The Annals of Statistics
1990, Vol. 18, No. 1, 90–120

EMPIRICAL LIKELIHOOD RATIO CONFIDENCE REGIONS[1]

BY ART OWEN

Stanford University

An empirical likelihood ratio function is defined and used to obtain confidence regions for vector valued statistical functionals. The result is a nonparametric version of Wilks' theorem and a multivariate generalization of work by Owen. Cornish–Fisher expansions show that the empirical likelihood intervals for a one dimensional mean are less adversely affected by skewness than are those based on Student's t statistic.

An effective method is presented for computing empirical profile likelihoods for the mean of a vector random variable. The method is a reduction by convex duality to an unconstrained minimization of a convex function on a low dimensional domain. Algorithms exist for finding the unique global minimum at a superlinear rate of convergence. A byproduct is a noncombinatorial algorithm for determining whether a given point lies within the convex hull of a finite set of points.

The multivariate empirical likelihood regions are justified for functions of several means, such as variances, correlations and regression parameters and for statistics with linear estimating equations. An algorithm is given for computing profile empirical likelihoods for these statistics.

1. Introduction. Let $X_1, X_2, \ldots$ be independent random vectors in $\mathbb{R}^p$, for $p \geq 1$, with common distribution function F_0. The empirical distribution

$$F_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

is well known to be the nonparametric maximum likelihood estimate of F_0 based on $X_1, \ldots, X_n$. Here δ_x denotes a point mass at x. The likelihood function that F_n maximizes is

$$L(F) = \prod_{i=1}^{n} F\{x_i\},$$

where $F\{x_i\}$ is the probability of the set $\{x_i\}$ under F, x_i is the observed value of X_i and F is any probability measure on $\mathbb{R}^p$. This motivates the term nonparametric m.l.e. for the estimate $T(F_n)$ of the parameter $T(F_0)$, where T is a statistical functional.

In some cases the empirical likelihood ratio function

$$R(F) = L(F)/L(F_n) \tag{1.1}$$

can be used to construct nonparametric confidence regions and tests for $T(F_0)$.

Received January 1988; revised January 1989.
[1]Research supported by NSF Grant DMS-86-00235.
AMS 1980 *subject classification*. Primary 62E20.
Key words and phrases. Bootstrap, confidence set, empirical likelihood, likelihood ratio test, nonparametric likelihood.

Consider sets of the form

$$C = \{T(F) | R(F) \geq r\}.$$

This article gives conditions under which sets like C may be used as confidence regions for $T(F_0)$. Under such conditions a test of $T(F_0) = t$ rejects when $t \notin C$, that is, when no distribution F with $T(F) = t$ has likelihood $L(F) \geq rL(F_n)$.

The central result is for the mean of X. Clearly some restrictions on F are needed, or else $C = \mathbb{R}^p$ whenever $r < 1$. To see this, let $F = \varepsilon\delta_x + (1 - \varepsilon)F_n$. For small enough $\varepsilon > 0$ we have $R(F) \geq r$. But then, as x ranges through $\mathbb{R}^p$, so does the mean of F, tracing out $C = \mathbb{R}^p$. The problem may be resolved by restricting to distributions F that are supported in a bounded set. It turns out to be possible to restrict attention 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 bounded support for F and because it reduces the construction of C to a finite dimensional problem. The theorem for the mean is:

THEOREM 1. *Let $X, X_1, X_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^p$, with $E(X) = \mu_0$ and $\operatorname{var}(X) = \Sigma$ of rank $q > 0$. For positive $r < 1$ let $C_{r,n} = \{\int X\,dF | F \ll F_n, R(F) \geq r\}$. Then $C_{r,n}$ is a convex set and*

$$\lim_{n\to\infty} P(\mu_0 \in C_{r,n}) = P\left(\chi^2_{(q)} \leq -2\log r\right).$$

Moreover if $E(\|X\|^4) < \infty$, then

$$\left|P(\mu \in C_{r,n}) - P\left(\chi^2_{(q)} \leq -2\log r\right)\right| = O(n^{-1/2}).$$

The proof of Theorem 1 is given in Section 2.

The $\chi^2_{(q)}$ random variable that appears in Theorem 1 is noteworthy. It is the same limit Wilks (1938) obtains in a parametric setting for a likelihood ratio test of an hypothesis imposing q constraints on the parameter. The rate attained is also that found by Wilks (1938) in the parametric case. Sharper rate estimates are discussed below. It is suggested in Section 2 that for small n, the $\chi^2_{(q)}$ distribution should be replaced by $(n-1)q/(n-q)$ times an $F_{q,n-q}$ distribution.

For $F \ll F_n$ the likelihood ratio (1.1) is that of a multinomial on the sample. If the support of F_0 is a finite set, then Theorem 1 is trivial. When F_0 has a continuous distribution and there are n observations, then with probability 1 there are $n - 1$ parameters in the multinomial distribution and p parameters of interest. It is somewhat surprising that with essentially equal numbers of parameters and observations the limit law is the same as in Wilks' theorem. Moreover, the multinomial family of distributions used is on a randomly selected set of points. As in Wilks' theorem, nuisance variables are "profiled out." Let F_μ maximize $R(F)$ subject to $F \ll F_n$ and $\int X\,dF = \mu$. Then $\mu \in C_{r,n}$ if and only if $R(F_\mu) \geq r$. The notation F_μ will be used below and the profile likelihood ratio $R(F_\mu)$ will sometimes be denoted $\mathscr{R}(\mu)$.

The restriction to $F \ll F_n$ is not as severe as it might seem at first. $C_{r,n}$ would not change if we restricted F to have support in the convex hull of $\{X_1, \ldots, X_n\}$.

It is easy to show that $R(F) \geq r$ implies that the distribution F place at least $1 - O(n^{-1})$ probability on the sample. Therefore when it is known that $X \in B$ with probability 1 for some bounded set B, changing the condition $F \ll F_n$ to $F \ll 1_B$ increases the diameter of $C_{r,n}$ by $O_p(n^{-1})$, which is small compared to the diameter $O_p(n^{-1/2})$.

The plan of the rest of this article is as follows. An example and a survey of related literature conclude Section 1. Section 2 contains a proof of Theorem 1 and a corollary that considers the behavior of empirical likelihood ratio tests at alternatives close to the true mean. A connection to Hotelling's T^2 emerges from the proof of Theorem 1 and is the basis for the claim that the F distribution should be a better reference than the χ^2. Some simulations are also cited to make this point. Section 3 discusses the computation of the profile empirical likelihood for a vector mean. A low dimensional dual problem is introduced. The dual problem is one of unconstrained convex programming and there are theorems guaranteeing the existence and uniqueness of a solution and the convergence to that solution of certain iterative methods. Section 4 considers extensions to functionals other than the mean. Simple linearization arguments extend Theorem 1 to a rich class of statistics. For $p = 1$ a Cornish–Fisher expansion in Section 5 shows that the empirical likelihood method may be thought of as a t test with a partial correction for skewness. In Section 6 some examples of empirical likelihood inference are given. Likelihood functions are considered for the standard deviation, the correlation coefficient and regression coefficients. A nested algorithm for computing the profile empirical likelihood is given and analyzed.

1.1. *Example.* For an illustration we use some data from Larsen and Marx (1986, page 440). Eleven male ducks, each a second generation cross between mallard and pintail, were examined. Their plumage was rated on a scale from 0 (completely mallardlike) to 20 (completely pintaillike) and their behavior was similarly rated on a scale from 0 (mallard) to 15 (pintail). Figure 1 shows these data, together with nested empirical likelihood confidence contours for the mean. The point with plumage = 14 and behavior = 11 is plotted with a circle of twice the area of the others, because it represents two ducks. The confidence contours are presented for nominal confidence levels: 0.50, 0.90, 0.95, 0.99, taken from $20/9$ times the $F_{2,9}$ distribution. An asterisk marks the sample mean.

In Figure 2, the same information is presented, with the contours now taken from a scaled $F_{2,9}$ distribution for Hotelling's T^2 statistic. These are parametric likelihood ratio contours assuming a bivariate normal distribution with unknown mean and variance.

A comparison to the bootstrap is natural here, though there does not yet seem to be a completely satisfactory way to construct bootstrap confidence regions in two or more dimensions. The difficulty lies in selecting the "central" $1 - \alpha$ fraction of B resampled points. Figure 3 is based on assigning each resampled mean the largest rank it attains in any linear projection. The $(1 - \alpha)B$ points with the smallest maximum rank determine the region. This method of ranking vectors is due to Donoho (1982) and Stahel (1981). If two points attained the

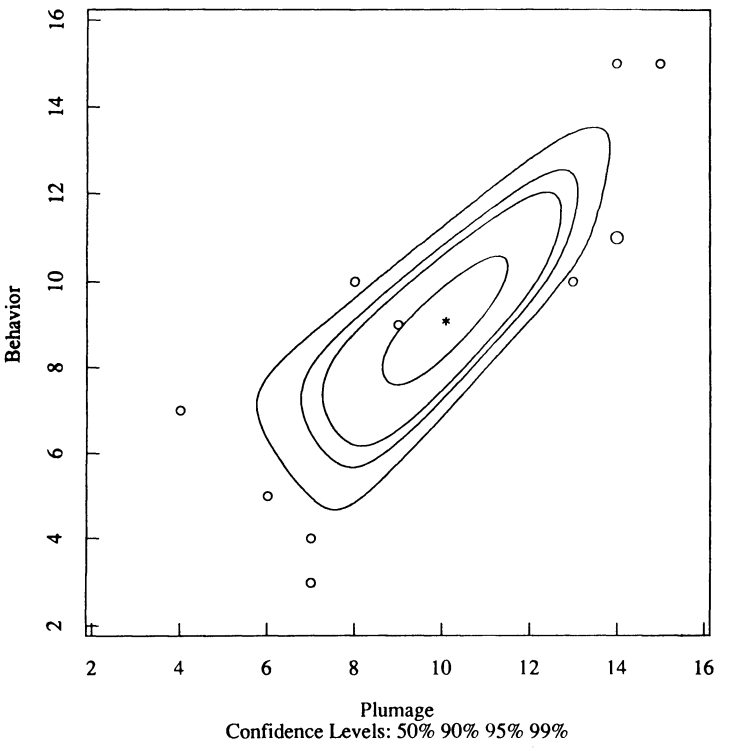

FIG. 1. *Empirical likelihood contours.*

same maximum rank, the tie was broken by considering the point attaining that rank over a greater angular range of projections to be farther from the center. The same confidence levels as were used in Figures 1 and 2 are shown. The regions are based on 1000 resampled means and 360 projections. They are smaller than the other regions and are quite angular. Hall (1987) estimates a density for resampled (studentized) means and selects a contour of the density containing $1 - \alpha$ of the resampled means. Unfortunately the contours are not convex unless the density estimate is "oversmoothed." One can expect these problems to be worse in higher dimensions. The regions in Figure 3 are also smaller than the corresponding regions in Figures 1 and 2.

