Statistics & Probability Letters 19 (1994) 147–151
North-Holland
27 January 1994

On the calculation of standard error for quotation in confidence statements

Song Xi Chen and Peter Hall
Centre for Mathematics and its Applications, Australian National University, Canberra, Australia

Received February 1993
Revised June 1993

Abstract: The variability of a parameter estimator $\hat{\theta}$ is often expressed by quoting, within parentheses, a value of standard error, say $\hat{\sigma}$. Typically the latter is calculated via an asymptotic formula, or by application of the bootstrap or jackknife. A common interpretation of the quotation "$\hat{\theta}(\hat{\sigma})$" is that the confidence interval $I=(\hat{\theta}-2\hat{\sigma},\ \hat{\theta}+2\hat{\sigma})$ covers the true value of θ with probability approximately 0.95. However, in problems where the allowable range of θ is sharply restricted by the context it is sometimes the case that one or other of the endpoints of I lies outside the allowable range. This happens because information about the allowable range is not adequately taken into account when computing the standard error. Examples include confidence statements about ratios of means. In the present paper we suggest a remedy for this problem. We propose a new, symmetrized version of Owen's empirical likelihood method, and use it to construct range-respecting, symmetric confidence intervals. These intervals suggest new formulae for standard error. An alternative approach may be based on a version of the bootstrap, although the latter is more expensive in computing time.

Keywords: Bootstrap; confidence interval; coverage; empirical likelihood; range respecting; standard error.

1. Introduction and methodology

A convenient and often-used way of making a confidence statement about the value of an unknown parameter, θ, is to quote the conclusions of data analysis in the form "$\hat{\theta}(\hat{\sigma})$", where $\hat{\theta}$ denotes a point estimator of θ and $\hat{\sigma}$ is a standard error. This statement is commonly interpreted to mean that the interval $I\equiv(\hat{\theta}-2\hat{\sigma},\ \hat{\theta}+2\hat{\sigma})$ covers the parameter value with probability 0.95. We therefore refer to the statement "$\hat{\theta}(\hat{\sigma})$" as the standard error form of a confidence interval. If the central limit theorem holds for $\hat{\theta}$ then asymptotically, as sample size n increases, I covers θ with probability 0.9544.

It can be rather difficult to compute a standard error which, when used in this way, does not produce a confidence statement that violates constraints on the known range of values that θ can assume. For example, part of I might lie outside $[0,\infty)$ when θ is a ratio of two positive means. These examples, which will be illustrated in Section 2, are instances where information about the range of the parameter θ may not be adequately taken into account when calculating standard error. Our aim in the present paper is to suggest ways in which proper account may be taken of range, thereby producing a range-respecting confidence interval $(\hat{\theta}-2\hat{\sigma},\ \hat{\theta}+2\hat{\sigma})$ having the appropriate nominal coverage, and at the same time giving a consistent standard error $\hat{\sigma}$. Our contributions are two-fold. First, we suggest that standard error be estimated directly from symmetric confidence intervals, rather than the other way around. Second, we propose new methods for constructing symmetric confidence intervals.

Correspondence to: Dr. P. Hall, Centre for Mathematics and its Applications, Australian National University, GPO Box 4, Canberra ACT 2601, Australia.

SSDI 0167-7152(93)E0096-C

The technique that we shall consider in greatest detail is based on Owen's (1988, 1990) empirical likelihood argument. This allows the construction, in a nonparametric setting, of a function l that shares many of the properties of loglikelihood ratio for parametric problems. This function is nonnegative and convex; is uniquely minimized at the bootstrap estimator, $\hat{\theta}$; has an asymptotic chi-squared distribution when evaluated at the parameter value (i.e. it satisfies Wilks' theorem); and, most importantly for our purposes, is infinite outside the allowable range of possible values of θ. In the usual application of empirical likelihood one employs l in much the same way as a parametric loglikelihood, to construct a confidence interval for the true value, θ_0 say, of θ. For example, if the asymptotic distribution of $l(\theta_0)$ were χ_1^2 (chi-squared with one degree of freedom), which would always be the case in the univariate contexts that are considered here, then to construct a confidence interval with nominal coverage α one would first refer to tables to find that value of $c > 0$ which satisfies

$$P\left(\chi_1^2 \leqslant c\right) = \alpha, \tag{1.1}$$

and then take the interval to be $J = \{\theta\colon l(\theta) \leqslant c\}$.

