, Mathew T, Ramachandran G. 2006. Generalized P-values and confidence intervals: a novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene* **3**: 642–650.

Krishnamoorthy K, Mathew TP. 2003. Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *Journal of Statistical Planning and Inference* **115**: 103–121.

Land CE. 1971. Confidence intervals for linear functions of the normal mean and variance. *Annals of Mathematical Statistics* **42**: 1187–1205.

Land CE. 1972. An evaluation of approximate confidence interval estimation methods for lognormal means. *Technometrics* **14**: 145–158.

Limpert E, Stahel WA, Abbt M. 2001. Log-normal distributions across the sciences: keys and clues. *BioScience* **51**: 341–352.

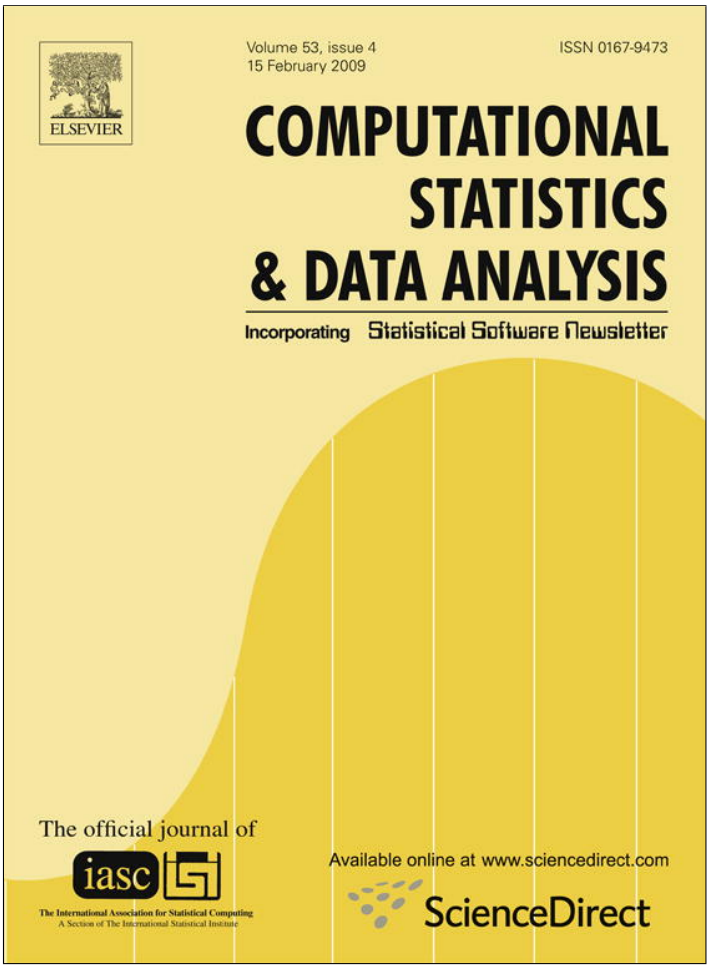
Schenker N. 1985. Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association* **80**: 360–361.

Wild P, Hordan R, Leplay A, Vincent R. 1996. Confidence intervals for probabilities of exceeding threshold limits with censored log-normal data. *Environmetrics* **7**: 247–259.

Zhou XH, Dinh P. 2005. Nonparametric confidence intervals for the one- and two-sample problems. *Biostatistics* **6**: 187–200.

Zou GY. 2007. Toward using confidence intervals to compare correlations. *Psychological Methods* **12**: 399–413.

Zou GY, Donner A. 2008. Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* **27**: http://dx.doi.org/10.1002/sim.3095



Computational Statistics and Data Analysis 53 (2009) 1080–1085

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A note on confidence interval estimation for a linear function of binomial proportions

Guang Yong Zou [a,b,*], Wenyi Huang [a], Xiaohe Zhang [c]

[a] Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada N6A 5C1

[b] Robarts Clinical Trials, Robarts Research Institute, University of Western Ontario, London, Ontario, Canada N6A 5K8

[c] Department of Medicine, McMaster University, Population Health Research Institute, Hamilton, Ontario, Canada L8L 2X2

ARTICLE INFO

Article history:
Received 23 May 2008
Received in revised form 24 September 2008
Accepted 25 September 2008
Available online 11 October 2008

ABSTRACT

The Wilson score confidence interval for a binomial proportion has been widely applied in practice, due largely to its good performance in finite samples and its simplicity in calculation. We propose its use in setting confidence limits for a linear function of binomial proportions using the method of variance estimates recovery. Exact evaluation results show that this approach provides intervals that are narrower than the ones based on the adjusted Wald interval while aligning the mean coverage with the nominal level.

1. Introduction

There exists a large literature on confidence interval estimation involving binomial proportions. For a single proportion, there are several choices. The first is given by adding to and subtracting from the maximum likelihood estimator the standard normal quantile multiplied by its estimated standard error. This procedure is commonly referred to as the Wald method. The second is the interval based on inverting the approximate normal test that uses the standard errors estimated at the lower and upper limits. This procedure is commonly referred to as Wilson score method (Wilson, 1927). The score confidence interval has now become very popular, especially after the expositions by Agresti and Coull (1998) and Newcombe (1998b). With an attempt to ease classroom teaching, Agresti and Coull (1998) suggested an adjusted Wald method by adding two successes and two failures and then using the Wald formula. Despite the terminology, the adjusted Wald method is actually an approximation of the score method.

The superior performance of the Wilson method has been carried over to cases of a difference between two proportions (Newcombe, 1998a) and a difference between two differences (Newcombe, 2001). It is interesting to note that this seemingly *ad hoc* procedure has become more popular than the rigorous score interval for a difference between two proportions (Mee, 1984; Miettinen and Nurminen, 1985; Gart and Nam, 1990), caused largely by the computation involved in obtaining the latter.

Due to its important practical value, confidence interval construction for a linear function of binomial proportions has received some attention recently (Price and Bonett, 2004; Tebbs and Roths, 2008). The purpose of this note is to extend the argument of Zou and Donner (2008) to a linear function of parameters, and in particular to binomial proportions. Since our main idea is to recover variance estimates from readily available confidence limits for single parameters, we refer to the approach as the MOVER, the method of variance estimates recovery. As shown below, the MOVER will not only shed some light to Newcombe (1998a) and Newcombe (2001) but also provide an alternative to Price and Bonett (2004) who proposed

* Corresponding address: Robarts Clinical Trials, Robarts Research Institute, P. O. Box 5015, 100 Perth Drive, London, Ontario, Canada N6A 5K8. Tel.: +1 519 663 3400x34092; fax: +1 519 663 3807.
E-mail address: gzou@robarts.ca (G.Y. Zou).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.033

a procedure for a linear function of proportions based on the adjusted Wald interval for single proportions (Agresti and Coull, 1998). We will show that the confidence interval for a linear function of binomial proportions based on the Wilson method is narrower than that of Price and Bonett (2004). We will not consider the approach by Tebbs and Roths (2008) because of its inherent drawbacks such as involved computation, restriction in parameter ranges, and undercoverage.

2. The MOVER and its application to linear functions of binomial proportions

Suppose we wish to construct an approximate $100(1-\alpha)\%$ two-sided confidence interval for $\theta_1+\theta_2$, where the estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are assumed to be independent. By the central limit theorem, the lower limit (L) is given by

$$L = \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2)}. \tag{1}$$

Inspired by the score method for interval estimation (Bartlett, 1953; Gart and Nam, 1990), we can estimate the variance needed for L at $\theta_1+\theta_2 = L$. This has at least one disadvantage that it is in general an iterative procedure, which can be an obstacle to wide application in practice as what happened to the score interval for a difference between two proportions. Therefore, we proceed with estimating the variance in the neighborhood of L.

Now, suppose that the $100(1-\alpha)\%$ two-sided confidence intervals (l_i, u_i) for single parameters θ_i, $i = 1, 2$ are available. Note that there is no need to specify the approaches taken to obtain (l_i, u_i). Among all the plausible parameter values of θ_1 provided by (l_1, u_1) and that of θ_2 provided by (l_2, u_2), $l_1 + l_2$ is usually closer to L than $\widehat{\theta}_1 + \widehat{\theta}_2$. As a result, for L, we can estimate $\mathrm{var}(\widehat{\theta}_1)$ at $\theta_1 = l_1$ and $\mathrm{var}(\widehat{\theta}_2)$ at $\theta_2 = l_2$.

Furthermore, we can recover the required variance estimates from $\hat{\theta}_i(l_i, u_i)$, $i = 1, 2$, as follows. By the central limit theorem and letting $z_{\alpha/2}$ be the upper $\alpha/2$ quantile of the standard Normal distribution, we have

$$l_i = \widehat{\theta}_i - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_i)},$$

which gives a variance estimate for $\widehat{\theta}_i$ at $\theta_i = l_i$ as

$$\widehat{\mathrm{var}}_l(\widehat{\theta}_i) = (\widehat{\theta}_i - l_i)^2/z^2_{\alpha/2}$$

and

$$u_i = \widehat{\theta}_i + z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_i)},$$

which gives a variance estimate at $\theta_i = u_i$ as

$$\widehat{\mathrm{var}}_u(\widehat{\theta}_i) = (u_1 - \widehat{\theta}_i)^2/z^2_{\alpha/2}.$$

Note that the recovered variance estimates $\widehat{\mathrm{var}}_l(\widehat{\theta}_i)$ and $\widehat{\mathrm{var}}_u(\widehat{\theta}_i)$ are different, except when the interval (l_i, u_i) is symmetric about $\widehat{\theta}_i$. Symmetric intervals are known to perform poorly in finite samples for most problems in practice. In fact, it was stated (Efron and Tibshirani, 1993, p. 180) that symmetry is the most serious error in confidence interval construction. The Wald interval for a binomial proportion is a perfect example. In contrast, the Wilson interval is asymmetric as a consequence of estimating variances at the lower and upper limits separately.

Plugging the recovered variance estimates into Eq. (1) results in

$$\begin{aligned} L &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2)} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{(\widehat{\theta}_1 - l_1)^2/z^2_{\alpha/2} + (\widehat{\theta}_2 - l_2)^2/z^2_{\alpha/2}} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (\widehat{\theta}_2 - l_2)^2}. \end{aligned}$$

Analogous steps with the notion that $u_1 + u_2$ is in the vicinity of U yield the upper limit U as

$$U = \widehat{\theta}_1 + \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (u_2 - \widehat{\theta}_2)^2}.$$

Rewriting $\theta_1-\theta_2$ as $\theta_1+(-\theta_2)$ and noting that the confidence limits for $-\theta_2$ are given by $(-u_2, -l_2)$, we obtain confidence limits for $\theta_1 - \theta_2$ as

$$L = \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2}$$

and

$$U = \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2}.$$

This confidence interval, apparently first presented by Howe (1974), has been applied by Newcombe (1998a) and by Newcombe (2001) to binomial proportions. There has been no analytic justification for its general applicability until recently (Zou and Donner, 2008).

Regarding $\theta_1+\theta_2$ and $\theta_1-\theta_2$ as $c_1\theta_1+c_2\theta_2$, where c_1 and c_2 are constants, we can rewrite the intervals as

$$L=c_1\widehat{\theta}_1+c_2\widehat{\theta}_2-\sqrt{[c_1\widehat{\theta}_1-\min(c_1l_1,c_1u_1)]^2+[c_2\widehat{\theta}_2-\min(c_2l_2,c_2u_2)]^2}$$

and

$$U=c_1\widehat{\theta}_1+c_2\widehat{\theta}_2+\sqrt{[c_1\widehat{\theta}_1-\max(c_1l_1,c_1u_1)]^2+[c_2\widehat{\theta}_2-\max(c_2l_2,c_2u_2)]^2}.$$

For a $100(1-\alpha)\%$ confidence interval for $\sum_{i=1}^{g}c_i\theta_i$, where $g>2$, an application of mathematical induction results in

$$\begin{cases} L=\sum_{i=1}^{g}c_i\widehat{\theta}_i-\sqrt{\sum_{i=1}^{g}\left[c_i\widehat{\theta}_i-\min(c_il_i,c_iu_i)\right]^2} \\ U=\sum_{i=1}^{g}c_i\widehat{\theta}_i+\sqrt{\sum_{i=1}^{g}\left[c_i\widehat{\theta}_i-\max(c_il_i,c_iu_i)\right]^2}. \end{cases} \tag{2}$$

Because L and U are derived using the recovered variance estimates, we can refer to the method as the MOVER, standing for method of variance estimates recovery. A further extension of the MOVER to incorporate dependence between $\widehat{\theta}_i$ and $\widehat{\theta}_j$ $(i\neq j)$ has been applied to measures of additive interaction in epidemiology (Zou, 2008).

We can now apply the confidence interval in (2) to linear functions of binomial proportions. Since there are at least three intervals for a single proportion, i.e., Wald, adjusted Wald (Agresti and Coull, 1998) and Wilson, we end up with three procedures for linear functions of binomial proportions.

Specifically, let Y_i $(i=1,2,\ldots,g)$ be independent binomial variates with parameters (n_i,p_i), and let $\widehat{p}_i=Y_i/n_i$ be the sample estimates for p_i. A linear function of binomial proportions may be defined as $\sum_{i=1}^{g}c_ip_i$, where the c_i are known constants. Using the equations in (2), the $100(1-\alpha)\%$ Wald confidence interval can be obtained by setting $\widehat{\theta}_i=\widehat{p}_i=Y_i/n_i$, $l_i=\widehat{p}_i-z_{\alpha/2}\sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$, and $u_i=\widehat{p}_i+z_{\alpha/2}\sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$.

The Wilson interval for $\sum_{i=1}^{g}c_ip_i$ may be obtained by setting $\widehat{\theta}_i=\widehat{p}_i=Y_i/n_i$,

$$l_i,u_i=\left(\widehat{p}_i+z^2_{\alpha/2}/(2n_i)\mp z_{\alpha/2}\sqrt{[\widehat{p}_i(1-\widehat{p}_i)+z^2_{\alpha/2}/(4n_i)]/n_i}\right)/(1+z^2_{\alpha/2}/n_i).$$

The adjusted Wald interval for $\sum_{i=1}^{g}c_ip_i$ (Price and Bonett, 2004) may be obtained by setting $\widehat{\theta}_i=\tilde{p}_i=(Y_i+2/k)/(n_i+4/k)$ (where k is the number of nonzero elements in c_i), $l_i=\tilde{p}_i-z_{\alpha/2}\sqrt{\tilde{p}_i(1-\tilde{p}_i)/n_i}$, and $u_i=\tilde{p}_i+z_{\alpha/2}\sqrt{\tilde{p}_i(1-\tilde{p}_i)/n_i}$. Note that the adjusted Wald method for a single proportion is an approximation of the Wilson score method for 95% interval, see Agresti and Coull (1998) for its motivation and derivation. We also must point out that this method has the potential to provide confidence limits that are out of parameter space.

It is fair to say that the superior performance of Newcombe (1998a) originates from that of the Wilson method for a single proportion (Agresti and Coull, 1998; Newcombe, 1998b). On the same token, we can postulate that applying Wilson interval for cases of more than two binomial proportions will be very competitive to that of Price and Bonett (2004).

To evaluate this claim, we conducted a numerical study to compare the performance of these two procedures in finite samples for 90%, 95%, and 99% two-sided confidence intervals, in terms of mean coverage, minimum coverage, and mean interval width as defined here.

For a $100(1-\alpha)\%$ interval (L,U) for $\sum_{i=1}^{g}c_ip_i$, the coverage is defined by

$$\text{Coverage}=100\sum_{y_1=0}^{n_1}\cdots\sum_{y_g=0}^{n_g}\prod_{i=1}^{g}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}I\left(L<\sum c_ip_i<U\right),$$

where $I(.)$ is an indicator function which takes values of 1 or 0 as the event in the brackets is true or not.

The expected interval width is defined as

$$\text{Width}=\sum_{y_1=0}^{n_1}\cdots\sum_{y_g=0}^{n_g}\prod_{i=1}^{g}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}(U-L).$$

We conducted the evaluation by first randomly sampling 1000 sets of p_i's from the uniform (0,1) distribution, and then applied the above two definitions to each set. We did not arbitrarily truncate the adjusted Wald confidence limits when they fell out of the parameter space. With respect to each method, we obtained the mean coverage, minimum coverage, and the mean interval width using these 1000 sets of values for coverage and width.

Table 1
Performance of the method of variance estimates recovery in constructing two-sided confidence intervals (CI) for a linear function of binomial parameters, $\sum_{i=1}^{3} c_i p_i$, using confidence limits for single proportions obtained by the adjusted Wald and Wilson methods. Entries in each row are based on 1000 sets of p_i's randomly sampled from uniform (0,1), and each set evaluated by exact calculation.

Group sizes	90% CI		95% CI	
$n_1/n_2/n_3$	Adjusted Wald	Wilson	Adjusted Wald	Wilson
$c = (1/3, 1/3, 1/3)$				
5/5/5	92.04 (80.18, 0.32)*	90.58 (82.60, 0.31)	96.12 (91.24, 0.38)	94.99 (86.34, 0.36)
5/5/10	91.65 (83.59, 0.29)	90.45 (85.36, 0.29)	95.95 (88.12, 0.35)	95.07 (89.55, 0.33)
5/10/15	91.53 (84.60, 0.25)	90.69 (87.42, 0.25)	95.86 (92.93, 0.30)	95.26 (90.65, 0.29)
5/10/20	91.63 (86.89, 0.25)	90.80 (87.37, 0.24)	95.88 (92.51, 0.29)	95.31 (91.02, 0.28)
5/15/20	91.57 (84.34, 0.23)	90.80 (84.39, 0.23)	95.83 (90.95, 0.28)	95.30 (90.51, 0.27)
5/20/20	91.53 (83.99, 0.23)	90.72 (87.15, 0.22)	95.73 (84.97, 0.27)	95.27 (91.50, 0.26)
$c = (1, -1/2, -1/2)$				
5/5/5	92.29 (80.22, 0.67)	90.82 (85.45, 0.65)	96.19 (86.68, 0.79)	95.14 (89.23, 0.75)
5/5/10	91.97 (84.26, 0.64)	90.84 (84.63, 0.62)	95.91 (89.26, 0.77)	95.27 (89.48, 0.72)
5/10/15	92.00 (82.28, 0.60)	91.06 (86.44, 0.58)	95.87 (86.43, 0.72)	95.31 (91.04, 0.67)
5/10/20	92.00 (82.74, 0.60)	91.06 (85.77, 0.57)	95.83 (87.82, 0.71)	95.33 (91.62, 0.66)
5/15/20	92.00 (82.28, 0.59)	91.06 (86.34, 0.56)	95.75 (88.50, 0.70)	95.27 (91.63, 0.65)
5/20/20	92.13 (81.69, 0.58)	91.13 (85.68, 0.55)	95.80 (86.75, 0.69)	95.27 (91.15, 0.64)
$c = (-1, 1/2, 2)$				
5/5/5	92.08 (79.06, 1.25)	90.94 (86.88, 1.20)	95.91 (86.89, 1.49)	95.19 (89.05, 1.39)
5/5/10	91.52 (85.49, 1.00)	90.67 (86.58, 0.98)	95.81 (91.28, 1.19)	95.24 (89.05, 1.15)
5/10/15	91.33 (85.15, 0.88)	90.64 (87.14, 0.87)	95.66 (91.82, 1.05)	95.29 (88.53, 1.02)
5/10/20	91.35 (85.32, 0.82)	90.72 (87.51, 0.81)	95.71 (92.89, 0.98)	95.32 (90.85, 0.95)
5/15/20	91.29 (82.82, 0.81)	90.65 (87.98, 0.80)	95.67 (92.39, 0.97)	95.23 (91.42, 0.94)
5/20/20	91.29 (85.58, 0.81)	90.65 (86.88, 0.80)	95.66 (90.11, 0.96)	95.30 (90.18, 0.93)
$c = (1, 1, -1)$				
5/5/5	92.04 (80.36, 0.95)	90.56 (81.21, 0.93)	96.19 (92.21, 1.13)	95.15 (86.22, 1.08)
5/5/10	91.74 (85.12, 0.88)	90.64 (85.70, 0.86)	95.96 (89.78, 1.04)	95.18 (89.86, 1.00)
5/10/15	91.49 (84.88, 0.76)	90.71 (87.33, 0.74)	95.78 (87.97, 0.90)	95.26 (90.45, 0.87)
5/10/20	91.49 (85.56, 0.74)	90.76 (86.80, 0.72)	95.80 (92.41, 0.88)	95.29 (90.54, 0.84)
5/15/20	91.42 (85.05, 0.70)	90.69 (86.68, 0.69)	95.70 (90.10, 0.84)	95.20 (91.24, 0.80)
5/20/20	91.59 (85.28, 0.68)	90.82 (87.43, 0.66)	95.82 (88.32, 0.81)	95.26 (90.77, 0.78)

∗ Mean coverage % (minimum coverage %, mean confidence interval width) based on 1000 sets of proportion parameters randomly sampled from uniform (0,1) distribution.

For linear functions of 3 binomial proportions, results in Table 1 show consistently that the intervals for linear functions based on the Wilson score method have mean coverage closer to the nominal levels, with narrow average width. For group sizes considered, the minimum coverage for the adjusted Wald can be as low as 79.06% for 90% nominal level, and 84.97% for 95% nominal level. For confidence interval based on the Wilson method, the minimum coverage can be as low as 81.21% for 90% nominal level, and 86.22% for 95% nominal level. Results from constructing confidence intervals for linear functions of 4 binomial proportions in Table 2 show again that the procedure based on Wilson score method performed better in terms of mean coverage and interval width, as well as minimum coverage. For example, the minimum coverage for the adjusted Wald can be as low as 77.16% for 90% nominal level, compared to that of 83.23% for Wilson score method. Similar trends were observed with nominal level of 99% (results not shown). One possible explanation for our results is that the adjusted Wald method was proposed to approximate the Wilson score method at 95% level, on the rationale that the middle point of Wilson interval is a weighted average of $\widehat{p}$ and 0.5, and that $1.96^2 \approx 4$ (Agresti and Coull, 1998, p. 122).

3. Examples

In the light of the above numerical results, we now compare confidence intervals using two examples from Price and Bonett (2004).

Example 1. This data set arose from a study in which rats are fed with different types of diets. The diets are controlled by two factors, namely fiber and fat. Each rat is observed to determine if it has developed a tumor during the study period. The outcome of the experiment is summarized in Table 3 (each group had 30 rats). It is of interest to construct confidence intervals for the main effects of fiber and fat, as well as their interaction. Here we can obtain the 95% confidence intervals using the MOVER for the linear functions of proportions. The results are shown in Table 3, which shows that the intervals obtained using the Wilson method for single proportions are narrower than those using the adjusted Wald method for single proportion. This is consistent with the results in our evaluation study. In fact, the Wilson method based intervals are all contained in that based on the adjusted Wald method for single proportions in this moderate size study.

Example 2. This example arose from the Framingham heart study. As an alternative to conventional generalized linear model with logistic link function, Price and Bonett (2004) approached the problem with a linear function of binomial

Table 2
Performance of the method of variance estimates recovery in constructing two-sided confidence intervals (CI) for a linear function of binomial parameters, $\sum_{i=1}^{4} c_i p_i$, using confidence limits for single proportions obtained by the adjusted Wald and Wilson methods. Entries in each row are based on 1000 sets of p_i's randomly sampled from uniform (0,1), and each set evaluated by exact calculation.

Group sizes	90% CI		95% CI	
$n_1/n_2/n_3/n_4$	Adjusted Wald	Wilson	Adjusted Wald	Wilson
$c = (1/4, 1/4, 1/4, 1/4)$				
5/5/5/5	91.27 (81.65, 0.28)*	90.24 (83.23, 0.27)	95.63 (92.16, 0.33)	94.86 (88.56, 0.31)
5/5/10/10	91.03 (87.49, 0.24)	90.33 (85.65, 0.24)	95.48 (93.03, 0.29)	95.07 (90.18, 0.28)
5/5/15/15	91.01 (87.61, 0.23)	90.53 (86.71, 0.22)	95.45 (92.91, 0.27)	95.13 (91.33, 0.26)
5/5/15/20	91.11 (88.33, 0.22)	90.67 (87.18, 0.22)	95.50 (93.10, 0.26)	95.25 (91.75, 0.26)
5/10/15/20	90.84 (86.26, 0.20)	90.57 (87.66, 0.20)	95.33 (90.52, 0.24)	95.27 (91.64, 0.23)
$c = (-1, 1, -1, 1)$				
5/5/5/5	91.33 (85.72, 1.10)	90.29 (82.48, 1.08)	95.70 (92.91, 1.31)	95.09 (88.82, 1.25)
5/5/10/10	91.03 (83.19, 0.96)	90.37 (85.96, 0.95)	95.49 (92.68, 1.15)	95.09 (90.56, 1.11)
5/5/15/15	91.11 (87.96, 0.91)	90.67 (86.77, 0.89)	95.50 (92.51, 1.08)	95.24 (91.47, 1.04)
5/5/15/20	91.08 (87.41, 0.89)	90.71 (87.05, 0.88)	95.50 (92.45, 1.06)	95.26 (91.74, 1.02)
5/10/15/20	90.86 (87.58, 0.81)	90.51 (87.68, 0.80)	95.37 (91.86, 0.97)	95.17 (91.40, 0.94)
$c = (1/3, 1/3, 1/3, 1)$				
5/5/5/5	91.33 (77.16, 0.63)	90.87 (86.18, 0.61)	95.29 (84.48, 0.75)	95.15 (89.89, 0.70)
5/5/10/10	90.82 (86.51, 0.49)	90.52 (87.43, 0.49)	95.23 (90.73, 0.59)	95.16 (91.24, 0.57)
5/5/15/15	90.64 (87.56, 0.44)	90.29 (87.33, 0.43)	95.22 (93.25, 0.52)	95.10 (91.60, 0.51)
5/5/15/20	90.74 (88.44, 0.40)	90.25 (86.91, 0.40)	95.34 (93.10, 0.48)	95.07 (90.91, 0.47)
5/10/15/20	90.48 (88.76, 0.39)	90.29 (87.89, 0.38)	95.12 (93.65, 0.46)	95.15 (92.17, 0.45)
$c = (-3, -1, 1, 3)$				
5/5/5/5	91.34 (83.03, 2.44)	90.89 (83.44, 2.38)	95.49 (90.28, 2.90)	95.20 (89.81, 2.76)
5/5/10/10	90.94 (83.95, 2.14)	90.80 (87.55, 2.10)	95.23 (88.85, 2.55)	95.32 (91.50, 2.44)
5/5/15/15	91.01 (82.16, 2.01)	90.80 (87.95, 1.96)	95.24 (86.35, 2.40)	95.22 (90.66, 2.29)
5/5/15/20	90.92 (83.77, 1.96)	90.77 (88.24, 1.90)	95.13 (86.91, 2.33)	95.19 (91.50, 2.22)
5/10/15/20	91.08 (83.32, 1.91)	91.02 (88.56, 1.86)	95.15 (85.88, 2.27)	95.36 (91.06, 2.16)

* Mean coverage % (minimum coverage %, mean confidence interval width) based on 1000 sets of proportion parameters randomly sampled from uniform (0,1) distribution.

Table 3
Confidence intervals for effects of factors in the diet–tumor study.

Fiber	Fat	$\widehat{p}_i$	c_i		
			Fiber × Fat	Fiber	Fat
Yes	High	20/30	1	1/2	1/2
	Low	14/30	−1	1/2	−1/2
No	High	27/30	1	−1/2	1/2
	Low	19/30	−1	−1/2	−1/2
Interval for $\sum c_i p_i$:					
		Adj Wald	−0.3806, 0.2516	−0.3516, 0.0355	0.0677, 0.3839
		Wilson	−0.3790, 0.2386	−0.3459, 0.0375	0.0691, 0.3773

Table 4
Framingham heart study.

Systolic BP	Number of subjects	Number with heart disease
115	156	3
121	252	17
131	284	12
141	271	16
151	139	12
161	85	8
176	99	16
190	43	8

proportions. Specifically, if the population proportion of heart disease is considered a linear function of systolic blood pressure, the slope is $\sum c_i p_i$, which is a linear function of the proportions p_i of heart disease of systolic blood pressure groups, where $c_i = (x_i - \sum x_i/g)/\sum(x_i - \sum x_i/g)^2$ and x_i is the value of the quantitative factor in group i. Using the data in Table 4, we obtained the 95% confidence interval for the population slope using the adjusted Wald method as 0.0010 to 0.0032, comparable to that of using the Wilson method as 0.0012 to 0.0034 in such a large study.

4. Concluding remarks

The confidence interval for a general linear function of binomial proportions introduced here is a simple application of a more general idea presented by Zou and Donner (2008). The basic idea is to recover variance estimates needed for linear functions of proportions from the confidence limits for single proportions. Since the Wilson interval procedure has been strongly recommended for single proportions (Agresti and Coull, 1998; Newcombe, 1998b; Santner, 1998), it is thus natural to extend it to linear functions of binomial proportions. By use of the MOVER, we have provided a very competitive procedure to that of Price and Bonett (2004), whose procedure can be seen as an application of the MOVER based on the adjusted Wald method for single proportions. The MOVER has also provided an analytic justification for Newcombe (1998a, 2001).

It should also be noted that the derivation of the MOVER relies only on the validity of confidence limits for single parameters such that variance estimates can be recovered by normal distributions. The direct implication is that one can apply the MOVER to linear functions of other discrete distribution parameters, e.g., Poisson rates (Stamey and Hamilton, 2006; Tebbs and Roths, 2008), and linear functions of normal mean and variance, e.g., lognormal means (Zou and Donner, 2008).

Acknowledgments

The authors gratefully acknowledge the comments from two anonymous reviewers which led to the insight that the adjusted Wald method is actually an approximation of Wilson score confidence interval for a single binomial proportion. Guang Yong Zou is a recipient of the Early Researcher Award, Ontario Ministry of Research and Innovation, Canada. His work was also partially supported by an Individual Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Agresti, A., Coull, B., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. American Statistician 52, 119–126.
Bartlett, M.S., 1953. Approximate confidence intervals. 2. More than one unknown parameter. Biometrika 40, 306–317.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, New York.
Gart, J.J., Nam, J.M., 1990. Approximate interval estimation of the difference in binomial parameters - correction for skewness and extension to multiple tables. Biometrics 46, 637–643.
Howe, W.G., 1974. Approximate confidence limits on the mean of $X + Y$ where X and Y are two tabled independent random variables. Journal of the American Statistical Association 69, 789–794.
Mee, R.W., 1984. Confidence bounds for the difference between two probabilities. Biometrics 40, 1175–1176.
Miettinen, O., Nurminen, M., 1985. Comparative analysis of two rates. Statistics in Medicine 4, 213–226.
Newcombe, R.G., 1998a. Interval estimation for the difference between independent proportions: Comparison of eleven methods. Statistics in Medicine 17, 873–890.
Newcombe, R.G., 1998b. Two-sided confidence intervals for the single proportions: Comparison of seven methods. Statistics in Medicine 17, 857–872.
Newcombe, R.G., 2001. Estimating the difference between differences: Measurement of additive scale interaction for proportions. Statistics in Medicine 20, 2885–2893.
Price, R.M., Bonett, D.G., 2004. An improved confidence interval for a linear function of binomial proportions. Computational Statistics & Data Analysis 45, 449–456.
Santner, T.J., 1998. Teaching large-sample binomial confidence intervals. Teaching Statistics 20, 20–23.
Stamey, J., Hamilton, C., 2006. A note on confidence intervals for a linear function of Poisson rates. Communications in Statistics–Simulation and Computation 35, 849–856.
Tebbs, J.M., Roths, S.A., 2008. New large-sample confidence intervals for a linear combination of binomial proportions. Journal of Statistical Planning and Inference 138, 1884–1893.
Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22, 209–212.
Zou, G.Y., 2008. On the estimation of additive interaction using the four-by-two table and beyond. American Journal of Epidemiology 168, 212–224.
Zou, G.Y., Donner, A., 2008. Construction of confidence limits about effect measures: A general approach. Statistics in Medicine 27, 1693–1702.

