

Empirical likelihood

This chapter develops empirical likelihood inference through a nonparametric likelihood ratio function. The result is an approach using a parametric family that is a multinomial distribution on all n observed data values. The focus is on setting confidence intervals for the mean of a scalar random variable. Later chapters extend the approach to other tasks.

2.1 Nonparametric maximum likelihood

We begin by defining the empirical cumulative distribution function, and showing that it is a nonparametric maximum likelihood estimate (NPMLE).

For a random variable $X \in \mathbb{R}$, the cumulative distribution function (CDF) is the function $F(x) = \Pr(X \leq x)$, for $-\infty < x < \infty$. We use $F(x-)$ to denote $\Pr(X < x)$ and so $\Pr(X = x) = F(x) - F(x-)$. The notation $1_{A(x)}$ represents the value 1 if the assertion $A(x)$ is true, and 0 otherwise.

Definition 2.1 Let $X_1, \dots, X_n \in \mathbb{R}$. The empirical cumulative distribution function (ECDF) of X_1, \dots, X_n is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq x}$$

for $-\infty < x < \infty$.

Definition 2.2 Given $X_1, \dots, X_n \in \mathbb{R}$, assumed independent with common CDF F_0 , the nonparametric likelihood of the CDF F is

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

Definition 2.2 reflects a very literal interpretation of the notion of likelihood. The value $L(F)$ is the probability of getting exactly the observed sample values X_1, \dots, X_n from the CDF F . One consequence is that $L(F) = 0$ if F is a continuous distribution. To have a positive nonparametric likelihood, a distribution F must place positive probability on every one of the observed data values.

Theorem 2.1 proves that the nonparametric likelihood is maximized by the ECDF. Thus the ECDF is the NPMLE of F .

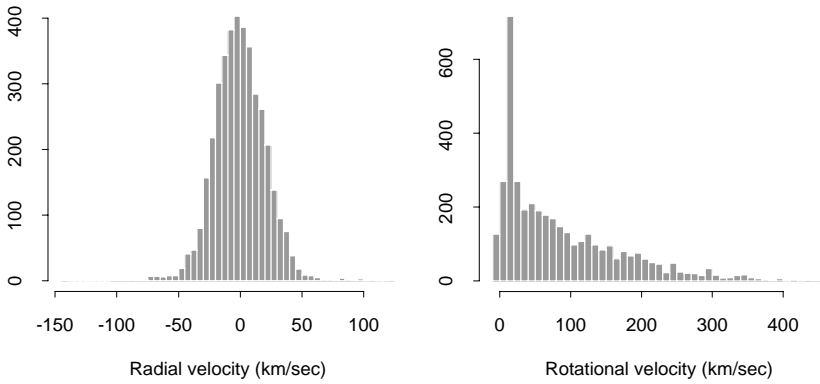


Figure 2.1 These histograms display the velocities of 3932 stars, from Hoffleit & Warren (1991), as described in the text.

Theorem 2.1 Let $X_1, \dots, X_n \in \mathbb{R}$ be independent random variables with a common CDF F_0 . Let F_n be their ECDF and let F be any CDF. If $F \neq F_n$, then $L(F) < L(F_n)$.

Proof. Let $z_1 < z_2 < \dots < z_m$ be the distinct values in $\{X_1, \dots, X_n\}$, and let $n_j \geq 1$ be the number of X_i that are equal to z_j . Let $p_j = F(z_j) - F(z_{j-1})$ and put $\hat{p}_j = n_j/n$. If $p_j = 0$ for any $j = 1, \dots, m$, then $L(F) = 0 < L(F_n)$, so we suppose that all $p_j > 0$, and that for at least one j , $p_j \neq \hat{p}_j$. Now $\log(x) \leq x - 1$ for all $x > 0$ with equality only when $x = 1$. Therefore

$$\begin{aligned} \log \left(\frac{L(F)}{L(F_n)} \right) &= \sum_{j=1}^m n_j \log \left(\frac{p_j}{\hat{p}_j} \right) \\ &= n \sum_{j=1}^m \hat{p}_j \log \left(\frac{p_j}{\hat{p}_j} \right) \\ &< n \sum_{j=1}^m \hat{p}_j \left(\frac{p_j}{\hat{p}_j} - 1 \right) \\ &\leq 0, \end{aligned}$$

and so $L(F) < L(F_n)$. \square

Figure 2.1 shows histograms of the radial and rotational velocities of some stars from the bright star catalogue. Stars rotate around the center of our galaxy, with a velocity that depends in part, upon their distance from the center. The radial velocity of a star is the speed with which it appears to be moving away from us, with negative values for stars getting closer. The rotational velocity of a star is its velocity, perpendicular to the line connecting it to the sun.

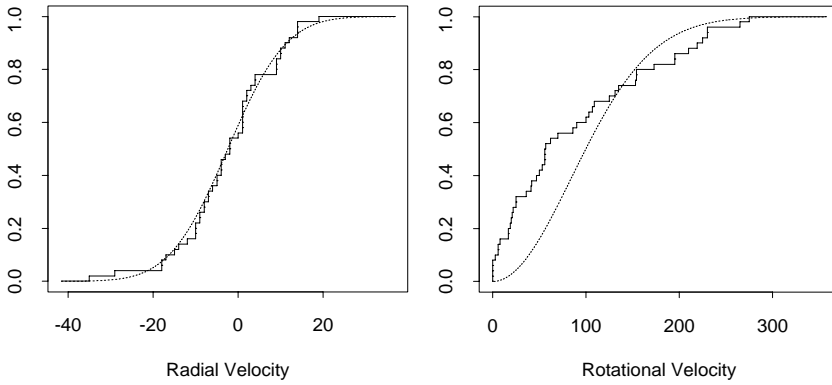


Figure 2.2 Each plot compares an empirical (solid) and parametric (dashed) CDF for the velocities of 50 stars. Radial velocities are compared to a normal distribution on the left. Rotational velocities are compared to a scaled square root of a $\chi_{(2)}^2$ distribution on the right.

The step functions in Figure 2.2 shows the NPMLE $F_n(x)$, for velocities of the first 50 stars in the data set. This sample size reduction was made so that the bumpy nature of the NPMLE would be visually apparent. The left plot of Figure 2.2 shows a smooth curve based on the parametric model $X_i \sim N(\mu, \sigma^2)$, for radial velocities X_i . Under this model

$$F(x) = F(x; \mu, \sigma) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(z - \mu)^2}{2\sigma^2}\right) dz,$$

and the curves shown are $F(x; \hat{\mu}, \hat{\sigma})$ for parametric MLE's $\hat{\mu}$ and $\hat{\sigma}$.