The empirical likelihood interval is not centred on the value of $\hat{\theta}$, but that problem may be remedied by employing the function

$$m(\theta) = \tfrac{1}{2}\left\{l(\theta) + l(2\hat{\theta} - \theta)\right\} \tag{1.2}$$

in place of l. With c defined by (1.1), take $\hat{a} > 0$ to be the solution of the equation $m(\hat{\theta} + \hat{a}) = c$, and let the interval be $I = (\hat{\theta} - \hat{a},\ \hat{\theta} + \hat{a})$. We suggest adopting $c = 4$ and $\alpha = 0.9544$, and taking $\hat{\sigma} = \frac{1}{2}\hat{a}$ to be the standard error of $\hat{\theta}$. Then $\hat{\sigma}$ is root-n consistent for the true standard deviation σ, in the sense that $\hat{\sigma}/\sigma = 1 + O_p(n^{-1/2})$ as $n \to \infty$.

Since $l(\theta) = \infty$ if θ lies outside the allowable range, then, provided only that the range is contiguous, I will lie wholly within the allowable range. The interval I is not Bartlett correctable, although the method of bootstrap calibration does reduce coverage error from $O(n^{-1})$ to $O(n^{-2})$. This refinement, and other extensions of our method, will be discussed below. The claims made earlier in this paragraph, about root-n consistency of $\hat{\sigma}$, coverage accuracy of I and non-Bartlett correctability of I, may be established using arguments of DiCiccio, Hall and Romano (1991) and Hall and La Scala (1990).

One may improve the coverage accuracy of I from $O(n^{-1})$ to $O(n^{-2})$ by employing a bootstrap, rather than chi-squared, approximation to the distribution of $m(\theta_0)$. Specifically, resample with replacement from the sample χ, let l^* and $\hat{\theta}^*$ denote the versions of l and $\hat{\theta}$ computed for the resample rather than the sample, define $m^*(\theta) = \frac{1}{2}\{l^*(\theta) + l^*(2\hat{\theta}^* - \theta)\}$, let $\hat{c}$ denote the solution of the equation $P\{m^*(\hat{\theta}) \leqslant \hat{c} \mid \chi\} = 0.9544$, define $\tilde{a}$ by $m(\hat{\theta} + \tilde{a}) = \hat{c}$, and put $\tilde{\sigma} = \frac{1}{2}\tilde{a}$. Then $(\hat{\theta} - 2\tilde{\sigma},\ \hat{\theta} + 2\tilde{\sigma})$ covers θ with probability $0.9544 + O(n^{-2})$ and is contained within the range of allowable values of θ, and $\tilde{\sigma}/\sigma = 1 + O_p(n^{-1/2})$.

A non-standard, symmetrized version of the percentile bootstrap method may be used in place of empirical likelihood to construct symmetric, range-preserving confidence intervals for θ_0, from which a standard error may be calculated much as suggested earlier. The technique is as follows. Let $\hat{\theta}^*$ denote the version of $\hat{\theta}$ computed from a resample χ^* drawn randomly, with replacement, from the original sample χ. Write $\hat{x}_\alpha$ for the solution of the equation $P(\hat{\theta}^* \leqslant \hat{x}_\alpha \mid \chi) = \alpha$. The usual percentile method confidence interval for θ, with nominal level α, is $(\hat{x}_{(1-\alpha)/2},\ \hat{x}_{(1+\alpha)/2})$. However, it is not suitable for our purpose since it is not symmetric about $\hat{\theta}$. Instead, we use the interval $I = (\hat{x}_{\hat{\beta}},\ \hat{x}_{\alpha+\hat{\beta}})$, where $\hat{\beta} \in (0,\ 1-\alpha)$ is that unique function of the data such that $\hat{x}_{\alpha+\hat{\beta}} - \hat{\theta} = \hat{\theta} - \hat{x}_{\hat{\beta}}$. Since each $\hat{x}_\gamma$, $0 < \gamma < 1$, is in the allowable range of θ then, provided the range is contiguous, I is a subset of that range. We suggest taking $\alpha = 0.9544$ and $\hat{\sigma} = \frac{1}{4}(\hat{x}_{\alpha+\hat{\beta}} - \hat{x}_{\hat{\beta}})$, in which case $\hat{\sigma}$ is root-n consistent for the standard deviation of $\hat{\theta}$.