Computational Statistics and Data Analysis 53 (2009) 3755–3764

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Confidence interval estimation for lognormal data with application to health economics

Guang Yong Zou [a,b,*], Julia Taleban [a], Cindy Y. Huo [c]

[a] Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada N6A 5C1

[b] Robarts Clinical Trials, Robarts Research Institute, University of Western Ontario, London, Ontario, Canada N6A 5K8

[c] Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada M4N 3M5

ARTICLE INFO

Article history:
Received 27 September 2008
Received in revised form 27 March 2009
Accepted 27 March 2009
Available online 2 April 2009

ABSTRACT

There has accumulated a large amount of literature on confidence interval construction involving lognormal data owing to the fact that many data in scientific inquiries may be approximated by this distribution. Procedures have usually been developed in a piecemeal fashion for a single mean, a single mean with excessive zeros, a difference between two means, and a difference between two differences (net health benefit). As an alternative, we present a general approach for all these cases that requires only confidence limits available in introductory texts. Simulation results confirm the validity of this approach. Examples arising from health economics are used to exemplify the methodology.

1. Introduction

The lognormal distribution may be used to approximate right skewed data arising in a wide range of scientific inquires (Limpert et al., 2001). Traditional statistical analysis of such data has usually been focused on the means of log-transformed data, resulting in inferences expressed in terms of geometric means rather than the arithmetic means. However, there are many situations, including in environmental science (Parkhurst, 1998) and in occupational health research (Rappaport and Selvin, 1987), in which arithmetic means may provide more meaningful information. Consequently, there has accumulated a relatively large amount of literature regarding statistical methods for this type of data, including Aitchison and Brown (1957) and Crow and Shimizu (1988), with more articles being added rapidly to the literature (Chen, 1994; Taylor et al., 2002; Wu et al., 2002, 2003, 2006; Gill, 2004; Tian and Wu, 2006; Shen et al., 2006; Krishnamoorthy et al., 2006; Bebu and Mathew, 2008; Fletcher, 2008).

Since many health cost data may be positively skewed (Thompson and Barber, 2000; Briggs et al., 2002), the literature dealing with the analysis of lognormal data in this context has also increased substantially. This includes procedures for a one sample mean, a difference between two independent sample means, a difference between two dependent sample means, and additional zero values for each of these cases (Zhou, 2002). Recent advances include a method based on the Edgeworth expansion (Zhou and Dinh, 2005; Dinh and Zhou, 2006). It is worthwhile to note that this approach not only fails to provide adequate coverage rates but also lacks invariance in the sense that a confidence interval for $-\theta$ differs from $(-u, -l)$ when confidence interval for θ is given by (l, u). As a consequence, one may reach different conclusions depending on the labeling of groups in a comparative study. One could naturally suggest the bootstrap, but simulation results (Diciccio and Efron, 1996; Zhou and Dinh, 2005; Dinh and Zhou, 2006; Zou and Donner, 2008) suggest that it can fail in the case of

* Corresponding address: Robarts Clinical Trials, Robarts Research Institute, P. O. Box 5015, 100 Perth Drive, London, Ontario, Canada N6A 5K8. Tel.: +1 519 663 3400x34092; fax: +1 519 663 3807.
E-mail address: gzou@robarts.ca (G.Y. Zou).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.03.016

lognormal data. A possible explanation is that the lognormal mean is a function of a normal variance and some bootstrap intervals have been shown to fail in confidence interval construction for a normal variance (Schenker, 1985).

Recently, a procedure relying on the simulation of pivotal statistics, commonly referred to as a generalized confidence interval, has generated a series of articles on lognormal data (see, e.g., Krishnamoorthy and Mathew, 2003; Tian, 2005; Chen and Zhou, 2006; Krishnamoorthy et al., 2006; Tian and Wu, 2007a,b; Bebu and Mathew, 2008).

Instead of adopting a simulation approach to each of the situations summarized above (Zhou, 2002; Chen and Zhou, 2006), we extend a simple confidence interval procedure proposed by Zou and Donner (2008) to each of these scenarios. One advantage of our procedure is that it relies only on techniques readily available in introductory texts. We also discuss confidence interval estimation for net health benefit (NHB), an alternative to incremental cost-effectiveness ratio (Stinnett and Mullahy, 1998; Willan, 2001). By assuming a value for the willingness-to-pay for a unit of effectiveness, a positive NHB indicates the treatment is cost-effective. Detailed principles for cost-effectiveness analysis in health care can be found in textbooks (e.g. Drummond et al., 2005; Willan and Briggs, 2006).

The rest of the article is structured as follows. In Section 2 we present a general approach applicable to confidence interval estimation, which will be referred to as the MOVER, standing for the method of variance estimates recovery. We then apply the MOVER to obtain confidence intervals for a single lognormal mean, a single lognormal mean with excessive zeros, a difference between two lognormal means, and the net health benefit in Section 3. In Section 4, we compare the performance of our approach to some existing methods, particularly to generalized confidence intervals, using simulation studies. We provide examples using data from previously published studies in Section 5. The article concludes with some final remarks in Section 6.

2. Confidence interval estimation by the method of variance estimates recovery

The complication in constructing a confidence interval for the lognormal mean appears to have been due to the fact that it involves two parameters, as reflected by the remark that 'obtaining the confidence interval for the lognormal estimator is a non-trivial problem since it is a function of two transformed sample estimates' (Briggs et al., 2005, p. 422). However, the confidence limits for each individual parameter (the normal mean and variance) are simple to obtain. Our strategy is to 'recover' variance estimates from these limits and then to form approximate confidence intervals for functions of the parameters, using similar arguments to those of Zou and Donner (2008) and Zou et al. (2009a).

Suppose we wish to construct a $100(1-\alpha)\%$ two-sided confidence interval (L, U) for $\theta_1+\theta_2$, where the estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent. Using the central limit theorem, a lower limit (L) is given by

$$L = \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_1) + \mathrm{var}(\widehat{\theta}_2)},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. The limit L is not readily applicable because $\mathrm{var}(\widehat{\theta}_i)$ $(i = 1, 2)$ is unknown.

Now, suppose that a $100(1-\alpha)\%$ two-sided confidence interval for θ_i is given by (l_i, u_i). Among all the plausible parameter values of θ_1 provided by (l_1, u_1) and that of θ_2 by (l_2, u_2), we know L is in the neighborhood of $l_1 + l_2$. Inspired by the score interval approach (Bartlett, 1953), we proceed to estimate the variances needed for L at $\theta_1 + \theta_2 = l_1 + l_2$, i.e., when $\theta_1 = l_1$ and $\theta_2 = l_2$.

We have, by the central limit theorem,

$$l_i = \widehat{\theta}_i - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_i)},$$

which gives a variance estimate for $\widehat{\theta}_i$ at $\theta_i = l_i$ of

$$\widehat{\mathrm{var}}(\widehat{\theta}_i) = (\widehat{\theta}_i - l_i)^2/z^2_{\alpha/2}.$$

Therefore, the lower limit L for $\theta_1 + \theta_2$ is given by

$$\begin{aligned} L &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1) + \widehat{\mathrm{var}}(\widehat{\theta}_2)} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - z_{\alpha/2}\sqrt{(\widehat{\theta}_1 - l_1)^2/z^2_{\alpha/2} + (\widehat{\theta}_2 - l_2)^2/z^2_{\alpha/2}} \\ &= \widehat{\theta}_1 + \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (\widehat{\theta}_2 - l_2)^2}. \end{aligned} \tag{1}$$

Analogous steps with the notion that $u_1 + u_2$ is close to U, and the variance estimate at $\theta_i = u_i$ is

$$\widehat{\mathrm{var}}(\widehat{\theta}_i) = (u_1 - \widehat{\theta}_i)^2/z^2_{\alpha/2},$$

we obtain an upper limit U as

$$U = \widehat{\theta}_1 + \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (u_2 - \widehat{\theta}_2)^2}. \tag{2}$$

Rewriting $\theta_1-\theta_2$ as $\theta_1+(-\theta_2)$ and noting that the confidence limits for $-\theta_2$ are given by $(-u_2, -l_2)$, we obtain confidence limits for $\theta_1-\theta_2$ as (Zou and Donner, 2008)

$$\begin{cases} L=\widehat{\theta}_1-\widehat{\theta}_2-\sqrt{(\widehat{\theta}_1-l_1)^2+(u_2-\widehat{\theta}_2)^2} \\ U=\widehat{\theta}_1-\widehat{\theta}_2+\sqrt{(u_1-\widehat{\theta}_1)^2+(\widehat{\theta}_2-l_2)^2}, \end{cases} \tag{3}$$

where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are assumed to be independent, and (l_i, u_i), $i=1,2$, are the $100(1-\alpha)\%$ confidence limits for θ_1 and θ_2, respectively. We note that this procedure satisfies the invariance property in the sense that the confidence interval for $\theta_2-\theta_1$ is given by $[-U, -L]$, in contrast to those based on the Edgeworth expansion (Zhou and Dinh, 2005; Dinh and Zhou, 2006; Zhou and Qin, 2007).

We can extend the above results to cases where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are dependent. Let r be the estimated correlation coefficient between $\widehat{\theta}_1$ and $\widehat{\theta}_2$, then the limits in Eq. (3) may be directly extended by including covariance terms, $r(\widehat{\theta}_1-l_1)(u_2-\widehat{\theta}_2)$ and $r(u_1-\widehat{\theta}_1)(\widehat{\theta}_2-l_2)$, in the expressions given for L and U as follows

$$\begin{cases} L=\widehat{\theta}_1-\widehat{\theta}_2-\sqrt{(\widehat{\theta}_1-l_1)^2+(u_2-\widehat{\theta}_2)^2-2r(\widehat{\theta}_1-l_1)(u_2-\widehat{\theta}_2)} \\ U=\widehat{\theta}_1-\widehat{\theta}_2+\sqrt{(u_1-\widehat{\theta}_1)^2+(\widehat{\theta}_2-l_2)^2-2r(u_1-\widehat{\theta}_1)(\widehat{\theta}_2-l_2)}. \end{cases} \tag{4}$$

Note that if the sampling distribution for $\widehat{\theta}_i$ is symmetric, then $\widehat{\theta}_i-l_i=u_i-\widehat{\theta}_i$, which results in a symmetric confidence interval.

Since the essence of the procedure discussed above relies on recovering a variance from confidence limits, it may be regarded as the method of variance estimates recovery (MOVER). Note that the validity of the MOVER relies on the validity of the confidence limits for each of the two parameters.

Note that Eq. (3) is similar to modified large sample confidence intervals for variance components (Howe, 1974; Graybill and Wang, 1980; Burick and Graybill, 1992; Lee et al., 2004). Previous authors tend to justify their procedures by assuming the limits are of certain form and solving the limits by forcing the confidence coefficients to be exact under special conditions (see, e.g, Graybill and Wang, 1980; Lee et al., 2004). 'Square-and-add' is another term used for Eq. (3) when applied to proportions (Newcombe, 2001, p. 2889). We prefer to use the term MOVER because it reflects clearly that the key step of the method is to recover variance estimates.

3. Applying the MOVER to lognormal data

We now apply the MOVER to lognormal data. We also present some existing methods along the way.

3.1. One sample lognormal mean

Let y_j denote observations such that $x_j=\ln y_j \sim N(\mu,\sigma^2)$, $j=1,2,\ldots,n$, and denote the distribution of y_j by $\Lambda(\mu,\sigma^2)$. From the moment generating function for the normal distribution, the arithmetic mean of y_j is given by $\exp(\eta)$ with $\eta=\mu+\sigma^2/2$, which may be estimated by

$$\widehat{\eta}=\bar{x}+s^2/2,$$

where $\bar{x}$ and s^2 are the sample mean and variance respectively. Confidence interval estimation for $\exp(\eta)$ is equivalent to that for η. A confidence interval for η, commonly known as the Cox method in Land (1972), is then given by

$$\bar{x}+s^2/2\pm z_{\alpha/2}\sqrt{s^2/n+s^4/\{2(n-1)\}}. \tag{5}$$

Previous evaluation (Land, 1972) has confirmed that this method performs reasonably well over a wide range of parameter values. Mohn (1979) has also shown that it competes well with the likelihood ratio based approach in terms of coverage. This conclusion is consistent with those of follow-up evaluations (Zhou and Gao, 1997).

An alternative approach based on simulation is the so-called generalized confidence interval, discussed by Krishnamoorthy and Mathew (2003). This method involves simulating the pivotal statistic

$$T=\bar{x}-\frac{Z}{U/\sqrt{n-1}}\frac{s}{\sqrt{n}}+\frac{s^2}{2U^2/(n-1)},$$

where $Z\sim N(0,1)$ and $U^2\sim\chi^2_{n-1}$. The resulting $100(1-\alpha)\%$ limits are given by the $\alpha/2$ and $1-\alpha/2$ percentiles of T, yielding the generalized confidence interval (GCI).

Here we can apply the MOVER to set confidence limits for η by setting $\theta_1 = \mu$ and $\theta_2 = s^2/2$ in Eqs. (1) and (2). The resultant confidence limits are given by

$$\begin{cases} l = \bar{x} + \dfrac{s^2}{2} - \sqrt{z_{\alpha/2}^2 \dfrac{s^2}{n} + \left\{ \dfrac{s^2}{2} \left(1 - \dfrac{n-1}{\chi^2_{1-\alpha/2,n-1}} \right) \right\}^2} \\ u = \bar{x} + \dfrac{s^2}{2} + \sqrt{z_{\alpha/2}^2 \dfrac{s^2}{n} + \left\{ \dfrac{s^2}{2} \left(\dfrac{n-1}{\chi^2_{\alpha/2,n-1}} - 1 \right) \right\}^2}. \end{cases} \tag{6}$$

Underlying these limits is the well-known result that the $100(1-\alpha)\%$ confidence interval for σ^2 is given by $[(n-1)s^2/\chi^2_{1-\alpha/2,n-1}, (n-1)s^2/\chi^2_{\alpha/2,n-1}]$, where $\chi^2_{\alpha,df}$ is the $\alpha\%$ percentile from the chi-square distribution with df degrees of freedom.

3.2. One sample Δ-distribution mean

In practice, a subgroup of patients may incur zero cost, with the remainder of the cost data assumed to be approximated by the lognormal distribution, as in Diehr et al. (1999). The resulting distribution, commonly referred to as the Δ-distribution, is denoted by $\Delta(\delta, \mu, \sigma^2)$ and has mean $M = (1-\delta)\exp(\mu + \sigma^2/2)$, where δ is the probability that a patient has zero value (Aitchison and Brown, 1957; Owen and DeRouen, 1980; Pennington, 1983; Smith, 1988).

Tian (2005) has shown that the GCI procedure performed better than procedures based on likelihood theory by Zhou and Tu (2000). An iterative procedure based on an adjusted likelihood function has also recently appeared in Tian and Wu (2006).

Alternatively, by recognizing that $\ln M = \ln(1-\delta) + \mu + \sigma^2/2$ and defining $\theta_1 = \ln(1-\delta)$ and $\theta_2 = (\mu + \sigma^2/2)$, one can obtain $100(1-\alpha)\%$ confidence limits for M using the MOVER in Eqs. (1) and (2). An accurate confidence interval for δ is readily available (Wilson, 1927; Agresti and Coull, 1998) and is given by

$$\left[\widehat{\delta} + z_{\alpha/2}^2/(2n) \pm z_{\alpha/2} \sqrt{\left\{ \widehat{\delta}(1-\widehat{\delta}) + z_{\alpha/2}^2/(4n) \right\} /n} \right] \Big/ \left(1 + z_{\alpha/2}^2/n\right), \tag{7}$$

where n_0 is the number of zeros and $\widehat{\delta} = n_0/n$. The confidence interval for θ_2 can be obtained using Eq. (6).

3.3. Difference between two lognormal means

To find the confidence interval for a difference between two lognormal means, an approach based on the Edgeworth expansion has been proposed by Zhou and Dinh (2005). Denote y_{ij} as the observation arising from subject j in group i ($i = 1, 2$ and $j = 1, 2, \ldots, n_i$), with sample mean and variance given by $\bar{y}_i = \sum_j y_{ij}/n_i$ and $s_i^2 = \sum_j (y_{ij} - \bar{y}_i)^2/(n_i - 1)$, respectively. The estimated variance for $\bar{y}_1 - \bar{y}_2$ is given by $s^2 = s_1^2/n_1 + s_2^2/n_2$. The simple asymptotic interval for the group difference is then given by $\bar{y}_1 - \bar{y}_2 \pm z_{\alpha/2}s$. To take into account the skewness estimated by $\widehat{\gamma}_i = n_i \sum_j (y_{ij} - \bar{y}_i)^3 / \left[s_i^3 (n_i - 1)(n_i - 2) \right]$, a nonparametric confidence interval procedure for skewed data based on the Edgeworth expansion was suggested by Zhou and Dinh (2005). This confidence interval is given by

$$\begin{cases} L = \bar{y}_1 - \bar{y}_2 - \sqrt{N} G^{-1}(z_{1-\alpha/2}/\sqrt{N}) s \\ U = \bar{y}_1 - \bar{y}_2 - \sqrt{N} G^{-1}(z_{\alpha/2}/\sqrt{N}) s, \end{cases}$$

where $N = n_1 + n_2$, z_x is the upper x percentile from the standard normal distribution,

$$G^{-1}(x) = \left\{ 1 + 3\left(x - \frac{A}{6N} \right) \right\}^{1/3} - 1,$$

and

$$A = \frac{\sqrt{N} \left[s_1^3 \gamma_1 / n_1^2 - s_2^3 \gamma_2 / n_2^2 \right]}{s^3},$$

which is a simplified version of the expression given by Zhou and Dinh (2005).

To see whether this procedure possesses the invariance property, we can apply it to an example from Zhou and Dinh (2005) where it was of interest to compare the cost of diagnosis of depressed patients to that of non-depressed patients. The summary statistics for the non-depressed group are $n_1 = 108$, $\bar{y}_1 = 1646.53$, $s_1 = 4103.84$, $\gamma_1 = 5.41$, and for the depressed group are $n_2 = 103$, $\bar{y}_2 = 1344.58$, $s_2 = 1785.54$, $\gamma_2 = 2.55$. By this method, the 95% confidence interval for the difference between population means (non-depressed group minus depressed group) is $(-492, 1338)$, while the interval for the opposite difference (depressed group minus non-depressed group) is given by $(-1077, 657)$. This abnormality was first

noticed in the case of binary data (Lecoutre and Faure, 2007), which prompted a response that appears to have confused the properties of invariance and symmetry (Zhou and Qin, 2007). Since a lot has been published (see, e.g., Zhou, 2002; Zhou and Dinh, 2005; Dinh and Zhou, 2006) based on the idea of removing skewness by transformation (Hall, 1992), it would be of high practical value to effectively fix this abnormality.

Alternatively, we may apply the MOVER to differences between lognormal means (Zou et al., 2009b). Let $\exp(\eta_i)$ denote the arithmetic mean of population i with a random sample of observations yielding the confidence limits (l_i, u_i) $(i = 1, 2)$, which may be obtained using Eq. (6). It is then simple to obtain the confidence limits for $\exp(\eta_1) - \exp(\eta_2)$ using the MOVER in Eq. (3).

When two groups are not independent, with data from at least one group assumed to be lognormally distributed, the correlation between observations in the two groups must be taken into account. From previous results presented by Shimizu (1983), one can show that the correlation on the original scale is given by

$$r = \frac{\exp(\rho\sigma_1\sigma_2) - 1}{\sqrt{\left[\exp(\sigma_1^2) - 1\right]\left[\exp(\sigma_2^2) - 1\right]}}, \tag{8}$$

for bivariate lognormal data $\Lambda_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. That is, $(\ln y_1, \ln y_2)$ has a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Moreover, for the semi-lognormal distribution, with $(\ln y_1, y_2)$ following bivariate normal with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, the correlation is given by

$$r = \frac{\rho\sigma_1}{\sqrt{\exp(\sigma_1^2) - 1}}. \tag{9}$$

The sample estimate $\widehat{r}$ may be obtained by substitution of estimates $\widehat{\rho}$ and s_i^2 for ρ and σ_i^2, respectively. Note that expression (8) has previously been obtained by Mostafa and Mahmoud (1964). Special cases of both Eqs. (8) and (9) can also be found in Kowalski (1972).

We note that the procedure also provides a simple solution to interval estimation of a difference in mean cost when data contain zero values, discussed in Chen and Zhou (2006). This can be done by using the method in Section 3.2 for each group, then applying Eq. (3) to obtain a confidence interval for the difference. We also note that since a difference on the log-scale is equivalent to a ratio on the original scale, the procedure discussed here provides a simple alternative to an iterative procedure based on likelihood ratio (Wu et al., 2002, 2006).

3.4. Net health benefit: A difference between two differences

Cost-effectiveness analysis has traditionally focused on the ratio of the cost difference to the effectiveness difference, commonly known as incremental cost-effectiveness ratio (ICER). The ICER is the additional cost incurred when the test intervention delivers one unit of additional health benefit relative to the standard intervention. Interpretational and statistical difficulties associated with ICER have resulted in the net health benefit (*NHB*) being used as an alternative summary measure of the value for money of health-care programs (Stinnett and Mullahy, 1998; Willan, 2001; Hoch et al., 2002).

Let C_{ij} and E_{ij} denote the cost and effectiveness, respectively, for subject j who receives treatment i, respectively, for $i = 1, 2$ and $j = 1, 2, \ldots, n_i$. Assuming the (C_{ij}, E_{ij}) are independent and identically distributed bivariate random variables with means μ_{C_i}, μ_{E_i}, variances $\sigma_{C_i}^2, \sigma_{E_i}^2$ and covariance $\sigma_{C_iE_i}$, then the *NHB* is defined as $NHB = (\mu_{E_1} - \mu_{C_1}/\lambda) - (\mu_{E_2} - \mu_{C_2}/\lambda)$, where λ is the amount society is willing to pay for a unit of effectiveness (Stinnett and Mullahy, 1998). The sample estimate for *NHB* may then be obtained for a given λ by substituting estimates for μ_{E_i} and μ_{C_i}. Summary statistics may be calculated as $\bar{E}_i = \sum E_{ij}/n_i$, $s_{E1}^2 = \sum(E_{ij} - \bar{E}_i)^2/n_i$, $\bar{C}_i = \sum C_{ij}/n_i$, $s_{C1}^2 = \sum(C_{ij} - \bar{C}_i)^2/n_i$, and $s_{CiEi} = \sum(C_{ij} - \bar{C}_i)(E_{ij} - \bar{E}_i)/n_i$.

For bivariate data (C_{ij}, E_{ij}), a simple asymptotic (SA) confidence interval for *NHB* is given by

$$\widehat{NHB} \pm z_{\alpha/2}\widehat{\sigma}, \tag{10}$$

with the estimated variance for $\widehat{NHB}$ given by

$$\widehat{\sigma}^2 = \widehat{\mathrm{var}}(\widehat{NHB}) = \sum_{i=1}^{2} \widehat{\mathrm{var}}\left(\widehat{\mu_{Ei} - \mu_{Ci}/\lambda}\right) = \sum_{i=1}^{2} \left(s_{Ei}^2 + s_{Ci}^2/\lambda^2 - 2s_{CiEi}/\lambda\right)/n_i.$$

This procedure may require very large sample sizes to be valid because of its enforced symmetry. As pointed out above, the procedure based on the transformation of the Edgeworth expansion (Zhou and Dinh, 2005) lacks the invariance property. Dinh and Zhou (2006) have nonetheless applied it to the NHB. We summarize this procedure here for comparison. The $100(1-\alpha)\%$ confidence limits are given by

$$\begin{cases} L = \widehat{NHB} - \sqrt{N}G^{-1}(z_{1-\alpha/2}/\sqrt{N})\widehat{\sigma} \\ U = \widehat{NHB} - \sqrt{N}G^{-1}(z_{\alpha/2}/\sqrt{N})\widehat{\sigma}, \end{cases}$$

Table 1
The performance of the GCI approach[a] and the MOVER for constructing a two-sided 95% confidence interval for the lognormal mean based on 10,000 simulations ($\mu = -\sigma^2/2$). Each interval by GCI was based on 10,000 simulated pivotal quantities.

n	σ^2	GCI		MOVER	
		Cover (<, >)%[b]	Width	Cover (<, >)%	Width
5	0.5	93.99 (2.00, 4.01)	2.69	93.47 (3.19, 3.34)	2.54
	1.0	93.98 (1.92, 4.10)	4.78	94.65 (2.75, 2.60)	4.66
	1.5	94.15 (2.07, 3.78)	6.76	95.03 (2.92, 2.05)	6.67
	2.0	93.77 (2.36, 3.87)	8.82	95.10 (2.99, 1.91)	8.76
	2.5	94.08 (2.38, 3.54)	10.59	95.30 (2.89, 1.81)	10.55
	3.0	93.90 (2.09, 4.01)	12.74	95.35 (2.60, 2.05)	12.72
20	0.5	94.56 (2.35, 3.09)	0.76	94.24 (3.37, 2.39)	0.74
	1.0	94.90 (2.17, 2.93)	1.22	95.19 (2.87, 1.94)	1.20
	1.5	94.36 (2.40, 3.24)	1.64	94.76 (3.04, 2.20)	1.63
	2.0	94.39 (2.39, 3.22)	2.06	94.94 (2.89, 2.17)	2.04
	2.5	94.89 (2.14, 2.97)	2.44	95.24 (2.61, 2.15)	2.43
	3.0	94.97 (2.37, 2.66)	2.87	95.41 (2.82, 1.77)	2.86

[a] GCI: generalized confidence interval; MOVER: method of variance estimates recovery.
[b] < lower error rate, > upper error rate.

where $N = n_1 + n_2$, z_x is the upper x percentile from the standard normal distribution,

$$G^{-1}(x) = \left\{1 + 3\left(x - \frac{A}{6N}\right)\right\}^{1/3} - 1,$$

$$A = \frac{\sqrt{N}}{\widehat{\sigma}^3}(B_1 - B_2),$$

and

$$B_i = \sum_{j=1}^{n_i}\left\{\frac{C_{ij} - \bar{C}_i}{n_i\lambda} - \frac{E_{ij} - \bar{E}_i}{n_i}\right\}^3, \quad i = 1, 2.$$

Note that these expressions are simplified version of those given in Dinh and Zhou (2006, p. 580).

As an alternative, we can first use the MOVER in Section 2 to construct two separate confidence intervals for $\theta_1 = \mu_{E1} - \mu_{C1}/\lambda$ and $\theta_2 = \mu_{E2} - \mu_{C2}/\lambda$, with correlations between the cost and effectiveness measure within groups taken into account using Eq. (4). Application of Eq. (3) will then yield a confidence interval for the *NHB*. An advantage of our approach is that it reflects the asymmetry of the sampling distribution for $\widehat{NHB}$ by recovering variance estimates separately from the lower and upper limits for individual parameters.

4. Simulation studies

Since its derivation relies on asymptotic theory, we conducted three simulation studies to evaluate the performance of the MOVER as applied to lognormal data. For comparisons, we also included several existing methods. All simulations were conducted with 10,000 replications and implemented using SAS PROC IML.

For a single lognormal mean, we compared the MOVER with the GCI for setting 95% confidence limits for $\mu + \sigma^2/2$ with $n = 5$ and 20, and $\sigma^2 = 0.5$ to 3.0 by increments of 0.5. Without loss of generality, we set $\mu = -\sigma^2/2$. Each confidence interval using GCI in Table 1 was obtained from 10,000 simulated pivotal quantities using the algorithm in Krishnamoorthy and Mathew (2003). These results indicate that the MOVER performed as least as well as the GCI, even for sample sizes as small as 5.

We also compare the MOVER with the GCI, after correcting for a typographic error (Tian, 2005, p. 3227), in the case of the Δ-distribution. For this purpose, we considered the proportion of zero values $\delta = 0.1$ and 0.2, with $n = 15$, 25, and 50, and $\sigma^2 = 1.0$ to 3.0 by increments of 1.0. Again, we set $\mu = -\sigma^2/2$. Each GCI was obtained using 10,000 sets of simulated pivotal statistics. Simulation results in Table 2 show that both the MOVER and the GCI generally have confidence interval coverage close to the nominal 95% level, with the former consistently resulting in narrower interval width. We also note that MOVER tends to provide slightly unbalanced tail errors as compared to the GCI.

The last simulation study compared the MOVER to the Edgeworth expansion based method for the NHB (Dinh and Zhou, 2006). Simulation results are presented in Table 3. Consistent with Zhou and Dinh (2005), the Edgeworth expansion based method performs poorly, often worse than the method of point estimate plus/minus 1.96 times standard error. These results also show that the simple asymptotic method may provide reasonable overall coverage with a sample size of at least 500 per group, although with unbalanced error levels in the tails. On the other hand, the MOVER performs very well. Simulation results (not shown) using the same parameters as those in a previous study by Dinh and Zhou (2006) also show this procedure performs well.

Table 2
The performance of GCI[a] and the MOVER for constructing a two-sided 95% confidence interval for the mean of a Δ-distribution based on 10,000 simulations ($\mu = -\sigma^2/2$). Each interval by GCI was based on 10,000 simulated pivotal quantities.

δ	n	σ^2	GCI		MOVER	
			Cover (<, >) %[b]	Width	Cover (<, >) %	Width
0.1	15	1.0	95.53 (2.34, 2.13)	1.72	95.03 (3.60, 1.37)	1.65
		2.0	95.50 (2.26, 2.24)	2.85	95.22 (3.13, 1.65)	2.78
		3.0	94.94 (2.35, 2.71)	3.94	94.87 (2.90, 2.23)	3.88
	25	1.0	95.95 (2.03, 2.02)	1.22	95.21 (3.09, 1.70)	1.17
		2.0	95.21 (2.19, 2.60)	1.97	94.94 (2.87, 2.19)	1.93
		3.0	95.31 (2.26, 2.43)	2.71	95.09 (2.80, 2.11)	2.67
	50	1.0	95.79 (2.26, 1.95)	0.80	95.10 (3.02, 1.88)	0.78
		2.0	95.41 (2.37, 2.22)	1.29	95.16 (2.87, 1.97)	1.26
		3.0	94.87 (2.43, 2.70)	1.76	94.86 (2.69, 2.45)	1.73
0.2	15	1.0	95.99 (2.14, 1.87)	1.98	95.17 (3.30, 1.53)	1.87
		2.0	95.70 (2.02, 2.28)	3.23	95.41 (2.78, 1.81)	3.13
		3.0	94.93 (2.54, 2.53)	4.47	94.97 (3.11, 1.92)	4.38
	25	1.0	96.01 (2.14, 1.85)	1.36	95.17 (3.20, 1.63)	1.30
		2.0	95.70 (2.11, 2.19)	2.16	95.34 (2.79, 1.87)	2.10
		3.0	95.57 (2.17, 2.26)	2.96	95.27 (2.74, 1.99)	2.91
	50	1.0	95.56 (2.26, 2.18)	0.88	95.00 (2.99, 2.01)	0.85
		2.0	95.28 (2.36, 2.36)	1.39	94.95 (2.92, 2.13)	1.37
		3.0	95.39 (2.24, 2.37)	1.90	95.30 (2.58, 2.12)	1.87

[a] GCI: generalized confidence interval; MOVER: method of variance estimates recovery.
[b] < lower error rate, > upper error rate.

5. Examples

With the satisfactory performance of the MOVER in the simulation studies here and elsewhere (Zou and Donner, 2008; Zou et al., 2009b), we now illustrate the procedure using examples arising from health economics. We retain excessive decimal places until the end of the calculation, and then round to two decimal places.

