Empirical Likelihood Ratio Confidence Intervals for a Single Functional
Author(s): Art B. Owen
Source: *Biometrika*, Vol. 75, No. 2 (Jun., 1988), pp. 237-249
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.org/stable/2336172
Accessed: 03-03-2017 14:51 UTC

Biometrika (1988), **75**, 2, *pp.* 237–49
Printed in Great Britain

Empirical likelihood ratio confidence intervals for a single functional

BY ART B. OWEN
Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

SUMMARY

The empirical distribution function based on a sample is well known to be the maximum likelihood estimate of the distribution from which the sample was taken. In this paper the likelihood function for distributions is used to define a likelihood ratio function for distributions. It is shown that this empirical likelihood ratio function can be used to construct confidence intervals for the sample mean, for a class of M-estimates that includes quantiles, and for differentiable statistical functionals. The results are nonparametric extensions of Wilks's (1938) theorem for parametric likelihood ratios. The intervals are illustrated on some real data and compared in a simulation to some bootstrap confidence intervals and to intervals based on Student's t statistic. A hybrid method that uses the bootstrap to determine critical values of the likelihood ratio is introduced.

Some key words: Bootstrap confidence interval; Empirical likelihood ratio; Nonparametric confidence interval; Nonparametric likelihood.

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be independent observations from a distribution function F_0. The empirical distribution function F_n is often considered a nonparametric maximum likelihood estimate of F_0 because it maximizes

$$L(F) = \prod_{i=1}^{n} \{F(X_i) - F(X_i-)\}$$

over all distribution functions F. With this in mind, we define the empirical likelihood ratio function

$$R(F) = L(F)/L(F_n). \tag{1·1}$$

Suppose that interest centres on $T(F_0)$, where $T(.)$ is a statistical functional. The nonparametric maximum likelihood estimate of $T(F_0)$ is $T(F_n)$. The goal of this paper is to show that, under some reasonable conditions, sets of the form

$$\{T(F) | R(F) \geqslant c\} \tag{1·2}$$

may be used as confidence regions for $T(F_0)$.

Statisticians use parametric likelihood ratio functions to construct confidence intervals and perform tests. In some situations nuisance parameters make it hard to use the likelihood ratio, or its distribution may be difficult to find. Wilks (1938) shows that under mild regularity conditions $-2 \log R_0$ has an asymptotic $\chi^2_{(p)}$ distribution, where R_0 is the maximum of the likelihood ratio function subject to an hypothesis that places p restrictions

on the parameter vector, and the hypothesis is true. In this paper nonparametric versions of Wilks's theorem are proved. Attention is restricted to hypotheses of the form $T(F_0) = t$ which place a single restriction on F_0, and the results are used to provide an asymptotic justification of (1·2).

From the outset it is obvious that regions given by (1·2) will not always work. For example, they fail dramatically when F_0 is absolutely continuous and $T(F)$ is the number of points at which F jumps. For the mean, the set given by (1·2) is the real line for any $c < 1$, unless a restricted set of distributions is considered. To see this observe that, for $F = (1-\varepsilon)F_n + \varepsilon H$ and small enough $\varepsilon > 0$, we have $R(F) \geqslant c$ and the mean of F can be made arbitrarily large by choosing the distribution H appropriately. A natural restriction is to distributions with support in $[-M, M]$ for some finite positive M. It turns out to be possible to restrict to distributions with support in the sample, that is, to distributions $F \ll F_n$. This is convenient because the statistician might not be willing to specify a bound M and because it reduces the problem to one of finite dimension.

The following theorem, proved in § 3, shows that empirical likelihood ratio confidence intervals can be calculated for the mean.

THEOREM 1. *Let X, X_1, $X_2, \ldots$ be independent random variables with nondegenerate distribution function F_0 with $\int |x|^3 \, dF_0 < \infty$. For positive $c < 1$ let*

$$\mathscr{F}_{c,n} = \{F \mid R(F) \geqslant c, F \ll F_n\},$$

and define $X_{U,n} = \sup \int x \, dF$ and $X_{L,n} = \inf \int x \, dF$ with both extrema taken over $F \in \mathscr{F}_{c,n}$. Then as $n \to \infty$

$$\text{pr}\{X_{L,n} \leqslant E(X) \leqslant X_{U,n}\} \to \text{pr}(\chi^2_{(1)} \leqslant -2 \log c).$$

The extrema over $F \ll F_n$ are equivalent to extrema over F with support in $[X_{(1)}, X_{(n)}]$. If F_0 has support in $[-M, M]$, taking extrema over F supported in $[-M, M]$ makes little difference asymptotically, because $R(F) \geqslant c > 0$ implies that the complement of $\{X_1, \ldots, X_n\}$ can have F probability at most $O(n^{-1})$ and from Corollary 1 in § 3 it follows that $X_{U,n} - X_{L,n} = O_p(n^{-\frac{1}{2}})$.

For each n, the empirical likelihood ratio function coincides with the likelihood ratio function of a multinomial on the observed data. As n increases, the representative power of the multinomial improves. For discrete F_0, this fact is all one needs. For continuous F_0 however, the number of parameters in the multinomial is $n-1$ when there are n observations. Given that maximum likelihood estimates are often inconsistent when the number of nuisance parameters increases with n, it is perhaps surprising that the likelihood ratio intervals in Theorem 1 should inherit the coverage properties of finite parameter space likelihood ratio intervals.