Calibrating the percentile method is similar in both practice and effect to calibrating our modified version of empirical likelihood. It produces a standard error $\tilde{\sigma}$ which enjoys the property that $(\hat{\theta} - 2\tilde{\sigma},\ \hat{\theta} + 2\tilde{\sigma})$ covers θ with probability $0.9544 + O(n^{-2})$ and is contained within the range of allowable values of θ, and $\tilde{\sigma}/\sigma = 1 + O_p(n^{-1/2})$. Note that the percentile-t bootstrap is not suitable for this application, since it does not preserve range.

We favour the construction of nominal 95.44% intervals, if only because the quantities "$\hat{\theta} \pm 2\hat{\sigma}$" are used so often to describe a confidence statement. However, our methods obviously have many variants. At the risk of introducing a note of arbitrariness into our discussion, we mention two of them here. The confidence interval $(\hat{\theta} - \hat{\sigma}, \hat{\theta} + \hat{\sigma})$ has nominal coverage 0.6826, and is also suggested by the confidence statement "$\hat{\theta}(\hat{\sigma})$", as is the "95% interval" $(\hat{\theta} - 1.96\hat{\sigma},\ \hat{\theta} + 1.96\hat{\sigma})$. Thus, for example, we could take $c = 1$ and $\alpha = 0.6826$, or $c = 1.96$ and $\alpha = 0.95$, in the procedure described three paragraphs above, and construct the confidence interval accordingly.

The numerical study in Section 2 treats both the empirical likelihood and percentile bootstrap methods, and describes the effects of calibration on both. The coverage accuracies were substantially improved by calibration, with little to choose between them. However, the empirical likelihood method was much faster to implement then its percentile bootstrap counterpart, taking less than one twentieth the amount of time in both calibrated and uncalibrated forms. This is due to very fast convergence of Brent's method employed to compute empirical likelihood quantiles, relative to the greater computational expense of conducting simulation to achieve similar accuracy of the percentile bootstrap method.

Our idea — of computing standard error from a confidence interval, rather than the other way around — has applications that go well beyond