Example 1: Mean of Δ-distribution

To illustrate the calculations for a mean from the Δ-distribution, consider the data found in Zhou and Tu (2000) involving a diagnostic test charge of $n = 40$ patients. Among them $n_0 = 10$ patients had no diagnostic tests during the study period. For the remaining patients, cost may be approximated by the lognormal distribution, as in Zhou and Tu (2000). Analysis of the data on the log scale yields $\bar{x} = 6.8535$ and $s^2 = 1.8696$ (Tian, 2005, p. 3231). From Eq. (7), the 95% confidence interval for $\hat{\delta} = 0.25$ is (0.141187, 0.40194). Setting $\theta_1 = \ln(1-\delta)$ and $\theta_2 = -(\mu+\sigma^2/2)$ and applying Eq. (6) to $\hat{\theta}_2$, the point estimates (and 95% confidence interval) of θ_1 and θ_2 are -0.28768 (-0.51406, -0.15300) and 7.7883 (7.19140, 8.68761), respectively. The MOVER then yields the 95% confidence limits for $(1-\delta)\exp(\mu+\sigma^2/2)$ given by (955.50, 4491.55), comparable to the GCI limits of (959.87, 4652.22) on the basis of 5000 simulated pivotal statistics (Tian, 2005, p. 3231), also close to our own result of (970.81, 4687.37) with 10,000 sets of simulated pivotal statistics.

Example 2: Difference between two independent lognormal means

As an example of a difference between two means, consider a study from Zhou et al. (1997) where the effects of race on the cost of medical care for patients with type I diabetes is of particular interest. The log-transformed cost data obtained on 119 black patients yields $\bar{x}_1 = 9.06694$, $s_1^2 = 1.82426$, while for 106 white patients yields $\bar{x}_2 = 8.69306$, $s_2^2 = 2.69186$. By Eq. (6), the mean cost (95% confidence interval) for black patients is estimated as 21570.19 (15806.00, 31388.77), while for white patients it is 22902.28 (14842.03, 39722.09). The difference in mean costs (95% confidence interval) is then obtained from the MOVER as -1332.09 (-19112.18, 11371.14).

Example 3: Difference between two dependent lognormal means

To exemplify our method for a difference between two dependent lognormal means, consider data presented by Zhou et al. (2001) in which the sample estimates based on 98 patients who had outpatient costs pre/post a policy change are given by $\bar{x}_1 = 6.41$, $s_1^2 = 2.73$, $\bar{x}_2 = 6.50$, $s_2^2 = 3.48$, and $\widehat{\rho} = 0.45$. By Eq. (6) the costs (95% confidence interval) at two periods are 2380.34 (1511.15, 4266.23) and 3789.54 (2195.38, 7770.97), respectively. Eq. (8) yields $\widehat{r} = 0.141416$, thus the pre/post cost difference (95% confidence interval) is given by -1409.20 (-5362.49, 881.57) using the MOVER.

6. Conclusion

We have demonstrated that interval estimation involving lognormal data requires only the application of confidence interval procedures found in introductory textbooks. Thus, it may be unnecessary to avoid lognormal assumptions for

Table 3
The performance of three procedures for constructing a two-sided 95% confidence intervals for the net health benefit ($\lambda = 10$) based on 10,000 simulations.

n	ρ_1/ρ_2	Edgeworth		SA[a]		MOVER	
		Cover (<, >) %[b]	Width	Cover (<, >) %	Width	Cover (<, >) %	Width
$(\ln C_{ij}, E_{ij}) \sim N_2(\mu_{Ci}, \mu_{Ei}, \sigma^2_{Ci}, \sigma^2_{Ei}, \rho_i)$[c]							
50	−0.8/−0.8	90.91 (0.20, 8.89)	276.7	94.04 (1.13, 4.83)	258.7	94.84 (3.46, 1.70)	353.7
50	−0.8/−0.2	91.03 (0.30, 8.67)	275.0	94.05 (1.18, 4.77)	257.1	94.82 (3.37, 1.81)	352.3
50	−0.2/−0.8	90.99 (0.20, 8.81)	274.7	94.04 (1.00, 4.96)	256.8	94.75 (3.26, 1.99)	350.9
50	−0.2/0.4	91.09 (0.21, 8.70)	270.4	94.16 (1.11, 4.73)	252.7	95.03 (3.12, 1.85)	349.0
50	0.4/−0.8	91.41 (0.16, 8.43)	275.2	94.68 (0.94, 4.38)	257.3	95.02 (3.10, 1.88)	352.6
50	0.4/0.4	91.11 (0.13, 8.76)	271.0	94.36 (0.81, 4.83)	253.3	95.36 (2.83, 1.81)	348.6
100	−0.8/−0.8	92.00 (0.34, 7.66)	195.2	93.98 (0.98, 5.04)	189.6	95.13 (2.98, 1.89)	207.5
100	−0.8/−0.2	92.18 (0.35, 7.47)	196.8	94.23 (1.05, 4.72)	191.2	95.01 (2.94, 2.05)	207.1
100	−0.2/−0.8	92.29 (0.24, 7.47)	194.7	94.04 (0.81, 5.15)	189.1	94.98 (3.02, 2.00)	207.2
100	−0.2/0.4	91.70 (0.34, 7.96)	195.3	93.93 (0.85, 5.22)	189.8	94.99 (2.95, 2.06)	207.1
100	0.4/−0.8	91.68 (0.39, 7.93)	194.9	94.12 (0.91, 4.97)	189.4	95.34 (2.83, 1.83)	205.9
100	0.4/0.4	91.90 (0.29, 7.81)	193.3	94.14 (0.90, 4.96)	187.7	94.93 (3.13, 1.94)	205.8
500	−0.8/−0.8	93.20 (0.74, 6.06)	92.53	94.11 (0.90, 4.99)	92.15	94.82 (2.84, 2.34)	81.62
500	−0.8/−0.2	93.15 (0.67, 6.18)	91.45	94.07 (0.92, 5.01)	91.07	95.30 (2.58, 2.12)	81.42
500	−0.2/−0.8	92.93 (0.69, 6.38)	91.86	93.64 (0.94, 5.42)	91.48	94.88 (2.64, 2.48)	81.44
500	−0.2/0.4	92.84 (0.64, 6.52)	92.08	93.72 (0.93, 5.35)	91.70	94.74 (2.84, 2.42)	81.39
500	0.4/−0.8	93.36 (0.79, 5.85)	91.76	94.18 (1.01, 4.81)	91.38	95.24 (2.59, 2.17)	81.32
500	0.4/0.4	93.11 (0.69, 6.20)	91.82	94.06 (0.94, 5.00)	91.45	94.83 (2.84, 2.33)	81.29
$(\ln C_{ij}, \ln E_{ij}) \sim N_2(\mu_{Ci}, \mu_{Ei}, \sigma^2_{Ci}, \sigma^2_{Ei}, \rho_i)$							
50	−0.8/−0.8	89.75 (0.04, 10.21)	370.8	92.82 (0.72, 6.46)	347.6	95.09 (3.80, 1.11)	453.3
50	−0.8/−0.2	90.02 (0.07, 9.91)	367.8	92.61 (0.69, 6.70)	344.8	95.13 (3.61, 1.26)	451.0
50	−0.2/−0.8	89.53 (0.08, 10.39)	353.4	92.95 (0.64, 6.41)	331.6	94.83 (4.06, 1.11)	440.0
50	−0.2/0.4	89.57 (0.08, 10.35)	345.3	92.96 (0.70, 6.34)	324.0	94.55 (4.46, 0.99)	437.0
50	0.4/−0.8	88.10 (0.04, 11.86)	314.2	91.69 (0.61, 7.70)	295.3	95.24 (3.78, 0.98)	410.5
50	0.4/0.4	87.90 (0.02, 12.08)	301.9	91.56 (0.44, 8.00)	283.6	95.60 (3.56, 0.84)	404.0
100	−0.8/−0.8	90.75 (0.33, 8.92)	261.2	92.68 (0.98, 6.34)	254.0	94.89 (3.55, 1.56)	267.2
100	−0.8/−0.2	91.47 (0.23, 8.30)	262.0	93.61 (0.66, 5.73)	254.8	94.95 (3.56, 1.49)	267.3
100	−0.2/−0.8	90.80 (0.16, 9.04)	249.3	92.76 (0.62, 6.62)	242.6	94.93 (3.82, 1.25)	257.0
100	−0.2/0.4	91.02 (0.16, 8.82)	245.8	93.03 (0.62, 6.35)	239.1	95.10 (3.58, 1.32)	254.4
100	0.4/−0.8	90.00 (0.14, 9.86)	223.0	92.30 (0.57, 7.13)	217.1	95.48 (3.30, 1.22)	236.5
100	0.4/0.4	90.03 (0.18, 9.79)	218.7	92.22 (0.59, 7.19)	213.0	95.38 (3.34, 1.28)	232.9
500	−0.8/−0.8	93.84 (0.70, 5.46)	122.4	94.39 (1.02, 4.59)	121.9	95.13 (2.87, 2.00)	105.9
500	−0.8/−0.2	93.29 (0.78, 5.93)	121.0	93.89 (1.09, 5.02)	120.5	94.83 (3.17, 2.00)	105.6
500	−0.2/−0.8	93.20 (0.62, 6.18)	115.9	94.07 (0.83, 5.10)	115.4	95.13 (2.88, 1.99)	101.1
500	−0.2/0.4	93.63 (0.62, 5.75)	113.9	94.20 (0.92, 4.88)	113.4	95.21 (2.96, 1.83)	99.50
500	0.4/−0.8	92.87 (0.57, 6.56)	104.6	93.62 (0.78, 5.60)	104.3	95.65 (2.78, 1.57)	91.39
500	0.4/0.4	92.70 (0.47, 6.83)	102.8	93.62 (0.64, 5.74)	102.4	95.68 (2.72, 1.60)	89.67

[a] SA: simple asymptotic; MOVER: method of variance estimates recovery.
[b] < lower error rate, > upper error rate.
[c] $N_2(\mu_{Ci}, \mu_{Ei}, \sigma^2_{Ci}, \sigma^2_{Ei}, \rho_i)$ denotes a bivariate normal distribution. Parameters used are $\mu_{C1} = 8$, $\mu_{E1} = 4$, $\sigma^2_{C1} = 0.5$, $\sigma^2_{E1} = 2$, $\mu_{C2} = 6$, $\mu_{E2} = 3$, $\sigma^2_{C2} = 2.5$, $\sigma^2_{E2} = 1.0$.

simplicity (Nixon and Thompson, 2005, p. 1226), or to rely on simulation of pivotal statistics (Krishnamoorthy and Mathew, 2003; Tian, 2005; Chen and Zhou, 2006; Krishnamoorthy et al., 2006).

Interval estimation based on transformations using the Edgeworth expansion (Zhou and Dinh, 2005) lacks the invariance property of a confidence interval for a difference, and performs poorly for lognormal data (Zhou and Dinh, 2005; Dinh and Zhou, 2006). One may argue that this procedure is nonparametric and thus it is unfair to compare it with the MOVER. Our position is that to be nonparametric, a procedure must be able to provide valid results for data having a common distribution, such as the lognormal in the present context.

Although we have deliberately used examples from health economics, the approach described here is also suitable for lognormal data arising from such disciplines as economics and environmental science (Rappaport and Selvin, 1987; Crow and Shimizu, 1988; Krishnamoorthy et al., 2006; Fletcher, 2008; Zou et al., 2009b). We should also mention that the MOVER we described here can be readily applied to lognormal regression models (Bradu and Mundlak, 1970; El-Shaarawi and Viveros, 1997; Wu et al., 2006; Tian and Wu, 2007b; Shen and Zhu, 2008), because in concept this problem is identical to that of producing confidence intervals for lognormal means.

As a final note, we should emphasize that the procedures in Section 3 are only applicable to lognormal data. Whenever there is apparent evidence against the lognormal assumption, the procedures presented in this section should not be used. However, this is not an inherent deficiency of the MOVER. This point is supported by several applications and extensions beyond lognormal data (Zou and Donner, 2008). Zou (2008) presents a further extension to interval estimation for measures of additive interaction, which are functions of risk ratios having only asymptotic lognormal distributions. New applications to set confidence limits for differences between Pearson correlations and coefficients of determination (R^2) may be found in Zou (2007), where it has been identified that the MOVER fails in the case of very small increments in R^2.

Acknowledgment

Partial results in this article have been presented at the 28th Annal Meeting of the Society for Clinical Trials in Montreal, Quebec, Canada, May 20–May 23, 2007. Guang Yong Zou is a recipient of Early Researcher Award, Ontario Ministry of Research and Innovation, Canada. His work is also partially supported by an Individual Discovery Grant (2007 to 2012) from Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Agresti, A., Coull, B.A., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. American Statistician 52, 119–126.
Aitchison, J., Brown, J.A.C., 1957. The Log-normal Distribution. Cambridge University Press, Cambridge.
Bartlett, M.S., 1953. Approximate confidence intervals. 2. More than one unknown parameter. Biometrika 40, 306–317.
Bebu, I., Mathew, T., 2008. Comparing the means and variances of a bivariate log-normal distribution. Statistics in Medicine 27, 2684–2696.
Bradu, D., Mundlak, T., 1970. Estimation in lognormal linear models. Journal of the American Statistical Association 65, 198–211.
Briggs, A., Nixon, R., Dixon, S., Thompson, S.G., 2005. Parametric modelling of cost data: Some simulation evidence. Health Economics 14, 421–428.
Briggs, A.H., O'Brien, B.J., Blackhouse, G., 2002. Thinking out of the box: Recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. Annual Review of Public Health 23, 337–401.
Burick, R.K., Graybill, F.A., 1992. Confidence Intervals on Variance Components. Dekker, New York.
Chen, H., 1994. Comparison of lognormal population means. Proceedings of the American Mathematical Society 121, 915–924.
Chen, Y.H., Zhou, X.H., 2006. Interval estimates for the ratio and difference of two lognormal means. Statistics in Medicine 25, 4099–4113.
Crow, E.L., Shimizu, K., 1988. The Log-normal Distributions: Theory and Applications. Dekker, New York.
Diciccio, T.J., Efron, B., 1996. Rejoinder of "bootstrap confidence intervals". Statistical Science 11, 223–228.
Diehr, P., Yanez, D., Ash, A., Hornbrook, M., Lin, D.Y., 1999. Methods for analyzing health care utilization and cost. Annual Review of Public Health 20, 125–144.
Dinh, P., Zhou, X.H., 2006. Nonparametric statistical methods for cost-effectiveness analyses. Biometrics 62, 576–588.
Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L., 2005. Methods for the Economic Evaluation of Health Care Programmes, 3rd ed. Oxford University Press, New York.
El-Shaarawi, A.H., Viveros, R., 1997. Inference about the mean in log-regression with environmental applications. Environmetrics 8, 569–582.
Fletcher, D., 2008. Confidence intervals for the mean of the delta-lognormal distribution. Environmental and Ecological Statistics 15, 175–189.
Gill, P.S., 2004. Small-sample inference for the comparison of means of log-normal distributions. Biometrics 60, 525–527.
Graybill, F.A., Wang, C.M., 1980. Confidence intervals on nonnegative linear combinations of variances. Journal of the American Statistical Association 75, 869–873.
Hall, P., 1992. On the removal of skewness by transformation. Journal of the Royal Statistical Society, **B** 54, 221–228.
Hoch, J.S., Briggs, A.H., Willan, A.R., 2002. Something old, something new, something borrowed, something blue: A framework for the marriage of health econometrics and cost-effectiveness analysis. Health Economics 11, 415–430.
Howe, W.G., 1974. Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variable. Journal of the American Statistical Association 69, 789–794.
Kowalski, C.J., 1972. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. Applied Statistics 21, 1–12.
Krishnamoorthy, K., Mathew, T., 2003. Inferences on the means of lognormal distributions using generalized p-vlaues and generalized confidence intervals. Journal of Statistical Planning and Inference 115, 103–121.
Krishnamoorthy, K., Mathew, T., Ramachandran, G., 2006. Generalized p-values and confidence intervals: A novel approach for analyzing lognormally distributed exposure data. Journal of Occupational and Environmental Hygiene 3, 642–650.
Land, C.E., 1972. An evaluation of approximate confidence interval estimation methods for lognormal means. Technometrics 14, 145–158.
Lecoutre, B., Faure, S., 2007. A note on new confidence intervals for the difference between two proportions based on an Edgeworth expansion. Journal of Statistical Planning and Inference 137, 355–356.
Lee, Y.H., Shao, J., Chow, S.C., 2004. Modified large-sample confidence intervals for linear combinations of variance components: Extension, theory, and application. Journal of the American Statistical Association 99, 467–478.
Limpert, E., Stahel, W.A., Abbt, M., 2001. Log-normal distributions across the sciences: Keys and clues. BioScience 51, 341–352.
Mohn, E., 1979. Confidence estimation of measures of location in the log normal distribution. Biometrika 66, 567–575.
Mostafa, M.D., Mahmoud, M.W., 1964. On the problem of estimation for the bivariate lognormal distribution. Biometrika 51, 522–527.
Newcombe, R.G., 2001. Estimating the difference between differences: Measurement of additive scale interaction for proportions. Statistics in Medicine 20, 2885–2893.
Nixon, R., Thompson, S.G., 2005. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluation. Health Economics 14, 1217–1229.
Owen, W.J., DeRouen, T.A., 1980. Estimation of the mean for lognormal data containing zeros and left-censored values, with application to the measurement of worker exposure to air contaminants. Biometrics 36, 707–719.
Parkhurst, D.F., 1998. Arithmetic versus geometric means for environmental concentration data. Environmental Science & Technology 88, 92A–98A.
Pennington, M., 1983. Efficient estimators of abundance, for fish and plankton surveys. Biometrics 39, 281–286.
Rappaport, S.M., Selvin, S., 1987. A method for evaluating the mean exposure from a lognormal distribution. American Industrial Hygiene Journal 48, 374–379.
Schenker, N., 1985. Qualms about bootstrap confidence intervals. Journal of the American Statistical Association 80, 360–361.
Shen, H., Brown, L.D., Zhi, H., 2006. Efficient estimation of log-normal means with application to pharmacokinetic data. Statistics in Medicine 25, 3023–3028.
Shen, H., Zhu, Z., 2008. Efficient mean estimation in log-normal linear models. Journal of Statistical Planning and Inference 138, 552–567.
Shimizu, K., 1983. UMVU estimation for covariance and first product moments of transformed variables. Communications in Statistics 12, 1661–1674.
Smith, S.J., 1988. Evaluating the efficiency of the Δ-distribution mean estimator. Biometrics 44, 485–493.
Stinnett, A., Mullahy, J., 1998. Net health benefit: A new framework for the analysis of uncertainty in cost-effectiveness analysis. Medical Decision Making 18, S68–S80.
Taylor, D.J., Kupper, L.L., Muller, K.E., 2002. Improved approximate confidence intervals for the mean of a log-normal random variable. Statistics in Medicine 21, 1443–1459.
Thompson, S.G., Barber, J.A., 2000. How should cost data in pragmatic randomised trials be analysed? British Medical Journal 320, 1197–1200.
Tian, L., Wu, J., 2006. Confidence intervals for the mean of lognormal data with excess zeros. Biometrical Journal 48, 149–156.
Tian, L., Wu, J., 2007a. Inferences on the common mean of several log-normal populations: The generalized variable approach. Biometrical Journal 49, 944–951.
Tian, L., Wu, J., 2007b. Inferences on the mean response in a log-regression model: The generalized variable approach. Statistics in Medicine 26, 5180–5188.
Tian, L.L., 2005. Inferences on the mean of zero-inflated lognormal data: The generalized variable approach. Statistics in Medicine 24, 3223–3232.
Willan, A.R., 2001. Analysis, sample size, and power for estimating incremental net health benefit from clinical trial data. Controlled Clinical Trials 22, 228–237.
Willan, A.R., Briggs, A.H., 2006. Statistical Analysis of Cost-effectiveness Data. Wiley, New York.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22, 209–212.
Wu, J., Jiang, G., Wong, A.C.M., Sun, X., 2002. Likelihood analysis for the ratio of means of two independent log-normal distributions. Biometrics 58, 463–469.
Wu, J., Wong, A.C.W., Wei, W., 2006. Interval estimation of the mean response in a log-regression model. Statistics in Medicine 25, 2125–2135.
Wu, J.R., Wong, A.C.M., Jiang, G.Y., 2003. Likelihood-based confidence intervals for a log-normal mean. Statistics in Medicine 22, 1849–1860.
Zhou, X.H., 2002. Inferences about population means of health care costs. Statistical Methods in Medical Research 11, 327–339.
Zhou, X.H., Dinh, P., 2005. Nonparametric confidence intervals for the one- and two-sample problems. Biostatistics 6, 187–200.
Zhou, X.H., Gao, S., 1997. Confidence intervals for the lognormal mean. Statistics in Medicine 16, 783–790.
Zhou, X.H., Gao, S., Hui, S.L., 1997. Methods for comparing the means of two independent log-normal samples. Biometrics 53, 1129–1135.
Zhou, X.H., Li, C., Gao, S., Tierney, W.M., 2001. Methods for testing equality of means of health care costs in a paired design study. Statistics in Medicine 20, 1703–1720.
Zhou, X.H., Qin, G.S., 2007. A supplement to: "A new confidence interval for the difference between two binomial proportions of paired data". Journal of Statistical Planning and Inference 137, 357–358.
Zhou, X.H., Tu, W., 2000. Confidence intervals for the mean of diagnostic test charge data containing zeros. Biometrics 56, 1118–1125.
Zou, G.Y., 2007. Toward using confidence intervals to compare correlations. Psychological Methods 12, 399–413.
Zou, G.Y., 2008. On the estimation of additive interaction by use of the four-by-two table and beyond. American Journal of Epidemiology 168, 212–224.
Zou, G.Y., Donner, A., 2008. Construction of confidence limits about effect measures: A general approach. Statistics in Medicine 27, 1693–1702.
Zou, G.Y., Huang, W., Zhang, X., 2009a. A note on confidence interval estimation for a linear function of binomial proportions. Computational Statistics and Data Analysis 53, 1080–1085.
Zou, G.Y., Huo, C.Y., Taleban, J., 2009b. Simple confidence intervals for lognormal means and their differences with environmental applications. Environmetrics 20, 172–180.

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1041–1053
Published online 13 January 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3523

Confidence interval construction for a difference between two dependent intraclass correlation coefficients

Chinthanie F. Ramasundarahettige[1], Allan Donner[1,2] and G. Y. Zou[1,2,*,†]

[1]*Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5C1*
[2]*Robarts Clinical Trials, Robarts Research Institute, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5K8*

SUMMARY

Inferences for the difference between two dependent intraclass correlation coefficients (ICCs) may arise in studies in which a sample of subjects are each assessed several times with a new device and a standard. The ICC estimates for the two devices may then be compared using a test of significance. However, a confidence interval for a difference between two ICCs is more informative since it combines point estimation and hypothesis testing into a single inference statement. We propose a procedure that uses confidence limits for a single ICC to recover variance estimates needed to set confidence limits for the difference. An advantage of this approach is that it provides a confidence interval that reflects the underlying sampling distribution. Simulation results show that this method performs very well in terms of overall coverage percentage and tail errors. Two data sets are used to illustrate this procedure. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: reliability; Fisher's z-transformation; skewness

1. INTRODUCTION

The intraclass correlation coefficient (ICC) is commonly used to assess the reliability of measurements when observations approximately follow a normal distribution [1, 2]. As a result, an extensive literature has been accumulated on inferences for this parameter, with extensive reviews provided by Donner [3] and McGraw and Wong [4]. With few exceptions [5–7], relatively little research has appeared on the problem of comparing two correlated ICCs computed from the same sample

*Correspondence to: G. Y. Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5C1.
†E-mail: gzou@robarts.ca

Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada
Contract/grant sponsor: Ontario Ministry of Research and Innovation

Received 18 July 2008
Accepted 3 December 2008

of subjects. However, such problems arise frequently, for example, when it is of interest to estimate the difference in reliability between two observers or devices, as in a study [8] aimed at evaluating the equivalence of two approaches designed to measure the size of a patients' brain ventricle relative to that of the patients' skull (VBR). In this study, two VBR measurements were obtained on each of the 50 patients using an automated pixel count based on the image displayed on a television screen or a hand-held planimeter projecting an X-ray image. As a second example, Gomez *et al.* [9] reported on a study in which five repeated measurements of broadband ultrasound attenuation (in dB/MHz) on the right heel of 34 subjects were obtained using 2 scanning devices (BEAM and UBIS3000). The purpose here was to compare the performance of the new device (the BEAM scanner) with the performance of the currently available scanner (UBIS3000).

In this paper, we use a general method for confidence interval estimation proposed by Zou and Donner [10] to obtain confidence limits about a difference between two dependent ICCs. The advantage of this method is that it accounts for the left-skewness of the sampling distribution of an ICC. Since the central idea here is to recover variance estimates from confidence limits about a single ICC, we refer to this method as the method of variance estimates recovery (MOVER). Monte Carlo simulation is used to assess the performance of this method using confidence limits for a single ICC obtained from the application of five different procedures: (1) the simple asymptotic approach, (2) application of Fisher's z-transformation, (3) application of the inverse hyperbolic tangent, (4) application of a modified Fisher's z-transformation, and (5) an exact method based on the F-distribution. The evaluation criteria considered are overall coverage, tail errors, and confidence interval width. Two examples are presented.

2. NOTATION AND TERMINOLOGY

Suppose two devices each measuring the same sample of subjects, with the first device yielding k_1 measurements and the second device yielding k_2 measurements on each subject. Adopting the notation in Donner and Zou [6], let

$$\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ik_1}, X_{i,k_1+1}, X_{i,k_1+2}, \ldots, X_{i,k_1+k_2})$$

denote measurements on the ith subject, $i = 1, 2, \ldots, n$, where $X_{i1}, X_{i2}, \ldots, X_{ik_1}$ are the measurements obtained by the first device (or observer) and $X_{i,k_1+1}, X_{i,k_1+2}, \ldots, X_{i,k_1+k_2}$ are the measurements obtained by the second device (observer). We assume that the following model holds:

$$\mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_1 \mathbf{1}_{k_1}^{\mathrm{T}}, \mu_2 \mathbf{1}_{k_2}^{\mathrm{T}})$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} [(1-\rho_1)\mathbf{I}_{k_1} + \rho_1 \mathbf{J}_{k_1}]\sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \mathbf{J}_{k_1 \times k_2} \\ \rho_{12}\sigma_1\sigma_2 \mathbf{J}_{k_2 \times k_1} & [(1-\rho_2)\mathbf{I}_{k_2} + \rho_2 \mathbf{J}_{k_2}]\sigma_2^2 \end{pmatrix}$$

In these expressions $\mathbf{1}_k$ is a column vector with all the k elements equal to 1, $\mathbf{I}_{k_1}$ is a $k_1 \times k_1$ identity matrix, while $\mathbf{J}_k$ and $\mathbf{J}_{k_1 \times k_2}$ are $k \times k$ and $k_1 \times k_2$ matrices with all elements equal to 1. Thus, the model assumes that the k_g measurements taken by the first device have common mean μ_g, common variance σ_g^2, and common intraclass correlation ρ_g, $g = 1, 2$. The interclass correlation coefficient ρ_{12} between any pair of observations X_{ij} $(j = 1, 2, \ldots, k_1)$ and $X_{i,k_1+j'}$ $(j' = 1, 2, \ldots, k_2)$ is also

DOI: 10.1002/sim

assumed constant across all subjects in the population. These assumptions imply that the X_{ij} for a given device are interchangeable.

Elston [11] derived the maximum likelihood estimators $\hat{\rho}_1$, $\hat{\rho}_2$ and $\hat{\rho}_{12}$ of ρ_1, ρ_2 and ρ_{12}, respectively. Briefly, one obtains $\hat{\rho}_1$ and $\hat{\rho}_2$ by computing the Pearson product–moment correlation overall possible pairs of measurements that can be constructed within devices 1 and 2, respectively, while $\hat{\rho}_{12}$ is similarly obtained by computing the correlation overall possible nk_1k_2 pairs $(X_{ij}, X_{i,k_1+j'})$.

Another frequently applied approach requires calculation of the ANOVA (analysis of variance) estimator for ρ_g, $g=1,2$, given by

$$\hat{\rho}_g = \frac{\mathrm{MSA}_g - \mathrm{MSW}_g}{\mathrm{MSA}_g + (k_g - 1)\mathrm{MSW}_g}$$

where MSA_g and MSW_g are the mean-square errors among and within subjects, respectively, as derived from an ANOVA on the measurements taken by device g. Note that the ANOVA estimator and the maximum likelihood estimator are virtually indistinguishable from each other in the context of reliability studies, where the number of measurements per subject tends to be constant.

3. REVIEW OF CONFIDENCE INTERVALS FOR ICCS

As mentioned above, our approach to interval estimation of $\rho_1 - \rho_2$ is to recover variance estimates from the confidence limits for a single ICC. The simplest such approach for single ρ is to apply the central limit theorem in conjunction with Slutzky's theorem, yielding

$$l, u = \hat{\rho} \mp z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\rho})} \tag{2}$$

where $\hat{\rho}$ is the sample estimate of ρ, $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, and

$$\widehat{\mathrm{var}}(\hat{\rho}) = \frac{2(nk-1)(1-\hat{\rho})^2[1+(k-1)\hat{\rho}]^2}{k^2(k-1)n(n-1)} \tag{3}$$

where n is the number of subjects in the study, k is the number of measurements on each subject (usually referred to as class size). This variance formula is a special case of Smith [12], who derived a formula for the case of variable class sizes. We refer to this method as the simple asymptotic method, which enforces symmetry on the sampling distribution for $\hat{\rho}$ and thus cannot be expected to perform well in most practical situation.

A well-known approach [13, p. 221] that may be used to account for the left-skewed sample distribution of the ICC is Fisher's z-transformation, given by

$$z(\hat{\rho}) = 0.5 \ln \frac{1+(k-1)\hat{\rho}}{1-\hat{\rho}}$$

which is distributed asymptotically as normal distribution

$$\mathrm{N}\left(0.5 \ln \frac{1+(k-1)\rho}{1-\rho}, \frac{k}{2(k-1)(n-2)}\right)$$

It has been suggested that [14, p. 566] the inverse hyperbolic tangent transformation (sometimes also referred to as arctanh transformation) may perform better than Fisher's z-transformation. This

DOI: 10.1002/sim

procedure entails first obtaining confidence limits on the transformed scales $0.5\ln\{[1+\rho]/(1-\rho)\}$, followed by inversion to obtain confidence limits on the original scale. The resulting limits are based on the large sample distribution of

$$Z=0.5\ln\frac{1+\hat{\rho}}{1-\hat{\rho}}$$

which may be taken as

$$\mathrm{N}\left(0.5\ln\frac{1+\rho}{1-\rho},\frac{2(nk-1)[1+(k-1)\hat{\rho}]^2}{k^2(k-1)n(n-1)}\right)$$

The variance estimate in this expression is derived using the delta method with variance formula for $\hat{\rho}$ given in (3).

Konishi [15] observed that Fisher's z-transformation cannot simultaneously normalize the sampling distribution and stabilize variance for the case of $k>2$, and thus proposed the modification $\sqrt{(k-1)/(2k)}\ln\{[1+(k-1)\rho]/(1-\rho)\}$. The resulting confidence interval is based on the large sample distribution of

$$Z=\sqrt{\frac{k-1}{2k}}\ln\frac{[1+(k-1)\hat{\rho}]}{1-\hat{\rho}}$$

which may be taken as

$$\mathrm{N}\left(\sqrt{\frac{k-1}{2k}}\ln\frac{[1+(k-1)\rho]}{1-\hat{\rho}}+\frac{7-5k}{n\sqrt{18k(k-1)}},\frac{1}{n}\right)$$

One can also construct an exact confidence interval for ρ based on the F-distribution [16, p. 659]. This confidence interval is given by

$$l=[k\hat{\rho}+(1-F_{1-\alpha/2})(1-\hat{\rho})]/[k-(k-1)(1-F_{1-\alpha/2})(1-\hat{\rho})]$$
$$u=[k\hat{\rho}+(1-F_{\alpha/2})(1-\hat{\rho})]/[k-(k-1)(1-F_{\alpha/2})(1-\hat{\rho})]$$

where F_q is the qth quantile of the F-distribution with degrees of freedom of $n-1$ and $n(k-1)$.