Computation of the empirical likelihood ratio confidence interval for the mean is considered in § 2. In § 4 empirical likelihood ratio confidence intervals are found for certain M-estimates including quantiles. A generalization to statistical functionals admitting a Fréchet derivative is made in § 5. A simulation comparing empirical likelihood ratio confidence intervals to bootstrap confidence intervals and intervals based on the central limit theorem is reported in § 6. A hybrid method in which the chi-squared critical value is replaced by one obtained by bootstrapping is introduced in § 6.

Empirical likelihood ratios were first used by Thomas & Grunkemeier (1975). Their application was to survival probabilities estimated by the Kaplan–Meier curve. The Kaplan–Meier curve can be obtained as a nonparametric maximum likelihood estimate of the survival function from censored data. Thomas & Grunkemeier provide a heuristic

argument to show that empirical likelihood ratio intervals for a survival probability based on the $\chi^2_{(1)}$ distribution have asymptotically correct coverage levels. Unlike delta-method intervals, Thomas & Grunkemeier's intervals can be asymmetric and they never include values outside [0, 1]. This is especially appealing for survival probabilities near 0 or 1.

We conclude this section with an example, based on Data Set 9 of Stigler (1977). The data are taken from 20 of Newcomb's measurements of the passage time of light. Most of the observations are between 20 and 30, but there is a notable outlier at −44. Figure 1 shows the function

$$r(x) = \sup\left\{R(F) \middle| \int vF(dv) = x,\ F \ll F_n\right\} \tag{1·3}$$

for these data. The horizontal lines shown are at the asymptotically justified 90% and 95% levels for R. Also shown is the analogous function of x with the mean in (1·3) replaced by the median. This curve takes a jump at each observation, a larger jump where some observations are tied. The computations were also done for Huber's M-estimate, defined in Example 2 of § 4. This curve, not shown, is quite smooth on the scale of Fig. 1 and has nearly the same centre and width as the curve for the median. The constant c in Huber's M-estimate was fixed at 6 which is 1·5 times the sample median of the absolute deviations of the light measurements from their median. A parametric likelihood ratio curve from a normal location family is a normal density with mean 21·75 scaled to have a maximum value of unity. There is such a curve for each value of the variance. A natural choice here has standard deviation 3·8, the sample maximum likelihood estimate of the standard deviation.

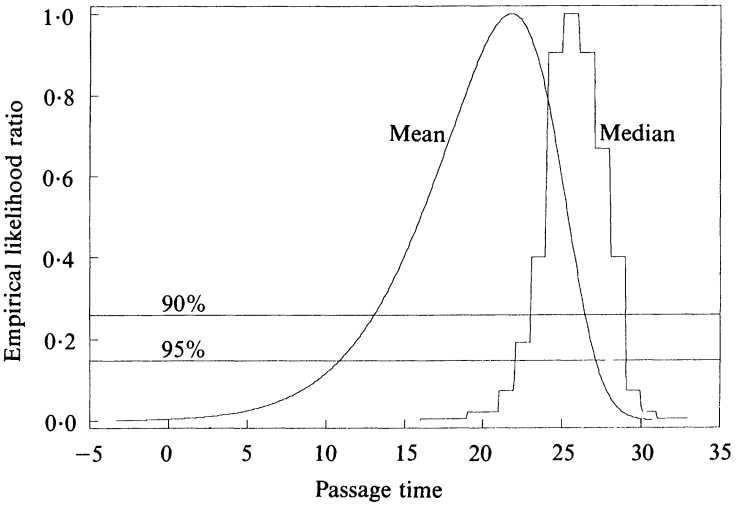

Fig. 1. Maximized empirical likelihood ratio functions for mean and median passage times of light.

2. Computing intervals for the mean

We begin with a device that simplifies the consideration of ties among the X_i. Let F be a distribution for X values and suppose

$$w_i \geqslant 0, \quad \sum_{j:X_j = X_i} w_j = F(X_i) - F(X_i-) \quad (i = 1, \ldots, n). \tag{2·1}$$

The w_i have the form of probabilities attached to observations instead of X values. Now define

$$\tilde{L}(F, w) = \prod_{i=1}^{n} w_i,$$

where w is a vector whose components w_i satisfy (2·1). The maximal value of $\tilde{L}$ is attained when $F = F_n$ and $w_1 = \ldots = w_n = 1/n$. With this in mind define

$$\tilde{R}(F, w) = \prod_{i=1}^{n} nw_i.$$

The functions $\tilde{L}$ and $\tilde{R}$ are observation based likelihood and likelihood ratio functions.

LEMMA 1. *For any* $c \in [0, 1]$

$$\{F: R(F) \geqslant c\} = \{F: \tilde{R}(F, w) \geqslant c \text{ some } w \text{ satisfying } (2\cdot1)\},$$

where F is a distribution function.