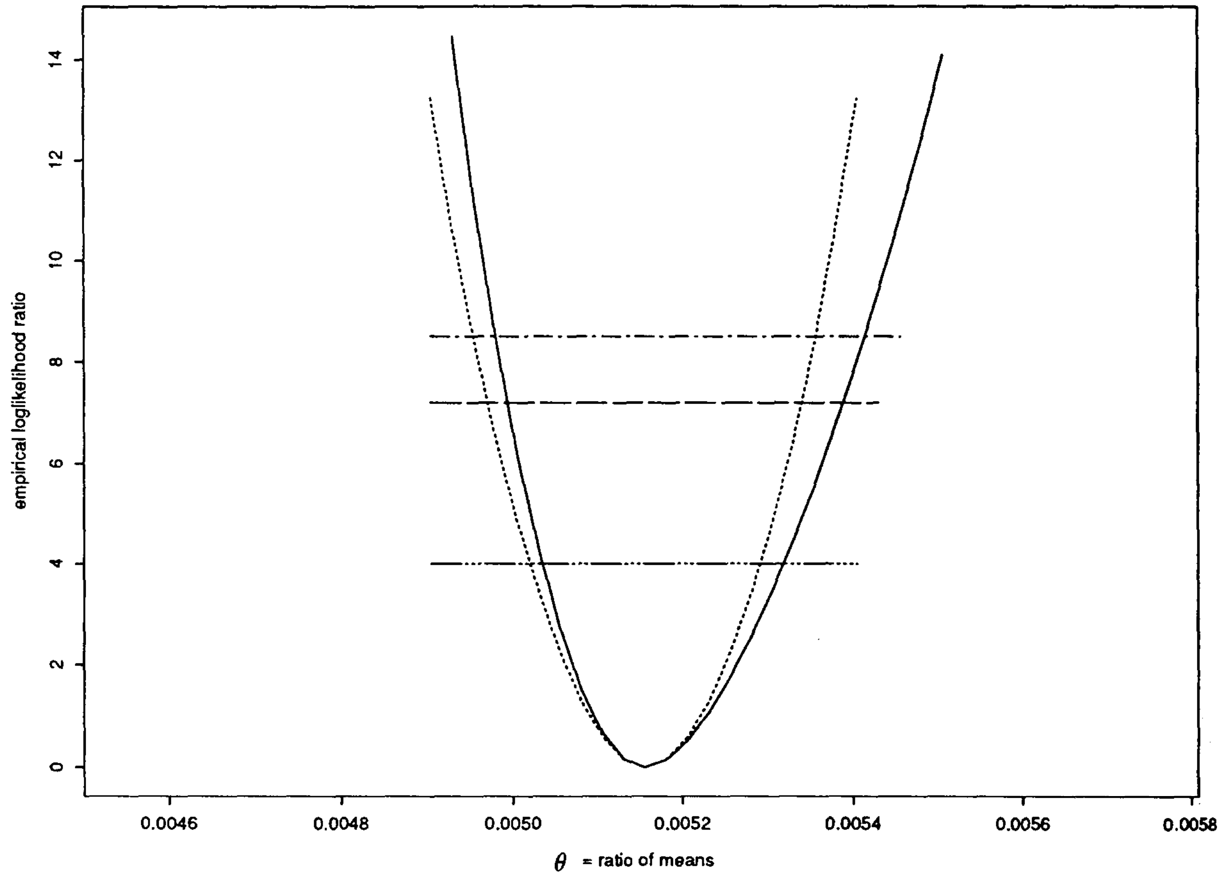

Figure 1. Empirical loglikelihood for a ratio of means, based on the law school data of Efron (1979). The curve represented by the unbroken line is the usual empirical loglikelihood ratio curve, $l(\theta)$, and the curve represented by the dotted line is our symmetrised version, $m(\theta)$. The horizontal line –···–··– is drawn at ordinate $c = 4$, the line ——— is drawn at ordinate 7.2, and the line ·–·–· is drawn at ordinate 8.5.

that of solving the problem of respecting range, addressed in this paper. For example, variants of our idea may be employed to construct accurate confidence statements, in standard error from, for a correlation coefficient and a quantile. The latter is a particularly awkward problem to solve by other means, since the true standard deviation of $\hat{\theta}$ is very nearly proportional to the sampling density at θ, and so direct calculation of standard error typically involves density estimation. This problem may be solved very effectively by using the ideas in this paper together with either a confidence interval based on interpolated order statistics, or one based on a smoothed version of empirical likelihood.

We conclude this section with an example, based on the now-classic law school data of Efron (1979). These data consist of paired observations, and we consider them in the context of estimating the ratio of the means of the two marginal populations. The numerical value of the ratio of the smaller to the larger sample mean is $\hat{\theta} = 0.005156$. The unbroken and dotted curves in Figure 1 depict respectively the usual empirical loglikelihood ratio $l(\theta)$ and the symmetrised empirical loglikelihood ratio $m(\theta)$ for these data. Of course, both have their minimum at 0.005156. Horizontal lines have been drawn at three potential values of c. The lowest line, corresponding to $c = 4$, is obtained by referring to chi-squared tables for one degree of freedom; see (1.1). The middle line, at $c = 7.2$, is derived by employing a bootstrap approximation to the distribution of $m(\theta_0)$, as described six paragraphs above. The highest line, at $c = 8.5$, is obtained from a bootstrap approximation to the distribution of $l(\theta_0)$; this is not of use to us in our procedure, but is included for the sake of comparison. The width of a confidence interval based on the middle value of c is obtained by finding the distance between the two arms of the dotted curve at the ordinate $c = 7.2$, and equals 0.000254. This is the value of $2\tilde{a}$ in the discussion six paragraphs above. Our estimate of standard error is simply one quarter of the width of that confidence interval, and so equals 0.00006363. Thus, our "standard error form" of the confidence statement is "5.156×10^{-3} (6.4×10^{-5})".

2. Numerical work

Here we summarise numerical work carried out to determine small-sample properties of various approaches to constructing confidence statements in standard error form. For the sake of brevity we consider here only the case where the parameter of interest is the ratio of two means. Our aim was

Table 1
Characteristics of 0.9544-level empirical likelihood bootstrap and Normal approximation confidence intervals for a ratio of two means, when $\theta = 1/m$, $m = 20$ or 40, and $n = 15$.