It may be surprising to find that radial velocities are nearly normally distributed. This would happen if the velocities of the stars relative to the sun had a spherical Gaussian distribution. In that case the rotational velocities would be the square roots of scaled $\chi_{(2)}^2$ distributions. The right plot in Figure 2.2 shows that such a model fits the data poorly. The parametric CDF in that plot is that of the square root of a $\chi_{(2)}^2$ distribution scaled to have mean equal to the sample mean squared rotational velocity.

In parametric models, when $\hat{\eta}$ is the MLE of η , and we are interested in η through some function $\theta(\eta)$, the MLE of θ is $\hat{\theta} = \theta(\hat{\eta})$. In the nonparametric setting, we suppose that we are interested in F through $\theta = T(F)$, where T is a real-valued function of distributions. The true unknown parameter is $\theta_0 = T(F_0)$. Proceeding by analogy, we take the NPMLE of θ to be $\hat{\theta} = T(F_n)$. Thus if we are interested in the mean $\theta_0 = \int x dF_0(x)$ of X when X has the distribution F_0 , then by analogy, the NPMLE of θ_0 is the mean of F_n . This mean is of course

$\bar{X} = (1/n) \sum_{i=1}^n X_i$. For a subset $A \subset \mathbb{R}$, the NPMLE of $\Pr(X \in A)$ is the sample fraction of X_i in A .

For the radial velocities, the parametric MLE leads to an estimate of 0.589 for $\Pr(X < 0)$, while the nonparametric one gives an estimate of 0.560, in close agreement. Both parametric and nonparametric MLE's estimate $E(X)$ as -2.42 . It is visually apparent that the parametric and nonparametric MLE's of tail probabilities for the rotational velocities differ sharply.

Either the parametric or nonparametric MLE can be best, depending on our goals and some assumptions on the data. If we are interested primarily in the probability that $X < x_0$ then the parametric MLE is likely to be best when the true distribution is close to the parametric distribution. If the underlying distribution does not follow the parametric one, then the NPMLE will ordinarily be better, at least for large enough n .

For this data set, the parametric model gives a reasonable fit for the radial velocities, but not for the rotational ones. Empirical CDFs are not very good at showing differences in the tails of a distribution. A QQ plot of all 3932 radial velocities shows a nearly normal distribution, but with heavier than normal tails.

2.2 Nonparametric likelihood ratios

In parametric inference we may base hypothesis tests and confidence regions on the likelihood ratio. If $L(\eta)$ is much smaller than $L(\hat{\eta})$, then we reject the hypothesis that $\eta_0 = \eta$, and exclude η from our confidence region for η_0 . Wilks's theorem provides that $-2 \log(L(\eta_0)/L(\hat{\eta}))$ tends to a chisquared distribution as $n \rightarrow \infty$, under mild regularity conditions, allowing us to decide just how small $L(\eta)$ must be in order for η to get rejected. The degrees of freedom in the chisquared distribution are usually equal to the dimension of the set of η values. When we want a confidence region for θ we take the image of a confidence region for η . That is

$$\{\theta(\eta) \mid L(\eta) \geq cL(\hat{\eta})\},$$

where the threshold c is chosen using Wilks's theorem, with degrees of freedom equal to the dimension of the set of θ values.

We may also use ratios of the nonparametric likelihood as a basis for hypothesis tests and confidence intervals. For a distribution F , define

$$R(F) = \frac{L(F)}{L(F_n)},$$

through the nonparametric likelihood $L(F)$ of [Definition 2.2](#). We proceed by analogy with parametric likelihood. Suppose that we are interested in a parameter $\theta = T(F)$ for some function T of distributions. This F is a member of a set \mathcal{F} of distributions. In some cases we may take \mathcal{F} to be the set of all distributions on \mathbb{R} . More often, we use a smaller set of distributions. Define the profile likelihood

ratio function:

$$\mathcal{R}(\theta) = \sup \{R(F) \mid T(F) = \theta, F \in \mathcal{F}\}. \quad (2.1)$$

Empirical likelihood hypothesis tests reject $H_0 : T(F_0) = \theta_0$, when $\mathcal{R}(\theta_0) < r_0$ for some threshold value r_0 . Empirical likelihood confidence regions are of the form

$$\{\theta \mid \mathcal{R}(\theta) \geq r_0\}. \quad (2.2)$$

In many settings, the threshold r_0 may be chosen using an empirical likelihood theorem (ELT), a nonparametric analogue of Wilks's theorem.

2.3 Ties in the data

If $X_i = X_j$ for $i \neq j$, we say that X_i and X_j are tied. Let us first consider data having no ties. If the distribution F places probability $p_i \geq 0$ on the value $X_i \in \mathbb{R}$, then $\sum_{i=1}^n p_i \leq 1$, and $L(F) = \prod_{i=1}^n p_i$ and so

$$R(F) = \frac{L(F)}{L(\hat{F}_n)} = \prod_{i=1}^n np_i. \quad (2.3)$$

For data possibly containing some ties, suppose that the distinct value z_j arises $n_j \geq 1$ times in the sample, and has probability p_j under F . Let k be the number of distinct values in the data. Then instead of (2.3) we find

$$R(F) = \prod_{j=1}^k \left(\frac{p_j}{\hat{p}_j} \right)^{n_j} = \prod_{j=1}^k \left(\frac{np_j}{n_j} \right)^{n_j}. \quad (2.4)$$

The theory of empirical likelihood is much simpler through equation (2.3) than equation (2.4). The computation can also be simpler with equation (2.3), though when the number of ties is enormous, so that $k \ll n$, equation (2.4) might lead to faster algorithms. Fortunately, we have the choice. If we use (2.3) instead of the true likelihood ratio (2.4) we get the same profile likelihood ratio function $\mathcal{R}(\theta)$. This holds for any family \mathcal{F} of distributions and for whatever function $T(F)$ is used to define θ .

To see this, we may apportion the probabilities p_j for a distribution F into observation specific weights $w_i \geq 0$ for $i = 1, \dots, n$. Choose the weights so that p_j is the sum of w_i over all i with $X_i = z_j$. Then a distribution putting weight w_i on observation X_i reproduces F , and hence any $T(F)$.