Proof. Suppose $R(F) \geqslant c$. Now put $\bar{w}_i = \{F(X_i) - F(X_i-)\}/t_i$, where $1 \leqslant t_i = \text{card}\{X_j \mid X_j = X_i\}$. Then $\bar{w}_i$ satisfy (2·1) and

$$\tilde{R}(F, \bar{w}) = \prod_{i=1}^{n} n\bar{w}_i = \prod_{i=1}^{n} t_i\bar{w}_i \Big/ \prod_{i=1}^{n} (t_i/n) = R(F) \geqslant c.$$

Conversely suppose $\tilde{R}(F, w) \geqslant c$ where the components of w satisfy (2·1). Then

$$R(F) = \tilde{R}(F, \bar{w}) \geqslant c\tilde{R}(F, \bar{w})/\tilde{R}(F, w) = c \prod_{i=1}^{n} \bar{w}_i \Big/ \prod_{i=1}^{n} w_i \geqslant c.$$

The final inequality follows by noting that in any set of tied observations the $\bar{w}_i$ and the w_i have the same sum, and since the $\bar{w}_i$ are equal in the set, their product cannot be less than that of the w_i. □

In applications below both sets of distributions in Lemma 1 will be intersected with some other sets, such as $\{F \ll F_n\}$ or $\{T(F) = t\}$.

It follows that the upper limits of the confidence intervals in Theorem 1 are of the form

$$X_{U,n} = \sup \sum_{i=1}^{n} w_i X_i,$$

where w_i are constrained by

$$w_i \geqslant 0, \quad \sum w_i = 1, \quad \prod nw_i \geqslant c. \tag{2·2}$$

Similarly $X_{L,n}$ is an infimum constrained by (2·2). Both extrema occur at values w_i that satisfy $\Pi\, nw_i = c$.

Next we derive an expression for the weights that give rise to the constrained extrema of the mean. Using Lagrange multipliers, let

$$G = \sum w_i X_i + \lambda_1(1 - \sum w_i) + \lambda_2\{\log c - \sum \log (nw_i)\}.$$

Setting $\partial G/\partial w_i = 0$ yields $w_i = \lambda_2/(X_i - \lambda_1)$.

For any value of λ_1, we may obtain λ_2 as a normalizing constant. Each $\lambda_1 < X_{(1)}$ corresponds to $X_{L,n}$ for some value of c, and each $\lambda_1 > X_{(n)}$ corresponds to $X_{U,n}$ for some c. Other values of λ_1 produce at least one w_i outside the unit interval.

The appropriate λ_1 for any value $c \in (0, 1)$ can be calculated by a zero finding algorithm such as Newton's method. It may be more convenient to pick a grid of λ_1 values and compute the corresponding value of w_i, R and $X_{L,n}$ or $X_{U,n}$. Other values may be found by interpolation. The author has found

$$X_{(1)} - 2^j(X_{(n)} - X_{(1)}), \quad X_{(n)} + 2^j(X_{(n)} - X_{(1)}) \quad (j = -5, -4, \ldots, 10)$$

to be satisfactory values for λ_1. This method is quite convenient for use with statistical languages like *S*, minitab or GLIM.

Efron (1981, eqn (11.8)) obtains the same weights w_i in his discussion of the non-parametric tilting bootstrap. Efron treats the weights as a one parameter family through F_n. Replacing likelihood by Kullback–Liebler distance he gets an exponential family in which the bootstrap distribution of the mean can be fitted. By contrast, we use an asymptotic determination to form intervals.

An alternative strategy is to compute for each x of interest the maximized likelihood ratio $r(x)$ given by (1·3) and find the two points at which $r(x) = c$. This approach underlies the proof of Theorem 1 in § 3, and the algorithm for finding empirical likelihood ratio intervals for M-estimates described in § 4.

3. Proof of Theorem 1

Here we prove Theorem 1, which is stated in § 1.

Proof of Theorem 1. By Lemma 1 we may assume $X_{U,n} = \sup \Sigma\, w_i X_i$ and $X_{L,n} = \inf \Sigma\, w_i X_i$ with both extrema taken over w_i satisfying (2·2).

We also assume without loss of generality that $E(X) = 0$. Because F_0 is not degenerate and has mean 0, it follows that as $n \to \infty$

$$X_{(1)} < 0 < X_{(n)} \tag{3·1}$$

all but finitely often, with probability one. Henceforth assume (3·1). Then $R_0 \equiv r(0)$ exists where $r(.)$ is given by (1·3), and $X_{L,n} \leqslant 0 \leqslant X_{U,n}$ if and only if $c \leqslant R_0$. Therefore we need only show that $-2 \log R_0 \to \chi^2_{(1)}$ in distribution.

To get an expression for R_0 let

$$G = \sum \log n w_i + \gamma(1 - \sum w_i) + n\lambda(0 - \sum w_i X_i).$$

Setting $\partial G/\partial w_i = 0$ one obtains $w_i = 1/(\gamma + n\lambda X_i)$ and summing $w_i \partial G/\partial w_i$ shows that $\gamma = n$. It follows that $\log R_0 = -\Sigma \log(1 + \lambda_0 X_i)$, where λ_0 is a root of

$$0 = n^{-1} \sum X_i/(1 + \lambda X_i) \equiv g(\lambda). \tag{3·2}$$

The desired root is in $J_n = (-X_{(n)}^{-1}, -X_{(1)}^{-1})$ for otherwise some w_i are outside the unit interval. From (3·1) it follows that $g'(\lambda) < 0$, strictly, in J_n. Because g has limits ∞ and $-\infty$ at the ends of J_n, there is a unique zero λ_0 of g in J_n. The Hessian of $\Sigma \log n w_i$ is negative-definite so the stationary point of G is a constrained relative maximum of $\Sigma \log(n w_i)$. The only other candidates for the supremum are the boundary points for which some $w_i = 0$. Therefore λ_0 provides the unique constrained supremum R_0.