1.2. *Related literature.* That F_n is the m.l.e. of F_0 was already well known by Kiefer and Wolfowitz (1956). They considered consistent estimation by maximum likelihood in various random effects settings. In a remark on page 893 they introduce an important method for defining m.l.e.s in undominated families of distributions. Kaplan and Meier (1958) give a derivation of the product limit estimator of the survival function as a nonparametric m.l.e. and Johansen (1978) shows that the product limit estimator is an m.l.e. in Kiefer and Wolfowitz's sense in an undominated class of point process models. More recently Bailey

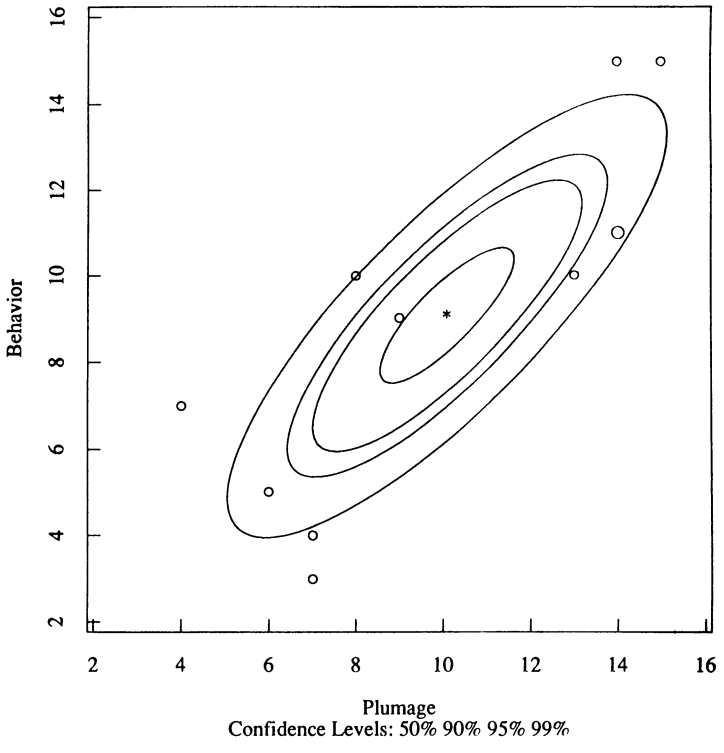

FIG. 2. *Normal likelihood contours.*

(1984) shows that in the absence of tied failure times and of time dependent covariates, nonparametric maximum likelihood applied to the Cox model yields the usual partial likelihood estimate for the regression parameter β, Tsiatis' (1981) estimate for the cumulative hazard function $\Lambda(t)$ and an information matrix for β and any particular $\Lambda(t_0)$ that can be inverted to form asymptotic confidence ellipses in the usual way. Vardi (1985) has used nonparametric maximum likelihood to estimate distribution functions in the presence of selection bias.

The first use of an empirical likelihood ratio function to set confidence intervals appears to be Thomas and Grunkemeier (1975). Their application was to survival probabilities estimated by the Kaplan–Meier curve. Thomas and 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 the usual intervals based on Greenwood's formula, Thomas and 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. Cox and Oakes (1984, Section 4.3) independently obtain the same intervals. A univariate version of Theorem 1 appears in Owen (1985) and a sharper univariate version appears in Owen (1988a). DiCiccio,

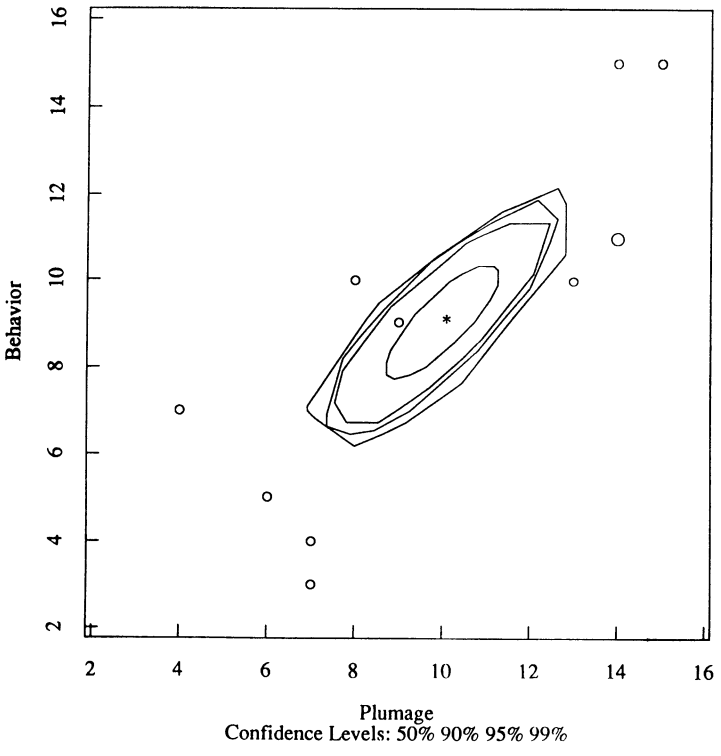

FIG. 3. *Bootstrap contours.*

Hall and Romano (1988) have shown that the error in Theorem 1 is $O(n^{-1})$ if the assumptions justifying Edgeworth expansions are met and that a Bartlett factor reduces the error to $O(n^{-2})$. The slower rate $O(n^{-1/2})$ still obtains for one sided problems. DiCiccio and Romano (1988) consider corrections to the signed root of the empirical likelihood ratio that allow construction of one sided confidence intervals with coverage error $O(n^{-1})$.

Empirical likelihood has parallels in the bootstrap literature. The Bayesian bootstrap (BB) of Rubin (1981) generates reweighted empirical distributions $\sum_{i=1}^{n} g_i \delta_{x_i}$, where the g_i are positive random variables with unit sum. In the simplest case they follow a unit Dirichlet distribution and may be sampled by taking the n gaps formed by 0, 1 and $n-1$ independent $U[0,1]$ random variables. To apply the BB to a functional T one computes $T\left(\sum_{i=1}^{n} g_i \delta_{x_i}\right)$ for many resampled vectors g. Rubin gives an example in which the histogram of 1000 BB correlation coefficients is similar to, but smoother than, a histogram of 1000 ordinary bootstrap correlation coefficients. The Bayesian argument behind the BB employs a finite support set $\{d_1, \ldots, d_k\}$ for F_0, a vector θ of probabilities with $\theta_j = F_0\{d_j\}$ and a prior for θ proportional to

$$\prod_{j=1}^{k} \theta_j^{p_j} \tag{1.2}$$

on the k variable unit simplex. The posterior for θ is then proportional to

$$\prod_{j=1}^{k} \theta_j^{p_j + n_j}, \tag{1.3}$$

where n_j is the number of observations equal to d_j. Rubin favors an improper prior with all $p_j = -1$, for then $\theta_j = 0$ with (improper) posterior probability 1 for any unobserved d_j. The advantage is that one need not know the value of d_j when $n_j = 0$. Empirical likelihood corresponds roughly to a noninformative prior with $p_j = 0$ in (1.2), and likelihood ratio function given by (1.3). While the BB samples from (1.3), empirical likelihood profiles (1.3) and uses an asymptotic calibration. Thus we have in a nonparametric setting, the familiar result that the likelihood function is a posterior for a noninformative prior. The correspondence is not exact since profiles of (1.3) would use $\theta_j > 0$ for $n_j = 0$ when d_j is outside of the convex hull of the data.

A variant of the nonparametric tilting bootstrap of Efron (1981, Section 11) uses the same family of multinomial distributions as empirical likelihood. Suppose $p = 1$ and define distributions $G_t \ll F_n$ with $G_t\{X_i\} \propto \exp(tX_i)$ for $-\infty < t < \infty$. This produces a one dimensional exponential family of discrete distributions passing through F_n at $t = 0$. The mean values of the G_t lie in the interval $(X_{(1)}, X_{(n)})$. Let $\mu_t = \int X\, dG_t$. The nonparametric tilting bootstrap interval for $E(X)$ has as its lower α limit, the value μ_t such that a bootstrap sample from G_t has fraction α of its resampled means larger than the sample mean $\overline{X}$. Upper limits and central intervals may be obtained similarly. One need not sample from many different family members because resampled tail areas from G_t can be obtained by exponentially tilting resampled tail areas from $G_0 = F_n$. A member G_t of this family minimizes the Kullback–Liebler distance from G to F_n among $G \ll F_n$ with $\int X\, dG = \mu_t$. The empirical log likelihood ratio is the Kullback–Liebler distance from F_n to $F \ll F_n$. Minimizing this distance for $F \ll F_n$ and $\int X\, dF = \mu$ yields F_μ. Efron parametrizes this family through $H_t \ll F_n$ with $H_t\{X_i\} \propto (1 + tX_i)^{-1}$ and suggest using bootstrap tail areas to form confidence intervals for $E(X)$. Since the family H_t is not exponential, tilting arguments cannot be exploited to reduce the computation. The families H_t and F_μ are reparametrizations of each other. Efron suggests one sided resampling inference from the members of H_t to form a confidence interval. Empirical likelihood forms a likelihood ratio function in the family F_μ or equivalently in H_t and profiles it. Efron notes that in the parametric family G_t the Cramér–Rao bound for $E(X)$ at $t = 0$ is $n^{-2}\Sigma(X_i - \overline{X})^2$ so that the reduction to a one dimensional family is not spuriously helpful. The same least favorable property holds in H_t. DiCiccio and Tibshirani (1986) also employ a least favorable family of distributions.

Finally we consider some results on multinomial likelihoods. Hoeffding (1965) shows that for multinomial distributions, tests based on the likelihood ratio are asymptotically optimal in the Bahadur sense. Tusnady (1977) extends Hoeffding's work by considering a sequence of multinomial families using finite partitions of the sample space that get finer as n increases. Tusnady shows that the likelihood

ratio tests on these families are asymptotically optimal (in the Bahadur sense) under some regularity conditions on the sequence of partitions. Tusnady does not indicate how to choose an optimal set of partitions. One of his regularity conditions is that the number of sets $m(n)$ in the partition satisfy $m(n)\log(n)/n \to 0$ as the sample size $n \to \infty$, so his results do not apply here where $m(n) = n - 1$. Berk and Jones (1979) use discrete likelihoods to test whether a sample is from the uniform distribution. They show that their test has a larger Bahadur slope than does any weighted Kolmogorov test at any alternative. Their test statistic is the most significant of n binomial p values, one at each of the order statistics. Their limit law is not χ^2, but is related to the extreme value distribution. Both Tusnady (1977) and Berk and Jones (1979) formulate their results through the Kullback–Leibler distance from the empirical measure to a hypothesized set of measures.