4. INTERVAL ESTIMATION FOR A DIFFERENCE BETWEEN TWO DEPENDENT ICCS

We first assume that $\hat{\rho}_1$ and $\hat{\rho}_2$ are statistically independent. By the central limit theorem, a $100(1-\alpha)$ per cent confidence interval for $\rho_1-\rho_2$ is given by

$$\begin{aligned}L&=\hat{\rho}_1-\hat{\rho}_2-z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\rho}_1)+\mathrm{var}(\hat{\rho}_2)}\\U&=\hat{\rho}_1-\hat{\rho}_2+z_{\alpha/2}\sqrt{\mathrm{var}(\hat{\rho}_1)+\mathrm{var}(\hat{\rho}_2)}\end{aligned}\tag{4}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. The confidence limits in (4) are not yet applicable without the appropriate variance estimates, which we denote by $\widehat{\mathrm{var}}(\hat{\rho}_g)$, $g=1,2$.

DOI: 10.1002/sim

Dropping the subscript g, suppose that now we are given two-sided $100(1-\alpha)$ per cent confidence limits (l, u) for ρ, as obtained using one of the five procedures presented in Section 3. Since the limits l and u are not necessarily symmetric about $\hat{\rho}$, we must have

$$l = \hat{\rho} - z_{\alpha/2}\sqrt{\text{var}(\hat{\rho})}$$

which yields

$$\widehat{\text{var}}(\hat{\rho}) = \frac{(\hat{\rho}-l)^2}{z^2_{\alpha/2}}$$

under $\rho \approx l$. Similarly,

$$u = \hat{\rho} + z_{\alpha/2}\sqrt{\text{var}(\hat{\rho})}$$

and thus

$$\widehat{\text{var}}(\hat{\rho}) = \frac{(u-\hat{\rho})^2}{z^2_{\alpha/2}}$$

under $\rho \approx u$.

Among the plausible values of ρ_1 provided by (l_1, u_1) and ρ_2 by (l_2, u_2), the lower limit L for $\rho_1 - \rho_2$ can be seen to be close to $l_1 - u_2$ and the upper limit U close to $u_1 - l_2$. In the spirit of score-type confidence intervals [17], we can use the variance estimates needed to obtain L in (4) under the condition of $\rho_1 \approx l_1$ and $\rho_2 \approx u_2$, i.e.

$$\widehat{\text{var}}(\hat{\rho}_1) + \widehat{\text{var}}(\hat{\rho}_2) = \frac{(\hat{\rho}_1 - l_1)^2}{z^2_{\alpha/2}} + \frac{(u_2 - \hat{\rho}_2)^2}{z^2_{\alpha/2}}$$

Substituting into (4), the lower limit L for $\rho_1 - \rho_2$ is given by

$$L = \hat{\rho}_1 - \hat{\rho}_2 - \sqrt{(\hat{\rho}_1 - l_1)^2 + (u_2 - \hat{\rho}_2)^2} \tag{5}$$

Similar steps lead to the upper limit U for $\rho_1 - \rho_2$ given by

$$U = \hat{\rho}_1 - \hat{\rho}_2 + \sqrt{(u_1 - \hat{\rho}_1)^2 + (\hat{\rho}_2 - l_2)^2} \tag{6}$$

Furthermore, by definition, we have

$$\text{cov}(\hat{\rho}_1, \hat{\rho}_2) = \text{corr}(\hat{\rho}_1, \hat{\rho}_2)\sqrt{\text{var}(\hat{\rho}_1)\text{var}(\hat{\rho}_2)}$$

which can be used to extend (5) and (6) to the case of dependent ICCs. Therefore, a $100(1-\alpha)$ per cent confidence interval for $\rho_1 - \rho_2$ is given by

$$\begin{aligned} L &= \hat{\rho}_1 - \hat{\rho}_2 - \sqrt{(\hat{\rho}_1 - l_1)^2 - 2\,\widehat{\text{corr}}(\hat{\rho}_1, \hat{\rho}_2)(\hat{\rho}_1 - l_1)(u_2 - \hat{\rho}_2) + (u_2 - \hat{\rho}_2)^2} \\ U &= \hat{\rho}_1 - \hat{\rho}_2 + \sqrt{(u_1 - \hat{\rho}_1)^2 - 2\,\widehat{\text{corr}}(\hat{\rho}_1, \hat{\rho}_2)(u_1 - \hat{\rho}_1)(\hat{\rho}_2 - l_2) + (\hat{\rho}_2 - l_2)^2} \end{aligned} \tag{7}$$

DOI: 10.1002/sim

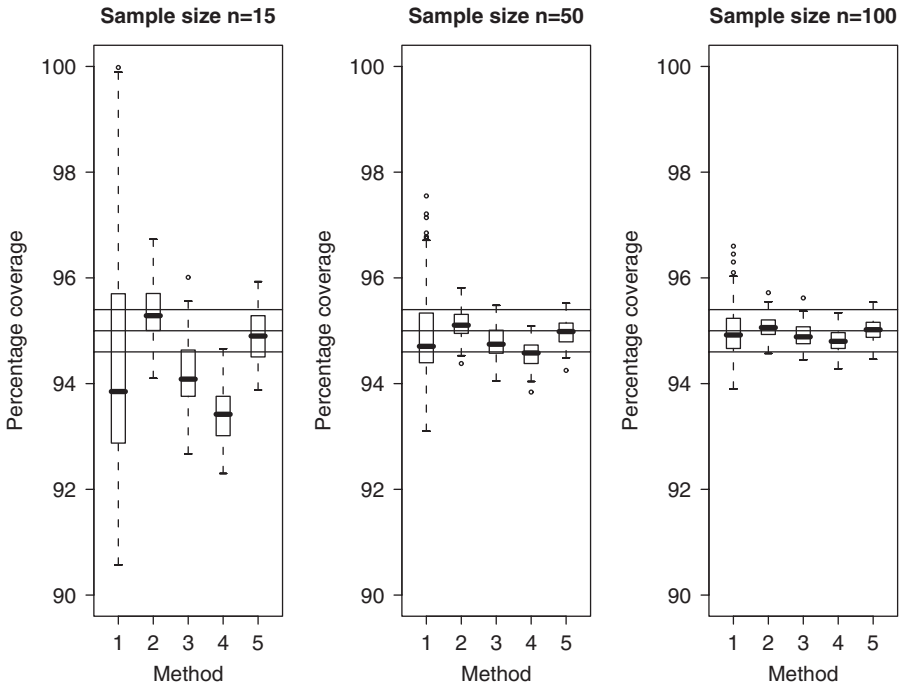


Figure 1. Empirical coverage per cent (based on 10 000 runs) of 95 per cent nominal confidence intervals for a difference between two dependent ICCs using variance estimates recovered from confidence limits for single ICCs, as obtained from (1) simple asymptotic, (2) Fisher's z transformation, (3) arctanh transformation, (4) Konish modified Fisher z-transformation, and (5) F-distribution. Each boxplot was drawn from coverage percentages of 128 $(=4(k_1,k_2)\times 4\rho_1\times 4(\rho_1-\rho_2)\times 2\rho_{12})$ parameter combinations. Horizontal lines in each plot are 94.6, 95, and 95.4 per cent, respectively.

where

$$\widehat{\mathrm{corr}}(\hat{\rho}_1,\hat{\rho}_2)=\frac{\sqrt{k_1k_2(k_1-1)(k_2-1)}}{[1+(k_1-1)\hat{\rho}_1][1+(k_2-1)\hat{\rho}_2]}\hat{\rho}_{12}^2 \tag{8}$$

As an example of computing $\hat{\rho}_{12}$, suppose that for a given subject the measurements taken by the first device are 2, 4, 3, and by the second device are 7, 6, 9 ($k_1=k_2=3$). We create the pairs for this subject as (2, 7), (2, 6), (2, 9), (4, 7), (4, 6) (4, 9), (3, 7), (3, 6), and (3, 9) and proceed in similar manner for all remaining subjects. The estimator $\hat{\rho}_{12}$ is then obtained by computing the Pearson product moment correlation between the first and second elements in all resulting pairs.

Note that the simple asymptotic confidence interval for a difference can be shown to be a consequence of using the simple asymptotic limits (2) to recover the variance estimates. This indicates that the validity of the MOVER relies on that of the confidence limits for single ICCs.

DOI: 10.1002/sim

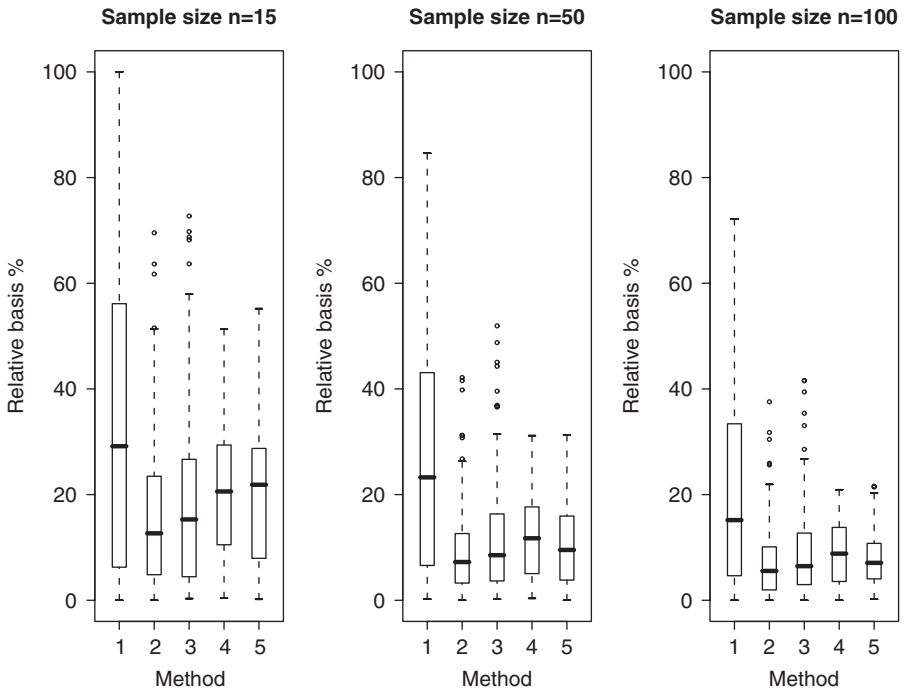


Figure 2. Imbalance of tail errors, quantified by relative bias per cent [100|MR−ML|/(MR+ML)], of 95 per cent nominal confidence intervals for a difference between two dependent ICCs using variance estimates recovered from confidence limits for single ICCs, as obtained from (1) simple asymptotic, (2) Fisher's z-transformation, (3) arctanh transformation, (4) Konish modified Fisher z-transformation, and (5) F-distribution. Each boxplot was drawn from coverage percentages of 128 $(=4(k_1,k_2)\times 4\rho_1\times 4(\rho_1-\rho_2)\times 2\rho_{12})$ parameter combinations.

5. SIMULATION STUDY

As the derivation of the MOVER and its application to interval estimation for a difference between two dependent ICCs rely largely on the central limit theorem, its theoretical properties are intractable in finite samples. We therefore used simulation to evaluate the performance of four procedures resulting from using the five sets of confidence limits described in Section 3.

The parameters for the simulation study include the total number of subjects (n), the number of measurements taken by each device (k_1, k_2), and values for ρ_1, ρ_2 and ρ_{12}. We selected the values of ρ_1 and ρ_2 based on suggested benchmark values [18] and considered $n=15, 50, 100$ and $(k_1,k_2)=(2,2),(2,4),(4,4)$, and (6, 6). The parameter values for ρ_1 were set to be 0.6, 0.8, 0.9 and 0.95, and for $\rho_1-\rho_2$ to be 0–0.3 with increment of 0.1. We considered ρ_{12} to be $0.5\sqrt{\rho_1\rho_2}-0.05$ and $\sqrt{\rho_1\rho_2}-0.05$ to satisfy the requirement of a positive definite variance–covariance matrix Σ. Finally, without loss of generality, we set $\mu_1=\mu_2=0$ and $\sigma_1=\sigma_2=1$, where μ_1, μ_2 and σ_1^2, σ_2^2 are the common mean and common variance of the measurements obtained by device 1 and device 2, respectively.

For each of 384 parameter combinations $[3n\times 4(k_1,k_2)\times 4\rho_1\times 4(\rho_1-\rho_2)\times 2\rho_{12}]$, we generated 10 000 runs from a multivariate normal distribution with correlation structure defined by (1).

DOI: 10.1002/sim

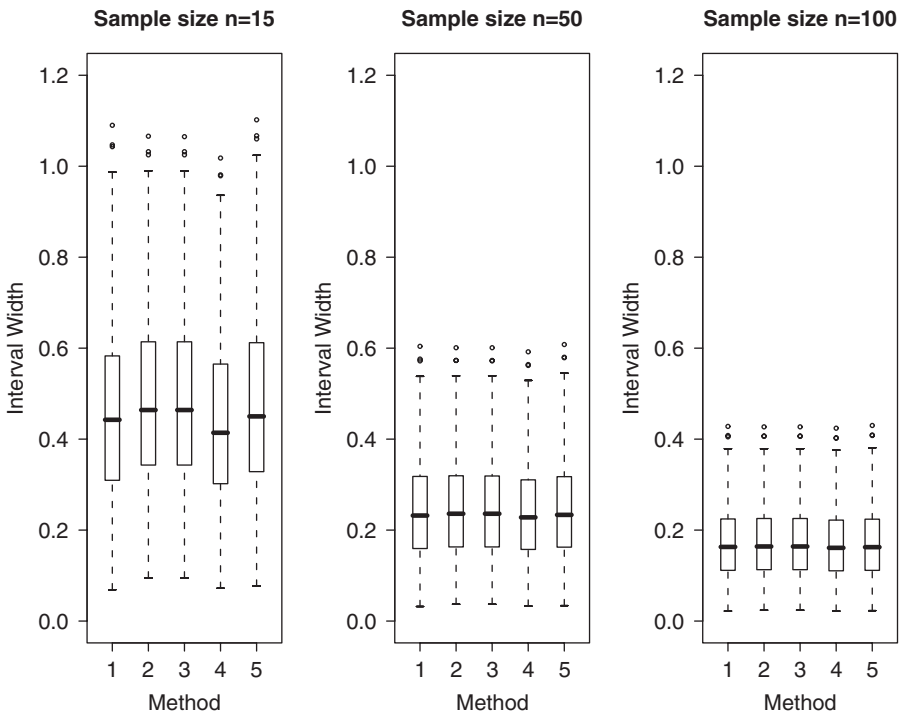


Figure 3. Confidence interval width of 95 per cent nominal confidence intervals for a difference between two dependent ICCs using variance estimates recovered from confidence limits for single ICCs, as obtained from (1) simple asymptotic, (2) Fisher's z-transformation, (3) arctanh transformation, (4) Konish modified Fisher z-transformation, and (5) F-distribution. Each boxplot was drawn from means of interval width for 128 $(=4(k_1,k_2)\times 4\rho_1\times 4(\rho_1-\rho_2)\times 2\rho_{12})$ parameter combinations.

The selected number of runs was based on a 0.4 per cent margin of error, i.e. we expected the empirical coverage to vary between 94.6–95.4 per cent for a two-sided 95 per cent confidence interval. Following advice given by Efron [19] on the evaluation of confidence interval methods, we also considered the balance between the left and right tail errors as the second criterion. Interval width was considered as the third criterion. The empirical coverage percentage was estimated by the relative frequency out of 10 000 intervals that contained the true parameter value. We also recorded the number of intervals that missed from the left (ML), occurring when the interval is completely to the left of the parameter value, and from the right (MR). The relative bias per cent was then obtained as

$$\frac{|\mathrm{MR}-\mathrm{ML}|}{\mathrm{MR}+\mathrm{ML}}\times 100$$

Figure 1 shows clearly that the simple asymptotic method provides erratic overall coverage, especially when $n=15$ and 50. The procedure based on Fisher's z-transformation has a tendency to provide coverage percentages that are greater than the nominal level, while those based on the arctanh transformation and Konish's modified z-transformation tends to show undercoverage. The

DOI: 10.1002/sim

trends are most obvious when $n=15$ and 50. Overall, the procedure based on the F-distribution performs very well, even at $n=15$. Figure 2 shows that the simple asymptotic method results in lop-sided intervals, a problem has been reduced by use of the three transformations considered. Confidence intervals constructed using variance estimates recovered from the F-distribution method are seen to be slightly more balanced. Figure 3 demonstrates that, even when compared with methods having poorer coverage levels, the MOVER method utilizing the F-distribution remains competitive in terms of expected interval width.

Table I presents comparative performance of the procedures based on the Fisher z-transformation with that based on the F-distribution at sample sizes of 15 and 50. It is clear that the former performs very well when the number of replicates made by each device is 2, but for replicates of 6, it tends to provide lop-sided intervals for nonzero differences between two ICC's. In contrast, the procedure based on F-distribution provides virtually balanced intervals. The numerical values of all 384 parameter combinations are available from the authors.

6. EXAMPLES

We now consider the examples discussed in Section 1. For the first example, the data derived from computer-aided tomographic scans of the heads of 50 patients are available in Donner and Zou [6] and Dunn [20, Chapter 5]. The two devices used in the study were an automated pixel count (PIX) based on the image displayed on a television screen and a hand-held planimeter (PLAN) based on a projection of the X-ray image. As reported in [6], we have $\hat{\rho}_1=0.994$, $\hat{\rho}_2=0.731$, and $\hat{\rho}_{12}=0.652$. By (8), $\widehat{\mathrm{corr}}(\hat{\rho}_1,\hat{\rho}_2)=0.246$. Table II shows the various confidence intervals for the difference in reliability between these two devices, which are all fairly wide. However, consistent with the simulation results, the confidence interval based on the simple asymptotic method differs considerably from other methods.

We present this example mainly for illustrative purpose. As pointed out by Dunn [20] it may be an oversimplification to conclude immediately that the pixel method is considerably more reliable than the planimetry-based method. This is because an inspection of the raw data shows that the planimeter-based measurements appear to be much less prone to gross errors, something that may be better understood by using more explicit modeling methods.

Table II also shows the results for the study comparing the reliability of two quantitative ultrasound devices in the diagnosis of osteoporosis. On the basis of observations of broadband ultrasound attenuation, Giraudeau *et al.* [7, p. 169] reported that $\hat{\rho}_1=0.982$ for the BEAM scanner, $\hat{\rho}_2=0.948$ for the UBIS 3000, and $\hat{\rho}_{12}=0.915$ that gives $\widehat{\mathrm{corr}}(\hat{\rho}_1,\hat{\rho}_2)=0.434$. Although the difference in point estimates is very small (0.034), the results obtained from using five different methods to recover variance estimates still shown considerable variability. However, in general, one may conclude that there is little difference between the standard device UBIS-3000 and the new BEAMER Scanner. Sole reliance on the result obtained from hypothesis testing ($P<0.001$) could reach a different conclusion.

7. CONCLUDING REMARKS

We have presented a new approach to confidence interval estimation for a difference between two dependent intraclass correlation coefficients (ICCs). The method is based on a recovery of

DOI: 10.1002/sim

Table I. Performance of two procedures (based on 10 000 runs) setting two-sided 95 per cent confidence intervals for a difference between intraclass correlation coefficients.

			$k_1 = k_2 = 2$			
			Fisher z-transformation		F-distribution	
Sample size n	ρ_1, ρ_2	ρ_{12}^*	Cover (L, R)$^+$ per cent	Width	Cover (L, R) per cent	Width
15	0.90, 0.90	1	95.22 (2.36, 2.42)	0.361	94.84 (2.56, 2.60)	0.376
	0.90, 0.90	2	95.52 (2.18, 2.30)	0.317	95.11 (2.40, 2.49)	0.331
	0.90, 0.80	1	94.96 (2.01, 3.03)	0.499	94.62 (2.06, 3.32)	0.519
	0.90, 0.80	2	95.90 (1.51, 2.59)	0.447	95.60 (1.56, 2.84)	0.465
	0.90, 0.70	1	95.26 (2.03, 2.71)	0.615	94.98 (2.09, 2.93)	0.638
	0.90, 0.70	2	95.74 (1.58, 2.68)	0.564	95.38 (1.68, 2.94)	0.585
	0.90, 0.60	1	95.17 (1.96, 2.87)	0.705	94.78 (2.04, 3.18)	0.730
	0.90, 0.60	2	95.95 (1.71, 2.34)	0.659	95.60 (1.80, 2.60)	0.683
	0.95, 0.95	1	94.94 (2.56, 2.50)	0.197	94.56 (2.78, 2.66)	0.206
	0.95, 0.95	2	95.48 (2.31, 2.21)	0.176	95.14 (2.49, 2.37)	0.184
	0.95, 0.85	1	95.38 (1.82, 2.80)	0.367	95.13 (1.85, 3.02)	0.382
	0.95, 0.85	2	95.70 (1.74, 2.56)	0.335	95.52 (1.78, 2.70)	0.348
	0.95, 0.75	1	95.36 (2.23, 2.41)	0.510	94.99 (2.31, 2.70)	0.530
	0.95, 0.75	2	95.62 (2.05, 2.33)	0.479	95.24 (2.14, 2.62)	0.497
	0.95, 0.65	1	95.24 (2.00, 2.76)	0.620	94.85 (2.08, 3.07)	0.642
	0.95, 0.65	2	95.45 (2.10, 2.45)	0.590	95.12 (2.16, 2.72)	0.611
50	0.90, 0.90	1	94.71 (2.51, 2.78)	0.159	94.60 (2.58, 2.82)	0.161
	0.90, 0.90	2	95.03 (2.30, 2.67)	0.134	94.97 (2.32, 2.71)	0.136
	0.90, 0.80	1	95.48 (2.00, 2.52)	0.233	95.42 (2.02, 2.56)	0.236
	0.90, 0.80	2	95.05 (2.51, 2.44)	0.207	94.98 (2.54, 2.48)	0.209
	0.90, 0.70	1	95.08 (2.13, 2.79)	0.306	95.00 (2.14, 2.86)	0.310
	0.90, 0.70	2	95.33 (1.96, 2.71)	0.279	95.28 (1.96, 2.76)	0.282
	0.90, 0.60	1	95.00 (2.50, 2.50)	0.371	94.95 (2.50, 2.55)	0.375
	0.90, 0.60	2	95.26 (2.06, 2.68)	0.345	95.16 (2.10, 2.74)	0.349
	0.95, 0.95	1	95.00 (2.58, 2.42)	0.083	94.93 (2.64, 2.43)	0.084
	0.95, 0.95	2	95.14 (2.22, 2.64)	0.069	95.08 (2.27, 2.65)	0.070
	0.95, 0.85	1	94.97 (2.39, 2.64)	0.171	94.88 (2.42, 2.70)	0.173
	0.95, 0.85	2	95.13 (2.23, 2.64)	0.154	95.01 (2.24, 2.75)	0.155
	0.95, 0.75	1	95.26 (2.11, 2.63)	0.254	95.15 (2.14, 2.71)	0.257
	0.95, 0.75	2	95.18 (2.31, 2.51)	0.237	95.08 (2.32, 2.60)	0.240
	0.95, 0.65	1	95.11 (2.22, 2.67)	0.327	95.05 (2.25, 2.70)	0.330
	0.95, 0.65	2	95.06 (2.15, 2.79)	0.312	94.90 (2.20, 2.90)	0.315
			$k_1 = k_2 = 6$			
			Fisher z-transformation		F-distribution	
			Cover (L, R) per cent	Width	Cover (L, R) per cent	Width
15	0.90, 0.90	1	94.92 (2.45, 2.63)	0.239	94.55 (2.62, 2.83)	0.214
	0.90, 0.90	2	96.30 (1.83, 1.87)	0.172	94.68 (2.48, 2.84)	0.141
	0.90, 0.80	1	94.74 (2.84, 2.42)	0.329	93.95 (2.16, 3.89)	0.299
	0.90, 0.80	2	95.92 (2.96, 1.12)	0.247	95.35 (1.76, 2.89)	0.215
	0.90, 0.70	1	95.01 (3.00, 1.99)	0.395	94.09 (2.09, 3.82)	0.367
	0.90, 0.70	2	95.97 (3.05, 0.98)	0.314	95.79 (1.55, 2.66)	0.288
	0.90, 0.60	1	94.97 (3.17, 1.86)	0.441	94.33 (2.00, 3.67)	0.416

DOI: 10.1002/sim

Table I. *Continued*.

	0.90, 0.60	2	96.45 (2.69, 0.86)	0.364	95.93 (1.42, 2.65)	0.343
	0.95, 0.95	1	94.90 (2.62, 2.48)	0.134	94.52 (2.81, 2.67)	0.119
	0.95, 0.95	2	96.73 (1.60, 1.67)	0.095	94.54 (2.65, 2.81)	0.077
	0.95, 0.85	1	95.31 (2.99, 1.70)	0.248	94.42 (2.07, 3.51)	0.227
	0.95, 0.85	2	95.01 (4.23, 0.76)	0.199	95.20 (2.21, 2.59)	0.179
	0.95, 0.75	1	95.14 (3.35, 1.51)	0.339	94.40 (2.13, 3.47)	0.316
	0.95, 0.75	2	95.14 (3.93, 0.93)	0.289	95.53 (1.92, 2.55)	0.270
	0.95, 0.65	1	94.66 (3.72, 1.62)	0.401	93.94 (2.40, 3.66)	0.381
	0.95, 0.65	2	94.83 (4.23, 0.94)	0.356	95.35 (2.14, 2.51)	0.339
50	0.90, 0.90	1	94.74 (2.74, 2.52)	0.110	94.63 (2.79, 2.58)	0.106
	0.90, 0.90	2	95.55 (2.22, 2.23)	0.070	95.13 (2.45, 2.42)	0.065
	0.90, 0.80	1	94.95 (2.71, 2.34)	0.160	94.69 (2.35, 2.96)	0.155
	0.90, 0.80	2	95.35 (2.86, 1.79)	0.114	95.13 (1.97, 2.90)	0.109
	0.90, 0.70	1	95.10 (2.67, 2.23)	0.202	94.77 (2.14, 3.09)	0.197
	0.90, 0.70	2	95.55 (2.81, 1.64)	0.157	95.52 (1.87, 2.61)	0.153
	0.90, 0.60	1	95.24 (2.66, 2.10)	0.234	95.03 (2.20, 2.77)	0.230
	0.90, 0.60	2	95.09 (3.21, 1.70)	0.191	95.05 (2.23, 2.72)	0.188
	0.95, 0.95	1	94.60 (2.64, 2.76)	0.058	94.51 (2.67, 2.82)	0.056
	0.95, 0.95	2	95.63 (2.24, 2.13)	0.037	94.92 (2.62, 2.46)	0.034
	0.95, 0.85	1	94.80 (3.00, 2.20)	0.119	94.61 (2.29, 3.10)	0.116
	0.95, 0.85	2	94.71 (3.76, 1.53)	0.094	94.79 (2.53, 2.68)	0.091
	0.95, 0.75	1	94.68 (3.12, 2.20)	0.173	94.61 (2.32, 3.07)	0.170
	0.95, 0.75	2	94.90 (3.61, 1.49)	0.148	95.23 (2.30, 2.47)	0.145
	0.95, 0.65	1	95.17 (3.03, 1.80)	0.213	95.18 (2.28, 2.54)	0.210
	0.95, 0.65	2	94.95 (3.53, 1.52)	0.190	95.20 (2.39, 2.41)	0.188

*Value of 1 denotes $\rho_{12}=0.5(\sqrt{\rho_1\rho_2}-0.05)$ and 2 denotes $\rho_{12}=\sqrt{\rho_1\rho_2}-0.05$. Such ρ_{12} was used to ensure positive definite variance–covariance matrix. $^{+}$L: the interval lies completely below the parameter and R: the interval lies completely above the parameter.

variance estimates from the confidence limits for a single ICC. In the spirit of score-type confidence intervals, the method accounts naturally for the skewness of the sampling distributions of the separate ICCs.

In addition to its use as a tool for primary data analysis, the method presented here can be applied to secondary data analysis in which one does not have access to the raw data. This presents an alternative to indirect procedures that are sometimes used such as judging significance at the 5 per cent level for a difference in reliability between two devices based on whether two 84 per cent confidence intervals overlap [21].

The method presented here may also be easily extend to unequal class sizes. Since the MOVER derives its validity from that of confidence limits for a single ICC, all that needed then is to have a reliable procedure for an ICC arising from unequal class sizes. For this purpose, Thomas and Hultquist [22] have proposed using the harmonic mean of the class size and unweighted sum of squares of class means in the calculations. Further research [23] has shown that this procedure performs well for construction of confidence limits about a single ICC.

We stress that the validity of the MOVER relies on that of the confidence limits for single ICCs, which in turn rests on the assumption of approximate normality. Thus, when this assumption becomes unreasonable, alternative confidence interval methods should be considered. One

DOI: 10.1002/sim

Table II. Two-sided 95 per cent confidence intervals for a difference between intraclass correlation coefficients constructed using the method of variance estimates recovery from four approaches to single intraclass correlation coefficients.

Method for single ρ	$l_1\ {}_{\hat{\rho}_1}\ u_1$	$l_2\ {}_{\hat{\rho}_2}\ u_2$	$L\ {}_{(\hat{\rho}_1-\hat{\rho}_2)}\ U$ by equation (7)
Example 1 [6, 8]			
$n=50, k_1=k_2=2$, $\widehat{\text{corr}}(\hat{\rho}_1,\hat{\rho}_2)=0.246$			
Simple asymptotic	$0.991\ {}_{0.994}\ 0.997$	$0.600\ {}_{0.730}\ 0.860$	$0.135\ {}_{0.264}\ 0.393$
Fisher's z	$0.989\ {}_{0.994}\ 0.997$	$0.569\ {}_{0.730}\ 0.837$	$0.158\ {}_{0.264}\ 0.424$
Arctanh	$0.990\ {}_{0.994}\ 0.997$	$0.572\ {}_{0.730}\ 0.836$	$0.159\ {}_{0.264}\ 0.421$
Modified Fisher's z	$0.990\ {}_{0.994}\ 0.997$	$0.599\ {}_{0.730}\ 0.838$	$0.157\ {}_{0.264}\ 0.414$
F-distribution	$0.989\ {}_{0.994}\ 0.997$	$0.570\ {}_{0.730}\ 0.837$	$0.158\ {}_{0.264}\ 0.423$
Example 2 [7, 9]			
$n=34, k_1=k_2=5$, $\widehat{\text{corr}}(\hat{\rho}_1,\hat{\rho}_2)=0.434$			
Simple asymptotic	$0.972\ {}_{0.982}\ 0.992$	$0.921\ {}_{0.948}\ 0.975$	$0.013\ {}_{0.034}\ 0.055$
Fisher's z	$0.969\ {}_{0.982}\ 0.990$	$0.913\ {}_{0.948}\ 0.969$	$0.019\ {}_{0.034}\ 0.064$
Arctanh	$0.969\ {}_{0.982}\ 0.989$	$0.913\ {}_{0.948}\ 0.969$	$0.019\ {}_{0.034}\ 0.064$
Modified Fisher's z	$0.971\ {}_{0.982}\ 0.990$	$0.917\ {}_{0.948}\ 0.970$	$0.018\ {}_{0.034}\ 0.060$
F-distribution	$0.971\ {}_{0.982}\ 0.990$	$0.917\ {}_{0.948}\ 0.971$	$0.017\ {}_{0.034}\ 0.060$

possibility is the bootstrap, although evaluation of this approach is beyond the scope of the present study.

