STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:639–651
Published online 12 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2346

Confidence intervals of effect size in randomized comparative parallel-group studies

Jianrong Wu[1,*,†], Guoyong Jiang[2] and Wei Wei[3]

[1] *Department of Biostatistics, St. Jude Children's Research Hospital, 332 North Lauderdale St. Memphis, TN 38105, U.S.A.*
[2] *Department of Biometrics, Cephalon, Inc., West Chester, PA 19380, U.S.A.*
[3] *Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.*

SUMMARY

We have presented a new likelihood-based approach for constructing confidence intervals of effect size that are applicable to small samples. We also conduct a simulation study to compare the coverage probability of the new likelihood-based method with other three methods proposed by Hedges and Olkin and Kraemer and Paik. Simulation studies show that the confidence interval generated by the modified signed log-likelihood ratio method possesses essentially exact coverage probabilities even for small samples, although the coverage probabilities are consistently but slightly less than the nominal level. The methods are also applied to two examples. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: confidence interval; effect size; r^*-formula; sample size; signed log-likelihood ratio; small sample

1. INTRODUCTION

In biomedical research, examining statistical significance may not be the best way to review findings from a comparative experiment where an innovative treatment is compared with a control (placebo or standard treatment). A very small, even trivial, difference can turn out to be statistically significant in a very large study. Instead, a more relevant question is whether an observed difference is large enough to matter.

The most common index of effect magnitude is effect size, defined as the difference between means of the treatment and control groups divided by the standard deviation [1].

[*] Correspondence to: Jianrong Wu, Department of Biostatistics, St. Jude Children's Research Hospital, 332 North Lauderdale St. Memphis, TN 38105, U.S.A.
[†] E-mail: jianrong.wu@stjude.org

Contract/grant sponsor: National Cancer Institute; contract/grant number: CA21765
Contract/grant sponsor: American Lebanese Syrian Associated Charities

Received 4 October 2004
Accepted 16 June 2005

It provides a simple but useful measure of how valuable a treatment really is. Many influential medical journals specifically request to include the estimated effect size and its confidence interval in a manuscript as one of the prerequisites for publication [2].

Because the effect size is a standardized mean difference, it is a dimensionless measurement of effect magnitude. In particular, it facilitates the comparison of different innovative treatments. For example, if, under similar experimental conditions, treatment A has an effect size of 0.5 on an efficacy measure, and treatment B has an effect size of 0.8 on the same efficacy measure, it is believed that treatment B is more efficacious than treatment A on that efficacy measure. Furthermore, effect size is extensively employed in a meta-analysis for expressing and for combining the results of studies that assess the effectiveness of an innovative treatment.

Point estimates of effect size have been derived and widely applied [1, 3, 4]. Although large sample asymptotic confidence intervals have also been proposed, they may not be sufficiently accurate for small samples. Exact confidence intervals have also been explored by using the exact distribution of an effect size estimator, but the effect size appears in the non-centrality parameter of a non-central t-distribution. As a result, it is computationally intractable to calculate an exact confidence interval for the effect size.

The primary goal of this article is to develop a new likelihood-based approach for constructing confidence intervals of effect size that are applicable to small samples. Moreover, these confidence intervals can be implemented without much difficulty. In Section 2, we revisit the currently available approaches for constructing confidence intervals of effect size. In Section 3, we explicitly derive new likelihood-based interval estimators. In Section 4, we examine two examples. In Section 5, we report some simulation results to compare the small sample performance of the proposed methods to competing approaches. The determination of sample size is given in Section 6, and the article concludes with some discussion in Section 7.

2. APPROXIMATE INTERVAL ESTIMATORS

Suppose, in a comparative, parallel-group study, $x_1, \ldots, x_n$ are observations from experimental units receiving an innovative treatment, and they are independently and identically distributed as $\mathrm{N}(\mu_1, \sigma^2)$; $y_1, \ldots, y_m$ are observations from those receiving a control, and they are independently and identically distributed as $\mathrm{N}(\mu_2, \sigma^2)$. An effect size of the treatment relative to the control is defined as

$$\delta = (\mu_1 - \mu_2)/\sigma$$

A natural estimator of the effect size δ is $g = (\bar{x} - \bar{y})/s$, where $\bar{x}$ and $\bar{y}$ are the sample means for the treatment and control groups, respectively, and s is the pooled sample standard deviation. It is, however, a biased estimator.