The existence of $\int |X|^3\, dF_0$ is equivalent to $\Sigma \operatorname{pr}(|X_n|^3 > n) < \infty$ and by the Borel–Cantelli lemma it follows that $|X_n| < n^{1/3}$ all but finitely often, with probability 1. This in turn implies that

$$\max_{1 \leqslant i \leqslant n} |X_i| < n^{1/3} \tag{3·3}$$

all but finitely often with probability 1, since $n^{1/3}$ eventually exceeds the largest of the finite collection of $|X_k|$'s that exceed $k^{1/3}$. Therefore we assume that (3·3) holds.

Now pick $q<\frac{1}{2}$. We show that $\lambda_0 = O_p(n^{-q})$. Consider

$$g(n^{-q}) = n^{-1}\sum X_i/(1+n^{-q}X_i) = \bar{X} - n^{-1-q}\sum X_i^2/(1+n^{-q}X_i)$$
$$\leqslant \bar{X} - n^{-q}S^2/(1+n^{-q}n^{1/3}),$$

where $S^2 = n^{-1}\Sigma X_i^2$. Equivalently

$$n^{\frac{1}{2}}g(n^{-q}) \leqslant n^{\frac{1}{2}}\bar{X} - n^{\frac{1}{2}}S^2/(n^q + n^{1/3}). \tag{3·4}$$

The first term on the right-hand side of (3·4) has a limiting normal distribution. The quantity subtracted goes to ∞ almost surely, so it follows that $\text{pr}\{g(n^{-q})>0\}\to 0$ and a similar result for $g(-n^{-q})$ yields $\lambda_0 = O_p(n^{-q})$.

Let $h = g^{-1}$ and apply Taylor's theorem:

$$\lambda_0 = h(0) = h(\bar{X}) + (0-\bar{X})h'(\xi) = -\bar{X}h'(\xi),$$

where $|\xi|\leqslant|\bar{X}|$. Now $h'(\xi) = 1/g'(\eta)$, where $\eta = h(\xi)$ and $|\eta|\leqslant|\lambda_0|$. Therefore $\lambda_0 = r_0\bar{X}/S^2$, where

$$r_0 = \frac{S^2}{-g'(\eta)} = \frac{\sum X_i^2}{\sum\{X_i/(1+\eta X_i)\}^2}.$$

Since $|\eta|\leqslant|\lambda_0| = O_p(n^{-q})$, we have $\eta X_i = O_p(n^{-q})O(n^{1/3}) = O_p(1)$ and hence $r_0\to 1$ in probability. Therefore $\lambda_0 = O_p(n^{-\frac{1}{2}})$.

Finally, we make the following Taylor expansion for all sufficiently large n:

$$-2\log R_0 = 2\sum\log(1+\lambda_0 X_i) = 2\sum\lambda_0 X_i - (\lambda_0 X_i)^2/2 + \eta_i \quad (|\eta_i|<|\lambda_0 X_i|^3)$$
$$= 2n\bar{X}^2 r_0/S^2 - nS^2(\bar{X}r_0/S^2)^2 + \sum\eta_i = (2r_0 - r_0^2)n\bar{X}^2/S^2 + \sum\eta_i,$$

where

$$|\sum\eta_i| \leqslant |\lambda_0|^3\sum|X_i|^3 = O_p(n^{-\frac{1}{2}}), \quad 2r_0 - r_0^2 = 1 + o_p(1),$$

and $n\bar{X}^2/S^2\to\chi^2_{(1)}$ in distribution, by the central limit theorem. □

COROLLARY 1. *Under the conditions of Theorem* 1, *let* F_0 *have mean* μ *and variance* σ^2 *and let* $r(x)$ *be given by* (1·3). *Let* $s^2 = n^{-1}\Sigma(X_i-\bar{X})^2$ *and let* τ *be any real constant. Then*

$$-2\log r(\bar{X} + \tau s n^{-\frac{1}{2}})\to\tau^2$$

in probability, and

$$-2\log r(\mu + \tau\sigma n^{-\frac{1}{2}})\to\chi^2_{(1)}(\tau^2),$$

the noncentral chi-squared distribution, with noncentrality parameter τ^2 *as* $n\to\infty$.

The proof of Theorem 1 applies here with minor modifications.

From Corollary 1, we see that the confidence intervals are asymptotically the same as those obtained by using an approximation based on the central limit theorem for the mean of a sample. It remains to see whether there is a useful difference in small samples. Based on the simulation reported in § 6, there does appear to be such a difference.

Corollary 1 sheds some light on the asymptotic relative efficiency of inferences based on the empirical likelihood ratio. The following discussion is not rigorous and assumes that suitable regularity conditions hold. Suppose that F_0 belongs to a parametric family indexed by $\mu = E(X)$. Then, by comparing the curvature of the empirical log likelihood ratio function to that of the parametric log likelihood ratio function we obtain an asymptotic relative efficiency of $\{i_\mu \operatorname{var}(X)\}^{-1}$, where i_μ is the Fisher information for μ in X. The same efficiency is obtained by inferences based on the central limit theorem.