2. Proof of Theorem 1. We begin with a device that allows us to proceed as if there were no ties among the X_i. Let F be a distribution on $\mathbb{R}^p$ and suppose

$$w_i \geq 0, \qquad \sum_{j:\, X_j = X_i} w_j = F\{X_i\} \tag{2.1}$$

for $1 \leq i, j \leq n$. 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 $w_1 = w_2 = \cdots = w_n = n^{-1}$. 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* $r \in [0, 1]$,

$$\{F | R(F) \geq r\} = \{F | \tilde{R}(F, w) \geq r \text{ some } w \text{ satisfying } (2.1)\}.$$

PROOF. See Owen (1988a). □

In applications of Lemma 1, both sets of distributions are intersected with sets such as $\{F \ll F_n\}$ or $\{T(F) = t\}$. The consequence is that the regions $\{T(F) | R(F) \geq r\}$ are the same as $\{T(F) | \prod nw_i \geq r\}$, where F and w are related through (2.1). Therefore we may use

$$R(F) = \prod nw_i \tag{2.2}$$

in place of (1.1).

The next lemma will be used to guarantee that the mean of a distribution is eventually an interior point of the convex hull of a sample from that distribution.

LEMMA 2. *Let F_0 be a distribution on $\mathbb{R}^p$ with mean μ_0 and finite covariance matrix Σ of full rank p. Let Ω be the set of unit vectors in $\mathbb{R}^p$. Then for $X \sim F_0$,*

$$\inf_{\theta \in \Omega} P((X - \mu_0)'\theta > 0) > 0.$$

PROOF. Without loss of generality $\mu_0 = 0$. Suppose that there exists a sequence θ_n such that $P(X'\theta_n > 0) < n^{-1}$. Then by compactness of Ω there is a subsequence $\theta_n^* \to \theta_0 \in \Omega$. Let $E = \{X | X'\theta_0 > 0\}$. Then

$$1_{X'\theta_n^* > 0} \to 1_{X'\theta_0 > 0}$$

pointwise in E by continuity in θ of $X'\theta$. Now by Lebesgue's dominated convergence theorem (Royden 1968, page 88),

$$\begin{aligned} P(X'\theta_0 > 0) &= \int_E 1_{x'\theta_0 > 0}\, dF_0(X) \\ &= \lim_{n \to \infty} \int_E 1_{X'\theta_n^* > 0}\, dF_0(X) \\ &\le \lim_{n \to \infty} P(X'\theta_n^* > 0) \\ &= 0. \end{aligned}$$

Since $X'\theta_0$ has mean zero we also have $P(X'\theta_0 < 0) = 0$ so that $X'\theta_0 = 0$ a.s. $[F_0]$. But this contradicts the assumption that Σ has full rank. □

Lemma 3 provides some strong order bounds used in the proof of Theorem 1 and Lemma 4 sharpens one of them to obtain a rate.

LEMMA 3. *Let $Y_i \ge 0$ be i.i.d. random variables and define $Z_n = \max_{1 \le i \le n} Y_i$. If $E(Y_1^2) < \infty$, then*

$$Z_n = o(n^{1/2}) \tag{2.3}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} Y_i^3 = o(n^{1/2}),$$

both with probability 1 as $n \to \infty$.

PROOF. Since $E(Y_1^2) < \infty$ we have $\Sigma_1^\infty P(Y_1^2 > n) < \infty$, which implies that $\Sigma P(Y_n > n^{1/2}) < \infty$ and hence by the Borel–Cantelli lemma $Y_n > n^{1/2}$ finitely often with probability 1. But $Y_n > n^{1/2}$ finitely often implies $Z_n > n^{1/2}$ finitely

often. By the same argument $Z_n > An^{1/2}$ finitely often for any $A > 0$. Therefore

$$\limsup Z_n/n^{1/2} \leq A \tag{2.4}$$

with probability 1. Inequality (2.4) holds simultaneously with probability 1 for any countable set of values for A, so $Z_n = o(n^{1/2})$ with probability 1, establishing (2.3).

For the second assertion,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i^3 \leq \frac{Z_n}{n}\sum_{i=1}^{n} Y_i^2 = o(n^{1/2})$$

by (2.3) and the strong law of large numbers applied to ΣY_i^2. □

LEMMA 4. *Let* $Y_i \geq 0$ *be i.i.d. random variables and define* $Z_n = \max_{1 \leq i \leq n} Y_i$. *If* $E(Y_1^3) < \infty$ *and* $A > 0$, *then*

$$P\left(Z_n > An^{1/2}\right) = O(n^{-1/2}).$$

PROOF.

$$\begin{aligned} n^{1/2}P\left(Z_n > An^{1/2}\right) &\leq n^{3/2}P\left(Y_1 > An^{1/2}\right) \\ &\leq n^{3/2}E\left(Y_1^3\right)/\left(An^{1/2}\right)^3 \\ &= A^{-3}E\left(Y_1^3\right). \end{aligned}$$

□

PROOF OF THEOREM 1. If $q < p$ the problem may be reparametrized in terms of a q dimensional linear transformation of the X_i, so without loss of generality we may assume $q = p$.

Convexity of $C_{r,n}$ follows trivially from Jensen's inequality.

First we establish the limit law of $P(\mu_0 \in C_{r,n})$. Finiteness of Σ implies by Lemma 3 that

$$Z_n \equiv \max_{1 \leq i \leq n} \|X_i - \mu_0\| = o(n^{1/2}) \tag{2.5}$$

and that

$$\frac{1}{n}\sum \|X_i - \mu_0\|^3 = o(n^{1/2}), \tag{2.6}$$

both with probability 1 as $n \to \infty$.

By Lemma 2, with Ω the set of unit vectors in $\mathbb{R}^p$,

$$0 < \varepsilon \equiv \inf_{\theta \in \Omega} P\{(X - \mu_0)'\theta > 0\}$$

and by a generalization of the Glivenko–Cantelli theorem to uniform convergence over half spaces,

$$\sup_{\theta \in \Omega} |P\{(X - \mu_0)'\theta > 0\} - P_n\{(X - \mu_0)'\theta > 0\}| \to 0 \quad \text{a.s.},$$

where P_n denotes probability for $X \sim F_n$. It follows that

$$\inf_{\theta \in \Omega} P_n\{(X - \mu_0)'\theta > 0\} > \frac{\epsilon}{2}, \tag{2.7}$$

all but finitely often with probability 1.

We shall assume that n is large enough that (2.7) holds. A consequence of (2.7) is that μ_0 is contained in the convex hull of $\{X_1, \ldots, X_n\}$ as an interior point. Therefore

$$\mathcal{R}(\mu_0) = \sup\left\{R(F) \middle| \int X\,dF = \mu_0,\ F \ll F_n\right\} \tag{2.8}$$

exists and is positive. Since $\mu_0 \in C_{r,n}$ if and only if $\mathcal{R}(\mu_0) \geq r$, we need only show that

$$-2\log \mathcal{R}(\mu_0) \to \chi^2_{(p)}$$

in distribution as $n \to \infty$.

By Lemma 1 we may identify $\{F | F \ll F_n\}$ with the simplex of vectors $w \in \mathbb{R}^n$ with

$$w_i \geq 0, \qquad \sum w_i = 1.$$

In this formulation

$$\mathcal{R}(\mu_0) = \sup \prod nw_i, \tag{2.9}$$

where the supremum is over w in the simplex satisfying

$$\sum w_i(X_i - \mu_0) = 0.$$

Since R is continuous in F and $\{F \ll F_n\} \cap \{\int X\,dF = \mu_0\}$ is compact, it follows that the supremum $\mathcal{R}(\mu_0)$ is attained. A unique distribution attains the supremum, for otherwise a convex combination of two distinct distributions attaining the supremum would have mean μ_0 and a likelihood ratio strictly larger than $\mathcal{R}(\mu_0)$, which is a contradiction. We needed $\mathcal{R}(\mu_0) > 0$ to force uniqueness.

This unique distribution may be found via Lagrange multipliers. We maximize $R(F) = \prod nw_i$ subject to the p linear constraints $\sum w_i(X_i - \mu_0) = 0$. The result is

$$F_{\mu_0}\{X_i\} = w_i = \frac{1}{n}\frac{1}{1 + \lambda'(X_i - \mu_0)}, \tag{2.10}$$

where the multiplier $\lambda \in \mathbb{R}^p$ satisfies

$$0 = \frac{1}{n}\sum \frac{X_i - \mu_0}{1 + \lambda'(X_i - \mu_0)} \equiv g(\lambda). \tag{2.11}$$

Now we show that $\|\lambda\| = O_p(n^{-1/2})$. Write $\lambda = \rho\theta$ where $\rho \geq 0$ and $\|\theta\| = 1$. We need the positivity of $1 + \rho\theta'(X_i - \mu_0)$ below, but said positivity follows

easily from $w_i \le 1$. Now

$$
\begin{aligned}
0 &= \|g(\rho\theta)\| \\
&\ge |\theta' g(\rho\theta)| \\
&= \frac{1}{n}\left|\theta'\left(\sum (X_i - \mu_0) - \rho\sum \frac{(X_i - \mu_0)\theta'(X_i - \mu_0)}{1 + \rho\theta'(X_i - \mu_0)}\right)\right| \\
&\ge \frac{\rho}{n}\theta'\sum \frac{(X_i - \mu_0)(X_i - \mu_0)'}{1 + \rho\theta'(X_i - \mu_0)}\theta - \frac{1}{n}\left|\sum_{j=1}^{p} e_j'\sum (X_i - \mu_0)\right| \\
&\ge \frac{\rho\theta' S\theta}{1 + \rho Z_n} - \frac{1}{n}\left|\sum_{j=1}^{p} e_j'\sum (X_i - \mu_0)\right|,
\end{aligned}
\tag{2.12}
$$

where e_j is the unit vector in the jth coordinate direction and

$$
S = \frac{1}{n}\sum (X_i - \mu_0)(X_i - \mu_0)'. \tag{2.13}
$$

Now $\theta' S\theta \ge \sigma_p + o_p(1)$ where $\sigma_p > 0$ is the smallest eigenvalue of Σ. The second term in (2.12) is $O_p(n^{-1/2})$ by the central limit theorem. It follows that

$$
\frac{\rho}{1 + \rho Z_n} = O_p(n^{-1/2})
$$

and therefore by (2.5),

$$
\rho = \|\lambda\| = O_p(n^{-1/2}). \tag{2.14}
$$

Let $\gamma_i = \lambda'(X_i - \mu_0)$ where λ is the root of (2.11). Then by (2.14) and (2.5),

$$
\max_{1 \le i \le n} |\gamma_i| = O_p(n^{-1/2})o(n^{1/2}) = o_p(1). \tag{2.15}
$$