An unbiased estimator is $d = J(N-2)g$, where $N = n + m$ and the correction factor $J(s)$ is a constant that is tabulated in Reference [4, p. 80]. For a large value of s, $J(s)$ can be approximated by

$$J(s) = 1 - \frac{3}{4s - 1}$$

In this article, the above approximation is used for the calculation of d. It has been known that d has the asymptotic normal distribution with the mean δ and variance [4]

$$\sigma^2_\infty(\delta)=\frac{n+m}{nm}+\frac{\delta^2}{2(n+m)}$$

Hence the variance of d can be approximately estimated by

$$\hat{\sigma}^2(d)=\frac{n+m}{nm}+\frac{d^2}{2(n+m)}$$

A $100(1-\alpha)$ per cent confidence interval $(\delta_\mathrm{L},\delta_\mathrm{U})$ for δ is therefore given by

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

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution.

Because the variance of d depends on the unknown parameter δ, one can use a variance-stabilizing transformation of d as follows [4]:

$$h(d)=\sqrt{2}\,\sinh^{-1}\frac{d}{a}=\sqrt{2}\,\log\left(\frac{d}{a}+\sqrt{\frac{d^2}{a^2}+1}\right)$$

where $\sinh^{-1}$ is the inverse hyperbolic sine function, and

$$a=\sqrt{4+2(n/m)+2(m/n)}$$

For simplicity of notation, let $h=h(d)$ and $\eta=h(\delta)$. It is known that $\sqrt{N}(h-\eta)$ has an approximate standard normal distribution. Therefore, a $100(1-\alpha)$ per cent confidence interval $(\eta_\mathrm{L},\eta_\mathrm{U})$ for η is given by

$$\eta_\mathrm{L}=h-z_{\alpha/2}/\sqrt{N},\quad \eta_\mathrm{U}=h+z_{\alpha/2}/\sqrt{N}$$

The confidence limits η_L and η_U can be inverted to produce a confidence interval $(\delta_\mathrm{L},\delta_\mathrm{U})$ for δ by finding the values δ_L and δ_U that correspond to η_L and η_U as follows:

$$\delta_\mathrm{L}=h^{-1}(\eta_\mathrm{L}),\quad \delta_\mathrm{U}=h^{-1}(\eta_\mathrm{U}) \tag{2}$$

where $h^{-1}(x)=a\,\sinh(x/\sqrt{2})$.

Another variance-stabilizing transformation was suggested by Kraemer and Paik [5] and Kraemer [6]. Let

$$r=d/(d^2+v)^{1/2},\quad \rho=\delta/(\delta^2+v)^{1/2}$$

where $v=N(N-2)/(nm)$. It has been shown that the variate

$$\sqrt{N-2}u/(1-u^2)^{1/2}$$

has an approximate Student's t-distribution with $N-2$ degrees of freedom, where $u=(\rho-r)/(1-r\rho)$. The confidence limits ρ_L and ρ_U can be obtained based on u_L and u_U as follows:

$$\rho_\mathrm{L}=(u_\mathrm{U}-r)/(u_\mathrm{U}r-1),\quad \rho_\mathrm{U}=(u_\mathrm{L}-r)/(u_\mathrm{L}r-1)$$

Finally, we can obtain the confidence interval for δ by transforming ρ_{L} and ρ_{U} back to δ_{L} and δ_{U} as follows:

$$\delta_{\mathrm{L}} = \rho_{\mathrm{L}}\sqrt{v}/(1-\rho_{\mathrm{L}}^2)^{1/2}, \quad \delta_{\mathrm{U}} = \rho_{\mathrm{U}}\sqrt{v}/(1-\rho_{\mathrm{U}}^2)^{1/2} \tag{3}$$

3. LIKELIHOOD-BASED INTERVAL ESTIMATORS

Let $\ell(\theta)=\ell(\theta;x,y)$ be the joint log-likelihood function based on observed samples $x=(x_1,\ldots,x_n)$ and $y=(y_1,\ldots,y_m)$, where $\theta=(\delta,\lambda)$, with δ being a scalar parameter of interest, and λ being a vector nuisance parameter. It is also well known that the signed log-likelihood ratio statistic

$$r \equiv r(\delta) = \mathrm{sgn}(\hat{\delta}-\delta)\{2[\ell(\hat{\theta})-\ell(\hat{\theta}_\delta)]\}^{1/2} \tag{4}$$

is asymptotically distributed as the standard normal distribution, where $\hat{\theta}=(\hat{\delta},\hat{\lambda})$ is the maximum likelihood estimator of θ, and $\hat{\theta}_\delta=(\delta,\hat{\lambda}_\delta)$ is the constrained maximum likelihood estimator of θ for a given δ. Therefore, a $100(1-\alpha)$ per cent confidence interval for δ based on the signed log-likelihood ratio statistic is given by

$$\{\delta : |r(\delta)| \leqslant z_{\alpha/2}\} \tag{5}$$

Note that this method has the first order of accuracy only, and hence it can be quite inaccurate when the sample size is small.