	$\theta = 0.05$					$\theta = 0.025$				
	Coverage	% of out range	Std error est.	Av. int. length	Std devn. of int. length	Coverage	% of out range	Std error est.	Av. int. length	Std devn. of int. length
Symmetric e.l.	0.821	0	0.0267	0.0535	0.006	0.843	0	0.0136	0.0272	0.0003
Calibrated symmetric e.l.	0.925	0	0.0424	0.0849	0.001	0.929	0	0.0214	0.0429	0.0006
Symmetric percentile	0.857	0	0.0304	0.0608	0.0007	0.872	0	0.0155	0.0310	0.0004
Calibrated symmetric percentile	0.909	0	0.0362	0.0723	0.006	0.907	0	0.0184	0.0368	0.0016
Percentile-t	0.824	0.1	0.0280	0.0560	0.0007	0.854	0.1	0.0141	0.0282	0.0004
Normal approx. with bootstrap var. est.	0.861	2.8	0.0311	0.0623	0.0008	0.870	1.8	0.0159	0.0318	0.0004
Normal approx. with Asymptotic var. est.	0.868	3.0	0.0322	0.0643	0.0008	0.875	3.0	0.0165	0.0329	0.0005
e.l.	0.885	0	NA	0.0624	0.0008	0.894	0	NA	0.0320	0.0004
Calibrated e.l.	0.934	0	NA	0.0941	0.0019	0.937	0	NA	0.0494	0.0011
Percentile	0.878	0	NA	0.0740	0.0007	0.890	0	NA	0.0313	0.0004
Iterated percentile	0.922	0	NA	0.0608	0.0063	0.924	0	NA	0.0377	0.0017

to determine the coverage accuracy of the methods suggested in Section 1, and in particular to assess their performance relative to more conventional procedures.

We generated independent and identically distributed bivariate random variables (X_i, Y_i), $1 \leqslant i \leqslant n$, and present our results in the case $n = 15$. The dependence of the two components of the pairs was determined by setting $Y_i = X_i + Z_i$, where X_i and Z_i were independent χ_1^2 and χ_{m-1}^2 random variables respectively, drawn using the routines of Press et al. (1989). The unknown parameter here is $\theta = E(X_i)/E(Y_i) = 1/m$ and $m = 20$ or 40. For a nominal coverage of 0.9544 we calculated the following symmetric confidence intervals: symmetric uncalibrated (i.e. based on the chi-squared approximation), empirical likelihood, bootstrap-calibrated symmetric empirical likelihood, symmetric percentile bootstrap, calibrated (i.e. iterated) symmetric percentile bootstrap, symmetric percentile-t, Normal approximation using the bootstrap variance estimate, and Normal approximation using the asymptotic variance estimate. In each case we computed the standard error estimate as one quarter of the width of the interval. For the sake of comparison we also calculated more conventional, asymmetric confidence intervals based on usual uncalibrated empirical likelihood, usual bootstrap-calibrated empirical likelihood, usual percentile bootstrap and usual calibrated percentile bootstrap methods.

Our results are summarised in Table 1. Each entry in that table was computed by averaging 1000 different, independently drawn samples. In each bootstrap algorithm there were 499 simulations at the first level of resampling. For those methods involving the double bootstrap, i.e. the two forms of calibrated percentile bootstrap intervals, 199 simulations were performed at the second level of resampling.

The main conclusions to emerge from our analysis are as follows.

(1) Up to 3% of confidence intervals based on Normal approximation or on the percentile-t bootstrap run outside the range $(0,\infty)$, whereas of course none of the symmetric empirical likelihood or symmetric percentile method intervals transgresses this range.

(2) Standard errors obtained by Normal approximation or percentile-t bootstrap methods are much smaller than those from calibrated symmetric empirical likelihood or calibrated symmetric percentile methods. This results in consistent undercoverage by confidence intervals based on Normal approximation or the percentile-t bootstrap. Sometimes the undercoverge is severe.

(3) Bootstrap calibration does substantially improve the coverage accuracy of intervals based on symmetric empirical likelihood or the symmetric percentile method.

Acknowledgements

The helpful, constructive comments of a referee are gratefully acknowledged.

References

DiCiccio, T., P. Hall and J. Romano (1991), Empirical likelihood is Bartlett correctable, *Ann. Statist.* **19**, 1053–1061.

Efron, B. (1979), Computers and the theory of statistics: thinking the unthinkable, *SIAM Rev.* **21**, 460–480.

Hall, P. and B. La Scala (1990), Methodology and algorithms of empirical likelihood, *Internat. Statist. Rev.* **58**, 109–127.

Owen, A.B. (1988), Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* **75**, 237–249.

Owen, A.B. (1990), Empirical likelihood confidence regions, *Ann. Statist.* **18**, 90–120.

Press, W.H., B.F. Flannery, S.A. Teukolsky and W.T. Vertterling (1989), *Numerical Recipes: the Art of Scientific Computing* (Cambridge University Press, Cambridge).