For many parametric families $\operatorname{var}(X) = i_\mu^{-1}$ which leads to an asymptotic relative efficiency of 1 for the empirical method. Some special cases are the binomial, Poisson, exponential and normal location families. The result holds for sufficiently regular one-parameter exponential families in which the mean of X is the natural parameter. The asymptotic relative efficiency is 0·5 in the double exponential location family and is approximately 0·91 in the logistic location family. In neither of these last two families is the sample mean the maximum likelihood estimate of the population mean.

COROLLARY 2. *Let*

$$E_n = \operatorname{pr}\{X_{L,n} \leqslant E(X) \leqslant X_{U,n}\} - \operatorname{pr}(\chi^2_{(1)} \leqslant -2 \log c)$$

be the error of approximation in Theorem 1. *If* $|X| \leqslant M < \infty$ *with probability* 1 *then* $E_n = O_p(n^{-\frac{1}{2}})$, *as* $n \to \infty$ *and if* $E(|X|^p) < \infty$ *for* $3 \leqslant p < \infty$ *then* $E_n = O_p(n^{-\frac{1}{2}+1/p})$ *as* $n \to \infty$.

Proof. By the Berry–Esseen theorem $n\bar{X}^2/S^2 \to \chi^2_{(1)}$ with errors of order $O_p(n^{-\frac{1}{2}})$. In the proof of Theorem 1, $\Sigma\, \eta_i = O_p(n^{-\frac{1}{2}})$ so it only remains to consider the order of $r_0 - 1$. If $|X| \leqslant M < \infty$ then $|\eta X_i| \leqslant M|\eta_i| = O_p(n^{-\frac{1}{2}})$ whence $r_0 - 1 = O_p(n^{-\frac{1}{2}})$. If a pth absolute moment of X exists we may assume that $\max |X_i| < n^{1/p}$ all but finitely often and hence that $|\eta X_i| = O_p(n^{-\frac{1}{2}})O(n^{1/p}) = O_p(n^{-\frac{1}{2}+1/p})$ which implies that $r_0 - 1 = O_p(n^{-\frac{1}{2}+1/p})$. □

Note that the rate achieved for bounded random variables is the same as Wilks (1938) obtains for parametric likelihood ratios.

4. *M*-ESTIMATES

Theorem 1 extends readily to certain M-estimates. An M-estimate is a statistical functional defined as a root $\tau = T(F)$ of

$$\int \psi(X, \tau) F(dX) = 0. \tag{4·1}$$

Conditions must be imposed on $\psi(x, t)$ to guarantee existence of a solution to (4·1). Further conditions may be adopted to provide a unique solution to (4·1), or a tie-breaking rule such as 'infimum of the roots' or 'closest root to the median of F' may be adopted to select a root. We will impose conditions on ψ expressed through families of univariate functions $\psi_{\cdot t}$ and $\psi_{x\cdot}$ given by

$$\psi_{\cdot t}(x) = \psi(x, t) = \psi_{x\cdot}(t). \tag{4·2}$$

THEOREM 2. *Let* $T(F)$ *be a solution of* (4·1) *and let* $X_1, X_2 \ldots$ *be independent random variables with common distribution* F_0. *Assume that* $\psi(x, t)$ *satisfies*:

(i) $T(F_0) = \tau$ *exists and is unique,*
(ii) $\psi_{\cdot\tau}(x)$ *measurable,*
(iii) $\operatorname{var}\{\psi(X_1, \tau)\} > 0$,
(iv) $E\{|\psi_{\cdot\tau}(X)|^3\} < \infty$.

For positive $c<1$ and R given by (1·1) *let* $\mathcal{F}_{c,n}=\{F \mid R(F)\geqslant c,\ F\ll F_n\}$ *and*

$$S_{c,n}=\bigcup_{F\in\mathcal{F}_{c,n}}\left\{t\left|\int\psi(x,t)F(dx)=0\right.\right\}.$$

Then $\operatorname{pr}\{T(F_0)\in S_{c,n}\}\to\operatorname{pr}(\chi^2_{(1)}\leqslant -2\log c)$ *as* $n\to\infty$.

If ψ also satisfies

(v) $\psi_{x_\cdot}(t)\leqslant\psi_{x_\cdot}(s)$ *whenever* $t\geqslant s$

for all x in the support of F_0 then $S_{c,n}$ is an interval.

Proof. Let $Z_i=\psi(X_i,\tau)$, where $\tau=T(F_0)$ is the unique root of (4·1) for $F=F_0$ assumed in (i). Conditions (i)–(iv) yield that the Z_i satisfy the conditions of Theorem 1 with $E(Z_i)=0$. Condition (ii) ensures that the Z_i are random variables.