We define the likelihood of F in terms of these weights as $\prod_{i=1}^n w_i$. When there are ties, this likelihood value is not unique. The value θ enters a confidence region if and only if for some F having $T(F) = \theta$, the largest value of $\prod_{i=1}^n w_i$ exceeds a threshold. So we only need to consider the maximum of $\prod_{i=1}^n w_i$ over weights generating the p_j . This maximum arises when $w_i = p_{j(i)}/n_{j(i)}$, with $j(i)$ determined by $X_i = z_{j(i)}$.

The maximum of $\prod_{i=1}^n w_i$ for a given F is

$$\prod_{j=1}^k \left(\frac{p_j}{n_j} \right)^{n_j} = L(F) \times \prod_{j=1}^k n_j^{-n_j}.$$

When we use nonparametric likelihoods through ratios such as $L(F)/L(F_n)$ the factor $\prod_{j=1}^k n_j^{-n_j}$ cancels. Thus we may proceed computationally and theoretically as if there were no ties, writing

$$R(F) = \prod_{i=1}^n n w_i, \tag{2.5}$$

where $w_i \geq 0$, $\sum_{i=1}^n w_i \leq 1$, and F puts probability $\sum_{j: X_i = X_j} w_j$ on X_i . Equations (2.5) and (2.3) are, of course, equivalent.

Equation (2.5) describes a function on the n -dimensional set of weights

$$\mathbb{S}_{n-1} = \left\{ (w_1, \dots, w_n) \mid w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}. \tag{2.6}$$

The set \mathbb{S}_{n-1} is called the probability simplex, or simply the simplex. Because $w_1 + \dots + w_n = 1$, the weight set is actually $n - 1$ dimensional and so the subscript is $n - 1$ not n . For $n = 3$ the allowable points (w_1, w_2, w_3) are interior to the equilateral triangle with corners at $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. Figure 2.3 shows contours of $R(F)$ within this triangle. Empirical likelihood confidence regions are usually constructed as the image under a statistical function $T(F)$, of the interior of an $n - 1$ dimensional contour of $R(F)$.

There is nothing special about one-dimensional data in the arguments above. Ties can be ignored for $X_i \in \mathbb{R}^d$, for any $d \geq 1$. In settings more complicated than n IID observations, where we wish to prove that ties can be ignored, we return to this approach of putting probabilities p_j on the distinct observed values and weights w_i on the data points.

It is intuitively reasonable that we should ignore ties. Suppose that we generate tie-breaker random variables $U_i \sim U(0, 1)$ independently of each other and of the X_i . Now form the observations $(X_i, U_i) \in \mathbb{R}^{d+1}$. Because the U_i have a continuous distribution, there will be no ties among the U_i , and hence none among the (X_i, U_i) . Now consider a function T on the distribution of (X_i, U_i) pairs, where T completely ignores the U values. Because empirical likelihood ignores ties, we get the same confidence regions for T on the (X_i, U_i) pairs as we do on the X_i data alone. Any other answer would be unreasonable. We know that the U_i contain no information and so their presence should not change our answer.

2.4 Multinomial on the sample

A natural starting point for nonparametric inference is the mean of a scalar random variable, which we take up here. Developing empirical likelihood confi-

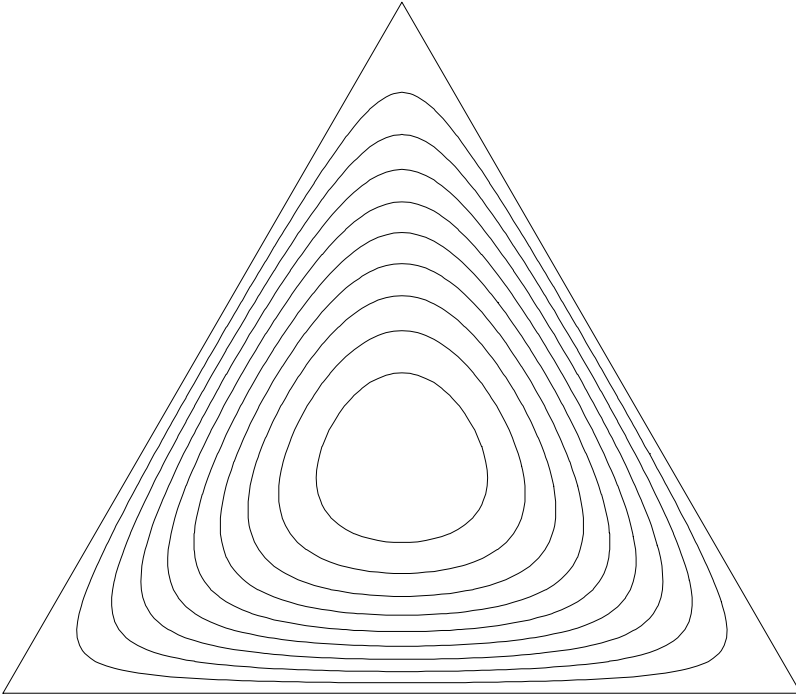


Figure 2.3 Shown are contours of the empirical likelihood ratio function $R(F)$ for the case of $n = 3$ observations. The likelihood ratio values plotted are 0.1 to 0.9 by steps of 0.1. The bounding triangle has $R(F) = 0$, and the maximum value of $R(F)$ is 1.0 at the center of the triangle.

dence intervals by analogy with parametric likelihood methods gives a degenerate confidence interval for the mean. To see this, consider the distribution $F = (1 - \varepsilon)F_n + \varepsilon\delta_x$ where δ_x is a distribution taking the value x with probability one, and x is not one of the observed X_i values. The likelihood ratio for this F is

$$R(F) = \frac{\prod_{i=1}^n (1 - \varepsilon)/n}{\prod_{i=1}^n 1/n} = (1 - \varepsilon)^n,$$

and the mean is $(1 - \varepsilon)\bar{X} + \varepsilon x$. For any threshold $r_0 < 1$, taking a small enough ε makes $R(F) > r_0$. Then sending x to $\pm\infty$ causes the empirical likelihood ratio confidence interval for the mean to have infinite length.

We can eliminate this problem by changing the set \mathcal{F} of candidate distributions. If X is known to be a bounded random variable, with $-\infty < A \leq X \leq B < \infty$ then by taking \mathcal{F} to be the set of all distributions for which X satisfies these

bounds, a nondegenerate confidence interval results. The practical difficulty is that even if we know that X is bounded, we might not know a good bound to use in practice. For example, we might be convinced that the height of a human being is a bounded random variable, yet it might not be easy to specify an upper bound to use in practice.