Expanding (2.11),

$$
\begin{aligned}
0 = g(\lambda) &= \frac{1}{n}\sum (X_i - \mu_0)\big(1 - \gamma_i + \gamma_i^2/(1 - \gamma_i)\big) \\
&= \overline{X} - \mu_0 - S\lambda + \frac{1}{n}\sum (X_i - \mu_0)\gamma_i^2/(1 - \gamma_i),
\end{aligned}
\tag{2.16}
$$

where $\overline{X}$ is the sample mean of $X_1, \ldots, X_n$ and S is given by (2.13). The final term in (2.16) has norm bounded by

$$
\frac{1}{n}\sum \|X_i - \mu_0\|^3 \|\lambda\|^2 |1 - \gamma_i|^{-1} = o(n^{1/2})O_p(n^{-1})O_p(1) = o_p(n^{-1/2}),
$$

using (2.6), (2.14) and (2.15). Therefore we may write

$$
\lambda = S^{-1}\big(\overline{X} - \mu_0\big) + \beta, \quad \text{where} \quad \|\beta\| = o_p(n^{-1/2}). \tag{2.17}
$$

By (2.15) we may expand

$$
\log(1 + \gamma_i) = \gamma_i - \gamma_i^2/2 + \eta_i,
$$

where, for some finite $B > 0$,

$$P\big(|\eta_i| \le B|\gamma_i|^3, 1 \le i \le n\big) \to 1$$

as $n \to \infty$. Substituting (2.10) and (2.17) in (2.9),

$$\begin{aligned}
-2\log \mathcal{R}(\mu_0) &= -2\sum \log nw_i \\
&= 2\sum \log(1+\gamma_i) \\
&= 2\sum \gamma_i - \sum \gamma_i^2 + 2\sum \eta_i \\
&= 2n\lambda'\big(\overline{X} - \mu_0\big) - n\lambda' S\lambda + 2\sum \eta_i \\
&= 2n\big(\overline{X} - \mu_0\big)'S^{-1}\big(\overline{X} - \mu_0\big) + 2n\beta'\big(\overline{X} - \mu_0\big) \\
&\quad - n\big(\overline{X} - \mu_0\big)'S^{-1}\big(\overline{X} - \mu_0\big) - 2n\beta'\big(\overline{X} - \mu_0\big) \\
&\quad - n\beta' S^{-1}\beta + 2\sum \eta_i \\
&= n\big(\overline{X} - \mu_0\big)'S^{-1}\big(\overline{X} - \mu_0\big) - n\beta' S^{-1}\beta + 2\sum \eta_i,
\end{aligned}$$

where, as $n \to \infty$,

$$n\big(\overline{X} - \mu_0\big)'S^{-1}\big(\overline{X} - \mu_0\big) \to \chi^2_{(p)} \tag{2.18}$$

in distribution and

$$n\beta' S^{-1}\beta = o_p(1)$$

and

$$\begin{aligned}
\Big|2\sum \eta_i\Big| &\le 2B\|\lambda\|^3 \sum \|X_i - \mu_0\|^3 \\
&= 2BO_p(n^{-3/2})o_p(n^{3/2}) \\
&= o_p(1).
\end{aligned}$$

This establishes the limit law for $P(\mu_0 \in C_{r,n})$. Now we establish that the rate of convergence is $O_p(n^{-1/2})$ assuming $E\|X\|^4 < \infty$. Using only $E\|X\|^3 < \infty$, the convergence in (2.18) is at the rate $n^{-1/2}$ by the Berry–Esseen theorem. We need to show that

$$n\beta' S^{-1}\beta = O_p(n^{-1/2}) \tag{2.19}$$

and

$$\sum \eta_i = O_p(n^{-1/2}) \tag{2.20}$$

and that the convergence in (2.5), (2.7) and (2.15) is fast enough so that the events they allowed us to neglect in finding the limit law have probability $O(n^{-1/2})$. We do not need a rate on (2.6) since it will not be used to establish (2.19) and (2.20) and the limit law used it only to bound the last term of (2.16). A sharper bound is used below for that term.

If $E\|X\|^3 < \infty$ then the strong law of large numbers applies to $\Sigma\|X_i - \mu_0\|^3$ and so the final term in (2.16) is

$$O(1)O_p(n^{-1})O_p(1) = O_p(n^{-1}).$$

Therefore $\|\beta\| = O_p(n^{-1})$ and the term in (2.19) is $O_p(n^{-1})$. Similarly

$$\left|\sum \eta_i\right| = O_p(n^{-3/2})O(n) = O_p(n^{-1/2}),$$

establishing (2.20).

Lemma 4 applied to $\|X_i - \mu_0\|^{4/3}$ implies that

$$P(Z_n > n^{3/8}) = O(n^{-1/2}) \tag{2.21}$$

so that a result stronger than (2.5) holds at the rate $n^{-1/2}$. By the Vapnik–Chervonenkis inequality [Gaenssler (1983), page 28] the probability in (2.7) converges to 0 exponentially fast.

To establish a rate for (2.15) we need a rate of convergence on $\|\lambda\|$. We work with the last term in (2.12):

$$\begin{aligned} P\left(\frac{1}{n}e_j'\sum(X_i - \mu_0) > n^{-3/8}\right) &= P\left(\left(\frac{1}{n}e_j'\sum(X_i - \mu_0)\right)^4 > n^{-3/2}\right) \\ &\le E\left(\left(\frac{1}{n}e_j'\sum(X_i - \mu_0)\right)^4\right)n^{3/2} \\ &= O(n^{-1/2}) \end{aligned}$$

and so

$$P(\|\lambda\| \ge n^{-3/8}) = O(n^{-1/2}). \tag{2.22}$$

Note that the smallest eigenvalue of S is larger than $\sigma_p/2$ with probability $1 - O(n^{-1/2})$ by a Chebyshev argument on the components of S.

The version of (2.15) with a rate is

$$\begin{aligned} P(\max|\gamma_i| > 0.25) &\le P(n^{3/8}\|\lambda\| > 0.5) + P(n^{-3/8}Z_n > 0.5) \\ &= O(n^{-1/2}), \end{aligned}$$

by easy extensions of (2.21) and (2.22). The theorem is proved. □

The leading term in the expansion of $-2\log\mathscr{R}(\mu_0)$ is nearly Hotelling's T^2. The distinction is that S is not the usual sample variance–covariance matrix. After some algebra,

$$n(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0) = T^2\left(1 - \frac{1}{n-1} + \frac{T^2}{n}\right)^{-1} = T^2 + O_p(n^{-1}).$$

For normally distributed X_i, the distribution of $(n-p)T^2/((n-1)p)$ is $F_{p,n-p}$. The scaled F would be replaced by $\chi^2_{(p)}$ if Σ could be used in place of S in T^2. The use of the F distribution instead of the χ^2 compensates for the estimation of Σ from a finite sample. This suggests that the F distribution might provide a

better reference point than the χ^2 for $-2\log \mathcal{R}(\mu_0)$. A simulation (Owen, 1988b) for samples of size 3 through 20 and a variety of sampling distributions bears this out for $p = 1$.

Now consider the limiting distribution of $-2\log \mathcal{R}(\mu)$ for μ within $O_p(n^{-1/2})$ of μ_0. We have the following corollary to the proof of Theorem 1:

COROLLARY 1. *Under the conditions of Theorem* 1 *for any* $\tau \in \mathbb{R}^p$,

$$-2\log \mathcal{R}\bigl(\mu_0 + n^{-1/2}\Sigma^{1/2}\tau\bigr) \to \chi^2_{(q)}\bigl(\|\tau\|^2\bigr)$$

in distribution, where q *is the rank of* Σ *and* $\|\tau\|^2$ *is a noncentrality parameter and*

$$-2\log \mathcal{R}\bigl(\overline{X} + n^{-1/2}\hat{\Sigma}^{1/2}\tau\bigr) \to \|\tau\|^2$$

in probability where $\hat{\Sigma} = (1/n)\Sigma(X_i - \overline{X})(X_i - \overline{X})'$.

PROOF. The proof is essentially the same as that given above for Theorem 1. □

The first result in Corollary 1 shows that the power of a level α empirical likelihood ratio test is asymptotically $P(\chi^2_{(q)}(\|\tau\|^2) > \chi^2_{(q),1-\alpha})$ on the ellipsoid $\mu = \mu_0 + n^{-1/2}\hat{\Sigma}^{1/2}\tau$ generated by $\tau \in \mathbb{R}^p$ with $\|\tau\|$ constant. The second result shows the same quadratic limit, centered on $\overline{X}$.

3. Computing intervals for the mean. We need to compute the profile empirical likelihood function $\mathcal{R}(\mu) = R(F_\mu)$ for various candidates μ for $E(X)$. Tests of $E(X) = \mu$ may be based on the magnitude of $\mathcal{R}(\mu)$. Confidence regions for $E(X)$ can be constructed by computing $\mathcal{R}(\mu)$ on a grid of μ values and applying a contouring algorithm.

By an argument in Section 2 it follows that for μ in the convex hull of $X_1, \ldots, X_n$,

$$F_\mu\{X_i\} = \frac{1}{n}\frac{1}{1 + \lambda'(x_i - \mu)}, \tag{3.1}$$

where the multiplier $\lambda = \lambda(\mu) \in \mathbb{R}^p$ exists and is uniquely determined by

$$0 = \sum \frac{X_i - \mu}{1 + \lambda'(X_i - \mu)} \equiv g(\lambda). \tag{3.2}$$

It follows that

$$\mathcal{R}(\mu) = \prod\bigl(1 + \lambda'(X_i - \mu)\bigr)^{-1}. \tag{3.3}$$

For $p = 1$ it is easy to solve (3.2) with a safeguarded zero finding algorithm such as Brent's method (Press, Flannery, Teukolsky and Vetterling, 1986). Owen (1988a) uses Brent's method to maximize empirical likelihood ratios for certain M estimates and shows how to find an initial interval containing λ. The bisection algorithm used by safeguarded zero finders is unavailable for $p > 1$, so we reformulate the problem as one of minimization of a convex function.

Note that $-g$ is the gradient with respect to λ of

$$f(\lambda) = -\sum \log(1 + \lambda'(X_i - \mu)),$$

so the zero of g is a stationary point of f. We now show attention can be restricted to a convex domain for f and that under mild conditions, the domain is compact and f is strictly convex. By (3.1) we need only consider λ for which

$$1 + \lambda'(X_i - \mu) \geq 1/n, \qquad 1 \leq i \leq n, \tag{3.4}$$

so we may assume that λ is in the intersection D of n half spaces given by (3.4). D is convex and closed. Assuming that the points $X_1, \ldots, X_n$ span $\mathbb{R}^p$, D is also compact, its boundedness following from the assumption that μ is interior to the convex hull of the data. The Hessian of f is

$$H(\lambda) = \sum \frac{(X_i - \mu)(X_i - \mu)'}{[1 + \lambda'(X_i - \mu)]^2},$$

which is positive semidefinite on D because $1 + \lambda'(X_i - \mu) > 0$. Assuming that $X_1, \ldots, X_n$ span $\mathbb{R}^p$, the sample variance of the X_i has full rank, so H is positive definite on D and hence f is strictly convex. It follows that f has a unique global minimum on D and that f 's minimizer is the solution of (3.2).