ACKNOWLEDGEMENTS

The paper was developed from the first author's thesis submitted in partial fulfillment of the requirements for the degree of Master of Science, Department of Epidemiology and Biostatistics, University of Western Ontario, Canada, July 2008. The work of Drs Donner and Zou was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. Dr Zou is a recipient of the Early Researcher Award of Ontario Ministry of Research and Innovation, Canada.

REFERENCES

1. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 1966; **19**:3–11.
2. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* 1979; **86**:420–428.
3. Donner A. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* 1986; **54**:67–82.
4. McGraw KO, Wong S. Forming inference about some intraclass correlation coefficients. *Psychological Methods* 1996; **1**:30–46.
5. Alsawalmeh YM, Feldt LS. Testing the equality of two related intraclass reliability coefficients. *Applied Psychological Measurement* 1994; **18**:183–190.
6. Donner A, Zou GY. Testing the equality of dependent intraclass correlation coefficients. *The Statistician* 2002; **51**:367–379.
7. Giraudeau B, Porcher R, Mary JY. Power calculation for the likelihood ratio-test when comparing two dependent intraclass correlation coefficients. *Computer Methods and Programs in Biomedicine* 2005; **77**:165–173.
8. Turner SW, Toone BK, Brett-Jones JR. Computerized tomographic scan changes in early schizophrenia-preliminary finding. *Psychological Medicine* 1986; **16**:219–225.

DOI: 10.1002/sim

9. Gomez MA, Defontaine M, Giraudeau B, Camus E, Colin L, Laugier P, Patat F. In vivo performance of a matrix-based quantitative ultrasound imagine device dedicated to calcaneus investigation. *Ultrasound in Medicine and Biology* 2002; **28**:1285–1293.
10. Zou GY, Donner A. Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* 2008; **27**:1693–1702.
11. Elston RC. On the correlation between correlations. *Biometrika* 1975; **62**:133–140.
12. Smith CBA. On the estimation of intraclass correlation. *Annals of Human Genetics* 1956; **21**:363–373.
13. Fisher RA. *Statistical Methods for Research Workers* (14th edn). Hafner Publishing Company: New York, 1973.
14. Lachin JM. The role of measurement reliability in clinical trials. *Clinical Trials* 2004; **1**:553–566.
15. Konishi S. Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics* 1985; **37**:87–94.
16. Cochran WG. Errors of measurement in statistics. *Technometrics* 1968; **10**:637–666.
17. Bartlett MS. Approximate confidence intervals. 2. More than one unknown parameter. *Biometrika* 1953; **40**: 306–317.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–174.
19. Efron B. Second thoughts on the bootstrap. *Statistical Science* 2003; **18**:135–140.
20. Dunn GA. *Design and Analysis of Reliability Studies*: *The Statistical Evaluation of Measurement Errors*. Edward Arnold: London, U.K., 1989.
21. Robinson CN, van Aswegen A, Julious SA, Nunan TO, Thomson WH, Tindale WB, Tout DA, Underwood SR. The relationship between administered radiopharmaceutical activity in myocardinal perfusion scintigraphy and imaging outcomes. *European Journal of Nuclear Medicine and Molecular Imaging* 2008; **35**:329–335.
22. Thomas JD, Hultquist RA. Interval estimation for the unbalanced case of the one-way random effects model. *Annals of Statistics* 1978; **6**:582–587.
23. Donner A, Wells G. A comparison of confidence interval methods for the intraclass correlation coefficients. *Biometrics* 1986; **42**:401–412.

DOI: 10.1002/sim

Computational Statistics and Data Analysis 54 (2010) 55–64

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Confidence interval estimation under inverse sampling

G.Y. Zou *

Department of Epidemiology and Biostatistics, Robarts Clinical Trials of Robarts Research Institute, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada N6A 5C1
Department of Epidemiology and Biostatistics, School of Public Health, Southeast University, Nanjing, Jiangsu Province, PR China

ARTICLE INFO

Article history:
Received 7 December 2008
Received in revised form 16 June 2009
Accepted 9 July 2009
Available online 18 July 2009

ABSTRACT

In comparative studies of rare events, fixing group sizes may result in groups with zero events. To overcome this difficulty, one may adopt an inverse sampling design which fixes the number of events, resulting in random variables following the negative binomial distribution. This article presents a new approach to setting confidence intervals for effect measures under inverse sampling, using the variance estimates recovered from exact confidence limits for single negative binomial proportions. Exact numerical evaluation results demonstrate that the proposed procedure performs well.

1. Introduction

In comparative studies of rare events, a conventional design of pre-specifying group sizes may result in 2×2 tables with zero cell(s). As such data can complicate statistical inference, one may instead adopt an inverse sampling design in which sampling is continued until a pre-specified number of events are seen (Haldane, 1945). Unlike conventional binomial sampling, the numbers of non-events are random variables and follow negative binomial distribution.

Lui (2004) presents a comprehensive review of statistical inference procedures for inverse sampling throughout his text on statistical estimation of epidemiological risks, focusing on confidence interval estimation. Recent work on this topic includes confidence intervals for relative risk (Tian et al., 2008) and risk difference (Tang and Tian, 2009).

Most of these procedures rely explicitly on the likelihood function of a negative binomial distribution. Since the kernel of likelihood function of a negative binomial is identical to that of a binomial distribution, these procedures fail to account for the design explicitly. This falls right into the problem that 'when analyzing data, not taking into account how they were collected can lead to an inaccurate or erroneous conclusion' (Hirji, 2006, p. 50).

To fully capture the feature of inverse sampling, I propose to use the negative binomial distribution in constructing confidence intervals for common effect measures in 2×2 tables under inverse sampling. Specifically, I first obtain exact confidence limits for each comparison group based on the F-distribution (Lui, 1995; Casella and Berger, 1990, p. 449), and then use these limits to recover variance estimates needed for risk difference as done elsewhere (Zou and Donner, 2008; Zou, 2008; Zou et al., 2009a). Since the relative risk and odds ratio can be parameterized as differences on the log scale, this approach also lends itself naturally to these effect measures.

There is a widespread notion that exact confidence intervals for discrete distribution parameters are conservative, as illustrated by the memorable title that '[a]pproximate is better than "exact" for interval estimation of binomial proportions' (Agresti and Coull, 1998). However, I show below that the exact confidence interval based on the negative binomial distribution is always shorter than that from the binomial distribution. This implies that under inverse sampling,

* Corresponding address: Robarts Clinical Trials, Robarts Research Institute, P.O. Box 5015, 100 Perth Drive, London, Ontario, Canada N6A 5K8. Tel.: +1 519 663 5777x24092; fax: +1 519 931 5705.
E-mail address: gzou@robarts.ca.

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.07.010

Table 1
Notation and definition of effect measures under an inverse sampling design.

Group		Events	Non-events	Total	Risk estimate
1		r_1	$Y_1 \sim NBin(r_1, p_1)$	$n_1 = r_1 + Y_1$	$\widehat{p_1} = \frac{r_1}{n_1}$
2		r_2	$Y_2 \sim NBin(r_2, p_2)$	$n_2 = r_2 + Y_2$	$\widehat{p_2} = \frac{r_2}{n_2}$
Risk difference	$p_1 - p_2$				
Relative risk	$p_1/p_2 = \exp(\log p_1 - \log p_2)$				
Odds ratio	$\frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \exp[\log\{p_1/(1-p_1)\} - \log\{p_2/(1-p_2)\}]$				

confidence intervals for effect measures obtained by recovering variance estimates using the exact confidence limits may perform well.

The rest of the article is organized as follows. Section 2 presents the notation and terminology, followed by Section 3 where I present confidence interval procedures for single negative binomial proportions and risk difference, relative risk and odds ratio. In Section 4, I conduct an exact numerical evaluation to compare the new approach to some existing methods for setting confidence intervals for risk difference. The evaluation criteria considered are overall coverage, balance of tail errors, and confidence interval width. Section 5 presents an illustrative example. The article ends with a brief discussion.

2. Notation and terminology

Consider a study involving two comparative groups, with group 1 denoting exposed and group 0 unexposed. The key feature of the inverse sampling design is that the number of events r_i ($i = 1, 2$) is pre-specified. Let Y_i denote the number of non-events to ensure that r_i pre-specified events are observed. The probability mass function of Y_i is given by

$$f_{Y_i}(y_i) = \Pr(Y_i = y_i|p_i) = \binom{y_i + r_i - 1}{y_i} p_i^{r_i}(1 - p_i)^{y_i}, \quad y_i = 0, 1, 2, \ldots.$$

The observed data may be presented as in a 2×2 table (Table 1). For the development in this article, I have also re-written relative effect measures as differences on the log scale in Table 1. The kernel of the likelihood function in terms of p_1 and p_2 is given by

$$L = p_1^{r_1}(1 - p_1)^{y_1} p_2^{r_2}(1 - p_2)^{y_2},$$

which can be recognized to be identical to that of a conventional binomial sampling design. In other word, L does not capture completely the fact that the last observation in group i must be an event in inverse sampling.

3. Confidence interval estimation

3.1. Single negative binomial proportion

Lui (2004, pp. 8–10) presents three procedures for constructing $100(1 - \alpha)\%$ two-sided confidence intervals for a single negative binomial proportion. The presentation here uses the notation in Table 1. Let $z_{\alpha/2}$ denote the upper $\alpha/2$ quantile of the standard normal distribution. The first procedure is the Wald method, given by

$$\begin{aligned}(l_i^m, u_i^m) &= \widehat{p}_i \mp z_{\alpha/2}\sqrt{\frac{\widehat{p}_i^2(1 - \widehat{p}_i)}{r_i}} \\ &= \widehat{p}_i \mp z_{\alpha/2}\sqrt{\frac{\widehat{p}_i(1 - \widehat{p}_i)}{n_i}},\end{aligned}$$

which is identical to the Wald method for a single binomial proportion (Agresti and Coull, 1998). For cases where the interval (l_i^m, u_i^m) contains values outside the parameter space of (0, 1), the limits are usually truncated.

The second method is based on the uniformly minimum variance unbiased estimator (UMVUE), given by

$$(l_i^u, u_i^u) = \widehat{p}_i^u \mp z_{\alpha/2}\sqrt{\frac{\widehat{p}_i^u(1 - \widehat{p}_i^u)}{n_i - 2}},$$

where

$$\widehat{p}_i^u = \frac{r_i - 1}{r_i + y_i - 1} = \frac{r_i - 1}{n_i - 1}.$$

Note that this interval can only be calculated for $r_i > 2$. Again, one must truncate the limits when inadmissible values are contained in the interval.

The third interval presented by Lui (2004, p. 9) is derived directly from the negative binomial distribution, using its relationship with the binomial distribution (Lui, 1995; Casella and Berger, 1990, p. 449), and given by

$$l_i^e = \frac{1}{1 + (y_i + 1)/r_i \times F_{1-\alpha/2,2(y_i+1),2r_i}}$$

and

$$u_i^e = \frac{1}{1 + y_i/r_i \times F_{1-\alpha/2,2y_i,2r_i}},$$

where $F_{q,df1,df2}$ is the qth quantile of an F-distribution with degrees of freedom $df1$ and $df2$. For $y_i = 0$, u_i^e is set to 1.

Comparing (l_i^e, u_i^e) with the exact confidence interval for a binomial proportion from observed data of r_i events and y_i non-events (Agresti and Coull, 1998; Newcombe, 1998b), one can see that the lower limits are the same, but u_i^e here is smaller than the upper limit of a binomial proportion, which is given by

$$u = \frac{1}{1 + y_i/(r_i + 1) \times F_{\alpha/2,2y_i,2(r_i+1)}}.$$

Note that u is the upper limit for the case when there are $r_i + 1$ events and y_i non-events under inverse sampling. This property has previously been recognized for $r_i = 1$ (George and Elston, 1993). The evaluation below shows that the exact confidence interval for negative binomial proportions is very accurate in terms of coverage, in contrast to the case of binomial proportions (Agresti and Coull, 1998).

3.2. Difference between two independent negative binomial proportions

For the data layout in Table 1, common effect measures are the risk difference, the relative risk and the odds ratio. As shown in Table 1, all of them can be formed as a difference. Thus, I only focus on risk difference $p_1 - p_2$, after presenting a general method for a difference.

Zou and Donner (2008) have presented a general approach to confidence interval construction for the difference between two parameters. Specific applications of this approach have appeared in a variety of occasions, notably Method 10 of Newcombe (1998a). As shown by Zou and Donner (2008), the basis of this approach is to use readily available confidence limits to recover variance estimates. Zou (2008) referred to it as the MOVER (method of variance estimates recovery), and extended it to linear functions of parameters with dependent point estimates. With the exception of the increment in R^2, the coefficient of determination, in multiple linear regression (Zou, 2007), the MOVER has been successful in a wide range of applications (Zou and Donner, 2008; Zou et al., 2009a,b).

We now present a summary for $\theta_1 - \theta_2$, where the corresponding estimators $\widehat{\theta}_i$, $i = 1, 2$, are assumed to be independent. By the duality of confidence interval estimation and hypothesis testing that the confidence interval (L, U) for $\theta_1 - \theta_2$ contains all parameter values that cannot be rejected at the α level, for reasonable sample size, L is the parameter value that satisfies

$$\frac{\widehat{\theta}_1 - \widehat{\theta}_2 - L}{\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1) + \widehat{\mathrm{var}}(\widehat{\theta}_2)}} \approx z_{\alpha/2}$$

and U is the parameter value that satisfies

$$\frac{U - (\widehat{\theta}_1 - \widehat{\theta}_2)}{\sqrt{\widehat{\mathrm{var}}(\widehat{\theta}_1) + \widehat{\mathrm{var}}(\widehat{\theta}_2)}} \approx z_{\alpha/2}.$$

The performance of the confidence interval depends on how variance estimates are obtained. This is particularly important when the variance is related to underlying parameter values, as in the case of a binomial proportion $\widehat{p}$, where $\mathrm{var}(\widehat{p}) = p(1-p)/n$. A common approach is to plug-in the point estimate in the variance formula, resulting in a symmetric confidence interval which is known to have inferior performance in finite samples. One can alternatively adapt the idea in Bartlett (1953, p. 15) and estimate variances for L and U in the neighborhood of L and U. The idea of estimating variances at confidence limits has also been endorsed by Efron (1987, p. 175).

Now, among all the plausible parameter values of θ_1 provided by (l_1, u_1) and that of θ_2 by (l_2, u_2), the distance between $l_1 - u_2$ and L is smaller than that of $\widehat{\theta}_1 - \widehat{\theta}_2$ and L. Likewise, the distance between $u_1 - l_2$ and U is smaller than that of $\widehat{\theta}_1 - \widehat{\theta}_2$ and U. Thus, one can obtain variance estimates needed for L at $\theta_1 = l_1$ and $\theta_2 = u_2$ and that for U at $\theta_1 = u_1$ and $\theta_2 = l_2$. Compared with Bartlett (1953), the difference is that the variances here are estimated not exactly at, but in the neighborhoods of L and U.

By again the general principle that (l_i, u_i) contains all parameter values of θ_i that cannot be rejected at the α level, and the central limit theorem,

$$\frac{\widehat{\theta}_i - l_i}{\sqrt{\mathrm{var}(\widehat{\theta}_i)}} \approx z_{\alpha/2} \Longleftrightarrow \widehat{\mathrm{var}}(\widehat{\theta}_i) \approx \frac{(\widehat{\theta}_i - l_i)^2}{z_{\alpha/2}^2}$$

at $\theta_i = l_i$ and

$$\frac{u_i - \widehat{\theta}_i}{\sqrt{\text{var}(\widehat{\theta}_i)}} \approx z_{\alpha/2} \Longleftrightarrow \widehat{\text{var}}(\widehat{\theta}_i) \approx \frac{(u_i - \widehat{\theta}_i)^2}{z^2_{\alpha/2}}$$

at $\theta_i = u_i$. With these variance estimates, the lower limit for $\theta_1 - \theta_2$ is given by

$$\begin{aligned} L &\approx \widehat{\theta}_1 - \widehat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\widehat{\theta}_1) + \widehat{\text{var}}(\widehat{\theta}_2)} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 - z_{\alpha/2}\sqrt{(\widehat{\theta}_1 - l_1)^2/z^2_{\alpha/2} + (u_2 - \widehat{\theta}_2)^2/z^2_{\alpha/2}} \\ &= \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2}. \end{aligned} \tag{1}$$

Similar steps result in the upper limit as

$$U \approx \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2}. \tag{2}$$

Note that the objective here is to construct a confidence interval for $\theta_1 - \theta_2$, not simultaneous confidence intervals for θ_1, θ_2 and $\theta_1 - \theta_2$. The latter problem would require not the standard normal quantile $z_{\alpha/2}$, but quantile value, $z'_{\alpha/2}$, from a tri-variate normal distribution with mean vector (0, 0, 0) and variance–covariance given by

$$\begin{pmatrix} 1 & 0 & \text{var}(\widehat{\theta}_1)/\{\text{var}(\widehat{\theta}_1)[\text{var}(\widehat{\theta}_1 - \widehat{\theta}_2)]\}^{1/2} \\ \cdot & 1 & -\text{var}(\widehat{\theta}_2)/\{\text{var}(\widehat{\theta}_2)[\text{var}(\widehat{\theta}_1 - \widehat{\theta}_2)]\}^{1/2} \\ \cdot & \cdot & 1 \end{pmatrix}.$$

This is because $\text{cov}(\widehat{\theta}_1, \widehat{\theta}_1 - \widehat{\theta}_2) = \text{var}(\widehat{\theta}_1)$ and $\text{cov}(\widehat{\theta}_2, \widehat{\theta}_1 - \widehat{\theta}_2) = -\text{var}(\widehat{\theta}_2)$. It is interesting to note that the simultaneous confidence interval for $\theta_1 - \theta_2$ still have the same form as Eqs. (1) and (2), with (l_i, u_i) obtained using $z'_{\alpha/2}$.

Substituting into Eqs. (1) and (2) point estimates and the confidence limits for single negative binomial proportions discussed in Section 3.1 results in three procedures for $p_1 - p_2$. Specifically, with the maximum likelihood estimator for p_i, we have

$$(L^m, U^m) \approx \widehat{p}_1 - \widehat{p}_2 \mp z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}},$$

which is identical to the MLE method in Lui (1999) for inverse sampling and the Wald method in Agresti and Caffo (2000) for binomial sampling. Plugging in the UMVUE estimator for p_i results in what Lui (1999) referred to as the UMVUE, i.e.,

$$(L^u, U^u) \approx \widehat{p}^u_1 - \widehat{p}^u_2 \mp z_{\alpha/2}\sqrt{\frac{\widehat{p}^u_1(1-\widehat{p}^u_1)}{n_1 - 2} + \frac{\widehat{p}^u_2(1-\widehat{p}^u_2)}{n_2 - 2}}.$$

Despite Lui (1999) recommends the above two procedures, it is well-known that symmetrical confidence intervals in general perform poorly (Newcombe, 1998a; Agresti and Caffo, 2000). In fact, it has been stated that '[t]he most serious errors made by standard intervals are due to their enforced symmetry' (Efron and Tibshirani, 1993, p. 180). In response, Tang and Tian (2009) proposed an asymmetric confidence interval procedure based on the score statistic. However, these confidence limits must be obtained by an iterative algorithm, which could be the obstacle for its wide application. A similar comment can also be made on the confidence interval based on likelihood ratio test. In addition, Tang and Tian (2009) has shown that the likelihood ratio test based interval has inferior performance as compared to the score method. As an alternative, I apply the MOVER with the exact confidence limits, resulting in the following closed-form confidence interval for $p_1 - p_2$

$$\begin{cases} L^e \approx \widehat{p}_1 - \widehat{p}_2 - \sqrt{(\widehat{p}_1 - l^e_1)^2 + (u^e_2 - \widehat{p}_2)^2} \\ U^e \approx \widehat{p}_1 - \widehat{p}_2 + \sqrt{(u^e_1 - \widehat{p}_1)^2 + (\widehat{p}_2 - l^e_2)^2} \end{cases}$$

where l^e_i, u^e_i are the exact limits as given in Section 3.1.

3.3. Relative risk and odds ratio

Given a confidence interval for p, confidence intervals for $\log(p)$ and $\log\{p/(1-p)\}$ are readily available with the transformation principle (Daly, 1998). This principle states that given a confidence interval for θ as (l, u), the corresponding confidence interval for $f(\theta)$ is $[f(l), f(u)]$ if $f(\theta)$ is a monotone increasing function, or $[f(u), f(l)]$ if $f(\theta)$ is a monotone deceasing function.

With the transformation principle (Daly, 1998) and the limits from MOVER in Eqs. (1) and (2), it is straightforward to obtain confidence intervals for the relative risk and the odds ratio. This is because both measures can also be formed as

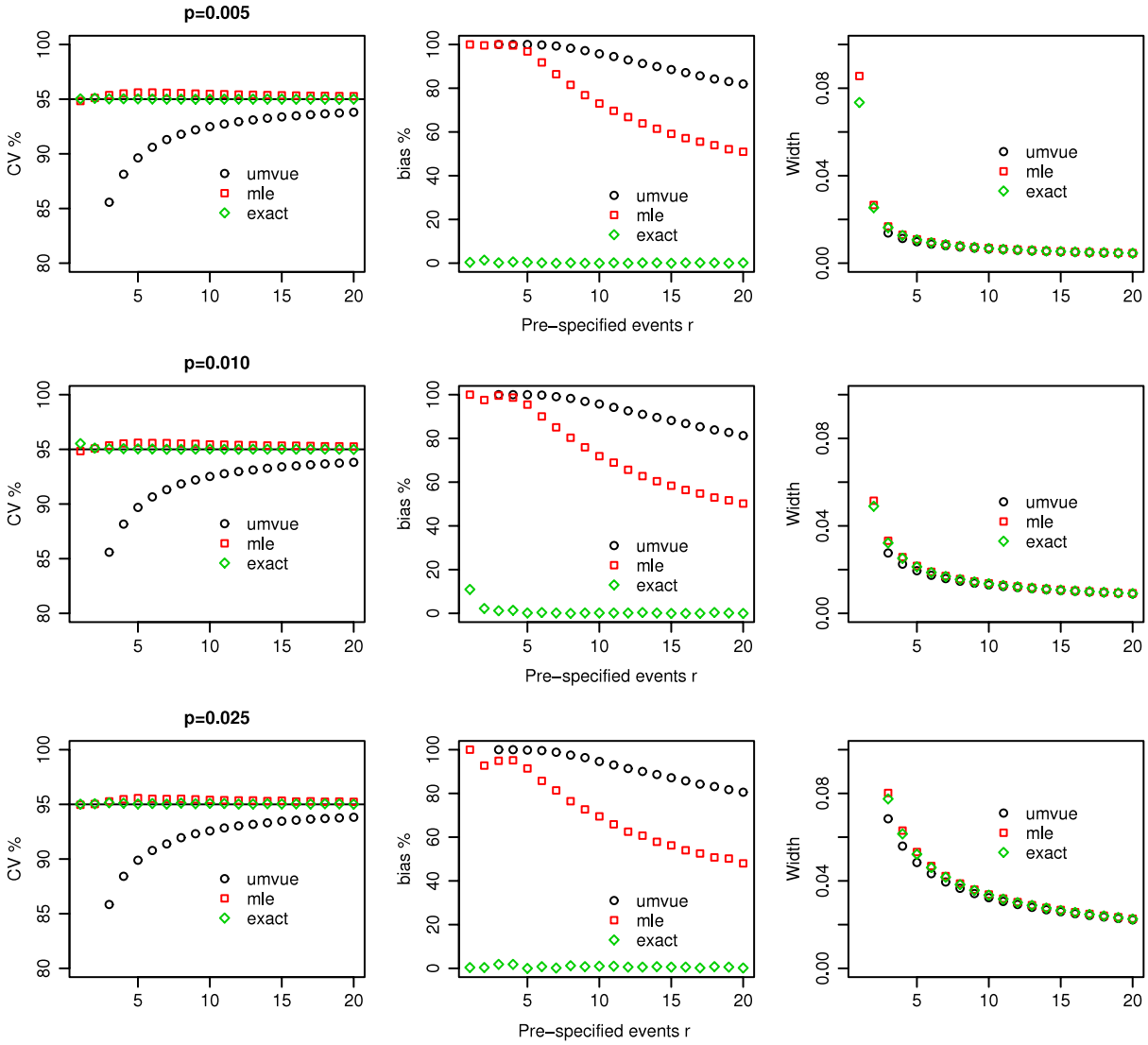


Fig. 1. Performance of 3 procedures in constructing two-sided 95% confidence intervals for single negative binomial proportions, based on exact numerical evaluation with extreme right tail truncated at 1.0E-5. Bias (%) is defined by percentage of absolute difference between left and right tail errors in terms of their sum. For the UMVUE, the evaluation was done for $r > 2$. p: event probability.

differences on the log scale (see Table 1). The performance of this approach to setting confidence limits for relative risk under binomial sampling has been evaluated by Zou and Donner (2008).

Since confidence intervals for p_i are a necessary step for effect measures, an additional advantage of the MOVER is that it promotes the reporting of group risks and associated precision, and thus puts the interpretation of each effect measure in its proper context.

4. Exact numerical evaluation

To evaluate the finite sample performance of the three confidence intervals for a single negative binomial proportion in Section 3.1, I conducted numerical evaluation by considering overall coverage percentage (CV%), and tail errors in terms of missing the parameter form its left (ML%) and right (MR%). I also computed the expect confidence interval width.

Given a confidence interval (l_i, u_i), these were defined as

$$CV\% = 100 \sum_{y=0}^{\infty} I[p \in (l_i, u_i)] f_{Y_i}(y_i).$$

$$ML\% = 100 \sum_{y=0}^{\infty} I[u_i < p] f_{Y_i}(y_i)$$

$$MR\% = 100 \sum_{y=0}^{\infty} I[l_i > p] f_{Y_i}(y_i).$$

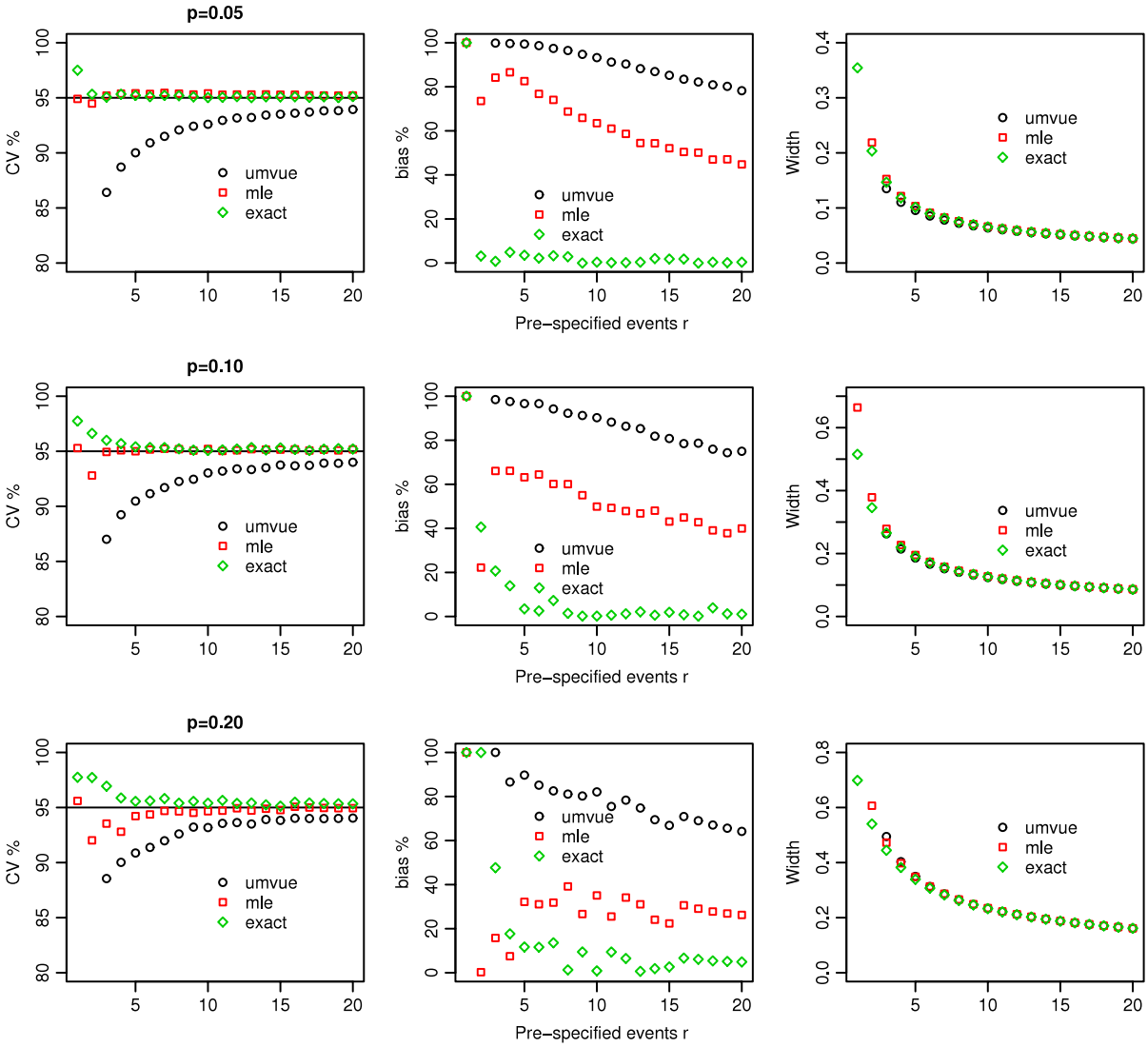


Fig. 2. Performance of 3 procedures in constructing two-sided 95% confidence intervals for single negative binomial proportions, based on exact numerical evaluation with extreme right tail truncated at 1.0E-5. Bias (%) is defined by percentage of absolute difference between left and right tail errors in terms of their sum. For the UMVUE, the evaluation was done for $r > 2$. p: event probability.

and

$$\text{Width} = \sum_{y=0}^{\infty} (u_i - l_i) f_{Y_i}(y_i),$$

where $I[\cdot]$ is an indicator function taking the value 1 if the condition in square brackets is satisfied, and 0 otherwise.

A similar approach was taken to evaluate the performance of the three procedures for $p_1 - p_2$ discussed in Section 3.2, with $f_{Y_i}(y_i)$ replaced by $f_{Y_1}(y_1) \times f_{Y_2}(y_2)$ in the above definitions.

Since it is impossible to sum up all possible values of y_i, I truncated the extreme right tail so that the errors of the calculated *CV%* are less than 0.00001 and 0.001 for single proportions and difference between proportions, respectively.

Figs. 1 and 2 show the performance of the three procedures for single negative binomial proportions for $r = 1$ to 20 by increments of 1 and $p = 0.005, 0.01, 0.025, 0.05, 0.1$, and 0.2. It is clear that, in addition to not being applicable to cases of $r = 1, 2$, the UMVUE performs poorly, with under nominal coverage, unbalanced tail errors and no advantage of interval width. The MLE (or Wald method) provides adequate coverage, but in a lop-sided fashion. In contrast to that of a binomial proportion, the exact confidence interval performs very well, with coverage very close to the nominal level and well-balanced tail errors.