If we are sampling a bounded random variable, then the sample minimum $A_n = \min_{1 \leq i \leq n} X_i$ and maximum $B_n = \max_{1 \leq i \leq n} X_i$ will approach the tightest possible bounds A and B , as n increases. We may obtain finite length confidence intervals by taking $\mathcal{F} = \mathcal{F}_n$ to be the set of distributions of random variables X for which $A_n \leq X \leq B_n$.

Now suppose that $F \in \mathcal{F}_n$ and $\int x dF(x) = \mu$. Let w_i be the weight that F places on observation X_i . When constructing the profile empirical likelihood function for the mean, we may suppose that $\sum_{i=1}^n w_i = 1$. If instead, we have $\sum_{i=1}^n w_i < 1$, then F puts probability $1 - \sum_{i=1}^n w_i > 0$ on the interval (A_n, B_n) exclusive of sample points there. This probability can be “reassigned” to data points in such a way that the new distribution \tilde{F} has the same mean as F but has $L(\tilde{F}) > L(F)$. This reassignment can, for example, be done by increasing the weights w_i for the largest and smallest sample values. The result is that we get the same profile empirical likelihood ratio function for the mean by taking \mathcal{F}_n to be the distributions with $\sum_{i=1}^n w_i = 1$ as we do taking \mathcal{F}_n to be all distributions over the interval $[A_n, B_n]$.

Using the distributions with $\sum_{i=1}^n w_i = 1$, we may write the profile empirical likelihood ratio function for the mean as

$$\mathcal{R}(\mu) = \max \left\{ \prod_{i=1}^n n w_i \mid \sum_{i=1}^n w_i X_i = \mu, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}$$

and the resulting empirical likelihood confidence region for the mean as

$$\{\mu \mid \mathcal{R}(\mu) \geq r_0\} = \left\{ \sum_{i=1}^n w_i X_i \mid \prod_{i=1}^n n w_i \geq r_0, w_i \geq 0, \sum_{i=1}^n w_i = 1 \right\}.$$

This region is an interval, as shown in Chapter 2.5.

Empirical likelihood inferences for the mean may be recognized as parametric likelihood inferences for the mean, using a data-determined parametric family. The parametric family involved is the multinomial distribution on the observed values of X_i .

For continuous F , this parametric family will have n parameters w_1, \dots, w_n . Since they sum to 1, we can reduce this to $n - 1$ parameters. Most asymptotic results for parametric likelihood are framed in a setting where there are a finite number of parameters and a sample size n tending to infinity. When the number of parameters grows with n , the parametric MLE might not even approach the true parameter value. By having the number of parameters grow as quickly as the sample size, empirical likelihood appears to be very different from parametric likelihoods.

When F is discrete, the empirical likelihood ratio function is that of a multinomial, but on the observed values only. If there are a finite number of possible values of X , then as n increases, eventually all the distinct values have been seen at least once, and empirical likelihood reduces to a parametric likelihood. If F is a discrete distribution for which infinitely many values have positive probability, then the empirical likelihood will be based on a random data-determined multinomial with an ever-increasing number of parameters.

Simply requiring a modestly large likelihood ratio forces all but a vanishingly small amount of the probability to be placed on the sample. [Lemma 2.1](#) quantifies this effect.

Lemma 2.1 *Suppose that distribution F places probability $p_0 = 1 - \sum_{i=1}^n w_i$ on the set $\mathbb{R} - \{x_1, \dots, x_n\}$, and that $\mathcal{R}(F) \geq r_0 > 0$. Then $p_0 \leq (1/n) \log(1/r_0)$.*

Proof. The largest possible value for $\mathcal{R}(F)$, under the problem constraints, arises with all weights equal to $(1 - p_0)/n$. Thus $r_0 \leq \mathcal{R}(F) \leq (1 - p_0)^n$, from which

$$\begin{aligned} p_0 &\leq 1 - r_0^{1/n} \\ &= 1 - \exp(n^{-1} \log r_0) \\ &\leq 1 - (1 + n^{-1} \log r_0) \\ &= \frac{1}{n} \log \left(\frac{1}{r_0} \right). \end{aligned}$$

□

Anticipating that a 95% confidence interval corresponds to $-2 \log(r_0)$ close to $\chi_{(1)}^{2, .95} = 3.84$, we consider $\log(1/r_0)/n = 1.92/n$. To contribute a point to the 95% confidence interval, a distribution has to put more than $1 - 1.92/n$ probability on the sample. Our restriction to distributions that reweight the data might push about $2/n$ more probability onto the sample than would otherwise have been there. Thus the empirical profile likelihood ratio function itself does most of the work in reducing the class of functions to those supported on the sample.

We have seen that the restriction to distributions that reweight the sample only changes $O(1/n)$ of the probability. This probability is small because confidence regions typically have diameter of order $n^{-1/2}$. But changes to $O(1/n)$ of the probability of F can have arbitrarily large effects on nonrobust statistics like the mean, so some clipping of the range of F is necessary. Clipping to the sample is perhaps the simplest choice.

In most settings, empirical likelihood is a multinomial likelihood on the sample. There are some exceptions, such as those where boundedness arises naturally in the structure of the problem and need not be imposed. For example, when we are interested in the mean of a bounded function of X such as $1_{X \geq 0}$ or $\sin(X)$, then \mathcal{F} can be the set of all distributions on \mathbb{R} . Similarly, inferences for the median of X do not require us to restrict the family of distributions.

2.5 EL for a univariate mean

Nondegenerate intervals still need to be calibrated, so that we can approximate a desired level of confidence such as 95%. The following univariate empirical likelihood theorem (ELT) is the basis for such calibration.

Theorem 2.2 (Univariate ELT) *Let X_1, \dots, X_n be independent random variables with common distribution F_0 . Let $\mu_0 = E(X_i)$, and suppose that $0 < \text{Var}(X_i) < \infty$. Then $-2 \log(\mathcal{R}(\mu_0))$ converges in distribution to $\chi_{(1)}^2$ as $n \rightarrow \infty$.*

Proof. See Exercises 2.4 and 2.5 for a sketch, Chapter 11.2 for a proof. \square

Two features of Theorem 2.2 are noteworthy. First, the chisquared limit is the same as we typically find for parametric likelihood models with one parameter. Second, there is no assumption that X_i are bounded random variables. They need only have a bounded variance, which constrains how fast the sample maximum and minimum can grow as n increases.