In this paper, we consider a modified signed log-likelihood ratio statistic, also known as the r^*-formula, which is introduced by Barndorff-Nielsen [7, 8] and has the form

$$r^* \equiv r^*(\delta) = r(\delta) + r(\delta)^{-1}\log\left\{\frac{u(\delta)}{r(\delta)}\right\} \tag{6}$$

where $u(\delta)$ is a statistic based on $\ell(\theta)$. It has been shown that $r^*(\delta)$ is approximately distributed as the standard normal distribution with a higher order of accuracy [7–9]. Hence, a $100(1-\alpha)$ per cent confidence interval based on $r^*(\delta)$ is

$$\{\delta : |r^*(\delta)| \leqslant z_{\alpha/2}\} \tag{7}$$

In general, the statistic $u(\delta)$ can be hard to obtain. However, it has been shown that if the log-likelihood function $\ell(\theta)=\ell(\theta;x,y)$ can be written as $\ell(\theta;t)$, where t is a minimum sufficient statistic with the same dimension as θ, then

$$u \equiv u(\delta) = \frac{|\ell_{;t}(\hat{\theta}) - \ell_{;t}(\hat{\theta}_\delta) \quad \ell_{\lambda;t}(\hat{\theta}_\delta)|}{|\ell_{\theta;t}(\hat{\theta})|}\left\{\frac{|j_{\theta\theta}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\delta)|}\right\}^{1/2} \tag{8}$$

where the sample space derivatives are defined as

$$\ell_{;t}(\theta) = \frac{\partial}{\partial t}\ell(\theta;t)$$

the mixed derivatives as

$$\ell_{\theta;t}(\theta) = \frac{\partial}{\partial\theta'}\ell_{;t}(\theta)$$

$j_{\theta\theta}(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta'$ is the observed information matrix, and $j_{\lambda\lambda}(\theta) = -\partial^2\ell(\theta)/\partial\lambda\partial\lambda'$ is the observed nuisance information matrix [10]. In recent years, various adjustments to $r(\delta)$ have been proposed to improve the accuracy of the signed log-likelihood ratio method. Reid [11] gave a detailed overview of this development. The modified signed log-likelihood ratio statistics have been widely used in small sample statistical inference. In particular, this method has been applied to life data analysis in Reference [12], ratio of two independent normal means in Reference [13], and ratio of means of two independent log-normal distributions in Reference [14].

Once we have r^*, we can obtain a confidence interval of δ from (7). Note that the amount of calculations involved in obtaining r^* from (6) is not substantially more than that in obtaining r from (4). The extra calculation required is simply about $u(\delta)$ that involves the sample space derivatives and mixed derivatives. These derivatives can be obtained analytically when $\ell(\theta; x, y) = \ell(\theta; t)$.

For the case studied here about effect size, the log-likelihood function $\ell(\theta)$ can be written as

$$\ell(\theta) = -(n+m)\log\sigma - \frac{1}{2\sigma^2}t_3 + n\left(\frac{\delta}{\sigma} + \frac{\mu}{\sigma^2}\right)t_1 + \frac{m\mu}{\sigma^2}t_2 - \frac{1}{2\sigma^2}\{n(\delta\sigma+\mu)^2 + m\mu^2\}$$

where $\delta = (\mu_1 - \mu_2)/\sigma$, $\mu = \mu_2$, $\lambda = (\mu, \sigma)$, and $t = (t_1, t_2, t_3) = (\bar{x}, \bar{y}, \sum x_i^2 + \sum y_j^2)$ is a minimum sufficient statistic with the same dimension as $\theta = (\delta, \mu, \sigma)$. It can be shown that the maximum likelihood estimator $\hat{\theta} = (\hat{\delta}, \hat{\mu}, \hat{\sigma})$ is

$$\hat{\mu} = t_2$$

$$\hat{\sigma}^2 = \frac{1}{n+m}(t_3 - nt_1^2 - mt_2^2)$$

$$\hat{\delta} = (t_1 - \hat{\mu})/\hat{\sigma}$$

Furthermore, for a fixed value of δ, the constrained maximum likelihood estimator $\hat{\lambda}_\delta = (\hat{\mu}_\delta, \hat{\sigma}_\delta)$ is defined by the following equations:

$$n\delta\hat{\sigma}_\delta + (n+m)\hat{\mu}_\delta = nt_1 + mt_2$$

$$(n+m)\hat{\sigma}_\delta^2 = t_3 - n\delta\hat{\sigma}_\delta t_1 - (nt_1 + mt_2)\hat{\mu}_\delta$$