Table 2 shows the comparative performance of confidence intervals for $p_1 - p_2$ when $r_1 = r_2 = r = 1, 2, 3, 5, 10$, using three intervals for single proportions to recover variance estimates under 45 parameter combinations ($3p_1 \times 3p_2 \times 5r$). From the observed coverage (%), and left and right tail errors (defined as missing the parameter from the left or right, respectively), as well as the expected interval width, it is clear that the MLE cannot be recommended for practical usage. On the other

Table 2
Comparative performance of three procedures for setting a 95% confidence interval for risk difference $p_1 - p_2$ under inverse sampling, based on exact numerical evaluation with extreme right tail truncated to ensure the coverage errors less than 0.001. The pre-specified numbers of events were set to $r_1 = r_2 = r$.

p_1	p_2	r	MLE		UMVUE		Exact	
			CV% (ML, MR)	Width	CV% (ML, MR)	Width	CV% (ML, MR)	Width
0.05	0.025	1	99.86 (0.05, 0)	0.5955			95.88 (4.02, 0.0)	0.4915
		2	98.93 (0.30, 0.67)	0.2710			94.53 (3.71, 1.66)	0.2572
		3	99.03 (0.54, 0.33)	0.1823	99.00 (0.89, 0.01)	0.1608	94.59 (3.44, 1.88)	0.1780
		5	97.64 (1.88, 0.39)	0.1201	97.77 (2.10, 0.04)	0.1107	94.72 (3.16, 2.03)	0.1190
		10	96.31 (2.72, 0.87)	0.0755	96.32 (3.23, 0.35)	0.0725	94.85 (2.91, 2.15)	0.0753
	0.050	1	99.90 (0, 0)	0.7098			98.69 (0.61, 0.61)	0.5824
		2	98.68 (0.61, 0.61)	0.3398			95.85 (2.03, 2.03)	0.3218
		3	99.24 (0.33, 0.33)	0.2312	99.88 (0.01, 0.01)	0.2066	95.45 (2.23, 2.23)	0.2259
		5	98.85 (0.53, 0.53)	0.1527	99.78 (0.06, 0.06)	0.1414	95.23 (2.34, 2.34)	0.1515
		10	97.32 (1.29, 1.29)	0.0958	98.29 (0.80, 0.80)	0.0921	95.11 (2.40, 2.40)	0.0957
	0.750	1	99.90 (0, 0)	0.8029			97.63 (0, 2.28)	0.6526
		2	98.05 (1.27, 0.58)	0.4065			95.49 (1.54, 2.88)	0.3820
		3	98.89 (0.65, 0.37)	0.2828	99.84 (0.04, 0.02)	0.2584	95.17 (1.83, 2.90)	0.2750
		5	98.21 (0.63, 1.07)	0.1897	99.44 (0.09, 0.37)	0.1774	95.07 (2.03, 2.80)	0.1878
		10	96.77 (1.09, 2.04)	0.1201	97.47 (0.53, 1.90)	0.1159	95.03 (2.17, 2.70)	0.1198
0.10	0.05	1	99.87 (0.04, 0)	0.8819			96.34 (3.57, 0)	0.7096
		2	97.18 (0.69, 2.04)	0.4685			95.12 (3.51, 1.27)	0.4366
		3	97.91 (0.95, 1.04)	0.3334	99.00 (0.81, 0.09)	0.3124	94.92 (3.32, 1.67)	0.3223
		5	97.02 (2.05, 0.84)	0.2276	97.66 (2.10, 0.14)	0.2155	94.92 (3.08, 1.91)	0.2246
		10	96.05 (2.71, 1.13)	0.1457	96.25 (3.14, 0.51)	0.1413	94.96 (2.86, 2.08)	0.1452
	0.10	1	99.91 (0, 0)	1.0296			99.70 (0.10, 0.10)	0.8237
		2	96.55 (1.68, 1.68)	0.5750			96.59 (1.66, 1.66)	0.5334
		3	97.85 (1.03, 1.03)	0.4161	99.74 (0.08, 0.08)	0.3985	95.87 (2.02, 2.02)	0.4015
		5	97.90 (1.00, 1.00)	0.2865	99.48 (0.21, 0.21)	0.2737	95.48 (2.21, 2.21)	0.2827
		10	96.83 (1.54, 1.54)	0.1837	97.93 (0.99, 0.99)	0.1787	95.25 (2.33, 2.33)	0.1832
	0.15	1	99.91 (0, 0)	1.1426			98.21 (0, 1.69)	0.9051
		2	95.81 (2.61, 1.48)	0.6695			96.47 (0.96, 2.48)	0.6155
		3	96.97 (1.82, 1.13)	0.4965	99.59 (0.22, 0.10)	0.4930	95.72 (1.53, 2.66)	0.4761
		5	97.00 (1.37, 1.53)	0.3492	98.91 (0.35, 0.64)	0.3397	95.41 (1.85, 2.65)	0.3434
		10	96.22 (1.51, 2.17)	0.2270	97.12 (0.80, 1.97)	0.2226	95.22 (2.07, 2.61)	0.2262
0.20	0.10	1	99.88 (0.03, 0)	1.2333			97.07 (2.84, 0)	0.9673
		2	95.61 (1.60, 2.70)	0.7507			96.36 (3.14, 0.41)	0.6839
		3	95.43 (1.77, 2.71)	0.5688	98.96 (0.74, 0.22)	0.5877	95.55 (3.10, 1.25)	0.5421
		5	95.70 (2.30, 1.91)	0.4089	97.23 (2.09, 0.58)	0.4065	95.33 (2.93, 1.65)	0.4009
		10	95.48 (2.70, 1.72)	0.2702	96.06 (2.96, 0.89)	0.2676	95.20 (2.78, 1.93)	0.2691
	0.20	1	99.92 (0, 0)	1.4056			99.92 (0, 0)	1.1014
		2	94.38 (2.77, 2.77)	0.8941			97.93 (0.99, 0.99)	0.8113
		3	94.79 (2.56, 2.56)	0.6912	99.28 (0.32, 0.32)	0.7389	96.69 (1.61, 1.61)	0.6565
		5	95.89 (2.01, 2.01)	0.5049	98.41 (0.75, 0.75)	0.5099	95.96 (1.98, 1.98)	0.4941
		10	95.88 (2.01, 2.01)	0.3364	97.13 (1.39, 1.39)	0.3352	95.56 (2.17, 2.17)	0.3350
	0.30	1	99.92 (0, 0)	1.5246			99.00 (0, 0.92)	1.1886
		2	94.77 (2.47, 2.67)	1.0023			98.14 (0.02, 1.76)	0.9047
		3	93.54 (3.84, 2.53)	0.7901	99.52 (0.04, 0.36)	0.8916	96.86 (0.84, 2.22)	0.7478
		5	94.57 (3.01, 2.34)	0.5910	97.29 (1.37, 1.25)	0.6155	96.09 (1.45, 2.37)	0.5776
		10	95.13 (2.40, 2.38)	0.4022	96.30 (1.52, 2.09)	0.4066	95.60 (1.85, 2.47)	0.4008

hand, the confidence interval procedure based on the exact limits for single proportions performed very well, particularly for $r \geq 3$. Similar conclusion can be drawn for cases of $r_1 = r_2/2$ based on the results in Table 3.

The performance of the procedure based on UMVUE is less clear-cut. As shown in Table 2 for $r_1 = r_2$, it resulted in coverage percentages that are greater than the nominal level. This is associated with narrower width when p_i's are small and wider width when p_i's are large, relative to confidence intervals based on the exact limits for single proportions. When $r_1 \neq r_2$, results in Table 3 show that the coverage of the procedure based on UMVUE can range from 92.76% to 99.65% for the parameter combination considered. Compared to the procedure based on exact limits for single proportions, this procedure can result in narrower confidence intervals for small p_i's and wider intervals for large p_i's. In addition to its restrictive applicability to only $r_i > 2$, these simulation results do not support the recommendation of the procedure based on UMVUE (Lui, 1999).

Further evaluations in terms of relative risk reached similar conclusion. Detailed results are available upon request.

5. Example

I now illustrate the methodology using an example from Hirji (2006, p. 134), which was analyzed by Tian et al. (2008) in terms of relative risk and Tang and Tian (2009) in terms of risk difference. The study involved sampling until 5 events

Table 3
Comparative performance of three procedures for setting a 95% confidence interval for risk difference $p_1 - p_2$ under inverse sampling, based on exact numerical evaluation with extreme right tail truncated to ensure the coverage errors less than 0.001. The pre-specified numbers of events were set to $r_1 = r_2/2$.

p_1	p_2	r_2	MLE		UMVUE		Exact	
			CV% (ML, MR)	Width	CV% (ML, MR)	Width	CV% (ML, MR)	Width
0.05	0.025	2	99.32 (0.59, 0)	0.4813			96.39 (3.52, 0)	0.4014
		4	96.82 (2.36, 0.72)	0.2346			94.97 (3.11, 1.82)	0.2217
		6	96.45 (3.11, 0.35)	0.1630	92.76 (7.13, 0.01)	0.1454	94.91 (2.95, 2.04)	0.1584
		10	96.08 (3.42, 0.40)	0.1102	93.68 (6.19, 0.03)	0.1022	94.93 (2.80, 2.17)	0.1087
		20	95.63 (3.41, 0.86)	0.0705	94.44 (5.16, 0.30)	0.0679	94.96 (2.67, 2.27)	0.0701
	0.05	2	99.41 (0.50, 0)	0.5367			97.57 (1.93, 0.40)	0.4579
		4	98.83 (0.39, 0.69)	0.2656			95.62 (2.21, 2.07)	0.2540
		6	98.65 (0.90, 0.35)	0.1857	99.42 (0.47, 0.01)	0.1680	95.35 (2.28, 2.27)	0.1821
		10	97.81 (1.62, 0.47)	0.1263	98.14 (1.72, 0.04)	0.1182	95.19 (2.33, 2.38)	0.1253
		20	96.61 (2.20, 1.09)	0.0811	96.85 (2.57, 0.48)	0.0784	95.09 (2.37, 2.44)	0.0809
	0.10	2	98.88 (1.03, 0)	0.5918			96.91 (1.37, 1.62)	0.5110
		4	98.71 (0.53, 0.66)	0.3014			95.41 (1.78, 2.72)	0.2899
		6	98.80 (0.74, 0.37)	0.2135	99.65 (0.24, 0.01)	0.1961	95.20 (1.91, 2.79)	0.2100
		10	98.03 (1.18, 0.70)	0.1467	99.01 (0.82, 0.07)	0.1388	95.10 (2.03, 2.77)	0.1460
		20	96.67 (1.70, 1.54)	0.0950	97.40 (1.59, 0.91)	0.0924	95.05 (2.15, 2.70)	0.0950
0.10	0.05	2	99.00 (0.91, 0)	0.7330			96.56 (3.34, 0)	0.5976
		4	95.01 (2.55, 2.34)	0.4089			95.46 (3.02, 1.42)	0.3804
		6	95.65 (3.13, 1.13)	0.2988	93.01 (6.79, 0.10)	0.2826	95.20 (2.89, 1.82)	0.2878
		10	95.67 (3.36, 0.88)	0.2088	93.86 (5.91, 0.13)	0.1990	95.11 (2.76, 2.04)	0.2052
		20	95.43 (3.33, 1.14)	0.1359	94.52 (4.95, 0.44)	0.1322	95.06 (2.64, 2.20)	0.1351
	0.10	2	98.62 (1.29, 0)	0.8184			98.27 (1.62, 0.01)	0.6842
		4	96.95 (0.93, 2.02)	0.4635			96.18 (2.08, 1.65)	0.4366
		6	97.52 (1.29, 1.09)	0.3404	99.01 (0.81, 0.09)	0.3252	95.68 (2.18, 2.04)	0.3307
		10	97.12 (1.83, 0.96)	0.2389	97.82 (1.92, 0.16)	0.2293	95.40 (2.26, 2.24)	0.2362
		20	96.26 (2.28, 1.36)	0.1560	96.63 (2.63, 0.64)	0.1524	95.21 (2.33, 2.36)	0.1556
	0.15	2	97.71 (2.20, 0)	0.8960			98.05 (0.93, 0.93)	0.7581
		4	96.80 (1.31, 1.80)	0.5220			96.03 (1.60, 2.28)	0.4943
		6	97.46 (1.30, 1.14)	0.3881	99.23 (0.59, 0.08)	0.3760	95.57 (1.79, 2.54)	0.3783
		10	97.15 (1.54, 1.21)	0.2753	98.52 (1.15, 0.24)	0.2669	95.33 (1.95, 2.63)	0.2728
		20	96.25 (1.90, 1.76)	0.1813	97.05 (1.80, 1.05)	0.1780	95.19 (2.09, 2.62)	0.1811
0.20	0.10	2	98.22 (1.69, 0)	1.0537			96.88 (3.03, 0)	0.8391
		4	93.79 (2.80, 3.32)	0.6620			96.75 (2.89, 0.26)	0.6042
		6	93.86 (3.10, 2.94)	0.5113	93.69 (6.13, 0.10)	0.5320	95.79 (2.76, 1.36)	0.4869
		10	94.64 (3.22, 2.04)	0.3746	94.01 (5.35, 0.55)	0.3750	95.48 (2.66, 1.77)	0.3667
		20	94.96 (3.16, 1.79)	0.2513	94.59 (4.52, 0.80)	0.2499	95.29 (2.57, 2.05)	0.2500
	0.20	2	97.75 (2.17, 0)	1.1688			98.85 (1.07, 0)	0.9595
		4	94.70 (2.03, 3.18)	0.7479			97.31 (1.80, 0.79)	0.6923
		6	95.04 (1.99, 2.88)	0.5808	98.13 (1.55, 0.24)	0.6060	96.39 (1.98, 1.55)	0.5577
		10	95.68 (2.18, 2.05)	0.4271	96.98 (2.32, 0.61)	0.4290	95.76 (2.14, 2.01)	0.4201
		20	95.56 (2.42, 1.93)	0.2874	96.15 (2.75, 1.01)	0.2863	95.45 (2.25, 2.21)	0.2866
	0.30	2	98.03 (1.89, 0)	1.2574			99.45 (0.10, 0.37)	1.0486
		4	93.92 (2.86, 3.13)	0.8242			97.35 (1.20, 1.36)	0.7688
		6	94.72 (2.43, 2.77)	0.6483	98.07 (1.60, 0.25)	0.6854	96.33 (1.49, 2.10)	0.6254
		10	95.37 (2.28, 2.26)	0.4821	97.20 (1.96, 0.75)	0.4889	95.89 (1.72, 2.31)	0.4757
		20	95.46 (2.29, 2.16)	0.3273	96.32 (2.26, 1.33)	0.3279	95.49 (1.98, 2.44)	0.3272

occurred in each comparison group. In the end, group 1 had 53 non-events while group 2 had 312 non-events. The data and corresponding calculations are shown in Table 4. Both the MLE and UMVUE provide statistically non-significant results, while the confidence intervals based on exact limits show clear evidence of a difference between the two comparison groups.

To further check on the validity of the three confidence intervals for risk difference, I conducted an evaluation study using the same approach as in Section 4 for parameter values of $p_1 = 5/(5+53)$, $p_2 = 5/(5+312)$, and $r_1 = r_2 = 5$. The coverage percentage (missing from left, missing from right) and expect confidence interval width for MLE and UMVUE are 94.92 (4.27, 0.71) 0.1759 and 91.79 (8.02, 0.09) 0.1656, respectively. It is clear that the MLE maintains an overall coverage close to the nominal level of 95%, but through a lop-sided fashion, while the UMVUE cannot even deliver an adequate coverage, let alone maintaining balanced tail errors. In contrast, the confidence interval based on the exact limits for single proportions yielded coverage (tail errors) and expected width of 94.83 (3.08, 2.00) and 0.1725, which are very consistent with the results in Section 4.

6. Discussion

I have presented a simple approach to setting asymmetrical confidence intervals for effect measures arising from an inverse sampling design. The basis of this approach is to use confidence limits for single proportions based on a negative

Table 4
Illustrative example of confidence interval estimation in a 2 × 2 table arising from inverse sampling design.

Group	r	y	Method	p $l_{\hat{\theta}}u$	⇒	$\log p$ $l_{\hat{\theta}}u$	$\text{logit} = \log\{p/(1-p)\}$ $l_{\hat{\theta}}u$
1	5	53	MLE	$.0140_{.0862}$ $.1584$		$-4.2687_{-2.4511}$ -1.8426	$-4.2546_{-2.3609}$ -1.6702
			UMVUE	$.0033_{.0702}$ $.1371$		$-5.7138_{-2.6564}$ -1.9870	$-5.7105_{-2.5836}$ -1.8396
			Exact	$.0286_{.0862}$ $.1700$		$-3.5543_{-2.4511}$ -1.7720	$-3.5253_{-2.3609}$ -1.5856
2	5	312	MLE	$.0021_{.0158}$ $.0295$		$-6.1658_{-4.1477}$ -3.5234	$-6.1637_{-4.1318}$ -3.4934
			UMVUE	$.0003_{.0127}$ $.0250$		$-8.1117_{-4.3662}$ -3.6889	$-8.1114_{-4.3534}$ -3.6636
			Exact	$.0051_{.0158}$ $.0321$		$-5.2785_{-4.1477}$ -3.4389	$-5.2734_{-4.1318}$ -3.4063
				⇓	⇓		⇓
				Confidence interval for effect measures by the MOVER $L = \widehat{\theta}_1 - \widehat{\theta}_2 - \sqrt{(\widehat{\theta}_1 - l_1)^2 + (u_2 - \widehat{\theta}_2)^2}$ $U = \widehat{\theta}_1 - \widehat{\theta}_2 + \sqrt{(u_1 - \widehat{\theta}_1)^2 + (\widehat{\theta}_2 - l_2)^2}$		
				⇓	⇓		⇓
				$p_1 - p_2$		$RR = \exp(\log p_1 - \log p_2)$	$OR = \exp(\text{logit}_1 - \text{logit}_2)$
			MLE	$-.0031_{.0704}$ $.1439$		$0.7983_{5.4554}$ 44.9002	$0.7965_{5.8761}$ 50.2488
			UMVUE	$-.0105_{.0575}$ $.1255$		$0.2413_{5.5279}$ 248.3025	$0.2388_{5.8697}$ 270.6266
			Exact	$.0105_{.0704}$ $.1549$		$1.4701_{5.4554}$ 20.4023	$1.4903_{5.8761}$ 23.3566

binomial distribution to recover the variance estimates needed for functions of proportions. This is justifiable by the fact that the exact confidence interval for a negative binomial proportion is very accurate, in contrast to the common belief that exact confidence intervals are conservative (Agresti and Coull, 1998; Newcombe, 1998b; Agresti and Gottard, 2007).

I have also presented a unified framework for all three common effect measures, in contrast to previous work of focusing on effect measures in a piecemeal fashion. See Lui (2004) for extensive literature, in addition to Tian et al. (2008) and Tang and Tian (2009). Moreover, as a necessary step, the approach proposed here provides confidence intervals for each group risk parameter. Rather than interpreting an effect measure in isolation, the proposed method helps to interpret different effect measures in their proper contexts, and thus avoid using risk difference against relative risk (Lui, 1999; Tang and Tian, 2009).

I did not discuss hypothesis testing procedures for two reasons. First, Kang (2008) has recently presented exact procedures for testing the three effect measures under inverse sampling. Further results can be found in Tang et al. (2007) and Tang et al. (2008). Second, confidence interval construction encompasses hypothesis testing because by definition a $100(1-\alpha)\%$ two-sided confidence interval contains all parameter values that cannot be rejected at the α level.

Acknowledgments

Dr Zou is a recipient of the 2007 Early Researcher Award, Ontario Ministry of Research and Innovation, Canada. His work was also partially supported by an Individual Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. American Statistician 54, 280–288.
Agresti, A., Coull, B.A., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. American Statistician 52, 119–126.
Agresti, A., Gottard, A., 2007. Nonconservative exact small-sample inference for discrete data. Computational Statistics and Data Analysis 51, 6447–6458.
Bartlett, M.S., 1953. Approximate confidence intervals. Biometrika 40, 12–19.
Casella, G., Berger, R.L., 1990. Statistical Inference. Duxbury, Belmont, CA.
Daly, L.E., 1998. Confidence limits made easy: Interval estimation using a substitution method. American Journal of Epidemiology 147, 783–790.
Efron, B., 1987. Better bootstrap confidence intervals. Journal of the American Statistical Association 82, 171–185.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.
George, V.T., Elston, R.C., 1993. Confidence limits based on the first occurrence of an event. Statistics in Medicine 12, 685–690.
Haldane, J.B.S., 1945. On a method of estimating frequencies. Biometrika 33, 222–225.
Hirji, K.F., 2006. Exact Analysis of Discrete Data. Chapman & Hall/CRC, New York.
Kang, S.H., 2008. Exact tests for epidemiological risks under independent negative binomial sampling for rare events in clinical trials. Drug Information Journal 42, 115–121.
Lui, K.J., 1995. Confidence limits for the population prevalence rate based on the negative binomial distribution. Statistics in Medicine 14, 1471–1477.
Lui, K.J., 1999. Interval estimation of simple difference under independent negative binomial sampling. Biometrical Journal 41, 83–92.
Lui, K.J., 2004. Statistical Estimation of Epidemiological Risk. John Wiley & Sons Ltd., New York.
Newcombe, R.G., 1998a. Interval estimation for the difference between independent proportions: Comparison of eleven methods. Statistics in Medicine 17, 873–890.
Newcombe, R.G., 1998b. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in Medicine 17, 857–872.
Tang, M.L., Liao, Y.J., Ng, H.K.T., 2008. On tests of rate ratio under standard inverse sampling. Computer Methods and Programs in Biomedicine 89, 261–268.
Tang, M.L., Liao, Y.J., Ng, H.K.T., Chan, P.S., 2007. Testing the non-unity of rate ratio under inverse sampling. Biometrical Journal 49, 551–564.
Tang, M.L., Tian, M., 2009. Asymptotic confidence interval construction for risk difference under inverse sampling. Computational Statistics and Data Analysis 53, 621–631.
Tian, M., Tang, M.L., Ng, H.K.T., Chan, P.S., 2008. Confidence intervals for the risk ratio under inverse sampling. Statistics in Medicine 27, 3301–3324.

Zou, G.Y., 2007. Toward using confidence intervals to compare correlations. Psychological Methods 12, 399–413.
Zou, G.Y., 2008. On the estimation of additive interaction by use of the four-by-two table and beyond. American Journal of Epidemiology 168, 212–224.
Zou, G.Y., Donner, A., 2008. Construction of confidence limits about effect measures: A general approach. Statistics in Medicine 27, 1693–1702.
Zou, G.Y., Huang, W., Zhang, X., 2009a. A note on confidence interval estimation for a linear function of binomial proportions. Computational Statistics and Data Analysis 53, 1080–1085.
Zou, G.Y., Taleban, J., Huo, C.Y., 2009b. Confidence interval estimation for lognormal data with application to health economics. Computational Statistics and Data Analysis 53, 3755–3764.

Research Article

Statistics in Medicine

Received 25 March 2009, Accepted 3 September 2009 Published online in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/.3751

Assessment of risks by predicting counterfactuals

G. Y. Zou[a,b,c]*†

Risk assessment is fundamental to most epidemiological and biomedical investigations. In this article, risks are assessed in terms of risk difference and risk ratio by predicting counterfactual outcomes. Models considered for binary outcomes are probit, logistic, and extreme-value regressions. New confidence intervals for the effect measures are proposed using the method of variance estimates recovery, and evaluated by a simulation study. A SAS macro is provided for the calculations. A risk ratio obtained using counterfactuals is also compared in the simulation with that directly estimated from the modified Poisson model to answer a recent concern about the validity of the latter approach. Two examples are used to illustrate the methods. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: causal effect; confounding; logistic regression; probit regression; odds ratio

1. Introduction

An ideal design to assess risk is to have all subjects in the entire study exposed to a risk factor of interest and the outcomes observed; then, turn back the clock, and observe the outcomes on the same subjects in the absence of the exposure. The risk is then quantified by a within-subject contrast of outcomes under two conditions. In reality, however, we cannot turn back the clock. Consequently, we can never observe the outcomes under the exposure for the subjects in the unexposed group, nor can we observe the outcomes in the absence of exposure for those subjects in the exposed group. These unobservable outcomes are referred to as counterfactuals.

With its formal roots in randomized controlled trials in agriculture, counterfactual theory, or the theory of potential outcomes, it has become a basis for causal inference in a wide range of disciplines [1], including epidemiology [2, 3] where confounding can be defined unambiguously using this theory [4, 5].

Counterfactual theory can also provide insight into effect measures. For example, although it is well known that the odds ratio is misleading when it is interpreted as a risk ratio [6], only in light of counterfactual theory is it clearer that the odds ratio from logistic regression is a biased estimator for the causal odds ratio, except under the assumption that all subjects have identical baseline risk [7, 8]. As a numerical example, consider 50 per cent of subjects in a study having a risk of 0.6 when exposed, and 0.2 when unexposed, which gives an odds ratio of 6.0. Suppose the remaining 50 per cent subjects have a risk of 0.035 when exposed. Under the assumption of constant individual odds ratio (the assumption of logistic regression), the risk when unexposed must then be 0.006. The overall risk for subjects in the entire study when exposed is 0.3175 $(=(0.6+0.035)/2)$ versus 0.103 $(=0.20+0.006)/2$ when unexposed, yielding an odds ratio of 4.05, which is less than 6.0, a result that would have been obtained by logistic regression. The property that the overall odd ratio differs from the constant stratum odds ratio has also been referred to as 'noncollapsibility' [5]. Applying the Zhang–Yu [9] formula will result in a risk ratio of 3.96, compared with 3.08 $(\frac{0.3175}{0.103})$. In contrast, with the assumption of constant individual risk ratio of 3.0, the second half of subjects when unexposed would have risk of $0.035/3$, resulting in an overall risk for all subjects when unexposed given by 0.1058 $(=(0.2+0.035/3)/2)$. Therefore, the overall risk ratio is given by $\frac{0.3175}{0.1058}$, identical to the constant individual risk ratio of 3.0. As this calculation is a special case of the

[a]*Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5C1*
[b]*Robarts Clinical Trials, Robarts Research Institute, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5K8*
[c]*Department of Epidemiology and Biostatistics, School of Public Health, Southeast University, Nanjing, People's Republic of China*
**Correspondence to: G. Y. Zou, Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ont., Canada N6A 5C1.*
†*E-mail: gzou@robarts.ca*

Contract/grant sponsor: Ontario Ministry of Research and Innovation
Contract/grant sponsor: Natural Sciences and Engineering Research Council (NSERC) of Canada

modified Poisson regression [10], it is fair to conclude that this model can avoid the noncollapsibility problem inherent in using the logistic regression model to obtain an adjusted odds ratio.

We can also apply counterfactual theory to the analysis of binary outcomes from prospective studies. Specifically, we can first fit the observed binary data to a generalized linear model, and then predict event probabilities in the presence and absence of exposure for each subject. Finally, we can use these predicted probabilities to construct estimates of parameters such as the risk difference and risk ratio. Although the odds ratio may also be constructed, we will avoid this practice on the grounds that consumers could confuse this odds ratio with those obtained directly from logistic regression.

The above steps have been adopted previously [11–14]. The method has also been termed model-based standardization [15]. Although Wald-type confidence intervals (point estimate ± 1.96 times standard error) for the resultant effect measures have been available [16, 17], the perceived complexity of this method has precluded wide application in practice. Recent literature has implemented the percentile bootstrap and bias-corrected percentile bootstrap [18–20]. To motivate the use of the percentile bootstrap, Ahren *et al.* [21, p. 1143] stated that 'there is typically no straightforward analytic estimates of the standard error available', which of course is false because the standard error based on the delta method has been available for quite some time [16, 17]. Moreover, without adequate evaluation following the steps suggested by Efron [22], one would naturally doubt the reliability of the bootstrap approach in the present context, since it is well known that neither approach is reliable as a general tool for setting confidence intervals [23, 24, Chapter 13].

As an alternative, we extend the steps described by Lee in 1981 [12], who used a logistic regression model to predict probabilities under counterfactual conditions. To predict counterfactuals, we considered the probit, logistic, and extreme-value regression models. In addition, closed-form confidence intervals for risk, risk difference, and risk ratio are proposed and evaluated using a simulation study.

We also examine the validity of the modified Poisson regression in estimating the risk ratio under the violation of the risk ratio homogeneity assumption. This is useful because the model has been used in many occasions in practice, but serious concern about its validity has been raised [18].

A SAS macro implementing the methods is provided and illustrated using two examples from the literature, with the first one showing adjustment for confounding and the second for improving efficiency.

2. Methods

2.1. Estimates of mean risks and their variances

Suppose we have information about the event status (Y:1 = yes, 0 = no), exposure status (E:1 = expose, 0 = unexposed), and $p-1$ covariates ($x_2,\ldots,x_p$) that need to be adjusted for when assessing the effect of the exposure on the outcome. Assume that the underlying probability of $Y_i=1$ for subject i ($i=1,2,\ldots,n$) is related to the exposure and covariates through

$$\Pr(Y_i=1)=F(lp_i)=F(\beta_0+\beta_1E_i+\beta_2x_{2i}+\cdots+\beta_{pi}x_{pi})$$

where $F(\cdot)$ is known as the link function, with the common choices of probit, logit, or cloglog. It can be recognized that these link functions corresponding to cumulative distribution functions (CDF) for the standard normal, logistic, and the extreme-value distributions. The derivative of $F(\cdot)$ with respect to its argument is called the probability density function (PDF), denoted here as $f(\cdot)$. The explicit forms of $F(\cdot)$ and $f(\cdot)$ along with their implementation in the SAS package are given in Table I.

Having fit a generalized model to the data and obtained the beta-coefficient estimates $(\widehat{\beta}_0,\widehat{\beta}_1,\ldots,\widehat{\beta}_p)$, we can obtain the estimates of the exposed linear predictor for all subjects, regardless of exposure status (i.e. $E_i=1$ for all i), as given by

$$\widehat{lp_{1i}}=\widehat{\beta}_0+\widehat{\beta}_1\times 1+\widehat{\beta}_2x_{2i}+\cdots+\widehat{\beta}_px_{pi}$$

Plugging $\widehat{lp_{1i}}$ into $F(\cdot)$ yields predicted risk for subject i when exposed. The average risk is then obtained by

$$\overline{p}_1=\frac{1}{n}\sum_{i=1}^{n_1}F(\widehat{lp_{1i}})+\frac{1}{n}\underbrace{\sum_{i=1}^{n_0}F(\widehat{lp_{1i}})}_{\text{Counter factuals}}=\frac{1}{n}\sum_{i=1}^{n}F(\widehat{lp_{1i}})$$

Table I. Function forms and implementation in SAS.

Link	$F(lp_i)$	SAS function	$f(lp_i)=\frac{\partial F(lp_i)}{\partial lp_i}$	SAS function
Probit	$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{lp_i}\exp(-x^2/2)\,dx$	`CDF('normal', lp_i)`	$\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$	`pdf('normal', lp_i)`
Logit	$\frac{1}{1+\exp(-lp_i)}$	`CDF('logistic', lp_i)`	$\frac{\exp(-lp_i)}{[1+\exp(-lp_i)]^2}$	`pdf('logistic', lp_i)`
Cloglog	$1-\exp[\exp(lp_i)]$	Not yet available	$\exp[lp_i-\exp(lp_i)]$	Not yet available

where n is the size of the entire study. The predicted probabilities are counterfactuals for the n_0 subjects whose exposure status are $E_i=0$. Similarly, we have ($E_i=0$ for all i)

$$\widehat{lp_{0i}}=\widehat{\beta}_0+\widehat{\beta}_1\times 0+\widehat{\beta}_2x_{2i}+\cdots+\widehat{\beta}_px_{pi}$$

and the average predicted risk when all subjects are not exposed is defined as

$$\overline{p}_0=\frac{1}{n}\underbrace{\sum_{i=1}^{n_1}F(\widehat{lp_{0i}})}_{\text{Counter factuals}}+\frac{1}{n}\sum_{i=1}^{n_0}F(\widehat{lp_{0i}})=\frac{1}{n}\sum_{i=1}^{n}F(\widehat{lp_{0i}})$$

The predicted probabilities are counterfactuals for the n_1 subjects whose exposure status are $E_i=1$.