Theorem 2.2 provides an asymptotic justification for tests that reject the value μ_0 at the α level, when $-2 \log \mathcal{R}(\mu_0) > \chi_{(1)}^{2,1-\alpha}$. The unrejected values of μ_0 form a $100(1 - \alpha)\%$ confidence region, with the same asymptotic justification. Details of the proof and some simulations both suggest that the $\chi_{(1)}^{2,1-\alpha}$ threshold should perhaps be replaced by $F_{1,n-1}^{1-\alpha}$. The $F_{1,n-1}$ distribution is the square of a $t_{(n-1)}$ distribution while the $\chi_{(1)}^2$ distribution is the square of a $N(0, 1)$ distribution. As $n \rightarrow \infty$, we have $t_{(n-1)} \rightarrow N(0, 1)$ in distribution, and so $F_{1,n-1}^{1-\alpha} - \chi_{(1)}^{2,1-\alpha} \rightarrow 0$. Thus, as n increases, the difference between the two calibrations disappears. The F calibration usually gives better results in simulations.

It is easy to see that the empirical likelihood confidence region for a mean is always an interval. If μ_1 and μ_2 are in the confidence region, then there are weights $w_{ij} \geq 0$, $i = 1, \dots, n$, $j = 1, 2$, with $\sum_{i=1}^n w_{ij} X_i = \mu_j$ and $\sum_{i=1}^n w_{ij} = 1$, and $-2 \sum_{i=1}^n \log(nw_{ij}) \leq \chi_{(1)}^{2,1-\alpha}$. Now for $0 < \tau < 1$, let $\mu_\tau = \mu_1 \tau + \mu_2(1 - \tau)$. The nonnegative weights $w_i = w_{i1} \tau + w_{i2}(1 - \tau)$ sum to 1, satisfy $\sum_{i=1}^n w_i X_i = \mu_\tau$, and

$$\begin{aligned} -2 \sum_{i=1}^n \log(nw_i) &= -2 \sum_{i=1}^n \log(\tau n w_{i1} + (1 - \tau) n w_{i2}) \\ &\leq -2 \left[\tau \sum_{i=1}^n \log(nw_{i1}) + (1 - \tau) \sum_{i=1}^n \log(nw_{i2}) \right] \\ &\leq \chi_{(1)}^{2,1-\alpha}. \end{aligned}$$

It follows that the empirical likelihood confidence region for the mean contains the line segment connecting any two of its points, and so it is an interval.

Figure 2.4 shows 29 determinations of the mean density of the earth, relative to water. These were made by Cavendish in 1798 and appear in Stigler (1977). The mean of Cavendish's values is 5.420, somewhat below the presently accepted

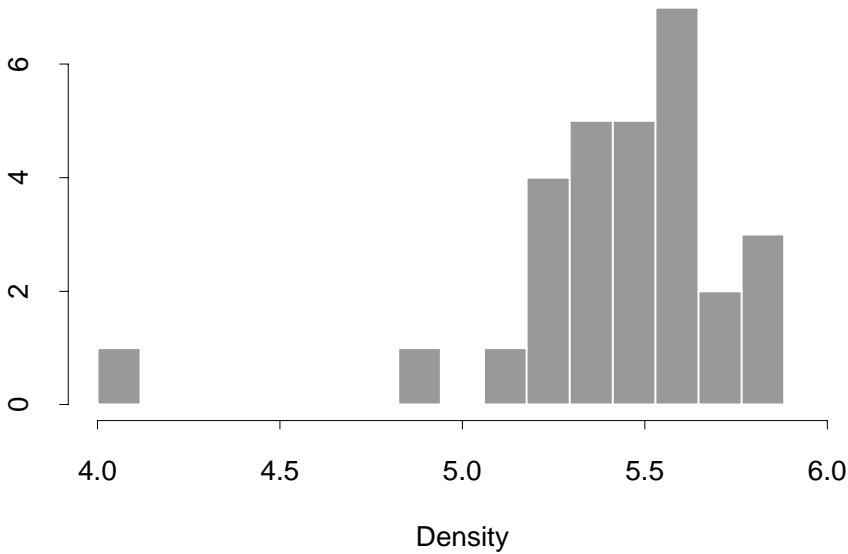


Figure 2.4 Shown are 29 of Cavendish's measurements of the mean density of Earth, relative to water. Source: Stigler (1977).

value of 5.517. Figure 2.5 shows the empirical likelihood ratio function for the mean of these data, with the modern value for the density of the earth marked. The modern value lies just inside the 95% empirical likelihood confidence interval, which extends from 5.256 to 5.521.

2.6 Coverage accuracy

A $100(1-\alpha)\%$ empirical likelihood confidence interval is formed by taking those values μ for which $-2 \log \mathcal{R}(\mu) \leq \chi_{(1)}^{2,1-\alpha}$, that is $\mathcal{R}(\mu) \geq \exp(-\chi_{(1)}^{2,1-\alpha}/2)$. The probability that μ_0 is in this interval approaches the nominal value $1 - \alpha$ as $n \rightarrow \infty$. That is, the coverage error

$$\Pr(-2 \log \mathcal{R}(\mu_0) \leq \chi_{(1)}^{2,1-\alpha}) - (1 - \alpha) \rightarrow 0$$

as $n \rightarrow \infty$. The mathematical analysis of coverage error is presented in some works described in the bibliographic notes at the end of this chapter. This section outlines the findings of those works.

Ideally, a confidence interval should have exactly the coverage $1 - \alpha$ for any n and any sampling distribution F_0 . As discussed in Chapter 2.11, no exact nonparametric confidence intervals exist for the sample mean. As a result nonparametric confidence intervals are asymptotic confidence intervals, as indeed are most para-