We now have the following dual problem: to maximize $\log R(F)$ over the simplex in n dimensions subject to the p constraints $\int X\, dF = \mu$ is to minimize f over D without constraints. Thus we have reduced the dimension and obtained a problem with no constraints except $\lambda \in D$, which can be removed as described below. Notice that $f = \log \mathscr{R}$ except that the former is written as a function of λ and the latter as a function of μ, since $\lambda = \lambda(\mu)$ in (3.3). Interestingly this makes the dual problem one of minimum likelihood. For values $\lambda \in D$ other than the solution of (3.2), F_μ given by (3.1) is not in general a probability measure.

The preceding result is a special case of convex duality. The minimization problem is one of convex programming. For a general discussion of these topics, see Pshenichny and Danilin (1978, Chapter 3).

There are several theorems that guarantee that the unique solution of (3.2) can be found.

Theorem 4.9 of Rheinboldt (1974, page 48) guarantees that certain damped Newton algorithms will converge to the solution of (3.2).

Theorem 5.2 of Rheinboldt (1974, page 62) guarantees superlinear convergence for the Davidon–Fletcher–Powell algorithm, provided the starting point λ_0 is one for which the level set $\{\lambda | f(\lambda) \leq f(\lambda_0)\}$ is compact. This holds for $\lambda_0 = 0$.

Theorem 4.14 of Rheinboldt (1974, page 51) guarantees convergence of three different one-variable-at-a-time algorithms. Such algorithms proceed by adjusting λ_j for $j = 1 + (i \bmod p)$, $i = 0, 1, 2 \ldots$, until convergence to suitable accuracy has been obtained. Indeed one could use Brent's method component by component on λ in an iterative manner to solve (3.2).

It is unusual to have such strong theoretical support for a minimization problem. More commonly one can only get local convergence results that guarantee convergence to a relative minimum from a starting point sufficiently close to the solution.

It is convenient to extend f from D to $\mathbb{R}^p$, while preserving convexity. Let $Q(\cdot)$ be the quadratic function of $\mathbb{R}$ that matches $\log(\cdot)$ and its first two derivatives at $1/n$. Now let $\log^*(x) = \log(x)$ for $x \geq 1/n$ and $\log^*(x) = Q(x)$ for $x < 1/n$. Finally put

$$f^*(\lambda) = -\sum \log^*(1 + \lambda'(X_i - \mu)).$$

By (3.4), $f^* = f$ on D and strict global convexity of f^* follows from the strict concavity of $\log^*$.

The convergence theorems describe the performance of the algorithms when computations are made with infinite precision and infinite sequences of steps are carried out. In practice one has to contend with finite approximations on both issues. It has been the author's experience that the computations are most easily made for μ near $\bar{X}$ and that as μ approaches the convex hull of the data the computation becomes more difficult. Algorithms may therefore be compared according to how small the log likelihood ratio must become before the algorithm encounters difficulty. A natural goal for computation is to be able to compute the log likelihood ratio down to values corresponding to confidence intervals with coverage well beyond that required in practice. For other values of μ the approximation $\mathcal{R}(\mu) = 0$ is adequate. For the duck data, the IMSL conjugate gradient routine ZXCGR applied to f^* allows computation of log likelihoods smaller than -50, which far exceeds the needs of any reasonable confidence regions for the mean.

When μ is outside of the convex hull of the data, there is no solution. In practice what happens is that the algorithm terminates at a large value of λ for which the slope of the logarithm is so small that the gradient is zero to the required precision. One can tell that this has happened because the w_i are not in the unit simplex. Typically they sum to less than 1. This provides an alternative way to decide whether a point is in the convex hull of a finite set of points: Maximize the empirical log likelihood ratio of the first point considered as the mean of the others, using the extension by $\log^*$ and inspect the solution. This may be more convenient than making a preliminary check of whether a given point is within the convex hull of the data, especially when the dimension of the data is higher than 2.

4. Extensions to other statistics. Many statistics are related in simple ways to sample means. In this section we consider three extensions from sample means to more general classes of statistics. They are all based on delta method arguments.

First we consider parameters which are functions of means. Examples include the variance of X, which is a function of the mean of (X, X^2), and the correlation between X and Y, which is a function of the mean of (X, Y, X^2, Y^2, XY). Apart from factors like $n/(n-1)$ the natural sample statistics in these examples are the analogous functions of sample means.

Let $Z_1, \ldots, Z_n$ be i.i.d. in $\mathbb{R}^p$ with mean μ_Z and sample mean $\bar{Z}$. Let H: $\mathbb{R}^p \to \mathbb{R}^q$ and consider the estimate $H(\bar{Z})$ of the parameter $H(\mu_Z)$, where H has a Frechet derivative $G \neq 0$ at μ_Z. That is, $H(Z) = H(\mu_Z) + G(Z - \mu_Z) +$

$o(\|Z - \mu_Z\|)$ as $\|Z - \mu_Z\| \to 0$. Under additional conditions Theorem 1 justifies confidence regions for $H(\mu_Z) + G(Z - \mu_Z)$ if μ_Z is known. These regions cannot be computed in practice, because they depend on μ_Z, but they can be shown to be very close to the empirical likelihood regions for H which do not require knowledge of μ_Z.

THEOREM 2. *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. random vectors in $\mathbb{R}^p$, $p \geq 1$, with mean μ_Z and finite variance Σ of rank p. Let H: $\mathbb{R}^p \to \mathbb{R}^q$ have Frechet derivative $G \neq 0$ at μ_Z. Define*

$$Z_{r,n} = \left\{ \int Z\, dF(Z) \mid F \ll F_n,\ R(F) \geq r \right\},$$

$$C_{r,n} = \{H(Z) \mid \qquad Z \in Z_{r,n}\},$$

$$C'_{r,n} = \{H(\mu_Z) + G(Z - \mu_Z) \mid Z \in Z_{r,n}\},$$

and let $\bar{Z} = (1/n)\sum_{i=1}^n Z_i$. Then as $n \to \infty$,

$$P(H(\mu_Z) \in C'_{r,n}) \to P\left(\chi^2_{(d)} \leq -2\log(r)\right),$$

where $d = \min(p, \operatorname{rank}(G))$ and

$$\sup_{Z \in Z_{r,n}} \|H(Z) - H(\mu_Z) - G(Z - \mu_Z)\| = o_p(n^{-1/2}).$$

PROOF. The linearization $H(\mu_Z) + G(Z - \mu_Z)$ has mean $H(\mu_Z)$ and variance $G\Sigma G$ of rank $d > 0$, so the first assertion follows by Theorem 1.

By Corollary 1, $\sup_{Z \in Z_{r,n}} \|Z - \mu_Z\| = O_p(n^{-1/2})$. The second assertion now follows from the definition of G. □

The regions $C'_{r,n}$ justified by Theorem 1 differ from the empirical likelihood regions $C_{r,n}$ by $o_p(n^{-1/2})$, which is asymptotically negligible compared to their diameter $O_p(n^{-1/2})$.

Essentially the same argument and conclusions may be made for Frechet differentiable statistical functionals $T(F)$. Owen (1988a, Theorem 3) proves this for distributions F on $\mathbb{R}$, where the derivative is defined in terms of the Kolmogorov norm on an appropriate space of distributions. The proof extends to distributions on $\mathbb{R}^p$ with the Kolmogorov norm replaced by a sup norm over half spaces of $\mathbb{R}^p$.

Theorem 2 justifies empirical likelihood intervals for the variance of X, assuming $E(X^4)$ exists. An example is given in Section 6. Owen (1988a) justifies intervals for the variance of X under the restrictive condition that X is bounded, which makes the variance a differentiable statistical functional. The multivariate empirical likelihood result thus makes it possible to introduce a nuisance parameter for the mean and obtain a better result for a one dimensional statistic like the variance.

Theorem 1 may be extended to M estimates. An M estimate is defined as a root $\tau = T(F)$ of

$$\int \psi(X, \tau)\, dF = 0, \tag{4.1}$$

where $X \sim F$. Usually X, τ and ψ take real values, but here we generalize to $X \in \mathbb{R}^p$ and $\tau, \psi(X, \tau) \in \mathbb{R}^q$. Conditions must be imposed on $\psi(X, \tau)$ to guarantee a solution to (4.1) and further conditions may be used to obtain a unique root. We will impose conditions through families of 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 3. *Let $T(F)$ be a solution of* (4.1) *and let $X_1, X_2, \ldots, X_n$ be i.i.d. in $\mathbb{R}^p$ with distribution F_0. Assume that $\psi\colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^q$ satisfies*:

$$T(F_0) = \tau \text{ exists and is unique}, \tag{4.3a}$$

$$\psi_{\cdot\tau} \text{ is measureable and} \tag{4.3b}$$

$$0 \neq \operatorname{var}(\psi(X_1, \tau)) \text{ exists}. \tag{4.3c}$$

For positive $r < 1$ and R given by (1.1), *let*

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

and

$$C_{r,n} = \bigcup_{F \in \mathscr{F}_{r,n}} \left\{ t \middle| \int \psi(X, t)\, dF = 0 \right\}.$$

Then as $n \to \infty$,

$$P(T(F_0) \in C_{r,n}) \to P\left(\chi^2_{(d)} \leq -2\log(r)\right),$$

where $d = \operatorname{rank}(\operatorname{var}(\psi(X_1, \tau)))$.

PROOF. Let $Z(X) = \psi(X, \tau)$ where $\tau = T(F_0)$ is the unique root of (4.1) for $F = F_0$, assumed in (4.3a). Condition (4.3b) on $\psi_{\cdot\tau}$ ensures that $Z(X_i) = \psi(X_i, \tau)$ are random vectors. Then (4.3c) yields, via Theorem 1, that as $n \to \infty$,

$$P(\tau \in C_{r,n}) = P(0 \in W_{r,n}) \to P\left(\chi^2_{(d)} \leq -2\log(r)\right),$$

where

$$W_{r,n} = \left\{ \int Z(X)\, dF \middle| F \in \mathscr{F}_{r,n} \right\}. \qquad \square$$

For $p = 1$, Owen (1988a, Theorem 2) shows that if the $\psi_{x\cdot}(t)$ is nonincreasing in t for F_0-almost all x, then the confidence sets $C_{r,n}$ are intervals. A natural conjecture is that if $\psi_{x\cdot}\colon \mathbb{R}^q \to \mathbb{R}^q$ is a monotone function (i.e., the gradient of a convex function), then $C_{r,n}$ is a convex set in $\mathbb{R}^q$. An example in Section 6 shows that the conjecture is false.

DiCiccio, Hall and Romano (1988) show that empirical likelihood regions for smooth functions of means have coverage errors on the order $O(n^{-1})$ and that Bartlett adjustment reduces this to $O(n^{-2})$. DiCiccio and Romano (1988) show that corrections to the signed root of the profile empirical likelihood ratio for a smooth function of the mean are normal to $O(n^{-1})$.