and the sample space derivatives $\ell_{;t}(\theta) = \partial\ell(\theta; t)/\partial t$ and mixed derivatives $\ell_{\theta;t}(\theta) = \partial^2\ell(\theta; t)/\partial t\partial\theta'$ are given by

$$\ell_{;t}(\theta) = \left(n\left(\frac{\delta}{\sigma} + \frac{\mu}{\sigma^2}\right), m\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)'$$

and

$$\ell_{\theta;t}(\theta)=\begin{pmatrix} \dfrac{n}{\sigma} & \dfrac{n}{\sigma^2} & -n\left(\dfrac{\delta}{\sigma^2}+\dfrac{2\mu}{\sigma^3}\right) \\ 0 & \dfrac{m}{\sigma^2} & -2m\dfrac{\mu}{\sigma^3} \\ 0 & 0 & \dfrac{1}{\sigma^3} \end{pmatrix}$$

The observed information matrix is given by

$$j_{\theta\theta}(\theta)=\begin{pmatrix} n & \dfrac{n}{\sigma} & \dfrac{n(t_1-\mu)}{\sigma^2} \\ \dfrac{n}{\sigma} & \dfrac{(n+m)}{\sigma^2} & j_{\mu\sigma}(\theta) \\ \dfrac{n(t_1-\mu)}{\sigma^2} & j_{\mu\sigma}(\theta) & j_{\sigma\sigma}(\theta) \end{pmatrix}$$

where

$$j_{\mu\sigma}(\theta)=\frac{1}{\sigma^3}\{2(nt_1+mt_2)-n\delta\sigma-2(n+m)\mu\}$$

$$j_{\sigma\sigma}(\theta)=-\frac{1}{\sigma^4}\{(n+m)\sigma^2-3t_3+n(2\delta\sigma+6\mu)t_1+6m\mu t_2-2n\delta\mu\sigma-3(n+m)\mu^2\}$$

4. TWO EXAMPLES

We illustrate the methods discussed in this paper using two existing data sets. The first example is a study on systolic blood pressure, whose data were given in Reference [15] and are reproduced in Table I for the pre- and post-data from experimental and control groups.

Table I. Systolic blood pressure data on pre and post experimental and control data from 40 hypertensives.

Experimental group										
Pre	134	135	135	136	145	147	148	150	151	153
	153	155	156	158	162	165	167	168	179	180
Post	130	131	135	136	136	138	124	126	104	142
	114	166	153	169	127	130	120	121	149	150
Control group										
Pre	139	140	141	143	151	152	152	153	153	154
	154	159	160	160	162	163	165	169	175	176
Post	130	131	144	146	128	156	161	162	160	131
	158	166	150	186	188	153	144	147	169	170

Table II. Two-sided 95 per cent confidence intervals of δ using various methods for the first example.

Method	95 per cent CI	Length
1	$(0.308, 1.618)$	1.310
2	$(0.326, 1.646)$	1.320
3	$(0.300, 1.728)$	1.428
r	$(0.351, 1.667)$	1.316
r^*	$(0.320, 1.635)$	1.315

Table III. Lamb's worm counts data.

Drug-treated sheep	18	43	28	50	16	32	13
Untreated sheep	40	54	26	63	21	37	39

The change from baseline (post–pre) for experimental and control groups will be used to estimate the effect size.

The Shapiro–Wilk test for the normality of the change from baseline data gives a p-value of 0.12 for the experimental group and a p-value of 0.17 for the control group, respectively. It presents evidence that the change from baseline data from both groups follow normal distributions. Since the F-test for equal variances gives a p-value of 0.22, the equal variances assumption is plausible.

In practice, if an increase represents an improvement, the effect size δ and its estimates are defined and calculated as in previous sections. However, if a decrease represents an improvement, δ and its estimates are defined and calculated as -1 times those in previous sections. Here, a decrease in systolic blood pressure represents an improvement. Therefore, the unbiased estimate of δ is 0.963. The 95 per cent confidence intervals of δ based on methods 1, 2, 3 as specified in (1), (2), (3), respectively, r and r^* are presented in Table II. All five methods give similar 95 per cent confidence intervals but the confidence limits obtained by methods 2 and r^* are much closer.