From Lemma 1

$$S_{c,n}=\{t\mid\textstyle\sum w_i\psi(X_i,t)=0,\ w_i \text{ satisfy (2·2)}\}. \tag{4·3}$$

If $\tau\in S_{c,n}$ then

$$\sup\{\textstyle\prod nw_i\mid\sum w_iZ_i=0,\ w_i\geqslant 0,\ \sum w_i=1\}\geqslant c \tag{4·4}$$

and conversely, since $\prod nw_i$ is continuous and the set over which the supremum is taken is compact. But (4·4) is equivalent to

$$\inf\textstyle\sum w_iZ_i\leqslant 0\leqslant\sup\sum w_iZ_i$$

with both extrema taken over w_i satisfying (2·2). Therefore, because Z_i satisfy the conditions of Theorem 1

$$\lim_{n\to\infty}\operatorname{pr}\{T(F_0)\in S_{c,n}\}=\operatorname{pr}(\chi^2_{(1)}\leqslant -2\log c).$$

Now suppose (v) holds and that $t_1,\ t_2\in S_{c,n}$, where $t_1<s<t_2$. Then there exist w_{i1} and w_{i2} each satisfying (2·2) such that $\Sigma_i\, w_{ij}\psi(X_i,t_j)=0\ \ (j=1,2)$. We show that $s\in S_{c,n}$. Let $a_j=\Sigma_i\, w_{ij}\psi(X_i,s)$. Then $a_1\leqslant 0\leqslant a_2$ by (iii). If $a_1=0$ or $a_2=0$ then $s\in S_{c,n}$. Otherwise put $\lambda=a_2/(a_2-a_1)$ and $v_i=\lambda w_{i1}+(1-\lambda)w_{i2}$. It follows from Jensen's inequality that $\Sigma\, v_i 1_{x\leqslant X_i}\in\mathcal{F}_{c,n}$. But $\Sigma\, v_i\psi(X_i,s)=0$ so $s\in S_{c,n}$. Therefore $S_{c,n}$ is an interval. □

Example 1: Quantiles. Suppose F_0 has a unique γ quantile Q_γ where $\gamma\in(0,1)$. Then

$$\psi(x,t)=\begin{cases}1 & (x\leqslant t),\\ -\gamma/(1-\gamma) & (x>t)\end{cases}$$

determines Q_γ as the unique zero of (4·1) with $F=F_0$, and Theorem 2 shows that, for $c\in(0,1)$, $\{T(F)\mid R(F)\geqslant c,\ F\ll F_n\}$ is a confidence interval for Q_γ with asymptotic coverage $\alpha=\operatorname{pr}(\chi^2_{(1)}\leqslant -2\log c)$.

Example 2: Huber's location M-estimate. Let

$$\psi(x,t)=\begin{cases}c & (x-t\geqslant c),\\ x-t & (|x-t|<c),\\ -c & (x-t\leqslant -c),\end{cases}$$

and suppose that there is a unique solution to (4·1) for $F=F_0$. Then the likelihood ratio confidence regions are intervals and their asymptotic coverage is determined by the $\chi^2_{(1)}$ distribution.

We compute the likelihood ratio confidence region for $T(F)$ by first computing

$$r(t) = \sup \{\prod nw_i | \sum w_i \psi(X_i, t) = 0,\ w_i \geqslant 0,\ \sum w_i = 1\}$$

and finding the region $\{t | r(t) \geqslant c\}$. When this region is known to be an interval, a bisection algorithm can be used to find the endpoints once there are intervals in which the endpoints are known to lie. For interpretative purposes it would seem preferable to compute and plot a portion of $r(t)$. Put $Z_i = \psi(X_i, t)$. If $Z_{(1)} < 0 < Z_{(n)}$ then the set over which $r(t)$ is the supremum is nonempty. The maximizing value is $\Pi\, nw_i(\lambda)$, where

$$w_i(\lambda) = \{n(1 + \lambda Z_i)\}^{-1}$$

and λ is the unique root of

$$0 = n^{-1} \sum Z_i/(1 + \lambda Z_i)$$

in the interval $(-Z_{(n)}^{-1}, -Z_{(1)}^{-1})$. This all follows from the steps in the proof of Theorem 1. The value λ may be found numerically. A safeguarded zero-finding algorithm such as Brent's method (Press et al., 1986, pp. 251–4) works well. The author has found that useful bracketting values of λ can be found by solving $\{n(1 + \lambda z)\}^{-1} = 2$ with z taking the values $Z_{(1)}$ and $Z_{(n)}$.

5. Likelihood intervals for differentiable functionals

Theorem 1 justifies empirical likelihood ratio intervals for certain linear statistical functionals. Theorem 2 extends the result to M-estimates, which are defined through (4·1) in terms of linear statistical functionals involving ψ. In this section we show that functionals admitting a Fréchet derivative can be treated by applying Theorem 1 to the linear functionals given by their derivatives.

Definition. Let I be a closed interval subset of $\mathbb{R}$. Let $D(I)$ be the set of real functions on I that are continuous from the right and have limits from the left. Equip $D(I)$ with the sup norm:

$$\|f\| = \sup_{x \in I} |f(x)|.$$

A statistical functional $T: D(I) \to \mathbb{R}$ is said to be differentiable at $F_0 \in D(I)$ if there exists a bounded linear transformation $T_0': D(I) \to \mathbb{R}$ such that

$$\frac{|T(F) - T(F_0) - T_0'(F - F_0)|}{\|F - F_0\|} \to 0$$

as $\|F - F_0\| \to 0$.

The linear transformation T_0' is called the derivative of T at F_0.

The definition above is that of Frechet differentiability. Existence of a Frechet derivative is a fairly strong condition, and a popular alternative is the Hadamard or compact derivative. It is also possible to use spaces other than $D(I)$ for the domain of T. See Fernholtz (1983) for a discussion of differentiability of statistical functionals.

When the derivative T_0' exists we may write

$$T_0'(F - F_0) = \int \text{IC}(x, F_0, T) F(dx),$$

where $\text{IC}(x, F_0, T)$ is the influence curve of Hampel (1974).