5. Comparison to Johnson's t and Student's t. The proof of Theorem 1 justifies the empirical likelihood regions by showing that for fixed r they tend to regions based on the central limit theorem as $n \to \infty$. The extension in Theorem 2 shows that for certain nonlinear statistical functionals, an approximation by a linear functional justified by Theorem 1 makes an asymptotically negligible difference. In this section we show that there is reason to expect that the empirical likelihood regions should be better than intervals based on the central limit theorem, by incorporating an adjustment for skewness.

When $p = 1$, the normal theory intervals are obtained by referring the pivot

$$t = n^{1/2}(\bar{X} - \mu)s^{-1}$$

to the t_{n-1} distribution, where s is the usual sample standard deviation. Assuming the X_i have moments of all orders, Johnson (1978) develops a Cornish–Fisher expansion

$$\text{CF}(t) = Z_1 - n^{-1/2}\gamma/6 - n^{-1/2}\gamma Z_1^2/3 - n^{-1/2}AZ_1Z_2, \tag{5.1}$$

from which terms of order smaller than $n^{-1/2}$ are omitted. Here γ is the skewness of X_1, $A = (\kappa + 2 - \gamma^2)^{1/2}/2$, where κ is the kurtosis of X_1, and Z_1 and Z_2 are independent standard normal random variables. The presence of a constant term in (5.1) indicates bias in t and a nonzero coefficient for $Z_1^2 - 1$ indicates skewness in t.

Johnson then considers

$$t_1 = t + n^{-1/2}\gamma/6 + n^{-1/2}\gamma t^2/3,$$

which has Cornish–Fisher expansion

$$\text{CF}(t_1) = Z_1 - n^{-1/2}AZ_1Z_2. \tag{5.2}$$

Johnson's t_1 corrects t for bias and skewness associated with the skewness in X_1. Johnson proposes estimating γ by the sample skewness and using the resulting quantity $\hat{t}_1$. Based on some simulations, Johnson concludes that the accuracy of the t variable is improved.

To examine the empirical likelihood ratio, let $s_\mu^2 = n^{-1}\Sigma(X_i - \mu)^2$, $\gamma_\mu = n^{-1}\Sigma(X_i - \mu)^3/s_\mu^3$ and $t_\mu = n^{1/2}(\bar{X} - \mu)/s_\mu$. For $\|\mu - \mu_0\| = O_p(n^{-1/2})$ the expansion of $-2\log\mathscr{R}(\mu)$ to terms of order $n^{-1/2}$ is

$$\begin{aligned} -2\log\mathscr{R}(\mu) &= n(\bar{X} - \mu)^2/s_\mu^2 + \frac{2}{3}\sum(X_i - \mu)^3(\bar{X} - \mu)^3/s_\mu^6 \\ &= t_\mu^2 + 2n^{-1/2}t_\mu^3\gamma_\mu/3. \end{aligned}$$

This expansion is, to the order considered, the square of

$$t_r = t_\mu + n^{-1/2} t_\mu^2 \gamma_\mu / 3.$$

The difference between t_μ and t is of smaller order than $n^{-1/2}$ as is that between $n^{-1/2}\gamma_\mu$ and $n^{-1/2}\gamma$. Therefore the Cornish–Fisher expansion of the signed root of $-2\log \mathscr{R}(\mu)$ is, to the accuracy considered, the same as that of $t + n^{-1/2}t^2\gamma/3$, that is,

$$\mathrm{CF}(t_r) = Z_1 - n^{-1/2}\gamma/6 - n^{-1/2}AZ_1Z_2. \tag{5.3}$$

Skewness in X_1 results in skewness and bias in t. Examining (5.3) we see that the empirical likelihood method corrects the skewness but not the bias. It should therefore improve upon t in large samples, but not as much as Johnson's $\hat{t}_1$, which removes the skewness and bias. For central confidence intervals, the bias cancels but this does not happen for one sided intervals. In Owen (1988a) a simulation with X_i from the $\chi^2_{(1)}$ distribution and $n = 20$ shows that central 90% empirical likelihood intervals for the mean were much more accurate than those based on Student's t. Simulations in Owen (1988b) show an improvement over Student's t for samples of size 10 or more from skewed distributions. For very small samples (fewer than 10 points) Student's t is as good as or better than empirical likelihood. The empirical likelihood central intervals were also more accurate than some of the simpler bootstrap central intervals. The bias effects were evident in the one sided empirical likelihood intervals, which were quite inaccurate. This one sided inaccuracy is typical of likelihood based methods.

6. Other examples. In this section some examples of empirical likelihood inference based on Theorem 2 are given. Sections 6.1 and 6.2 consider the variance and product moment correlation, respectively. A two level (nested) algorithm is used for computing the profile empirical likelihoods. The inner level of the algorithm is the one discussed in Section 3 for the mean and the outer level profiles out nuisance parameters. It is shown in Section 6.3 that the outer level is, for large samples, a convex optimization. Some applications of Theorem 3 are discussed briefly in Section 6.4.

6.1. *Variance.* We consider first the variance σ^2 of a real random variable X. Abuse notation slightly and let F_σ maximize $R(F)$ subject to $F \ll F_n$ and $(\int X^2\,dF - (\int X\,dF)^2)^{1/2} = \sigma$. Let $F_{\mu,\sigma}$ maximize $R(F)$ subject to $F \ll F_n$, $\int X\,dF = \mu$ and $\int (X-\mu)^2\,dF = \sigma^2$. Since

$$R(F_\sigma) = \sup_\mu R(F_{\mu,\sigma}),$$

we can use a nested algorithm as follows: Compute $F_{\mu,\sigma}$ for σ and some candidate μ at the inner level, and optimize over μ at the outer level. The nested algorithm was used because a simpler algorithm based on iterative linear approximations failed. In that algorithm one alternates between maximizing the empirical likelihood for $E((X-\mu)^2) = \sigma^2$ assuming a "known" mean μ and updating μ by the mean of the likelihood maximizing distribution. The alternating

algorithm failed to compute even moderately extreme (0.90 to 0.99) upper confidence points for the variance.

We also use the notation $\mathscr{R}(\sigma) = R(F_\sigma)$ and $\mathscr{R}(\mu, \sigma) = R(F_{\mu,\sigma})$. The distinction from profile empirical likelihoods for a mean will be clear from context.

The inner optimization is done through the algorithm in Section 3. The outer optimization is done by any suitable one dimensional optimizer. It is convenient to use derivative information at the outer stage. It may be shown that for fixed σ,

$$\frac{d}{d\mu}\log \mathscr{R}(\mu, \sigma) = n\lambda_1,$$

where λ_1 is the Lagrange multiplier corresponding to the constraint $\int X\, dF_{\mu,\sigma} = \mu$ in the inner optimization.

Not every pair (μ, σ) is obtainable through reweighting the sample. One could require the outer level of optimization to consider only those values of μ compatible with the value of σ at hand, but it is far easier to extend the domain of the empirical log likelihood through the function log* as described in Section 3. With this extension

$$\frac{d}{d\mu}\log \mathscr{R}(\mu, \sigma) = n\lambda_1 \sum w_i(\lambda),$$

where

$$w_i(\lambda) = \frac{1}{n}\log^{*\prime}\left(1 + \lambda_1(X_i - \mu) + \lambda_2\left[(X_i - \mu)^2 - \sigma^2\right]\right).$$

Here $\lambda = (\lambda_1, \lambda_2)'$ is the vector of Lagrange multipliers from the inner level and log*′ is the derivative of log*.

That $d/d\mu \log \mathscr{R}(\mu, \sigma) = n\lambda_1$ can be used to assess properties of the outer optimization. Let Σ denote the variance of $(X - \mu, (X - \mu)^2 - \sigma^2)'$. Then for (μ, σ^2) with $O_p(n^{-1/2})$ of $(E(X), \mathrm{var}(X))$ and assuming $E(X^8) < \infty$,

$$\lambda = \Sigma^{-1}\begin{pmatrix} \overline{X} - \mu \\ \overline{V} - \sigma^2 \end{pmatrix} + O_p(n^{-1}),$$

where $V_i = (X_i - \mu)^2 - \sigma^2$ and $\overline{V} = n^{-1}\Sigma V_i$. This suggests that the second derivative of $\log \mathscr{R}(\mu, \sigma)$ with respect to μ should be negative and, hence, that $\log \mathscr{R}(\mu, \sigma)$ should be concave in μ for fixed σ. A general argument along these lines is made in Section 6.3.

Larsen and Marx (1986, page 332) give 19 estimated ages, in millions of years, of mineral samples collected in the Black Forest. The ages were estimated by potassium–argon dating. The variance of these measurements is of direct interest since it provides information on the precision of the dating method. A histogram of this data appears in Figure 4. The sample standard deviation is 27.1 million years and a normal theory 95% confidence interval is 20.1 to 40 million years.

The empirical likelihood ratio function was calculated for standard deviations in the range from 1.5 to 51.5 million years in steps of half a million years.

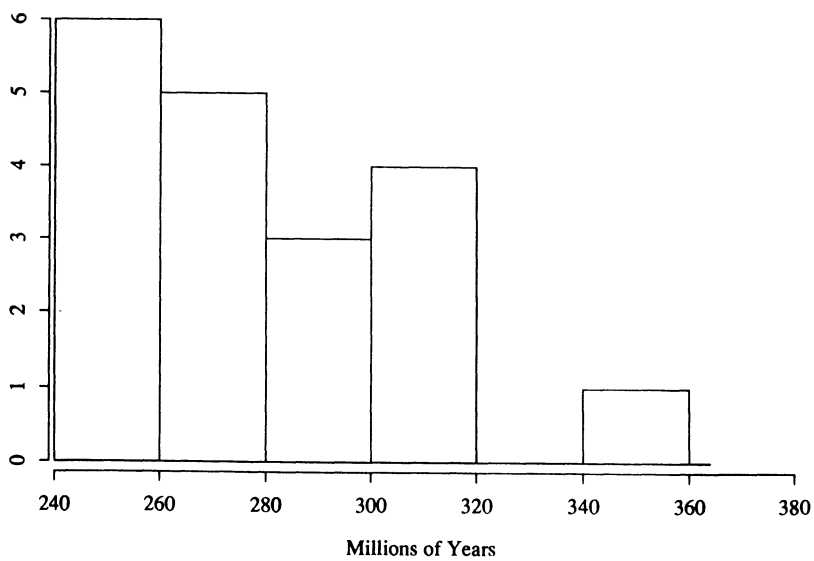

FIG. 4. *Potassium–argon dates.*

Computations were made for an increasing sequence of standard deviations starting near the maximum likelihood estimate and for a decreasing sequence starting there. This way the final values from each step could be used as starting values for the next. It took 2 minutes to make 102 likelihood evaluations on a microvax VaxStation II.