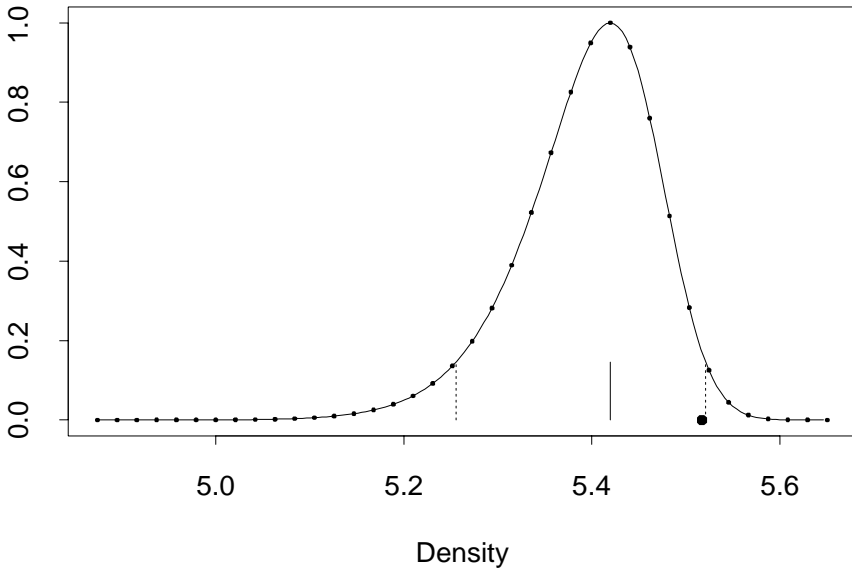


Figure 2.5 The solid curve shows the empirical likelihood ratio function for the mean of Cavendish's measurements of the density of the earth, relative to water. The modern value of 5.517 is shown as a solid reference point. Two short dotted reference bars delimit the 95% interval and a solid bar shows the sample mean. The points where empirical likelihood was computed are shown as small solid circles connected by interpolation as described in Chapter 2.9.

metric confidence intervals. The convention used here is that any confidence statement is an asymptotic one, unless explicitly stated otherwise.

Under mild moment conditions, the coverage error for empirical likelihood confidence intervals decreases to zero at the rate $1/n$ as $n \rightarrow \infty$. This is the same rate that typically holds for confidence intervals based on parametric likelihoods, the jackknife, and the simpler bootstrap methods. Even the standard confidence intervals, meaning $\bar{X} \pm Z^{1-\alpha/2}s$ where $s^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ and $\Pr(N(0,1) \leq Z^{1-\alpha}) = 1 - \alpha$, have this rate of coverage error. The coverage error in a nonparametric confidence interval for a univariate mean typically takes the form

$$\frac{1}{n} \varphi \left(Z^{1-\alpha/2} \right) \exp [A + B\gamma^2 + C\kappa] + O \left(\frac{1}{n^2} \right)$$

where φ is the standard normal density function. The quantities γ and κ are the skewness and kurtosis of X , introduced in Chapter 1.1. The constants A, B, C differ for the various confidence interval methods. By $n = 20$, this formula usually predicts the coverage error of a 95% confidence interval for the mean, to within 1

percentage point of its actual coverage error. The exceptions arise for very heavy tailed distributions, where the theory tends to greatly underestimate the coverage level.

For parametric likelihood intervals, the coverage error is typically $O(1/n)$ if the model is true, though it need not converge to 0 as n increases, if the model is not true. A Bartlett correction can be used to reduce parametric coverage errors to $O(n^{-2})$. A Bartlett correction also applies for empirical likelihood. Jackknife, bootstrap, and standard intervals are not Bartlett correctable.

2.7 One-sided coverage levels

For some applications a one-sided confidence interval, corresponding to a one-tailed test, is desired. We may know that $\mu \geq \mu'$ and so be interested in testing $\mu = \mu'$ versus $\mu > \mu'$. Or, when the consequences of small μ are benign while those of large μ are serious we might want a one-sided confidence interval of the form $(-\infty, U)$ for μ .

If (L, U) is a two-sided $100(1 - \alpha)\%$ empirical likelihood confidence interval for μ , then we might consider using $(-\infty, U)$ and (L, ∞) as one-sided $100(1 - \alpha/2)\%$ confidence intervals for the mean. The coverage error in these one-sided intervals decreases to zero, but only at the relatively slow rate $n^{-1/2}$ as $n \rightarrow \infty$. Chapter 13 presents methods to modify empirical likelihood, to achieve $O(n^{-1})$ coverage errors for one-sided confidence intervals. The modifications result in shifts of size Δ_L and Δ_R at the left and right endpoints of the interval, respectively, where Δ_L and Δ_R are data determined and equal or very nearly equal to each other.

Confidence intervals based on thresholding a parametric likelihood also typically have $O(n^{-1})$ two-sided coverage errors and $O(n^{-1/2})$ one-sided errors. They also can be modified to have $O(n^{-1})$ coverage errors for one-sided inference.

Every point inside a likelihood interval, parametric or empirical, is deemed to be more likely than every point outside it. When such intervals are modified to equalize coverage errors in the two tails, the result is that some points inside the new interval have lower likelihood than some other points outside of it.

Thus there is a trade-off between separating more likely from less likely parameter values, and equalizing tail errors. We can achieve one goal with high accuracy as $n \rightarrow \infty$, but then the other is attained at a slower rate. The examples in this book use confidence regions based on thresholding the empirical likelihood without employing any of the tail area equalization methods described in Chapter 13.

In multi-parameter problems, the goal of equalizing coverage errors between the tails becomes more challenging. Then the shape of the confidence region is more complicated than just a left and right distance from $\hat{\mu}$. Noncoverage events can happen in a continuum of directions, not just two. There is a sense in which empirical likelihood confidence regions have the right shape in higher dimensional problems. There again a shift is required. We must shift the whole set of

contours by some vector Δ . See the discussion of pseudo-likelihood in Chapter 13.3.

2.8 Power and efficiency

It is very important to have approximately the right coverage in a confidence interval, or hypothesis test, for otherwise the resulting inferences are not reliable. But there is also a need for efficiency. If a test does not have good power, or a confidence interval is too long, then the data have not been fully utilized. Accuracy and efficiency trade off in confidence interval problems, just as bias and variance often trade off in parameter estimation problems. An interval with nearly the right coverage but highly variable length is not useful. In the extreme case, consider an interval that is equal to all of \mathbb{R} with probability 0.95 and has length 0 with probability 0.05. It has exactly the right coverage but the corresponding test has only 5% power against any alternative.

One way to assess the power of empirical likelihood is through the curvature of \mathcal{R} at the NPMLE \bar{X} . For large sample sizes, $\log(\mathcal{R}(\mu)) \doteq -n\sigma_0^{-2}(\mu - \bar{X})^2/2$ for μ near \bar{X} , where $\sigma_0^2 = \text{Var}(X_i)$. The greater the (absolute) curvature in this quadratic, the shorter the confidence intervals for a given level of coverage, and hence the greater the power. It can be shown that

$$-2 \log \mathcal{R}(\mu_0 + \tau \sigma_0 n^{-1/2}) \rightarrow \chi_{(1)}^2(\tau^2), \quad (2.7)$$

in distribution, where τ^2 is a noncentrality parameter. This means that empirical likelihood inferences will have roughly the same power as parametric inferences, in a family with Fisher information equal to $1/\sigma_0^2$.