The second example is a study presented in Reference [16]. A sample of 14 worm-infected lambs was randomly divided into two groups. Seven were injected with an experimental drug and the remainder were left untreated. After a 6-month period, the lambs were slaughtered and worm counts were recorded as shown in Table III. The Shapiro–Wilk test for the normality of worm counts data from the drug-treated and untreated groups gives p-values of 0.75 and 0.51, respectively. It demonstrates that data from both groups follow normal distributions. Since the F-test for equal variances gives a p-value of 0.92, the equal variance assumption is plausible. Here, a decrease in worm counts represents an improvement. Therefore, the unbiased estimate of δ is 0.744. The 95 per cent confidence intervals based on methods 1, 2, 3, r and r^* are presented in Table IV. The confidence limits obtained by methods 2 and r^* are much closer than those obtained by methods 1, 3 and r.

From both examples, we observe the following findings. Hedges and Olkin's [4] interval in (2) and the r^*-interval are almost identical, while Hedges and Olkin's [4] interval in (1) shifts to the left side with the shortest interval length among the five intervals and the

Table IV. Two-sided 95 per cent confidence intervals of δ using various methods for the second example.

Method	95 per cent CI	Length
1	$(-0.340, 1.827)$	2.167
2	$(-0.313, 1.903)$	2.216
3	$(-0.376, 2.133)$	2.509
r	$(-0.235, 1.955)$	2.190
r^*	$(-0.311, 1.877)$	2.188

r-interval shifts to the right side. Kraemer and Paik's [5] interval in (3) covers all other four intervals with the largest interval length. These observations motivate us to study the coverage probabilities and other performance measures of the five methods in small samples through Monte Carlo simulations.

5. SIMULATION STUDY

In this section, we carry out simulation studies to compare the performance of the five approximate methods: confidence intervals as specified in (1), (2), (3), and confidence intervals based on the signed log-likelihood ratio r as specified in (5), and the modified signed log-likelihood ratio r^* as specified in (7), for constructing a two-sided 90 per cent confidence interval for the effect size δ in small samples. The performance of a method is judged using the following criteria:

1. Coverage probability: the percentage of a true parameter value falling within the intervals.
2. Coverage error: the absolute difference between the nominal level and coverage probability.
3. Upper/lower error probability: the percentage of a true parameter value falling above/below the intervals.
4. Average bias: the average of the absolute difference between upper and lower error probabilities and nominal levels.

The desired values for the coverage probability, coverage error, upper and lower error probabilities and average bias are 0.9, 0, 0.05, 0.05 and 0, respectively. These values reflect the desired properties of the exact coverage probability, accuracy and symmetry of the upper and lower error probabilities.

The sample sizes considered are $(n,m)=(5,5)$, (5,10) and (10,10), and the standard deviation σ ranges from 4 to 0.2. We set $(\mu_1,\mu_2)=(2,1)$, and hence δ ranges from 0.25 to 5.0. For each parameter configuration, we generated 10 000 random samples from the normal distribution. Because the nominal confidence level of the confidence intervals studied in the simulations is 0.9, the standard error of the simulated coverage probabilities on 10 000 random samples is 0.003. The simulated coverage probabilities, coverage errors, upper/lower error probabilities and average biases for each method are given in Tables V–VII, where the simulated average lengths are also given for each method.

Table V. Coverage probabilities, coverage errors, error probabilities and average biases and lengths of two-sided 90 per cent confidence intervals for various methods with $(n, m) = (5, 5)$.

δ	Method	Coverage probability	Coverage error	Upper error probability	Lower error probability	Average bias	Average length
0.25	1	**0.9125**	0.0125	0.0422	0.0453	0.0063	2.1439
	2	0.8981	0.0019	0.0513	0.0506	0.0009	2.1926
	3	**0.9253**	0.0253	0.0352	0.0395	0.0127	2.5207
	r	**0.8421**	0.0579	0.0722	0.0857	0.0289	2.1753
	r^*	0.8959	0.0041	0.0527	0.0514	0.0020	2.1729
0.5	1	**0.9121**	0.0121	0.0423	0.0456	0.0061	2.1685
	2	0.8981	0.0019	0.0542	0.0477	0.0032	2.2177
	3	**0.9273**	0.0273	0.0336	0.0391	0.0137	2.5834
	r	**0.8428**	0.0572	0.0643	0.0929	0.0286	2.2138
	r^*	0.8966	0.0034	0.0511	0.0523	0.0017	2.2105
0.75	1	**0.9108**	0.0108	0.0472	0.0420	0.0054	2.2085
	2	0.8969	0.0031	0.0467	0.0564	0.0049	2.2586
	3	**0.9316**	0.0316	0.0372	0.0312	0.0158	2.6841
	r	**0.8437**	0.0363	0.0570	0.0993	0.0282	2.2697
	r^*	0.8954	0.0046	0.0511	0.0535	0.0023	2.2651
1.0	1	0.9074	0.0074	0.0420	0.0506	0.0043	2.2630
	2	0.8957	0.0043	0.0575	0.0468	0.0054	2.3144
	3	**0.9365**	0.0365	0.0290	0.0345	0.0183	2.8193
	r	**0.8415**	0.0585	0.0516	0.1069	0.0293	2.3487
	r^*	0.8950	0.0050	0.0509	0.0541	0.0025	2.3422
2.0	1	**0.8903**	0.0097	0.0436	0.0661	0.0113	2.6042
	2	**0.8847**	0.0153	0.0651	0.0502	0.0076	2.6633
	3	**0.9544**	0.0544	0.0234	0.0222	0.0272	3.6195
	r	**0.8357**	0.0643	0.0341	0.1302	0.0481	2.8310
	r^*	0.8954	0.0046	0.0487	0.0559	0.0036	2.8138
5.0	1	**0.8445**	0.0555	0.0443	0.1112	0.0334	4.2767
	2	**0.8547**	0.0453	0.0726	0.0727	0.0226	4.3738
	3	**0.9751**	0.0751	0.0186	0.0063	0.0376	7.0456
	r	**0.8215**	0.0785	0.0152	0.1633	0.0740	5.0375
	r^*	0.8978	0.0022	0.0464	0.0558	0.0047	4.9843