Estimates for variances of $\overline{p}_1$ and $\overline{p}_0$, and their covariance are complicated by the fact that all terms in the expressions are correlated since they are all functions of estimated beta-coefficients $(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2,\ldots,\widehat{\beta}_p)$. The generalized estimating equations (GEE) approach is not applicable because the whole data set has become one cluster and GEE requires a large number of clusters [25]. However, straightforward calculation using the delta method (also known as the method of propagation of errors) yields,

$$\widehat{\text{var}}(\overline{p}_1)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}f(\widehat{lp_{1i}})f(\widehat{lp_{1j}})\widehat{\text{cov}}(\widehat{lp_{1i}},\ \widehat{lp_{1j}})$$

$$\widehat{\text{var}}(\overline{p}_0)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}f(\widehat{lp_{0i}})f(\widehat{lp_{0j}})\widehat{\text{cov}}(\widehat{lp_{0i}},\ \widehat{lp_{0j}})$$

and

$$\widehat{\text{cov}}(\overline{p}_1,\overline{p}_0)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}f(\widehat{lp_{1i}})f(\widehat{lp_{0j}})\widehat{\text{cov}}(\widehat{lp_{1i}},\ \widehat{lp_{0j}})$$

using the fact that the derivative of a CDF equals its PDF. These variance and covariance estimates can be obtained with estimates of beta-coefficients and their variance–covariance matrix provided by standard software output. To reduce the tedious and sometimes characterized as 'difficult' calculations [26, p. 276], we have developed a SAS macro (see the Appendix), which also implements the improved confidence intervals developed in the next section.

2.2. Confidence interval estimation

A confidence interval is more informative than a P-value, which is confounded by sample size. Thus, confidence interval construction has become a basic task for statistical inference. The Wald interval method of point estimate minus/plus a multiplier of standard error obtained using the delta method is known to perform poorly due to its enforced symmetry. In fact, Efron and Tibshirani [24, p. 180] stated 'exact intervals, when they exist, are often quite asymmetrical. The most serious errors made by standard intervals are due to their enforced symmetry.'

Instead of relying on the percentile bootstrap as did previous literature to obtain asymmetric intervals [18–21], we propose several alternatives for risk, risk difference, and risk ratio. In particular, we apply the MOVER, method of variances estimates recovery [27, 28], for the risk difference and risk ratio, after obtaining the logit-transformed and the Wilson confidence intervals for risks [29].

As documented by Zou and Donner [28], confidence intervals similar to the MOVER have been in use for quite some time, with a book-length application on variance components [30]. However, previous authors tend to justify their procedures by assuming the limits are of a certain form and solving the limits by forcing the confidence coefficients to be exact under special conditions [31, 32]. 'Square-and-add' is another term when it was applied to proportions [33]. We prefer to use the term MOVER because it reflects clearly that the key step of the method is to recover variance estimates needed for a linear combination of parameters. Q1

Two ideas went into the derivation of the MOVER [28]. First, confidence intervals constructed with variance estimates obtained at or close to the limits in general perform better than those with variances estimated at the point estimates [22, p. 175]. Respective examples are the Wilson method and the Wald method for a proportion. The idea of estimating variances at endpoints is related to the restricted Wald method in Maldonald and Greenland [34, Equation 9]. An apparent advantage of the MOVER approach is that it can account for asymmetry of sampling distributions by recovering variance estimates separately from lower and upper limits.

The second idea used in the MOVER is that variance estimates needed for linear combination of parameters are contained in the confidence limits for a single parameters and thus can be recovered. Zou and Donner [28] showed that the MOVER approach even outperforms many computational intensive methods, including six bootstrap methods for setting confidence limits for a lognormal mean.

The logit transformation, resulting in an asymmetric confidence interval, has previously been applied to risks [17] to provide confidence limits given by

$$l,u=\ln\frac{\overline{p}}{1-\overline{p}}\mp z\sqrt{\widehat{\text{var}}\left(\ln\frac{\overline{p}}{1-\overline{p}}\right)}$$

where z is the $\alpha/2$ upper quantile of the standard normal distribution. Anti-logit transformations of l and u yield confidence limits for p. The original Wilson method for a proportion inverts the approximate normal test that uses the standard errors estimated at the lower and upper limits [29, 35]. This method has now become very popular [36, 37], and also has been recommended in an introductory epidemiologic text [2, p. 145]. Interestingly, the Wilson confidence limits can also be obtained using the logit transformation [33] as

$$l,u=\ln\frac{\overline{p}}{1-\overline{p}}\mp 2\cdot\operatorname{arsinh}\left[\frac{z}{2}\sqrt{\widehat{\operatorname{var}}\left(\ln\frac{\overline{p}}{1-\overline{p}}\right)}\right]$$

where arsinh is the inverse hyperbolic sine function. Anti-logit transformations of l and u then yield confidence limits for p. It is shown that this interval is contained completely in that of the logit transformation [33].

One may also apply the log transformation to risk [16], yielding

$$l,u=\ln\overline{p}\mp z\sqrt{\operatorname{var}[\ln(\overline{p})]}$$

The anti-log transformation results in limits for p.

Confidence interval estimation for the risk difference p_1-p_0 starts with variance estimates for $\overline{p}_1$ and $\overline{p}_0$. It turns out that these variance estimates can be recovered from confidence limits for risks [27, 28]. Based on the MOVER, the confidence limits for p_1-p_0 are given by

$$L=\overline{p}_1-\overline{p}_0-\sqrt{(\overline{p}_1-l_1)^2+(u_0-\overline{p}_0)^2-2\widehat{\operatorname{corr}}(\overline{p}_1,\ \overline{p}_0)(\overline{p}_1-l_1)(u_0-\overline{p}_0)}$$

$$U=\overline{p}_1-\overline{p}_0+\sqrt{(u_1-\overline{p}_1)^2+(\overline{p}_0-l_0)^2-2\widehat{\operatorname{corr}}(\overline{p}_1,\ \overline{p}_0)(u_1-\overline{p}_1)(\overline{p}_0-l_0)}$$

where (l_1,u_1) and (l_0,u_0) are confidence intervals for p_1 and p_0, respectively, and

$$\widehat{\operatorname{corr}}(\overline{p}_1,\ \overline{p}_0)=\frac{\widehat{\operatorname{cov}}(\overline{p}_1,\overline{p}_0)}{\sqrt{\widehat{\operatorname{var}}(\overline{p}_1)\widehat{\operatorname{var}}(\overline{p}_0)}}$$

The validity of the MOVER relies heavily on the method used to obtain confidence limits for risks. Moreover, substituting (l_1,u_1) and (l_0,u_0) obtained by the Wald method will result in a Wald confidence interval.

One can also obtain a confidence interval for $\delta=p_1-p_0$ by applying the logit transformation to $(1+\delta)/2$ [17]. This approach is equivalent to applying Fisher's z-transformation to $\delta=p_1-p_0$ [38] since

$$\ln\frac{(1+\delta)/2}{1-(1+\delta)/2}=\ln\frac{1+\delta}{1-\delta}$$

While a confidence interval for the risk ratio can be obtained by applying the Wald procedure on the log scale, it can also be obtained using the MOVER on the log scale [27]. A simulation study has shown that the MOVER outperforms the Wald procedure in the case of simple proportions [27]. Briefly, denoting $\theta_1=\ln(p_1)$ and $\theta_0=\ln(p_0)$, confidence limits for p_1/p_0 can be obtained by taking the anti-log of L_θ and U_θ, where

$$L_\theta=\ln\overline{p}_1-\ln\overline{p}_0-\sqrt{(\ln\overline{p}_1-\ln l_1)^2+(\ln u_0-\ln\overline{p}_0)^2-2\widehat{\operatorname{corr}}(\overline{p}_1,\overline{p}_0)(\ln\overline{p}_1-\ln l_1)(\ln u_0-\ln\overline{p}_0)}$$

$$U_\theta=\ln\overline{p}_1-\ln\overline{p}_0+\sqrt{(\ln u_1-\ln\overline{p}_1)^2+(\ln\overline{p}_0-\ln l_0)^2-2\widehat{\operatorname{corr}}(\overline{p}_1,\overline{p}_0)(\ln u_1-\ln\overline{p}_1)(\ln\overline{p}_0-\ln l_0)}$$

Note that the correlation between $\ln\overline{p}_1$ and $\ln\overline{p}_0$ is equal to the correlation between $\overline{p}_1$ and $\overline{p}_0$.

3. Simulation study

We conducted a simulation study to evaluate the performance of the above methods in setting confidence intervals for a single risk, risk difference, and risk ratio. The data generation process was similar to that used previously [17]. Specifically, for subject i ($i=1,2,\ldots,n$) in each of 1000 simulation runs, we first generated three predictors (x_{1i},x_{2i},x_{3i}) as trivariate standard normal with common correlation ρ. We then obtained the binary exposure ($x_{1i}=E_i$) and a binary covariate (x_{2i}) by dichotomizing the first two variables at a mean of 0, and kept x_{3i} as a continuous variable. The probability of an event was obtained according to the standard normal CDF for a given set of beta-coefficients, $\Phi(\beta_0+\beta_1E_i+\beta_2x_{2i}+\beta_3x_{3i})$. The outcome y_i was then generated using a Bernoulli distribution with parameter π_i. These steps were repeated n times to obtain a data set which was analyzed using probit regression as implemented by `proc genmod` in SAS. Estimates of beta-coefficients and their variance-covariance estimates were then used to obtain confidence intervals for a single risk, risk difference, and risk ratio.

To check the concern raised previously [18], each data set was also analyzed using the modified Poisson regression for risk ratio, which was implemented using `sas proc genmod` with a log link and 'sandwich' error estimator [10].

All analyses were performed using the SAS macro as shown in the Appendix.

Empirical coverage percentage was estimated by the relative frequency of 1000 intervals which contained the parameter. Tail errors were estimated by calculating the frequencies of intervals lying completely to the left of the parameter value and those lying completely to the right of the parameter. Median width was also calculated.

We considered 32 parameter combinations. Specifically, sample sizes $n = 100$, 200, 500, and 1000, $\rho = 0.2$, and 0.5, and four sets of beta-coefficients: (−1.5, 0.5, −0.245, −0.4), (−1.2, 1.0, −0.245, −0.4), (−1.0, 0.5, −0.245, −0.4), and (−0.5, 1.0, −0.245, −0.4) were considered. These values were chosen to provide baseline risks ranging from 0.07 to 0.35 and risk ratios ranging from 1.87 to 3.34.

Table II shows simulation results for a single risk. These results indicate that the Wald procedure tends to provide coverage ranging from 89.9 to 94.0 per cent when the sample size is 100, and lop-sided errors even when the sample size is as large as 1000, while the log transformation provides lop-sided intervals with large width. These problems are less severe for intervals obtained with the logit transformation. Overall, the Wilson interval performs well. Similar trends can be observed in the case of the risk difference, as shown in Table III, which also demonstrates that the Fisher z-transformation for the risk difference performs well. Results in Table IV show that the Wald procedure applied on the log scale for the risk ratio has inferior performance as compared with the MOVER using the Wilson interval for single risks. In particular, the former tends to provide wider intervals with slightly unbalanced tail errors.

Table IV also demonstrates that modified Poisson regression performs reasonably well, both in terms of point estimates and confidence intervals. In theory, tighter intervals could be obtained with the binomial regression. We did not compare the binomial

Table II. Comparative performance of methods based on counterfactual prediction using a probit regression model in estimating risk and its 95 per cent two-sided confidence interval based on 1000 simulation runs*.

n	$\rho^{\dagger}$	$p_0^{\ddagger}$	Wald	ln	Logit	Wilson
100	0.25	0.07	89.9 (9.3,0.8) 0.125	94.1 (1.4,4.5) 0.146	94.4 (1.4,4.2) 0.140	94.0 (1.6,4.4) 0.133
		0.11	91.2 (8.1,0.7) 0.159	96.1 (0.7,3.2) 0.174	96.5 (0.9,2.6) 0.166	95.6 (1.5,2.9) 0.162
		0.15	91.6 (6.7,1.7) 0.181	94.8 (1.1,4.1) 0.193	94.8 (1.5,3.7) 0.185	94.6 (1.7,3.7) 0.181
		0.28	94.0 (4.4,1.6) 0.231	95.0 (0.9,4.1) 0.238	95.8 (1.7,2.5) 0.229	95.2 (2.1,2.7) 0.226
	0.50	0.07	90.7 (9.0,0.3) 0.121	95.9 (0.5,3.6) 0.138	96.4 (0.5,3.1) 0.133	95.3 (1.4,3.3) 0.127
		0.12	92.8 (6.5,0.7) 0.155	94.9 (1.1,4.0) 0.169	95.7 (1.2,3.1) 0.161	95.1 (1.7,3.2) 0.158
		0.15	93.0 (5.9,1.1) 0.178	94.8 (1.0,4.2) 0.189	95.6 (1.1,3.3) 0.182	95.1 (1.3,3.6) 0.178
		0.29	93.7 (3.7,2.6) 0.232	93.8 (1.6,4.6) 0.238	94.7 (2.1,3.2) 0.229	94.5 (2.2,3.3) 0.226
200	0.25	0.07	93.4 (5.4,1.2) 0.089	95.1 (0.7,4.2) 0.096	95.4 (0.9,3.7) 0.094	95.0 (1.2,3.8) 0.092
		0.11	93.6 (5.1,1.3) 0.112	95.4 (1.3,3.3) 0.118	95.1 (2.1,2.8) 0.115	94.9 (2.2,2.9) 0.113
		0.15	93.4 (4.8,1.8) 0.129	94.6 (1.6,3.8) 0.133	94.9 (2.0,3.1) 0.130	94.7 (2.0,3.3) 0.129
		0.28	95.2 (3.3,1.5) 0.165	95.9 (1.6,2.5) 0.167	95.7 (2.2,2.1) 0.164	95.6 (2.2,2.2) 0.163
	0.50	0.07	93.6 (5.6,0.8) 0.087	95.4 (1.0,3.6) 0.093	95.7 (1.2,3.1) 0.091	95.2 (1.6,3.2) 0.089
		0.12	93.5 (5.2,1.3) 0.110	95.0 (1.9,3.1) 0.114	94.8 (2.3,2.9) 0.112	94.6 (2.5,2.9) 0.111
		0.15	93.9 (4.9,1.2) 0.127	95.0 (1.7,3.3) 0.131	95.1 (2.2,2.7) 0.128	94.6 (2.4,3.0) 0.127
		0.29	93.9 (4.5,1.6) 0.165	94.5 (2.5,3.0) 0.167	94.7 (3.0,2.3) 0.164	94.7 (3.0,2.3) 0.163
500	0.25	0.07	93.8 (4.8,1.4) 0.057	95.9 (1.2,2.9) 0.059	96.1 (1.3,2.6) 0.058	95.8 (1.6,2.6) 0.058
		0.11	95.1 (3.2,1.7) 0.072	95.4 (1.1,3.5) 0.074	95.8 (1.2,3.0) 0.073	95.8 (1.2,3.0) 0.073
		0.15	95.3 (3.1,1.6) 0.082	95.4 (1.5,3.1) 0.083	95.5 (1.9,2.6) 0.082	95.4 (2.0,2.6) 0.082
		0.28	95.5 (2.8,1.7) 0.104	95.6 (2.1,2.3) 0.105	95.8 (2.4,1.8) 0.104	95.8 (2.4,1.8) 0.104
	0.50	0.07	93.8 (4.8,1.4) 0.056	95.6 (0.9,3.5) 0.057	95.7 (1.0,3.3) 0.057	95.4 (1.3,3.3) 0.056
		0.12	94.8 (3.6,1.6) 0.070	95.4 (1.6,3.0) 0.071	95.4 (1.7,2.9) 0.071	95.3 (1.7,3.0) 0.071
		0.15	95.1 (3.0,1.9) 0.080	95.4 (1.5,3.1) 0.081	95.3 (1.7,3.0) 0.081	95.2 (1.8,3.0) 0.080
		0.29	94.6 (3.7,1.7) 0.105	95.6 (2.4,2.0) 0.105	95.2 (3.0,1.8) 0.104	95.2 (3.0,1.8) 0.104
1000	0.25	0.07	95.5 (3.2,1.3) 0.041	95.3 (1.6,3.1) 0.041	95.3 (1.6,3.1) 0.041	95.3 (1.6,3.1) 0.041
		0.11	95.5 (2.9,1.6) 0.051	95.6 (1.7,2.7) 0.051	95.8 (1.8,2.4) 0.051	95.7 (1.9,2.4) 0.051
		0.15	94.7 (3.7,1.6) 0.058	95.6 (2.2,2.2) 0.058	95.5 (2.3,2.2) 0.058	95.5 (2.3,2.2) 0.058
		0.28	95.7 (3.1,1.2) 0.074	96.0 (2.1,1.9) 0.074	95.9 (2.4,1.7) 0.074	95.9 (2.4,1.7) 0.074
	0.50	0.07	95.4 (2.8,1.8) 0.039	95.5 (1.5,3.0) 0.040	95.6 (1.6,2.8) 0.040	95.6 (1.6,2.8) 0.040
		0.12	96.1 (2.6,1.3) 0.050	95.6 (1.5,2.9) 0.050	95.5 (1.6,2.9) 0.050	95.5 (1.6,2.9) 0.050
		0.15	95.5 (3.0,1.5) 0.057	95.7 (1.8,2.5) 0.057	95.8 (1.8,2.4) 0.057	95.7 (1.9,2.4) 0.057
		0.29	95.2 (2.8,2.0) 0.074	95.2 (2.3,2.5) 0.074	95.2 (2.5,2.3) 0.074	95.2 (2.5,2.3) 0.074

*Entries in columns 4–7 are in the format: estimated per cent coverage (per cent interval lying completely to the left, right of the true parameter value) median interval width.

†Correlation coefficient used to generate standard trivariate normal for the three predictors.

‡Average true value of risk. Four sets of beta-coefficients were (−1.5, 0.5, −0.245, −0.4), (−1.2, 1.0, −0.245, −0.4), (−1.0, 0.5, −0.245, −0.4), and (−0.5, 1.0, −0.245, −0.4).

Table III. Comparative performance of methods based on counterfactual prediction using a probit regression model in estimating risk difference and its 95 per cent two-sided confidence interval based on 1000 simulation runs*.

				MOVER		
n	$\rho^{\dagger}$	$d^{\ddagger}$	Wald	Logit	Wilson	Fisher z
100	0.25	0.08	93.8 (3.7,2.5) 0.241	96.4 (0.9,2.7) 0.265	95.3 (1.5,3.2) 0.254	94.2 (3.7,2.1) 0.240
		0.27	94.5 (2.4,3.1) 0.307	96.0 (2.2,1.8) 0.311	95.0 (2.5,2.5) 0.305	94.8 (2.8,2.4) 0.305
		0.13	92.6 (3.2,4.2) 0.309	94.1 (2.7,3.2) 0.314	93.4 (2.9,3.7) 0.308	93.4 (3.2,3.4) 0.307
		0.35	94.4 (2.3,3.3) 0.346	94.4 (3.4,2.2) 0.341	94.2 (3.5,2.3) 0.336	94.3 (3.2,2.5) 0.343
	0.50	0.08	93.7 (3.1,3.2) 0.265	96.0 (0.4,3.6) 0.288	95.2 (0.9,3.9) 0.277	93.9 (3.1,3.0) 0.263
		0.27	93.7 (2.8,3.5) 0.326	94.7 (2.5,2.8) 0.328	94.2 (2.8,3.0) 0.322	94.2 (3.0,2.8) 0.323
		0.13	94.1 (2.5,3.4) 0.330	95.3 (1.9,2.8) 0.334	94.7 (2.2,3.1) 0.327	94.7 (2.5,2.8) 0.327
		0.35	92.4 (3.3,4.3) 0.358	93.6 (3.9,2.5) 0.353	93.1 (4.1,2.8) 0.348	93.5 (3.6,2.9) 0.355
200	0.25	0.08	94.6 (3.3,2.1) 0.171	96.1 (1.9,2.0) 0.179	95.8 (2.0,2.2) 0.176	94.6 (3.3,2.1) 0.170
		0.27	95.5 (2.0,2.5) 0.218	96.1 (2.1,1.8) 0.219	96.0 (2.2,1.8) 0.217	96.1 (2.1,1.8) 0.217
		0.13	93.7 (2.5,3.8) 0.220	94.2 (2.3,3.5) 0.221	94.0 (2.5,3.5) 0.219	94.0 (2.6,3.4) 0.219
		0.35	94.7 (2.0,3.3) 0.245	95.0 (2.5,2.5) 0.243	94.9 (2.5,2.6) 0.241	95.2 (2.2,2.6) 0.244
	0.50	0.08	95.9 (2.2,1.9) 0.189	96.3 (1.0,2.7) 0.197	95.9 (1.3,2.8) 0.193	95.7 (2.4,1.9) 0.189
		0.27	95.2 (1.5,3.3) 0.232	95.6 (1.5,2.9) 0.232	95.6 (1.5,2.9) 0.230	95.6 (1.6,2.8) 0.231
		0.13	94.1 (2.8,3.1) 0.235	94.8 (2.4,2.8) 0.237	94.4 (2.6,3.0) 0.235	94.5 (2.9,2.6) 0.234
		0.35	94.5 (2.5,3.0) 0.253	94.2 (3.5,2.3) 0.251	93.8 (3.8,2.4) 0.250	94.4 (3.1,2.5) 0.252
500	0.25	0.08	94.7 (3.4,1.9) 0.109	95.3 (2.8,1.9) 0.111	95.1 (2.9,2.0) 0.110	94.7 (3.4,1.9) 0.108
		0.27	95.5 (1.8,2.7) 0.139	95.5 (2.1,2.4) 0.139	95.4 (2.1,2.5) 0.139	95.4 (2.1,2.5) 0.139
		0.13	94.1 (2.7,3.2) 0.139	94.1 (2.7,3.2) 0.140	94.1 (2.7,3.2) 0.139	94.1 (2.7,3.2) 0.139
		0.35	93.3 (2.7,4.0) 0.155	93.4 (3.6,3.0) 0.154	93.4 (3.6,3.0) 0.154	93.2 (3.5,3.3) 0.155
	0.50	0.08	95.5 (2.5,2.0) 0.120	95.3 (2.1,2.6) 0.122	95.1 (2.3,2.6) 0.121	95.5 (2.6,1.9) 0.119
		0.27	96.3 (1.5,2.2) 0.147	96.4 (1.5,2.1) 0.147	96.2 (1.6,2.2) 0.147	96.3 (1.6,2.1) 0.147
		0.13	95.2 (2.3,2.5) 0.149	95.4 (2.1,2.5) 0.150	95.3 (2.2,2.5) 0.149	95.1 (2.4,2.5) 0.149
		0.35	93.2 (3.2,3.6) 0.160	92.8 (3.8,3.4) 0.160	92.8 (3.8,3.4) 0.159	93.0 (3.6,3.4) 0.160
1000	0.25	0.08	95.4 (2.6,2.0) 0.077	95.6 (2.3,2.1) 0.078	95.6 (2.3,2.1) 0.077	95.4 (2.6,2.0) 0.077
		0.27	96.0 (2.1,1.9) 0.098	96.1 (2.2,1.7) 0.098	96.1 (2.2,1.7) 0.098	96.1 (2.2,1.7) 0.098
		0.13	93.9 (3.0,3.1) 0.098	93.9 (3.0,3.1) 0.099	93.9 (3.0,3.1) 0.098	94.0 (3.0,3.0) 0.098
		0.35	94.1 (3.1,2.8) 0.110	94.0 (3.3,2.7) 0.109	94.0 (3.3,2.7) 0.109	93.9 (3.3,2.8) 0.110
	0.50	0.08	95.4 (2.2,2.4) 0.085	95.7 (1.6,2.7) 0.085	95.6 (1.7,2.7) 0.085	95.3 (2.3,2.4) 0.085
		0.27	96.7 (1.6,1.7) 0.104	96.7 (1.7,1.6) 0.104	96.7 (1.7,1.6) 0.104	96.6 (1.8,1.6) 0.104
		0.13	94.3 (2.9,2.8) 0.106	94.9 (2.4,2.7) 0.106	94.7 (2.5,2.8) 0.105	94.4 (3.0,2.6) 0.105
		0.35	94.5 (2.7,2.8) 0.114	94.4 (3.1,2.5) 0.113	94.4 (3.1,2.5) 0.113	94.5 (3.0,2.5) 0.113

*Entries in columns 4 to 7 are in the format: estimated per cent coverage (per cent interval lying completely to the left, right of the true parameter value) median interval width.
†Correlation coefficient used to generate standard trivariate normal for the three predictors.
‡Average true value of risk difference. Four sets of beta-coefficients were (−1.5, 0.5, −0.245, −0.4), (−1.2, 1.0, −0.245, −0.4), (−1.0, 0.5, −0.245, −0.4), and (−0.5, 1.0, −0.245, −0.4).

regression model for the following reasons. First, previous simulation results have shown that both models perform similarly with categorical independent variables [10]. Second, the convergence problem has not been fixed entirely by the SAS macro of Spiegelman and Hertzmark [39]. In fact, it is stated that 'the modified Poisson estimates are used to start the iterations to obtain the log-binomial maximum likelihood estimates. These are the final estimates if convergence of the binomial likelihood is not obtained' [39, p. 200]. Third, it is not the purpose of this article to compare the efficiency between the modified Poisson and the log-binomial regression in estimating risk ratio. The purpose here is to provide an answer to the concern raised by Localio *et al.* [18] of using the modified Poisson model when estimating risk ratios.

Similar results (not shown) were obtained using the logistic regression as the predicting model for counterfactuals.

4. Illustrative examples

As a first illustrative example, consider the data in Table V involving 40 subjects [12]. Besides therapeutic regimes x_1 (with 1 indicating the new therapy and 0 the conventional therapy), information on the extent of disease (EOD with 0 denoting moderate and 1 severe) and age (years) are also available. The objective is to estimate the effect of a new therapy as compared with the

Table IV. Comparative performance of counterfactual prediction using a probit regression model and the modified Poisson regression in estimating risk ratio and its 95 per cent two-sided confidence interval based on 1000 simulation runs*.

Size				MOVER§		
n	ρ†	*RR*‡	Wald (ln¶)	Logit	Wilson	Modified Poisson
100	0.25	2.21-2.21-2.24	96.3 (1.9,1.8) 6.572	95.6 (2.4,2.0) 6.513	94.7 (2.8,2.5) 6.225	94.4 (4.9,0.7) 6.419
		3.40-3.50-3.48	97.3 (1.8,0.9) 6.167	96.7 (1.9,1.4) 6.208	96.1 (2.0,1.9) 6.042	96.7 (2.4,0.9) 5.991
		1.87-1.90-1.91	95.1 (2.6,2.3) 3.171	94.3 (2.9,2.8) 3.154	93.6 (3.1,3.3) 3.083	94.8 (2.9,2.3) 3.187
		2.23-2.23-2.22	95.2 (3.2,1.6) 2.111	95.3 (2.7,2.0) 2.144	94.7 (2.8,2.5) 2.116	94.7 (3.4,1.9) 2.052
	0.50	2.19-2.17-2.23	96.3 (1.3,2.4) 6.904	95.8 (1.5,2.7) 6.673	93.8 (2.4,3.8) 6.334	91.9 (5.2,2.9) 7.298
		3.35-3.32-3.41	96.2 (2.3,1.5) 6.004	95.8 (2.3,1.9) 6.010	94.9 (2.8,2.3) 5.854	95.8 (2.5,1.7) 6.046
		1.86-1.87-1.91	95.7 (2.0,2.3) 3.169	95.5 (2.2,2.3) 3.128	95.1 (2.4,2.5) 3.056	94.9 (2.2,2.9) 3.428
		2.22-2.20-2.19	93.7 (4.3,2.0) 2.155	94.0 (3.7,2.3) 2.183	93.6 (4.0,2.4) 2.151	94.1 (3.6,2.3) 2.137
200	0.25	2.21-2.25-2.24	96.7 (2.3,1.0) 4.163	96.4 (2.5,1.1) 4.156	95.7 (2.7,1.6) 4.067	96.5 (2.3,1.2) 4.217
		3.39-3.46-3.41	95.1 (3.1,1.8) 4.132	95.1 (3.1,1.8) 4.149	94.7 (3.1,2.2) 4.090	94.9 (3.2,1.9) 3.953
		1.87-1.90-1.91	94.7 (2.6,2.7) 2.132	94.6 (2.7,2.7) 2.130	94.4 (2.8,2.8) 2.108	94.8 (2.4,2.8) 2.140
		2.23-2.22-2.20	95.8 (2.1,2.1) 1.469	95.4 (2.0,2.6) 1.481	95.1 (2.2,2.7) 1.471	96.4 (2.0,1.6) 1.424
	0.50	2.19-2.21-2.23	96.7 (1.2,2.1) 4.240	96.6 (1.2,2.2) 4.194	95.7 (1.7,2.6) 4.104	95.5 (1.8,2.7) 4.573
		3.34-3.35-3.40	95.1 (2.4,2.5) 3.970	95.1 (2.3,2.6) 3.971	94.8 (2.4,2.8) 3.923	95.9 (1.7,2.4) 3.990
		1.85-1.86-1.89	95.0 (2.2,2.8) 2.149	94.6 (2.6,2.8) 2.143	94.3 (2.7,3.0) 2.122	94.2 (2.2,3.6) 2.272
		2.21-2.21-2.20	95.0 (3.1,1.9) 1.482	94.8 (2.8,2.4) 1.494	94.6 (2.9,2.5) 1.483	95.2 (2.6,2.2) 1.462
500	0.25	2.21-2.20-2.20	95.7 (2.9,1.4) 2.389	95.7 (2.9,1.4) 2.384	95.5 (3.0,1.5) 2.365	96.0 (2.3,1.7) 2.399
		3.39-3.42-3.37	96.1 (2.4,1.5) 2.486	96.1 (2.3,1.6) 2.488	95.9 (2.4,1.7) 2.476	96.0 (2.8,1.2) 2.361
		1.87-1.90-1.90	94.2 (2.8,3.0) 1.294	94.0 (2.8,3.2) 1.294	94.0 (2.8,3.2) 1.288	94.6 (2.9,2.5) 1.289
		2.23-2.23-2.21	94.2 (3.3,2.5) 0.925	94.0 (3.2,2.8) 0.928	93.9 (3.2,2.9) 0.925	94.0 (3.9,2.1) 0.894
	0.50	2.19-2.19-2.24	96.0 (1.9,2.1) 2.452	95.9 (1.9,2.2) 2.442	95.9 (1.9,2.2) 2.423	95.7 (1.7,2.6) 2.610
		3.34-3.36-3.36	95.6 (2.4,2.0) 2.434	95.4 (2.4,2.2) 2.435	95.2 (2.4,2.4) 2.425	96.0 (2.4,1.6) 2.373
		1.85-1.88-1.91	94.8 (3.1,2.1) 1.324	94.7 (3.2,2.1) 1.321	94.6 (3.2,2.2) 1.316	95.1 (2.2,2.7) 1.384
		2.21-2.22-2.21	94.5 (3.1,2.4) 0.938	94.6 (2.9,2.5) 0.941	94.6 (2.9,2.5) 0.938	93.9 (3.6,2.5) 0.916
1000	0.25	2.21-2.24-2.24	96.0 (2.5,1.5) 1.662	96.0 (2.5,1.5) 1.662	96.0 (2.5,1.5) 1.655	96.1 (2.3,1.6) 1.659
		3.40-3.42-3.37	95.8 (2.4,1.8) 1.745	95.7 (2.4,1.9) 1.746	95.7 (2.4,1.9) 1.742	94.9 (3.7,1.4) 1.668
		1.87-1.89-1.88	94.2 (2.6,3.2) 0.905	94.1 (2.6,3.3) 0.905	94.1 (2.6,3.3) 0.903	94.1 (2.6,3.3) 0.897
		2.23-2.23-2.21	95.3 (2.5,2.2) 0.655	95.3 (2.4,2.3) 0.656	95.3 (2.4,2.3) 0.655	94.3 (3.6,2.1) 0.631
	0.50	2.19-2.19-2.24	96.4 (1.7,1.9) 1.691	96.2 (1.9,1.9) 1.687	96.2 (1.9,1.9) 1.681	96.0 (1.4,2.6) 1.783
		3.34-3.35-3.35	96.4 (1.8,1.8) 1.716	96.5 (1.7,1.8) 1.716	96.4 (1.8,1.8) 1.712	96.1 (1.9,2.0) 1.669
		1.86-1.88-1.91	95.1 (2.1,2.8) 0.930	95.0 (2.1,2.9) 0.929	95.0 (2.1,2.9) 0.927	94.1 (2.0,3.9) 0.967
		2.22-2.22-2.21	95.1 (2.7,2.2) 0.663	95.1 (2.6,2.3) 0.664	95.1 (2.6,2.3) 0.663	95.3 (2.7,2.0) 0.644

*Entries in columns 4 to 7 are in the format: estimated per cent coverage (per cent interval lying completely to the left, right of the true parameter value) median interval width.
†Values of correlation among three predictors.
‡The first number is the average true value of risk ratio, followed by its estimates from methods of predicting counterfactuals and the modified Poisson regression, respectively.
§These limits are obtained by applying delta method on the log scale.
¶These confidence limits are obtained by applying the method of variance recovery on the log scale.

conventional therapy with respect to recovery from the disease. As discussed in Lee [12], there is a need to adjust for confounding in estimating the effect of the new therapy.