Figure 5 shows the empirical likelihood ratio function, together with the normal theory likelihood ratio function. The likelihood ratios are plotted against standard deviations in millions of years. The horizontal lines correspond to 90% and 95% empirical likelihood confidence intervals. Slightly different lines would be appropriate for the exact confidence regions based on a normal model. The empirical likelihood ratio curve has a shorter right tail and a very slightly longer left tail than the normal one. It is surprising how close the two curves are. The shorter right tail of the empirical curve seems natural given the apparent shortness of the tails in Figure 4. The sample kurtosis is 0.02 if one uses the normal maximum likelihood estimate of σ^2 as in Miller (1986, page 272) and -0.29 if one uses the unbiased estimate of σ^2. Since the sample maximum is 344 and the minimum is 243, the largest possible standard deviation for a reweighted sample is 51.5. The algorithm found an empirical log likelihood of -52.9 for a standard deviation of 51. The smallest standard deviation for which a meaningful solution was obtained was 3.5 and the corresponding empirical log likelihood was -51.8. These correspond to putative $\chi^2_{(1)}$ values larger than 100. It follows that for any confidence level of practical interest the empirical interval for the variance can be computed from this data. For standard deviations outside (3.5, 51) the modifications to the logarithm that make it possible to use generic optimizers lead to convergence to solutions for which the weights w_i sum to less

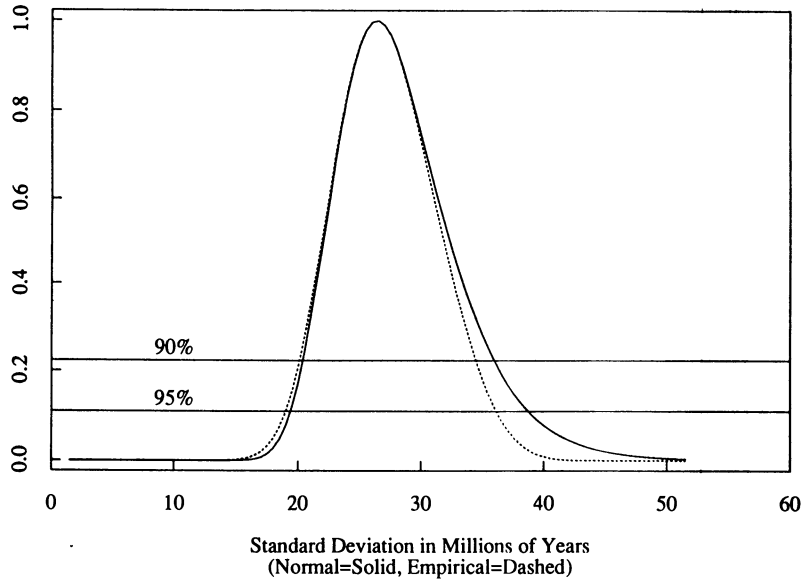

FIG. 5. *Profiled likelihood ratios.*

than 1. It made the computations more stable to divide the ages by 100 before computing the intervals.

The normal theory curve is exact if the observations are normally distributed and has a large sample justification if the kurtosis of the measurements is 0. The empirical likelihood curve has a large sample justification through Theorem 2 if the kurtosis is finite. Figure 6 shows a histogram of 1000 bootstrap replications of the standard deviation. The histogram has a location and scale comparable to those of the likelihood ratio curves.

6.2. *Correlation.* A similar nested algorithm works for the correlation ρ. The inner level consists of finding the likelihood of $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\sigma_x\sigma_y)$ as a mean for

$$\left(X, Y, (X-\mu_x)^2, (Y-\mu_y)^2, (X-\mu_x)(Y-\mu_y)\right).$$

The outer level consists of maximizing the result of the inner level over choices of $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)$. Using the dual problem, the inner optimization is over five variables and the outer is over four variables. The whole computation is done over a one dimensional grid of values for ρ. The four variables of the outer optimization must obey some constraints to be valid moments. Rather than check whether each trial point of the outer optimization is possible, it is easier to extend the inner function as described in Section 3. As before analytic derivatives are available for the outer optimization. Numerical performance is improved by centering and scaling both the X and Y variables.

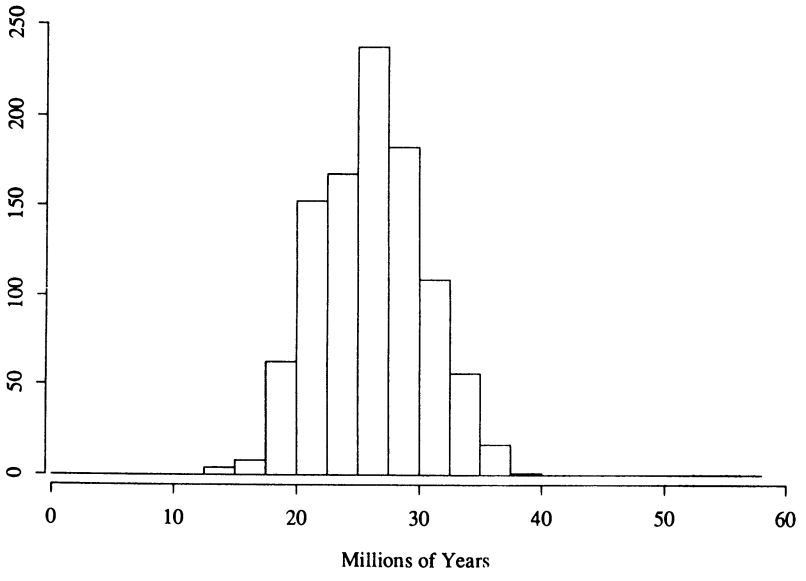

FIG. 6. *Bootstrap standard deviations.*

Larsen and Marx (1986, page 456) give 15 pairs of observations relating the frequency with which crickets chirp to the temperature. The data are plotted in Figure 7. The frequency is said to be in chirps per second and the temperature is given in degrees Fahrenheit. The sample correlation is 0.835.

The profile empirical likelihood ratio function for the correlation is plotted in Figure 8. Also shown is the normal theory profile likelihood ratio function. The empirical curve lies above the normal one. Figure 9 is a histogram of 1000 bootstrap replications of the correlation. The empirical likelihood ratio curve is very asymmetric, so it will yield inferences quite different from those based on an estimated standard deviation for ρ. The shape of the curve is similar to the bootstrap histogram. Theorem 2 justifies empirical likelihood for the correlation by a delta method argument. Thus we might expect to find a curve with a location and width determined by the sample correlation and a sample estimate of its standard deviation. That the shape of the curve is skewed in what is known to be the right direction for normal samples is reassuring and not predicted by the argument used in Theorem 2.

6.3. *Nested algorithm.* We show that the outer level of the optimization should be well behaved in a large class of problems, at least for large samples.

Generalize the approach illustrated above for the variance and correlation as follows. Let θ be a vector of parameters of interest. Let ν be a vector of nuisance parameters. Suppose that the estimating equation

$$\int m(Z, \nu, \theta)\, dF = 0$$

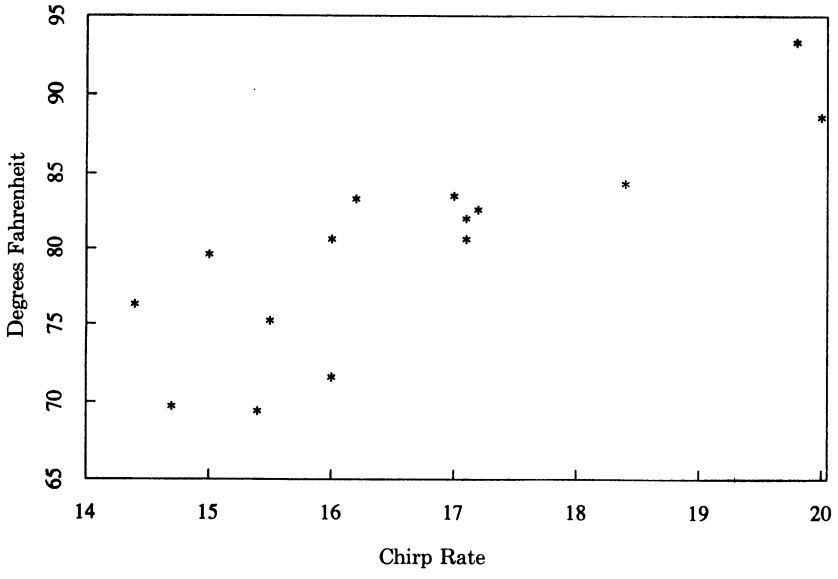

FIG. 7. *Temperature vs. chirp rate.*

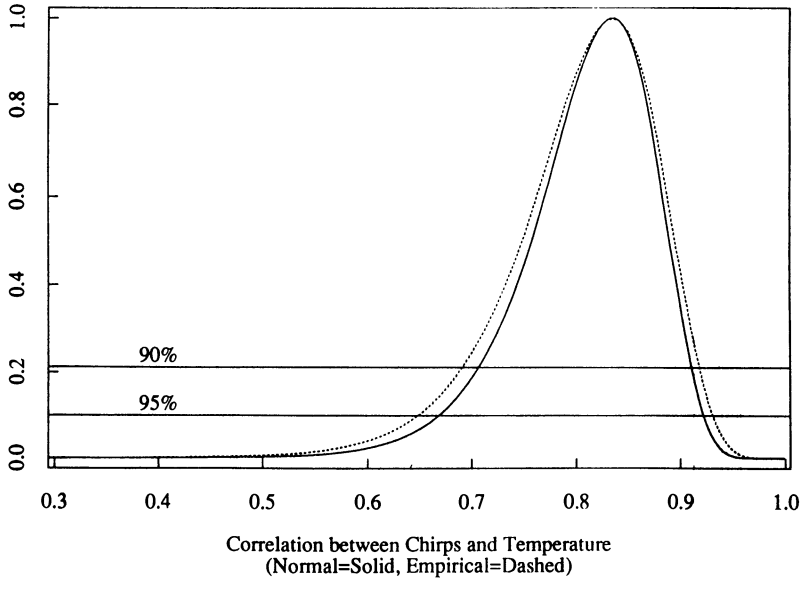

FIG. 8. *Profiled likelihood ratios.*

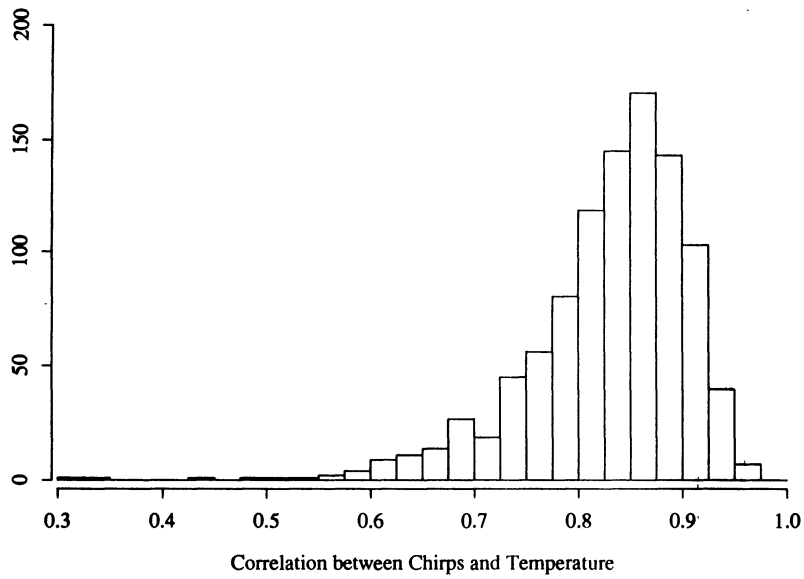

FIG. 9. *Bootstrap correlations.*

for vector m defines ν and θ implicitly as functionals of the distribution F of the vector Z. Consider θ to be fixed at a value to be tested. In the correlation example $Z = (X, Y)'$ and

$$m = \Big(X - \mu_x, Y - \mu_y, (X - \mu_x)^2 - \sigma_x^2,$$

$$(Y - \mu_y)^2 - \sigma_y^2, (X - \mu_x)(Y - \mu_y) - \rho\sigma_x\sigma_y\Big) .$$