Method 1: Hedges and Olkin's [4] interval in (1); method 2: Hedges and Olkin's [4] interval in (2); method 3: Kraemer and Paik's [5] interval in (3); the highlighted values are those which exceed three standard errors.

From Tables V–VII, we observe that the signed log-likelihood ratio method is liberal. Its coverage probabilities are lower than three standard errors below the nominal level (0.90) for all cases in the simulation study. Also, it has the largest coverage errors and average biases in all cases. Its error probabilities can be extremely inaccurate and asymmetric (0.015 *versus* 0.16) for large δ.

In contrast, method 3 is conservative. Its coverage probabilities are greater than three standard errors above the nominal level for all cases. Also, it has the second largest coverage

Table VI. Coverage probabilities, coverage errors, error probabilities and average biases and lengths of two-sided 90 per cent confidence intervals for various methods with $(n, m) = (5, 10)$.

δ	Method	Coverage probability	Coverage error	Upper error probability	Lower error probability	Average bias	Average length
0.25	1	0.9042	0.0042	0.0482	0.0476	0.0021	1.8393
	2	0.8947	0.0053	0.0547	0.0506	0.0026	1.8671
	3	**0.9123**	0.0123	0.0441	0.0436	0.0062	2.0297
	r	**0.8605**	0.0395	0.0639	0.0756	0.0197	1.8502
	r^*	0.8929	0.0071	0.0524	0.0547	0.0035	1.8494
0.5	1	0.9030	0.0030	0.0480	0.0490	0.0015	1.8576
	2	0.8938	0.0062	0.0561	0.0501	0.0031	1.8857
	3	**0.9149**	0.0149	0.0423	0.0428	0.0075	2.0722
	r	**0.8585**	0.0415	0.0602	0.0813	0.0207	1.8503
	r^*	0.8915	0.0085	0.0534	0.0551	0.0042	1.8495
0.75	1	0.9007	0.0007	0.0512	0.0481	0.0016	1.8879
	2	0.8912	0.0088	0.0507	0.0581	0.0044	1.9164
	3	**0.9177**	0.0177	0.0427	0.0396	0.0089	2.1418
	r	**0.8600**	0.0400	0.0544	0.0856	0.0200	1.9122
	r^*	0.8916	0.0084	0.0522	0.0562	0.0042	1.9105
1.0	1	0.8986	0.0014	0.0474	0.0540	0.0033	1.9295
	2	**0.8901**	0.0099	0.0600	0.0499	0.0050	1.9587
	3	**0.9238**	0.0238	0.0367	0.0395	0.0119	2.2360
	r	**0.8582**	0.0418	0.0511	0.0907	0.0209	1.9651
	r^*	**0.8905**	0.0095	0.0529	0.0566	0.0047	1.9625
2.0	1	**0.8894**	0.0106	0.0467	0.0639	0.0086	2.1941
	2	**0.8868**	0.0132	0.0619	0.0513	0.0066	2.2272
	3	**0.9451**	0.0451	0.0291	0.0258	0.0226	2.8043
	r	**0.8561**	0.0439	0.0354	0.1085	0.0366	2.2957
	r^*	0.8942	0.0058	0.0509	0.0549	0.0029	2.2884
5.0	1	**0.8646**	0.0354	0.0455	0.0899	0.0222	3.5274
	2	**0.8716**	0.0284	0.0665	0.0619	0.0142	3.5807
	3	**0.9723**	0.0723	0.0188	0.0089	0.0362	5.3095
	r	**0.8478**	0.0522	0.0208	0.1314	0.0553	3.8886
	r^*	0.8969	0.0031	0.0479	0.0552	0.0037	3.8655