Fitting the data to a probit model, we obtained $\widehat{\beta}_0 = 1.3807$ (intercept), $\widehat{\beta}_1 = 1.2689$ (therapy), $\widehat{\beta}_2 = -0.6764$ (EOD), and $\widehat{\beta}_3 = -0.0603$ (Age). To estimate the risk to be expected ($\overline{p}_1$) if all subjects had received the new therapy (regardless of their actual therapy groups), we compute

$$\Phi(\widehat{lp_{1i}}) = \Phi(\widehat{\beta}_0 + \widehat{\beta}_1 \times 1 + \widehat{\beta}_2 \times \mathrm{EOD}_i + \widehat{\beta}_3 \times \mathrm{Age}_i)$$

for all 40 subjects. As the first 20 subjects did not receive the new therapy, their predicted $\Phi(\widehat{lp_{1i}})$ are counterfactuals. The predicted counterfactual for the first subject is given by $\Phi[1.3807 + 1.2689 - 0.6764(0) - 0.0603(20)] = 0.9257$, and the values for the remaining

Table V. Illustrative example of estimating risks based on a probit regression model*.

Data					Predicted probability (Recovery)	
ID	Therapy	EOD	Age	Recovery	If treated ($\widehat{p}_{1i}$)†	If not treated ($\widehat{p}_{i0}$)‡
1	0	0	20	1	*0.9257*	0.5697
2	0	0	23	1	*0.8968*	0.4979
3	0	0	22	0	*0.9072*	0.5219
4	0	0	26	0	*0.8606*	0.4262
5	0	0	29	0	*0.8165*	0.3569
6	0	0	34	0	*0.7260*	0.2520
7	0	0	32	1	*0.7647*	0.2920
8	0	0	30	0	*0.8001*	0.3347
9	0	0	38	0	*0.6405*	0.1816
10	0	0	37	0	*0.6628*	0.1980
11	0	0	38	1	*0.6405*	0.1816
12	0	1	25	1	*0.6797*	0.2112
13	0	1	24	0	*0.7009*	0.2291
14	0	1	25	0	*0.6797*	0.2112
15	0	1	29	0	*0.5893*	0.1484
16	0	1	32	0	*0.5179*	0.1105
17	0	1	34	0	*0.4699*	0.0894
18	0	1	37	0	*0.3989*	0.0636
19	0	1	40	0	*0.3310*	0.0440
20	0	1	40	0	*0.3310*	0.0440
21	1	0	20	1	0.9257	*0.5697*
22	1	0	24	1	0.8856	*0.4739*
23	1	0	28	1	0.8321	*0.3796*
24	1	0	30	1	0.8001	*0.3347*
25	1	0	32	1	0.7647	*0.2920*
26	1	0	33	0	0.7457	*0.2716*
27	1	0	38	1	0.6405	*0.1816*
28	1	0	36	0	0.6845	*0.2152*
29	1	1	24	0	0.7009	*0.2291*
30	1	1	26	1	0.6578	*0.1942*
31	1	1	29	1	0.5893	*0.1484*
32	1	1	34	0	0.4699	*0.0894*
33	1	1	32	0	0.5179	*0.1105*
34	1	1	34	1	0.4699	*0.0894*
35	1	1	33	1	0.4939	*0.0995*
36	1	1	36	0	0.4223	*0.0715*
37	1	1	38	0	0.3758	*0.0564*
38	1	1	39	0	0.3531	*0.0499*
39	1	1	38	1	0.3758	*0.0564*
40	1	1	40	1	0.3310	*0.0440*
$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$		$\overline{p}_1$	$\overline{p}_0$
1.3807	1.2689	−0.6764	−0.0603		0.6344	0.2230

*_Italic_ numbers are predicted counterfactuals.

†$\widehat{p}_{1,1}=\Phi[\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_1\text{EOD}+\widehat{\beta}_3\text{Age}_1]=\Phi[1.3807+1.2689-0.6764(0)-0.0603(20)]=0.9257$, and the values the remaining 39 subjects are computed similarly by substituting the corresponding EOD and age values, and $\overline{p}_1=\frac{1}{40}(\widehat{p}_{1,1}+\widehat{p}_{1,2}+\cdots+\widehat{p}_{1,40})$.

‡$\widehat{p}_{0,1}=\Phi[\widehat{\beta}_0+\widehat{\beta}_1\text{EOD}+\widehat{\beta}_3\text{Age}_1]=\Phi[1.3807-0.6764(0)-0.0603(20)]=0.5697$, and the values the remaining 39 subjects are computed similarly by substituting the corresponding EOD and age values, and $\overline{p}_0=\frac{1}{40}(\widehat{p}_{0,1}+\widehat{p}_{0,2}+\cdots+\widehat{p}_{0,40})$.

39 subjects are computed similarly by substituting the corresponding EOD and age values with $\overline{p}_1=\frac{1}{40}(\widehat{p}_{1,1}+\widehat{p}_{1,2}+\cdots+\widehat{p}_{1,40})=0.6344$.

To estimate the risk to be expected ($\overline{p}_0$) if all subjects had not received the new therapy (regardless of their actual therapy groups), we compute

$$\Phi(\widehat{lp_{0i}})=\Phi(\widehat{\beta}_0+\widehat{\beta}_1\times 0+\widehat{\beta}_2\times\text{EOD}_i+\widehat{\beta}_3\times\text{Age}_i)$$

Table VI. Illustrative example of estimating risks, risk difference, and risk ratio using the method of counterfactual prediction and the modified Poisson regression.

		Link function		
		Probit	Cloglog	Logit
$\widehat{\beta}_0$ (Intercept)		1.3807	1.3304	2.2359
$\widehat{\beta}_1$ (Therapy)		1.2689	1.5066	2.0701
$\widehat{\beta}_2$ (EOD)		−0.6767	−0.7625	−1.0767
$\widehat{\beta}_3$ (Age)		−0.0603	−0.0706	−0.0984
Log likelihood		−21.9920	−22.0209	−22.0825
$\overline{p}_1$		0.6344	0.6316	0.6320
95 per cent CI	Wald	(0.4418, 0.8270)	(0.4447, 0.8185)	(0.4380, 0.8261)
	ln	(0.4683, 0.8594)	(0.4698, 0.8491)	(0.4649, 0.8592)
	Logit	(0.4307, 0.7992)	(0.4343, 0.7928)	(0.4272, 0.7983)
	Wilson	(0.4361, 0.7956)	(0.4393, 0.7895)	(0.4327, 0.7946)
$\overline{p}_0$	0.2230	0.2320	0.2262	
95 per cent CI	Wald	(0.0578, 0.3882)	(0.0620, 0.4020)	(0.0598, 0.3927)
	ln	(0.1063, 0.4678)	(0.1115, 0.4828)	(0.1084, 0.4721)
	Logit	(0.0996, 0.4268)	(0.1042, 0.4396)	(0.1015, 0.4307)
	Wilson	(0.1026, 0.4188)	(0.1073, 0.4315)	(0.1045, 0.4227)
Risk difference		0.4114	0.3996	0.4058
95 per cent CI	Wald	(0.1568, 0.6660)	(0.1479, 0.6512)	(0.1487, 0.6630)
	Logit	(0.1222, 0.6180)	(0.1144, 0.6045)	(0.1146, 0.6148)
	Wilson	(0.1318, 0.6133)	(0.1236, 0.6000)	(0.1242, 0.6101)
	Fisher z	(0.1301, 0.6314)	(0.1231, 0.6185)	(0.1221, 0.6282)
Risk ratio		2.8445	2.7222	2.7941
95 per cent CI	Wald(ln)	(1.2751, 6.3459)	(1.2379, 5.9862)	(1.2546, 6.2227)
	Logit	(1.3328, 6.5881)	(1.3014, 6.2436)	(1.3096, 6.4558)
	Wilson	(1.3635, 6.3957)	(1.3299, 6.0626)	(1.3397, 6.2687)
Modified Poisson		2.7256 (1.2664, 5.8660)		

for all 40 subjects. The last 20 subjects did receive the new therapy, implying their predicted $\Phi(\widehat{lp_{0i}})$ are counterfactuals. The predicted counterfactual for subject 21 is given by $\Phi[1.3807-0.6764(0)-0.0603(20)]=0.5697$, and the values the remaining 39 subjects are computed similarly by substituting the corresponding EOD and age values, with $\overline{p}_0=\frac{1}{40}(\widehat{p}_{0,1}+\widehat{p}_{0,2}+\cdots+\widehat{p}_{0,40})=0.2230$.

Confidence intervals for p_1, p_0, p_1-p_0 and p_1/p_0 are obtained using the SAS macro and are presented in Table VI. We also fitted extreme-value and logistic regression models for comparison. Overall, there are no major differences among the three models. Consistent with the simulation results, the Wald confidence intervals stand out as compared with the other methods.

It is also interesting to note that there are no material discrepancies between the modified Poisson regression method for estimating the risk ratio and the method of counterfactual prediction described above. This observation is consistent with the simulation results.

The same idea can also be applied to estimate a difference between two differences in proportions, i.e. interaction. As an illustration, suppose we are interested in the interaction between therapy and EOD. Continuing with the probit model, we can estimate risks of the following groups as

$$\overline{p}_{11}=\frac{1}{40}\sum_{i=1}^{40}\Phi(\widehat{\beta}_0+\widehat{\beta}_1+\widehat{\beta}_2+\widehat{\beta}_3\times \mathrm{Age}_i)=0.5279$$

$$\overline{p}_{10}=\frac{1}{40}\sum_{i=1}^{40}\Phi(\widehat{\beta}_0+\widehat{\beta}_1+0+\widehat{\beta}_3\times \mathrm{Age}_i)=0.7604$$

$$\overline{p}_{01}=\frac{1}{40}\sum_{i=1}^{40}\Phi(\widehat{\beta}_0+0+\widehat{\beta}_2+\widehat{\beta}_3\times \mathrm{Age}_i)=0.1310$$

and

$$\overline{p}_{00}=\frac{1}{40}\sum_{i=1}^{40}\Phi(\widehat{\beta}_0+0+0+\widehat{\beta}_3\times \text{Age}_i)=0.3132$$

Thus, the interaction is given by

$$\text{Interaction}=\overline{p}_{11}-\overline{p}_{10}-(\overline{p}_{01}-\overline{p}_{00})=-0.0503$$

Applying the MOVER with the Wilson interval for each of the four terms yields a 95 per cent confidence interval of (−0.1310, 0.0640), compared with that of (−0.1368, 0.0714), obtained from first applying the Fisher z-transformation to $p_{11}-p_{10}$ and $p_{01}-p_{00}$ separately, and second applying the MOVER approach.

As a second example, consider the data arising from a randomized experiment designed to determine whether increased reproduction reduces the longevity of male flies [40]. As done by previous authors [41], we focus here on comparing the risk of dying at 40 days between an experimental and a control group. In the experimental group, 25 males were provided with 8 receptive female flies every 2 days. In the control group, each of the 25 males was provided with 8 newly inseminated females every 2 days. A strong predictor for longevity is the length of thorax (in minimeters) which was well balanced by randomization. The full data set can be found elsewhere [42]. With 13 flies dead in the experimental group and 3 in the control group, the unadjusted odds ratio is 7.94 (95 per cent CI 1.88 to 33.50), compared with 47.00 (95 per cent CI 3.50 to 631.82) adjusted for length of thorax [41]. Note that the dramatic change in the odds ratio caused by adjusting balanced covariates is a well-known oddity of the odds ratio [43]. Application of the Zhang–Yu formula [9] in this case would result in two risk ratio estimates, resulting in misleading conclusions. This is a vivid illustration that adoption of an odds ratio simply cannot quantify the magnitude of effect, even with data from well-conducted randomization trials [8].

In contrast, using the risk ratio or risk difference has no such difficulty. The unadjusted risk ratio is 4.33 (95 per cent CI 1.40 to 13.37), which is similar to the adjusted risk ratio of 4.12 (95 per cent CI 1.56 to 10.86) obtained using the modified Poisson model. Moreover, these results are comparable to the risk ratio of 4.14 (95 per cent CI 1.89 to 9.61), obtained by predicting the counterfactuals with a logistic regression model (SAS codes are presented in the Appendix). Despite the lack of material differences between two risk ratio estimates, the narrower confidence interval of the adjusted estimate indicates efficiency gain. The estimated risk difference using the approach presented above is 0.39, with 95 per cent confidence intervals of 0.21 to 0.54 using the MOVER with the Wilson limits for separate risks, and 0.21 to 0.55 using the Fisher's z-transformation.

5. Discussion

This article has simplified the statistical calculations for risk assessment on the basis of counterfactual theory. Probit, logistic, and extreme-value regression models are discussed in the context of risk assessment. Using the method of variance estimates recovery, asymmetric confidence intervals for risk, risk difference, and risk ratio are derived. Simulation results showed the improvement as compared with those based on the delta method. For ease of practical application, a SAS macro has been developed. This should prove to be a useful alternative to the percentile bootstrap that was implemented by previous authors [18–21].

The advantage of assessing risks by predicting counterfactuals is that it does not require the homogeneity assumption. This approach has also been referred to as 'standardization' [14, 17]. An operational difficulty arises with the approach when the exposure of interest is continuous.

The simulation results have shown that applying the modified Poisson regression model in estimating the risk ratio is fairly reliable, even without the assumption of constant parameter values. This provides an empirical answer to a previous concern about this model [18]. This observation was confirmed in a study of wound healing in which the exposure is ankle-brachial index (high versus low) and the outcome is failure to heal in 24 months [18]. The modified Poisson model produced a risk ratio of 2.01 with 95 per cent confidence interval 1.51 to 2.67, compared with that of 2.04 (obtained by counterfactual prediction with a logistic regression model) with the 95 per cent confidence interval of 1.47 to 2.75 obtained using the bootstrap [18, p. 878]. This result enforces the previous suggestion that the modified Poisson model be applied to binary outcomes if the risk ratio is the parameter of interest [10, 39]. Further justification for the Poisson regression model may be found elsewhere [44].

We have also illustrated how to obtain the interaction alternative to that in previous literature, see, e.g. [2, Chapter 9]. We should note that the interaction here is defined as the difference of the differences in proportions, while previous literature defines it as a linear function of risk ratios [28]. It would be interesting to compare the utility of these two approaches, although the details are beyond the scope of this article.

The logistic regression model is predominantly used as a tool to obtain an adjusted odds ratio estimate, which is shown to be handicapped as a meaningful effect measure [7, 8]. The difficulty is also demonstrated in the second example, where the odds ratio leads to confusing results even in a well-controlled randomization trial. The property of the odds ratio changing with adjusting well-balanced factors is also referred to as 'noncollapsibility'. More discussion on the issue can be found elsewhere [5, 7, 44]. Although we deliberately downplayed the role of logistic regression model here, this is not to dismiss this model as a good tool in prediction problems. In fact, this type of application seems to be the primary goal in many classic articles that adopt the logistic regression model [11, 45].

It would be misleading to conclude the article without emphasizing that, as for any modeling exercise, the approach presented above assumes that there are no unmeasured confounding variables. Otherwise, the counterfacturals could not be predicted reliably, resulting in biased estimates. Therefore, wherever possible, rigorously executed randomized trials remain to be the gold standard in assessing effects or risks.

APPENDIX

This section provide the SAS macro call used to carry out the simulations and the example, followed by the macro.

```
**Data created from Hanley and Shapiro  (1994 Journal of Statistical Education 2(1) );
data Flies;
     input ID treatment thorax death40;
cards;
 1        0         0.64       1
 2        0         0.68       1
 3        0         0.68       0
 4        0         0.72       0
 5        0         0.72       0
 6        0         0.76       1
 7        0         0.76       0
 8        0         0.76       0
 9        0         0.76       0
10        0         0.76       0
11        0         0.80       0
12        0         0.80       0
13        0         0.80       0
14        0         0.84       0
15        0         0.84       0
16        0         0.84       0
17        0         0.84       0
18        0         0.84       0
19        0         0.84       0
20        0         0.88       0
21        0         0.88       0
22        0         0.92       0
23        0         0.92       0
24        0         0.92       0
25        0         0.94       0
26        1         0.64       1
27        1         0.64       1
28        1         0.68       1
29        1         0.72       1
30        1         0.72       1
31        1         0.74       1
32        1         0.76       1
33        1         0.76       0
34        1         0.76       0
35        1         0.78       1
36        1         0.80       1
37        1         0.80       1
38        1         0.82       1
39        1         0.82       0
40        1         0.84       1
41        1         0.84       1
42        1         0.84       0
43        1         0.84       0
44        1         0.88       0
45        1         0.88       0
46        1         0.88       0
47        1         0.88       0
48        1         0.88       0
49        1         0.88       0
50        1         0.92       0
;
 ************************************************************************
*   Note: Always remember to change data set name as well as variable names *
************************************************************************;
%CounterFactual(data =Flies,   /*data set to be analyzed*/
           yvar = death40,     /*outcome variabe*/
           trt  = treatment,      /* Exposure or treatment variable */
           xvar = thorax,       /* independent variables to be adjusted */
           link = logit,        /* link function used by PROC GENMOD: logit, probit, cloglog */
           alpha=0.05);   /* alpha level for confidence interval, default is 0.05*/
```

```
%macro CounterFactual(data=,   /*data set to be analyzed*/
           yvar =,     /*outcome variabe*/
           trt  =,      /* Exposure or treatment variable */
           xvar =,       /* independent variables to be adjusted */
           link =,        /* link function used in PROC GENMOD: logit, probit, cloglog */
           alpha=0.05    /* alpah level for confidence interval, default is 0.05*/
              );

%if &link= probit  %then %let dist='normal';
%else %if &link=logit %then %let dist='logistic';
%else %if &link=cloglog %then %let dist='extreme';

proc genmod desc data= &data;
      model &yvar = &trt &xvar/covb dist=bin link= &link;
      ods output ParameterEstimates=est(keep=estimate)
                 covB=cov(drop=rowname);
proc iml;
  start CI4P(Pest, Var, n);
      %let z=probit(1-&alpha/2);
       *Wald;
        W_L = pest - &z*sqrt(var) ;
        W_U = pest + &z*sqrt(var) ;
       *log, usd by Flanders and Rhodes (1987 J Chron Dis 40: 697-704);
        lOG_L = pest*exp(- &z*sqrt(var)/(pest));
        log_U = pest*exp(+ &Z*sqrt(var)/(pest));
        *logit;
 varlgt = var/( pest*(1-pest))**2 ;
 l=log(pest/(1-pest)) - &z*sqrt(varlgt);
         lgt_L = exp(l)/(1+exp(l));
         u=log(pest/(1-pest)) + &z*sqrt(varlgt);
         lgt_U = exp(u)/(1+exp(u));
        *Wilson, based on Newcombe (2001 Am Stat 55: 200-202);
         ll=log(pest/(1-pest)) - 2*arsinh(  &z/2 * sqrt(varlgt));
         Wln_L = exp(ll)/(1+exp(ll));
         uu=log(pest/(1-pest)) + 2*arsinh(  &z/2 * sqrt(varlgt));
         Wln_U = exp(uu)/(1+exp(uu));
 return(pest||W_L||W_U||log_L||log_U||lgt_L||lgt_U||wln_L||wln_U);
finish CI4P;

start MOVER(p1, l1, u1, p2, l2, u2, corr);
**Baded on Zou (2008 Am J Epidemiol 162: 212-224);
 point = p1-p2;
 L = p1- p2  - sqrt(max(0, (p1-l1)**2 -2*corr*(p1-l1)*(u2-p2) + (u2-p2)**2));
 U = p1- p2  + sqrt(max(0, (u1-p1)**2 -2*corr*(u1-p1)*(p2-l2) + (p2-l2)**2));
 return(point||L||U);
finish MOVER;

**Bring in data for prediction;
     use &data;
 read all var{&xvar} into X;
 n = nrow(X);
 m = ncol(X);

 X1 = J(n,2,1)||X;
 X0 = J(n,1,1)||J(n,1,0)||X;

     use cov;
 read all var _num_ into V;
 use est;
 read all var _num_ into beta;
 beta=beta[1:(m+2)];

 if &dist = 'extreme' then do;
    p_1 = 1-exp(-exp(X1 * beta));
        p_0 = 1-exp(-exp(X0 * beta));
        piece1 = exp(X1 * beta -exp(X1 * beta) )# X1;
        piece0 = exp(X0 * beta -exp(X0 * beta) )# X0;
 end;
     else do;
   p_1 = cdf(&dist,  X1 * beta);  **predicted prob, if exposed;
   p_0 = cdf(&dist,  X0 * beta);  **predicted prob, if unexposed;
       piece1 = pdf(&dist, X1 * beta)# X1;
       piece0 = pdf(&dist, X0 * beta)# X0;
 end;
```

```
     p1est = sum(p_1)/n;
 p0est = sum(p_0)/n;

   V1=0;  V0=0;  COV =0;
    do i =1 to n;
     do j=1 to n;
     Xi1 = X1[i,]; Xj1 = X1[j,];
 Xi0 = X0[i,]; Xj0 = X0[j,];
     if &dist = 'extreme' then do;
             V1  =  V1 + 1/N**2 * exp(Xi1 * beta - exp(Xi1 * beta) )*
                          exp(Xj1 * beta - exp(Xj1 * beta) )* (xi1*V*T(xj1));
             V0  =  V0 + 1/N**2 * exp(Xi0 * beta - exp(Xi0 * beta) )*
                          exp(Xj0 * beta - exp(Xj0 * beta) )* (xi0*V*T(xj0));
             COV  =  cov + 1/N**2 * exp(Xi1 * beta - exp(Xi1 * beta) )*
                          exp(Xj0 * beta - exp(Xj0 * beta) )* (xi1*V*T(xj0));
     end;
     else do;
         V1  =  V1 + 1/N**2*pdf(&dist, xi1* beta) *
                     pdf(&dist, xj1* beta) * (xi1*V*T(xj1));
         V0  =  V0 + 1/N**2*pdf(&dist, xi0* beta) *
                     pdf(&dist, xj0* beta) * (xi0*V*T(xj0));
         COV  =  COV  + 1/N**2* pdf(&dist, xi1* beta)*
                      pdf(&dist, xj0* beta) * (xi1*V*T(xj0));
       end;
    end; *END J;
   end;  *END I;

   ll = p1est-p0est - &z*sqrt(v1+v0-2*cov);
   uu =  p1est-p0est + &z*sqrt(v1+v0-2*cov);

    group1 = CI4P(p1est, V1, n); **CI for exposed risk;
    group0 = CI4P(p0est, V0, n); **CI for unexposed risk;

print '==== CI for risks ===';
name = {estimate Wald_L wald_U log_L log_U lgt_L lgt_U Wlsn_L Wlsn_U};
print group1[colname=name],
       group0[colname=name];

  **Confidenc einterval for difference;
  rho  = COV/sqrt(V1*V0);
  dWald = mover(group1[1], group1[2], group1[3], group0[1], group0[2], group0[3],  rho);
  dlgt  = mover(group1[1], group1[6], group1[7], group0[1], group0[6], group0[7],  rho);
  dWln  = mover(group1[1], group1[8], group1[9], group0[1], group0[8], group0[9],  rho);

  **The following is based on Zou and Donner (2004 Controlled Clin Trials 25: 3-12);
  vd = V1 + V0 - 2*cov;
  d = p1est - p0est;
  F_LL = log( (1+ d)/(1-d) ) - &z *2*sqrt(vd)/(1- d**2);
  F_L = (exp(F_LL) - 1)/(exp(F_LL) + 1);
  F_Uu = log( (1+ d)/(1-d) ) + &z *2* sqrt(vd)/(1-d**2);
  F_U = (exp(F_uu) - 1)/(exp(F_uu) + 1);
  Fisher = d||F_L||F_U;

print '===CI for difference==';
print dwald,   dlgt, dwln, Fisher;

** CI for Ratio;
 **log-delta;
RRest = p1est/p0est;
    var = V1/p1est**2 + V0/p0est**2 - 2 * cov/(p1est*p0est);
    Lower = RRest*exp(-&z*sqrt(var));
Upper = RRest*exp(+&z*sqrt(var));
    RlogDelta  = RRest||lower||upper;

* Zou and Donner (2008 Stat Med 27: 1693-1702);
    Rlgt  = exp(mover(log(group1[1]), log(group1[6]), log(group1[7]),
                log(group0[1]), log(group0[6]), log(group0[7]),  rho));
  RWln  = exp(mover(log(group1[1]), log(group1[8]), log(group1[9]),
                log(group0[1]), log(group0[8]), log(group0[9]),  rho));

print '=== CI for RR ===';
print  RlogDelta, Rlgt, Rwln;
quit;
 %mend  CounterFactual;
```

Acknowledgements

Dr Zou holds an Early Researcher Award of Ontario Ministry of Research and Innovation, Canada. This work was also partially supported by an Individual Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada. Helpful comments from Ms Julia Taleban are gratefully acknowledged.

References

1. Morgan SL, Winship C. *Counterfactuals and Causal Inference*. Cambridge University Press: New York, 2007.
2. Rothman KJ. *Epidemiology: An Introduction*. Oxford University Press, Cambridge University Press: Oxford, Cambridge, 2002.
3. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology* (3rd edn). Lippincott, Williams and Wilkins: New York, 2008.
4. Maldonado G, Greenland S. Estimating causal effects. *International Journal of Epidemiology* 2002; **31**:422–429.
5. Newman SC. Commonalities in the classical, collapsibility and counterfactual concepts of confounding. *Journal of Clinical Epidemiology* 2004; **57**:325–329. DOI: 10.1016/j.jclinepi.2003.07.014.
6. Schwartz LM, Woloshin S, Welch HG. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *New England Journal of Medicine* 1999; **341**:279–283.
7. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**:761–768.
8. Freedman DA. Randomization does not justify logistic regression. *Statistical Science* 2008; **23**:237–249. DOI: 10.1214/08-STS262.
9. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association* 1998; **280**:1690–1691.
10. Zou GY. A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* 2004; **159**:702–706. DOI: 10.1093/aje/kwh090.
11. Cornfield J. The university group diabetes program: a further statistical analysis of the mortality findings. *Journal of the American Medical Association* 1971; **217**:1676–1687.
12. Lee J. Covariance adjustment of rates based on the multiple logistic regression model. *Journal of Chronic Diseases* 1981; **34**:415–426.
13. Lane PW, Nelder JA. Analysis of covariance and standardization as instances of prediction. *Biometrics* 1982; **38**:613–621.
14. Wilcosky TC, Chambless LE. A comparison of direct adjustment and regression adjustment of epidemiologic measures. *Journal of Chronic Diseases* 1985; **38**:849–856.
15. Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. *American Journal of Epidemiology* 2004; **160**:301–305. DOI: 10.1093/aje/kwh221.
16. Flanders WD, Rhodes PH. Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. *Journal of Chronic Diseases* 1987; **40**:697–704.
17. Joffe MM, Greenland S. Standardized estimates from categorical regression models. *Statistics in Medicine* 1995; **14**:2131–2141.
18. Localio AR, Margolis DJ, Berlin JA. Relative risks and confidence intervals were easily computed indirectly from multivariable logistic regression. *Journal of Clinical Epidemiology* 2007; **60**:874–882. DOI: 10.1016/j.jclinepi.2006.12.001.
19. Kleinman LC, Norton EC. What's the risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. *Health Services Research* 2009; **44**:288–302. DOI: 10.1111/j.1475-67773.2008.00900.x.
20. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. *Journal of Clinical Epidemiology*; DOI: 10.1016/j.jclinepi.2008.11.004.
21. Ahern J, Hubbard A, Galea, S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *American Journal of Epidemiology* 2009; **169**:1140–1147. DOI: 10.1093/aje/kwp015.
22. Efron B. Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association* 1987; **82**:171–200.
23. Schenker N. Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association* 1985; **80**:360–361.
24. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
25. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
26. Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models: review and new applications. *American Statistician* 2004; **58**:272–279. DOI: 10.1198/000313004X5824.
27. Zou GY, Donner A. Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* 2008; **27**:1693–1702. DOI: 10.1002/sim.3095.
28. Zou GY. On the estimation of additive interaction by use of the four-by-two table and beyond. *American Journal of Epidemiology* 2008; **168**:212–224. DOI: 10.1093/aje/kwn104.
29. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**:209–212.
30. Burdick RK, Graybill FA. *Confidence Intervals on Variance Components*. Dekker: New York, 1992.
31. Howe WG. Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variable. *Journal of the American Statistical Association* 1974; **69**:789–794.
32. Graybill FA, Wang CM. Confidence intervals on nonnegative linear combinations of variances. *Journal of the American Statistical Association* 1980; **75**:869–873.
33. Newcombe RG. Logit confidence intervals and the inverse sinh transformation. *American Statistician* 2001; **55**:200–202.
34. Maldonada G, Greenland S. A comparison of the performance of model-based confidence intervals when the corrected model form is unknown: coverage of asymptotic means. *Epidemiology* 1994; **5**:171–182.
35. Wilks SS. Shortest average confidence intervals from large samples. *Annals of Mathematical Statistics* 1938; **9**:166–175.
36. Agresti A, Coull B. Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician* 1998; **52**:119–126.
37. Newcombe RG. Two-sided confidence intervals for the single proportions: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872.
38. Zou GY, Donner A. A simple alternative confidence interval for the difference between two proportions. *Controlled Clinical Trials* 2004; **25**:3–12. DOI: 10.1016/j.cct.2003.08.010.

39. Spiegelman D, Hertzmark E. Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology* 2005; **162**:199–200. DOI: 10.1093/aje/kwi188.
40. Partridge L, Farquhar M. Sexual activity and the life span of male fruit flies. *Nature* 1981; **294**:580–581.
41. Negassa A, Hanley JA. The effect of omitted covariates on confidence interval and study power in binary outcome analysis: a simulation study. *Contemporary Clinical Trials* 2007; **28**:242–248. DOI: 10.1016/j.cct.2006.08.007.
42. Hanley JA, Shapiro SH. Sexual activity and the life span of male fruit flies: a data set that gets attention. *Journal of Statistical Education* 1994; **2**(1). Q2
43. Hauck WW, Neuhaus JM, Kalbfleisch JD, Anderson S. A consequence of omitted covariates when estimating odds ratios. *Journal of Clinical Epidemiology* 1991; **44**:77–81.
44. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research* 2004; **13**:309–323. DOI: 10.1191/0962280204sm368ra.
45. Walker SH, Duncan DB. Estimation of probability of an event as a function of several independent variables. *Biometrika* 1967; **54**:167–179.