The parameter vectors are $\theta = (\rho)$ and $\nu = (\mu_x, \mu_y, \sigma_x, \sigma_y)'$. We wish to compute $\mathscr{R}(\theta) = \sup_\nu \mathscr{R}(\nu, \theta)$ in an obvious notation, using n i.i.d. observations Z_i. This may be done by maximizing over ν, the minimum over λ of

$$-\frac{1}{n}\sum_{i=1}^{n} \log(1 + \lambda' m(Z_i, \theta, \nu)).$$

It simplifies the notation to use m to denote $m(Z_i, \theta, \nu)$ and to use superscripts to denote components of λ and m. Indices that are repeated in a term are understood to be summed over. Thus the inner optimization results in

$$0 = \frac{1}{n}\sum \frac{m^j}{1 + \lambda^k m^k} \tag{6.1}$$

and this defines $\lambda = \lambda(\nu)$ implicitly, provided that the sample variance of m has full rank. Derivatives with respect to ν^r are denoted by a subscript r.

We now seek to maximize

$$\frac{1}{n}\log \mathcal{R}(\nu, \theta) = Q(\nu) = -\frac{1}{n}\sum \log(1 + \lambda^j m^j) \tag{6.2}$$

over ν with θ held fixed and $\lambda = \lambda(\nu)$. We will show that the Hessian of Q is a negative definite matrix plus $O_p(n^{-1/2})$, which suggests that the outer optimization should be well behaved.

We restrict attention to sequences of θ and ν vectors that are within $O_p(n^{-1/2})$ of the true parameter values and for which the sample mean of m is $O_p(n^{-1/2})$. It follows as in Section 2 that $\|\lambda\| = O_p(n^{-1/2})$. We assume that first and second derivatives of m with respect to components of ν have finite expectations and that the fourth moment of m is finite. Recall from the proof of Theorem 1 that the probability that $\max_i |m^j \lambda^j| < 0.25$, say, approaches 1 as $n \to \infty$.

Differentiating (6.1) with respect to ν^r yields

$$0 = \frac{1}{n}\sum \frac{m_r^j(1 + \lambda^k m^k) - m^j\left(\lambda_r^k m^k + \lambda^k m_r^k\right)}{\left(1 + \lambda^k m^k\right)^2}$$

and, since $\|\lambda\| = O_p(n^{-1/2})$,

$$O_p(n^{-1/2}) = \frac{1}{n}\sum \frac{m_r^j - m^j m^k \lambda_r^k}{\left(1 + \lambda^k m^k\right)^2}. \tag{6.3}$$

Differentiating (6.2) with respect to ν^r yields

$$\begin{aligned} Q_r &= -\frac{1}{n}\sum \frac{\lambda_r^j m^j + \lambda^j m_r^j}{1 + \lambda^k m^k} \\ &= -\frac{1}{n}\sum \frac{\lambda^j m_r^j}{1 + \lambda^k m^k}, \end{aligned}$$

by (6.1). Differentiating again

$$\begin{aligned} Q_{rs} &= -\frac{1}{n}\sum \frac{\left(\lambda_s^j m_r^j + \lambda^j m_{rs}^j\right)(1 + \lambda^k m^k) - \lambda^j m_r^j\left(\lambda_s^k m^k + \lambda^k m_s^k\right)}{\left(1 + \lambda^k m^k\right)^2} \\ &= O_p(n^{-1/2}) - \frac{1}{n}\sum \frac{\lambda_s^j m_r^j}{\left(1 + \lambda^k m^k\right)^2}. \end{aligned} \tag{6.4}$$

We can express (6.3) and (6.4) in matrix form as

$$O_p(n^{-1/2}) = \nabla \tilde{m} - \tilde{S}\nabla\lambda \tag{6.5}$$

and

$$Q_{rs} = O_p(n^{-1/2}) - (\nabla\lambda)'(\nabla\tilde{m}), \tag{6.6}$$

respectively, where

$$\tilde{S} = \frac{1}{n}\sum \frac{m(Z_i, \theta, \nu)m(Z_i, \theta, \nu)'}{\bigl(1 + \lambda(\nu)'m(Z_i, \theta, \nu)\bigr)^2},$$

the matrix $\nabla\tilde{m}$ has jr element

$$\frac{1}{n}\sum \frac{m_r^j(Z_i, \theta, \nu)}{\bigl(1 + \lambda(\nu)'m(Z_i, \theta, \nu)\bigr)^2}$$

and $\nabla\lambda$ denotes the matrix of partial derivatives of λ with respect to the components of ν.

The matrix $\tilde{S}$ is close to S, the sample variance–covariance matrix for m because $\lambda'm$ is uniformly bounded with high probability. Similarly $\nabla\tilde{m}$ is nearly the sample mean of ∇m, the matrix of partial derivatives of components of m with respect to those of ν. Combining (6.5) and (6.6), we express the Hessian of Q in vector notation as

$$O_p(n^{-1/2}) - (\nabla\tilde{m})'\tilde{S}^{-1}(\nabla\tilde{m}).$$

It follows that the Hessian of Q is a negative definite matrix of $O_p(1)$ plus a term of $O(n^{-1/2})$, provided that m has a variance of full rank and that the expected value of ∇m is of full rank. Since $\log \mathcal{R}(\nu, \theta) = nQ$, the matrix nQ_{rs} may be thought of as the information matrix for ν when estimated with known θ.

6.4. *M estimates and regression.* Owen (1988a) obtains a one variable version of Theorem 3. Applying it to the median results in a family of confidence intervals that reproduce those generated by the sign test. One potential application of Theorem 3 is to pairs of medians. The confidence regions for a pair of medians will not always be convex. As an extreme case, suppose that the sample is concentrated in $L = \{(x, y) | x \geq 0, y \geq 0, xy = 0\}$. Then any reweighted sample also has its marginal median vector in L, and so the empirical likelihood regions will be subsets of L and need not always be convex. Any resampled pair of medians would also lie in L, so bootstrap intervals might also be subsets of L. Noticing this and constructing appropriate confidence sets would require some special attention from the bootstrapper, while with empirical likelihood it is automatic.

Multiple linear regression may be treated by either Theorem 2 or Theorem 3. Suppose that $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ and that an i.i.d. sample of (X_i, Y_i) is available from a distribution on $\mathbb{R}^{p+1}$. Then the sample regression coefficients are continuous functions of certain means of responses, squares and cross-products and are subject to Theorem 2. Using Theorem 3, one can postulate that for some X_i, $\beta \in \mathbb{R}^p$ the vectors $(Y_i - X_i'\beta)X_i$ are independent and identically distributed with mean 0 and finite variance of full rank, and so for various β test whether the residuals might reasonably have mean 0 and be uncorrelated with the predictors.

When the X_i are fixed by design neither approach applies directly, because both require i.i.d. observations. In a forthcoming paper the author will consider this case. The method proceeds by replacing the appeal to an i.i.d. central limit theorem by one based on the Lindeberg condition.

Acknowledgments. I would like to thank Nils Hjort, Michael Steele, Bradley Efron and Tom DiCiccio for helpful comments. I also thank all those, including one referee, who brought references to my attention.

REFERENCES

BAILEY, K. R. (1984). Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model. *Ann. Statist.* **12** 730–736.

BERK, R. H. and JONES, D. H. (1979). Goodness-of-fit statistics that dominate the Kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* **47** 47–59.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

DICICCIO, T. J., HALL, P. J. and ROMANO, J. (1988). Bartlett adjustment for empirical likelihood. Technical Report No. 298, Dept. Statistics, Stanford Univ.

DICICCIO, T. J. and ROMANO, J. (1988). Nonparametric confidence limits by resampling methods and least favorable families. Technical Report No. 295, Dept. Statistics, Stanford Univ.

DICICCIO, T. J. and TIBSHIRANI, R. J. (1986). Approximating the profile likelihood through Stein's least favourable family. Technical Report No. 1, Dept. Statistics, Univ. Toronto.

DONOHO, D. L. (1982). Breakdown properties of multivariate location estimators. Qualifying paper, Harvard Univ.

EFRON, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.* **9** 139–172.

GAENSSLER, P. (1983). *Empirical Processes*. IMS, Hayward, Calif.

HALL, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika* **74** 481–493.

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–401.

JOHANSEN, S. (1978). The product limit estimator as maximum likelihood estimator. *Scand. J. Statist.* **5** 195–199.

JOHNSON, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *J. Amer. Statist. Assoc.* **73** 536–544.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

LARSEN, R. J. and MARX, M. L. (1986). *An Introduction to Mathematical Statistics and its Applications*. Prentice-Hall, Englewood Cliffs, N.J.

MILLER, R. G. (1986). *Beyond ANOVA, Basics of Applied Statistics*. Wiley, New York.

OWEN, A. B. (1985). Nonparametric likelihood ratio confidence intervals. Technical Report LCS 6, Dept. Statistics, Stanford Univ.

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

OWEN, A. B. (1988b). Small sample central confidence intervals for the mean. Technical Report 302, Dept. Statistics, Stanford Univ.

PSHENICHNY, B. N. and DANILIN, YU. M. (1978). *Numerical Methods in Extremal Problems*. Mir, Moscow.

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

RHEINBOLDT, W. C. (1974). *Methods for Solving Systems of Nonlinear Equations* **14**. SIAM, Philadelphia, Penn.

ROYDEN, H. L. (1968). *Real Analysis*. MacMillan, New York.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134.

STAHEL, W. A. (1981). Robuste Schatzungen: Infinitesimale Optimalitat und Schatzungen von Kovarianzmatrizen. Ph.D. Thesis, Technical Univ., Graz, Austria.

THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871.

TSIATIS, A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108.

TUSNADY, G. (1977). On asymptotically optimal tests. *Ann. Statist.* **5** 385–393.

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178–203.

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9** 60–62.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305