Method 1: Hedges and Olkin's [4] interval in (1); method 2: Hedges and Olkin's [4] interval in (2); method 3: Kraemer and Paik's [5] interval in (3); the highlighted values are those which exceed three standard errors.

errors and average biases in all cases. Its error probabilities are lower than the nominal level (0.05) and average biases are relatively large.

Meanwhile, methods 1 and 2 both perform well in terms of coverage probabilities and symmetric error probabilities for sample sizes exceeding 5 per group and effect sizes up to 1.0. Their average biases are also relatively small.

Not surprisingly, the modified signed log-likelihood ratio interval excels among the five studied here. It has nearly exact coverage probabilities, or equivalently, nearly zero coverage errors, and has the smallest average biases among the five methods in all the cases studied. In

Table VII. Coverage probabilities, coverage errors, error probabilities and average biases and lengths of two-sided 90 per cent confidence intervals for various methods with $(n,m)=(10,10)$.

δ	Method	Coverage probability	Coverage error	Upper error probability	Lower error probability	Average bias	Average length
0.25	1	0.9086	0.0086	0.0446	0.0468	0.0043	1.4957
	2	0.9012	0.0012	0.0508	0.0480	0.0014	1.5126
	3	**0.9141**	0.0141	0.0419	0.0440	0.0071	1.6067
	r	**0.8749**	0.0251	0.0583	0.0668	0.0125	1.5010
	r^*	0.8990	0.0010	0.0503	0.0507	0.0005	1.5006
0.5	1	0.9058	0.0058	0.0461	0.0481	0.0029	1.5126
	2	0.8997	0.0003	0.0523	0.0480	0.0021	1.5297
	3	**0.9174**	0.0174	0.0392	0.0434	0.0087	1.6444
	r	**0.8756**	0.0244	0.0539	0.0705	0.0122	1.5210
	r^*	0.8983	0.0017	0.0503	0.0514	0.0008	1.5204
0.75	1	0.9073	0.0073	0.0491	0.0436	0.0037	1.5404
	2	0.8987	0.0013	0.0474	0.0539	0.0033	1.5579
	3	**0.9219**	0.0219	0.0414	0.0367	0.0110	1.7056
	r	**0.8759**	0.0241	0.0489	0.0752	0.0132	1.5543
	r^*	0.8988	0.0012	0.0491	0.0521	0.0015	1.5533
1.0	1	0.9068	0.0068	0.0431	0.0501	0.0035	1.5786
	2	0.9007	0.0007	0.0535	0.0458	0.0038	1.5965
	3	**0.9288**	0.0288	0.0333	0.0379	0.0144	1.7882
	r	**0.8766**	0.0234	0.0437	0.0797	0.0180	1.5998
	r^*	0.9012	0.0012	0.0478	0.0510	0.0016	1.5983
2.0	1	0.8983	0.0017	0.0427	0.0590	0.0081	1.8191
	2	0.8943	0.0057	0.0576	0.0481	0.0048	1.8397
	3	**0.9488**	0.0488	0.0259	0.0253	0.0244	2.2785
	r	**0.8721**	0.0279	0.0350	0.0929	0.0289	1.8826
	r^*	0.8993	0.0007	0.0481	0.0526	0.0023	1.8782
5.0	1	**0.8787**	0.0213	0.0437	0.0776	0.0170	3.0040
	2	**0.8792**	0.0208	0.0648	0.0560	0.0104	3.0379
	3	**0.9725**	0.0725	0.0182	0.0093	0.0363	4.3901
	r	**0.8622**	0.0378	0.0243	0.1135	0.0446	3.2301
	r^*	0.8992	0.0008	0.0464	0.0544	0.0040	3.2167

Method 1: Hedges and Olkin's [4] interval in (1); method 2: Hedges and Olkin's [4] interval in (2); method 3: Kraemer and Paik's [5] interval in (3); the highlighted values are those which exceed three standard errors.

addition, its upper and lower error probabilities are more symmetric and accurate than those of the other methods in the most cases studied.