Asymptotic comparisons described in Chapter 3.17 show that, to first order, empirical likelihood has the same power as the bootstrap t against alternatives that are $O(n^{-1/2})$ distance from the null hypothesis. Surprisingly, empirical likelihood usually matches the power of parametric likelihood ratio tests to second order, as described in Chapter 3.17.

We might have expected good power properties for empirical likelihood, because likelihood ratio tests are known to be the most powerful tests in multinomial settings, with considerable generality regarding the hypothesis being tested, the alternative of interest, and the competing method under consideration. As Hoeffding (1965) writes:

If a given test of size α_n is “sufficiently different” from a likelihood ratio test then there is a likelihood ratio test of size $\leq \alpha_n$ that is considerably more powerful than the given test at “most” points in the set of alternatives when n is large enough, provided that $\alpha_n \rightarrow 0$ at a suitable rate.

The empirical likelihood setting is not as simple as a multinomial because the support set is random and may increase in cardinality with n . But a version of the universal power optimality of likelihood ratio tests has been established for empirical likelihood. These power results are of the large deviations kind, though they

do not necessarily require large sample sizes to be evident. They are described in Chapter 13.5.

Some simulation evidence exists to support these asymptotic results. Simulations can compare power of methods directly, or indirectly by measuring the length of confidence intervals. It has been observed empirically and theoretically that nonparametric confidence intervals tend to undercover, approaching their nominal coverage levels from below as $n \rightarrow \infty$. Simulations that ignore this phenomenon can assess coverage but are inconclusive regarding power and its trade-off with coverage. Comparisons of coverage alone favor methods with longer intervals (less power) while comparisons of interval length alone favor methods with more severe undercoverage.

Some simulations cited in Chapter 13.6 compare power after first doing a simulation to calibrate coverage levels. These found that empirical likelihood has better power than the other methods considered at most of the alternative hypotheses simulated. Another simulation, described in Chapter 2.11, compared methods by forcing them to use the same confidence interval length in each Monte Carlo sample. Empirical likelihood obtained competitive coverage whether it or the other method chose the interval length.

2.9 Computing EL for a univariate mean

Empirical likelihood inferences for the univariate mean require the following computational chores: To test whether $\mu = \mu_0$, we need to compute $\mathcal{R}(\mu_0)$. To set confidence limits for μ , we need to find the two values of μ that solve the equation $\mathcal{R}(\mu) = r_0$, given a threshold value r_0 . To plot the curve $\mathcal{R}(\mu)$, we need to compute $\mathcal{R}(\mu)$ at numerous points over the range of interest and then interpolate them. The computations are described at a high level, but with some practical details to ease the job of implementing them.

We begin by describing how to compute $\mathcal{R}(\mu)$. Let the ordered sample values be $X_{(1)} \leq \dots \leq X_{(n)}$. First we eliminate the trivial cases. If $\mu < X_{(1)}$ or $\mu > X_{(n)}$ then there are no weights $w_i \geq 0$ summing to 1 for which $\sum_i w_i X_i = \mu$. In such cases we take $\log \mathcal{R}(\mu) = -\infty$, and $\mathcal{R}(\mu) = 0$ by convention. Similarly if $\mu = X_{(1)} < X_{(n)}$ or $\mu = X_{(n)} > X_{(1)}$ we take $\mathcal{R}(\mu) = 0$, but if $X_{(1)} = X_{(n)} = \mu$, we take $\mathcal{R}(\mu) = 1$.

Now consider the nontrivial case, with $X_{(1)} < \mu < X_{(n)}$. We seek to maximize $\prod_i n w_i$, or equivalently $\sum_{i=1}^n \log(n w_i)$ over $w_i \geq 0$ subject to the constraints that $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n w_i X_i = \mu$. We write the latter constraint as $\sum_{i=1}^n w_i (X_i - \mu) = 0$. The objective function $\sum_{i=1}^n \log(n w_i)$ is a strictly concave function on a convex set of weight vectors. Accordingly a unique global maximum exists. We also know that the maximum does not have any $w_i = 0$, so it is an interior point of the domain.

We may proceed by the method of Lagrange multipliers. Write

$$G = \sum_{i=1}^n \log(nw_i) - n\lambda \sum_{i=1}^n w_i (X_i - \mu) + \gamma \left(\sum_{i=1}^n w_i - 1 \right),$$

where λ and γ are Lagrange multipliers. Setting to zero the partial derivative of G with respect to w_i gives

$$\frac{\partial G}{\partial w_i} = \frac{1}{w_i} - n\lambda (X_i - \mu) + \gamma = 0.$$

So

$$0 = \sum_{i=1}^n w_i \frac{\partial G}{\partial w_i} = n + \gamma,$$

from which $\gamma = -n$. We may therefore write

$$w_i = \frac{1}{n} \frac{1}{1 + \lambda (X_i - \mu)}.$$

The value of λ may be found by numerical search. We know that $\lambda = \lambda(\mu)$ solves

$$\frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{1 + \lambda (X_i - \mu)} = 0. \quad (2.8)$$

The left side of (2.8) equals $\bar{X} - \mu$ at $\lambda = 0$. It is strictly decreasing in λ , as may be found by differentiation. Monotonicity of (2.8) makes a bisection approach workable, but bisection is slow. Safeguarded search methods, like Brent's method or some versions of Newton's method, are preferable. They combine the reliability of bisection, with a superlinear rate of convergence to the solution.

To begin the search for $\lambda(\mu)$ we need an interval known to contain $\lambda(\mu)$. We know that every $w_i > 0$, and so every $w_i < 1$. A bracketing interval may be found by alternately setting to 1 the weight on the minimum and maximum observations. Thus we may start the search knowing that

$$\frac{1 - n^{-1}}{\mu - X_{(n)}} < \lambda(\mu) < \frac{1 - n^{-1}}{\mu - X_{(1)}}. \quad (2.9)$$

The algorithm then successively refines the interval for λ given by (2.9) until the endpoints agree to a user-specified tolerance. For example, the user might be satisfied if the two values of $\log(\mathcal{R}(\mu))$ from the endpoints of the interval for $\lambda(\mu)$ agree to within 10^{-6} .