THEOREM 3. *Let $X_1, \ldots, X_n$ be independent random variables with distribution F_0. Let I be a closed interval that contains the support of F_0, and let $T: D(I) \to \mathbb{R}$ be a statistical functional with derivative T_0' at F_0. Then for positive $c < 1$*

$$\sup |T(F) - T(F_0) - T_0'(F - F_0)| = o_p(n^{-\frac{1}{2}})$$

as $n \to \infty$, where the supremum is taken over $F \ll F_n$ satisfying $R(F) \geqslant c$.

Before proving Theorem 3 we give a technical lemma.

LEMMA 2. *For $n = 1, 2, \ldots$ let F_n be the empirical distribution based on real observations $X_1, \ldots, X_n$, and let G_n be a sequence of distributions with $G_n \ll F_n$ and $R(G_n) \geqslant c$ for positive $c < 1$. Then $D_n = \|F_n - G_n\| = O(n^{-\frac{1}{2}})$ as $n \to \infty$.*

Proof. The proof is a long and straightforward analysis which we sketch here.

One begins by reducing the problem to the case in which for each n and $X_1, \ldots, X_n$ only two different values of $G_n(X_i) - G_n(X_i-)$ occur. In this setting we have $D_n = nz(1-z)\varepsilon$ for $\varepsilon \geqslant 0$ and $0 < z < 1$. The reduction is such that $nz\varepsilon \leqslant 1$. Furthermore $R(G_n) = \{1 + n(1-z)\varepsilon\}^{nz}(1 + nz\varepsilon)^{n(1-z)}$.

It then follows from $R(G_n) \geqslant c$ that

$$D_n \leqslant (-2 \log c)^{\frac{1}{2}} n^{-\frac{1}{2}} \tag{5·1}$$

by taking logs, using $\log(x) < 2(x-1)/(x+1)$ for $0 < x < 1$ and $\log(x) \leqslant 1 + x$. □

Equation (5·1) provides an explicit upper bound for $n^{\frac{1}{2}}\|F_n - G_n\|$. The exponent of n cannot be reduced. Lemma 2 may be interpreted as providing a bound on the Kolmogorov-Smirnov distance between F_n and $G_n \ll F_n$ in terms of the Kullback-Liebler distance.

Proof of Theorem 3. Let $F \ll F_n$ satisfy $R(F) \geqslant c$. Then

$$\|F - F_0\| \leqslant \|F - F_n\| + \|F_n - F_0\| = O(n^{-\frac{1}{2}}) + O_p(n^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}})$$

using Lemma 2 and the well-known behaviour of $F_n - F_0$. By differentiability of T

$$|T(F) - T(F_0) - T_0'(F - F_0)| = o(\|F - F_0\|) = o_p(n^{-\frac{1}{2}}).$$ □

Theorem 3 implies that $\{T(F) \mid R(F) \geqslant c, F \ll F_n\}$ is within $o_p(n^{-\frac{1}{2}})$ of

$$\{T(F_0) + T_0'(F - F_0) \mid R(F) \geqslant c, F \ll F_n\}$$

which is an interval for $T(F_0)$ with asymptotic coverage given by Theorem 1. The intervals from Theorem 1 have width $O_p(n^{-\frac{1}{2}})$ so the difference in confidence sets is asymptotically negligible.

Example 1: Variance of a bounded random variable. Let $I = [-M, M]$ for some $M < \infty$, and let

$$\text{var}(F) = \int_I \{x - E(F)\}^2 F(dx)$$

be the variance functional for distributions supported in I, where $E(F) = \int_I xF(dx)$. Then var has influence $IC(x, \text{var}, F_0) = (x - E_0)^2 - \text{var}(F_0)$, where $E_0 = E(F_0)$, and var is easily

shown to be differentiable. By Theorem 3

$$\{\text{var}(F) \mid R(F) \geqslant c, F \ll F_n\} \tag{5·2}$$

is asymptotically close to

$$\left\{\int (x-E_0)^2 F(dx) \mid R(F) \geqslant c, F \ll F_n\right\} \tag{5·3}$$

for which Theorem 1 provides an asymptotic coverage level. While (5·3) shows that (5·2) is close to an interval with the right coverage it does not allow us to compute an interval because E_0 is unknown. A little algebra and an application of Lemma 2 shows that

$$\left\{\int (x-\bar{X})^2 F(dx) \mid R(F) \geqslant c, F \ll F_n\right\} \tag{5.4}$$

is also close to an interval whose coverage of var (F_0) may be taken from Theorem 1. The interval in (5·4) may be computed using the algorithm for the intervals for the mean, applied to $(X_i - \bar{X})^2$ where $\bar{X}$ is the sample mean. Such intervals are conservative because var $(F) \leqslant \int (x-\bar{X})^2 F(dx)$ for any F. □

6. Simulation results

Empirical likelihood ratio confidence intervals make very weak distributional assumptions and are justified by having asymptotically correct coverage levels. In this regard they are like bootstrap confidence intervals and also like intervals based on Student's t statistic. This section describes a simulation experiment carried out to compare empirical likelihood ratios with various bootstrap methods and the t intervals. The methods will be compared on the problem of estimating the mean of a sample of size 20 taken from the chi-squared distribution on one degree of freedom. This should be a hard problem for nonparametric methods, because it is similar to the problem of estimating the variance from a normal sample, which Schenker (1985) shows is hard.