Overall, both methods 1 and 2 and the modified signed log-likelihood ratio method perform much better than the liberal signed log-likelihood ratio method and the conservative method 3. Furthermore, the modified signed log-likelihood ratio interval is superior to the other intervals for all comparison criteria discussed in this paper. Therefore, based on the simulation results, we recommend the modified signed log-likelihood ratio method for use. Meanwhile, methods

1 and 2 have simple analytic solutions for confidence limits, and they also perform well for sample sizes exceeding 5 per group and effect sizes up to 1.0. Accordingly, methods 1 or 2 can also be used for sample sizes exceeding 5 per group and effect sizes up to 1.0. As a final note, the average lengths are very close for all methods except method 3 which has the largest average length.

We have also carried out simulation studies for sample sizes exceeding 10 per group. These results, though not reported here, essentially corroborate those of Tables V–VII, and are consistent with the simulation results presented in Reference [4, p. 87].

6. SAMPLE SIZE DETERMINATION

In many clinical trials, one of the primary objectives is to evaluate the efficacy of an innovative treatment. The estimation of effect size is a useful tool to achieve this objective. To ensure a certain degree of accuracy for effect size estimation, a sufficient number of subjects are needed. If we want to be $100(1-\alpha)$ per cent confident that the error will not exceed a specified amount E in estimating the effect size δ, the required sample size can be approximated based on Hedges and Olkin's [4] interval in (1) as

$$n = \frac{z_{\alpha/2}^2((1+\lambda)/\lambda) + (\delta^2/2(1+\lambda))}{E^2}$$

where $\lambda = m/n$.

7. DISCUSSION

In this paper, we have presented a new likelihood-based approach for constructing confidence intervals of effect size that are applicable to small samples. Simulation studies show that the confidence interval generated by the modified signed log-likelihood ratio method excels among the five methods studied in this paper and possesses essentially exact coverage probabilities even for small samples, although the coverage probabilities are consistently but slightly less than the nominal level. The calculations involved are straightforward because all the involved quantities can be derived explicitly. We recommend the confidence interval based on the modified signed log-likelihood ratio method for use. It can be applied to randomized parallel-group experiments, where the comparison of treatments is of primary interest. Meanwhile, some confidence intervals of effect size proposed by Hedges and Olkin [4], as in (1) and (2), perform well for sample sizes exceeding 5 per group and effect sizes up to 1.0. Typically, effect sizes rarely exceed 1.0. In addition, they have simple analytical solutions for confidence limits. Therefore, these confidence intervals can be used for studies with sample sizes exceeding 5 per group and moderate effect sizes.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge two anonymous reviewers for their valuable comments and suggestions that significantly improved the earlier version of this paper. The first author's research was supported in part by National Cancer Institute (NCI), support grant CA21765 and American Lebanese Syrian Associated Charities (ALSAC).

REFERENCES

1. Glass GV. Primary, secondary, and meta-analysis of research. *Educational Researcher* 1976; **5**:3–8.
2. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Medical Association* 2001; **285**:1987–1991.
3. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 1981; **6**:107–128.
4. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press: San Diego, 1985.
5. Kraemer HC, Paik M. A central t approximation to the noncentral t distribution. *Technometrics* 1979; **21**:357–360.
6. Kraemer HC. Theory of estimation and testing of effect sizes: use in meta-analysis. *Journal of Educational Statistics* 1983; **8**:93–101.
7. Barndorff-Nielsen OE. Inference on full and partial parameters, based on the standardized signed log-likelihood ratio. *Biometrika* 1986; **73**:307–322.
8. Barndorff-Nielsen OE. Modified signed log-likelihood ratio. *Biometrika* 1991; **78**:557–563.
9. Barndorff-Nielsen OE, Cox DR. *Inference and Asymptotics*. Chapman & Hall: London, 1994.
10. Fraser DAS, Reid N, Wu J. A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* 1999; **86**:249–264.
11. Reid N. Likelihood and higher-order approximations to tail areas: a review and annotated bibliography. *Canadian Journal of Statistics* 1996; **24**:141–166.
12. Wong ACM, Wu J. Practical small-sample asymptotics for distributions used in life-data analysis. *Technometrics* 2000; **42**:149–155.
13. Wu J, Jiang G. Small sample likelihood inference for the ratio of means. *Computational Statistics and Data Analysis* 2001; **38**:181–190.
14. Wu J, Jiang G, Wong ACM, Sun X. Likelihood analysis for the ratio of means of two independent log normal distributions. *Biometrics* 2002; **58**:144–150.
15. Kraemer HC, Andrews G. A non-parametric technique for meta-analysis. *Psychological Bulletin* 1982; **81**:404–412.
16. Ott RL. *An Introduction to Statistical Methods and Data Analysis* (4th edn). Duxbury Press: Belmont, 1998.