The simulation was based on 1000 samples of 20 $\chi^2_{(1)}$ random variables. These were obtained by squaring standard normal random variables produced by IMSL routine GGNML. From each sample a 90% empirical likelihood ratio confidence interval was computed, as was a Student's interval and the following bootstrap confidence intervals: percentile, bias-corrected percentile, bias-corrected accelerated percentile, and bootstrap t. The bootstrap intervals are discussed by Efron (1982), with the exception of bias-corrected accelerated intervals (Efron, 1987). A hybrid of bootstrap and empirical likelihood ratio confidence intervals was also tried. In the hybrid method the distribution of $-2 \log R_0$ assuming $F_0 = F_n$ is used in place of the limiting chi-squared distribution. The distribution of $-2 \log R_0$ for $F_0 = F_n$ is obtained by Monte Carlo; that is, a bootstrap distribution is used. This method will be called bootstrap calibrated empirical likelihood. For each sample 1000 bootstrap samples were drawn using IMSL routine GGUD. For comparison a parametric interval based on the 5th and 95th percentiles of the scaled $\chi^2_{(20)}$ distribution of the sample sum of squares was included.

The first column of Table 1 gives the observed coverage fraction for each method. Standard errors can be computed using the binomial formula, but it is more accurate to proceed as follows. Let X be an indicator that is 1 when the parametric interval contains the true mean and zero otherwise. Similarly let Y be the indicator of coverage for one of the other confidence interval methods. Then the coverage level for that method is

Table 1. *Simulation results*

Method	Observed central coverage	Estimated central coverage	Standard error	$N_{L>1}$	$N_{U<1}$	One-sided error
Empirical likelihood	0·872	0·879	0·011	7	121	114
Bootstrap calibrated empirical likelihood	0·906	0·913	0·010	5	89	84
Percentile	0·827	0·834	0·011	23	150	127
Bias corrected	0·829	0·836	0·010	36	135	99
Bias corrected accelerated	0·845	0·852	0·008	50	105	55
Bootstrap t	0·890	0·897	0·008	38	72	34
Ordinary t	0·839	0·846	0·011	13	148	135
Parametric	0·893	0·900	nil	51	56	7

$N_{L>1}$, the number of intervals in which the lower limit exceeds 1·0, the true mean of the $\chi^2_{(1)}$ distribution. Ideally should be 50 as should $N_{U<1}$, number of intervals in which upper limit less than 1. Measure of one-sided error given in right-hand column, $|N_{L>1}-50|+|N_{U<1}-50|$.

$0{\cdot}9+E(Y-X)$. The estimated values in Table 1 are based on the sample mean of $Y-X$ and the standard errors are based on the sample variance of $Y-X$. The estimated coverages have standard errors on the order of 1%.

Some readers may find it ironic that standard errors are used to assess the accuracy of these estimators instead of say, a bootstrap t interval for $Y-X$. But $Y-X$ is not very skew and 1000 independent replications of it are available, so the central limit approximation should be very good, and there should be correspondingly little to gain by more sophisticated interval estimates.

Of the nonparametric methods, the bootstrap t has the best estimated central coverage. The bootstrap calibrated empirical likelihood ratio interval method is the second closest and these two methods are the only ones within two standard errors of the desired coverage level. The uncalibrated empirical likelihood ratio intervals are approximately two standard errors off the desired level and all other nonparametric intervals are more than four standard errors off the desired level.

The noncoverage events observed in Table 1 are broken down according to the side of the true mean on which the interval fell. For all of the nonparametric methods noncoverage events are much more common for the upper confidence limit than for the lower. The bootstrap t method appears to have the best one-sided confidence intervals of any of the nonparametric methods considered. The two empirical likelihood methods have the worst lower endpoints. Intervals based on the ordinary t statistic and the percentile method have the worst upper endpoints.

Compared to the usual intervals based on the t statistic, the empirical likelihood intervals are slightly worse on the lower endpoint, better on the upper endpoint and much better on central coverage. The bootstrap t is the only bootstrap method to outperform the empirical likelihood methods on all three counts.

ACKNOWLEDGEMENTS

I would like to thank Peter Bickel and Bradley Efron for encouragement. I also thank Joseph Marhoul for comments that spurred me to simplify the proof of Theorem 1. I am grateful for the financial support of the National Science Foundation.

References

EFRON, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Can. J. Statist.* **9**, 139–72.

EFRON, B. (1982). *The Jackknife, the Bootstrap, and other Resampling Plans*, Conf. Series in Appl. Math., No. 38. Philadelphia: SIAM.

EFRON, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Am. Statist. Assoc.* **82**, 171–200.

FERNHOLZ, L. T. (1983). *von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics, No. 19. New York: Springer-Verlag.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69**, 383–93.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. & VETTERLING, W. T. (1986). *Numerical Recipes.* Cambridge University Press.

SCHENKER, N. (1985). Qualms about bootstrap confidence intervals. *J. Am. Statist. Assoc.* **80**, 360–1.

STIGLER, S. M. (1977). Do robust estimators work with real data? (with discussion). *Ann. Statist.* **5**, 1055–98.

THOMAS, D. R. & GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Am. Statist. Assoc.* **70**, 865–71.

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60–2.

[*Received April* 1987. *Revised October* 1987]