To set a confidence interval for μ , we need to locate upper and lower limits μ_+ and μ_- for which $\mathcal{R}(\mu_{\pm}) = r_0 \in (0, 1)$. We know that

$$X_{(1)} \leq \mu_- \leq \bar{X} \leq \mu_+ \leq X_{(n)}, \quad (2.10)$$

and these bounds can be used in two separate safeguarded searches. We could

search for the μ solving $\mathcal{R}(\mu) = r_0$ using a search to find $\mathcal{R}(\mu)$ at each candidate μ .

Such a nested search for μ is slower than necessary, though not necessarily slow in an absolute sense. A faster approach is to reformulate the problem as optimizing $\sum_{i=1}^n w_i X_i$ subject to the constraints $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n \log(nw_i) = \log(r_0)$. In this formulation we take the Lagrangian to be

$$G = \sum_{i=1}^n w_i X_i - \eta \left(\sum_{i=1}^n \log(nw_i) - \log(r_0) \right) - \gamma \left(\sum_{i=1}^n w_i - 1 \right).$$

Some calculus, like that above, shows that $w_i = \eta / (X_i - \gamma)$ and so

$$w_i = w_i(\gamma) = \frac{(X_i - \gamma)^{-1}}{\sum_{j=1}^n (X_j - \gamma)^{-1}}.$$

To find μ_- we solve $\sum_{i=1}^n \log(nw_i(\gamma)) = \log(r_0)$, searching for γ between $-\infty$ and $X_{(1)}$. To find μ_+ we search for γ between $X_{(n)}$ and ∞ .

The endpoints in the search for γ are more delicate than in the search for λ . One endpoint is infinite and the other gives an infinite value for \mathcal{R} . In practice we have to search first for endpoints near the ones given above before beginning the safeguarded search.

To display $\mathcal{R}(\mu)$ we need to compute it at several values. Let $\mu_{(i)} = \bar{X} + i\delta$ for some $\delta > 0$ and integer $i \geq 0$. A good strategy to compute the right side of the empirical likelihood ratio curve is to compute $\mathcal{R}(\mu_{(i)})$ for i increasing from 0, where $\mathcal{R}(\mu_{(0)}) = 1$, until $\log(\mathcal{R}(\mu_{(i)}))$ is too small to be of interest, but in any case stopping before $\mu_{(i)} > X_{(n)}$. For example a limit of $\log(\mathcal{R}) = -25.0$ corresponds to a nominal $\chi_{(1)}^2$ value of $-2 \times 25 = 50$. Such a χ^2 value in turn corresponds to a p -value of about 1.5×10^{-12} , and we seldom need to consider p -values smaller than this. When searching for $\lambda(\mu_{(i)})$, a good starting value is $\lambda(\mu_{(i-1)})$, and we may begin with $\lambda(\mu_0) = 0$. To compute the left side of the empirical likelihood ratio curve, we repeat the process above for i decreasing from 0 until $\mathcal{R}(\mu_{(i)})$ is very small.

If δ is too small then too many steps are required to compute the curve. If δ is too large, then not enough points appear in the curve. The value of δ can be found by trial and error. It is usually satisfactory to have about 20 of the profile points between the 95% confidence interval endpoints. Those endpoints are roughly $4sn^{-1/2}$ apart where s is the sample standard deviation. So $\delta = 0.2 \times sn^{-1/2}$ is usually reasonable. When n is small or the sample is very skewed it may be necessary to use a smaller value of δ .

Most plotting systems will connect the points $(\mu_{(i)}, \mathcal{R}(\mu_{(i)}))$ by straight lines. This can give an unsatisfactory appearance to the curve, often with a prominent triangular peak at the MLE. The function $\log(\mathcal{R}(\mu))$ is approximately quadratic around $\hat{\mu} = \bar{X}$. A better looking curve is obtained by fitting an interpolating spline through $(\mu_{(i)}, \log(\mathcal{R}(\mu)))$. If the spline curve has values (x_j, y_j) on a fine grid of x_j values, then a plot linearly interpolating the $(x_j, \exp(y_j))$ points usually gives

a reasonable version of $\mathcal{R}(\mu)$ versus μ . On rare occasions with badly spaced μ_i or unusual behavior in $\mathcal{R}(\mu_i)$ the spline can show a Gibbs effect in which the exponentiated spline produces likelihood ratios over 1.0. The remedy is to insert more likelihood evaluations, or to resort to linear interpolation.

2.10 Empirical discovery of parametric families

Suppose that F_0 is in fact inside a known parametric family. It is natural to wonder whether the empirical likelihood function can discover this fact and match the parametric inferences. It cannot. There is no unique parametric family through F_0 to discover.

Suppose, for example, that X_i are normally distributed with mean $\mu_0 = 1$ and variance $\sigma_0^2 = 1$. Then F_0 belongs to the following families among others:

$$\begin{aligned} N(\mu, 1), & \quad \mu \in \mathbb{R}, \\ N(\mu, \mu^2), & \quad \mu \in (0, \infty), \\ N(\mu, e^{1-\mu}), & \quad \mu \in \mathbb{R}, \\ N(1, \sigma^2), & \quad \sigma \in (0, 10). \end{aligned}$$

No sample from $N(1, 1)$, however large, can identify one of these models as the true parametric family. They are all equally true, and they have different consequences for how hard it is to learn μ from data. These parametric families are illustrated in [Figure 2.6](#).

A choice of a parametric family requires knowledge from outside of the sample. Such prior information may specify a set of distributions known to include F_0 , perhaps based on experience with previous data thought to be similar to the present data. Different investigators could reasonably have different prior information, select different parametric families, and so obtain different answers.

If side information is available, then it can often be used in empirical likelihood. If, for example, we know that $\text{Var}(X) = E(X)^2$ or that $\kappa = 0$, then these facts can be imposed directly as side constraints. It is not necessary to find a parametric family with those constraints built in. See [Chapter 3.10](#).

With empirical likelihood, we ordinarily assume no knowledge outside of the data. Therefore we expect empirical likelihood confidence intervals to be asymptotically at least as long as those for any reasonable parametric family containing F_0 . We would like empirical likelihood inferences to behave like parametric inferences in a least favorable family: one in which inference is not artificially easy. [Chapters 9.6 and 9.11](#) discuss connections between empirical likelihood and Stein's concept of least favorable families of distributions.

While empirical likelihood is expected to provide confidence intervals no narrower than a parametric family containing the true distribution F_0 , it is possible to find that empirical likelihood confidence intervals are sometimes narrower than parametric ones. This can easily happen if the parametric family is incorrect. For example, if $\kappa < 0$, then the normal theory confidence intervals for the variance will be longer than the empirical ones, for large enough n .

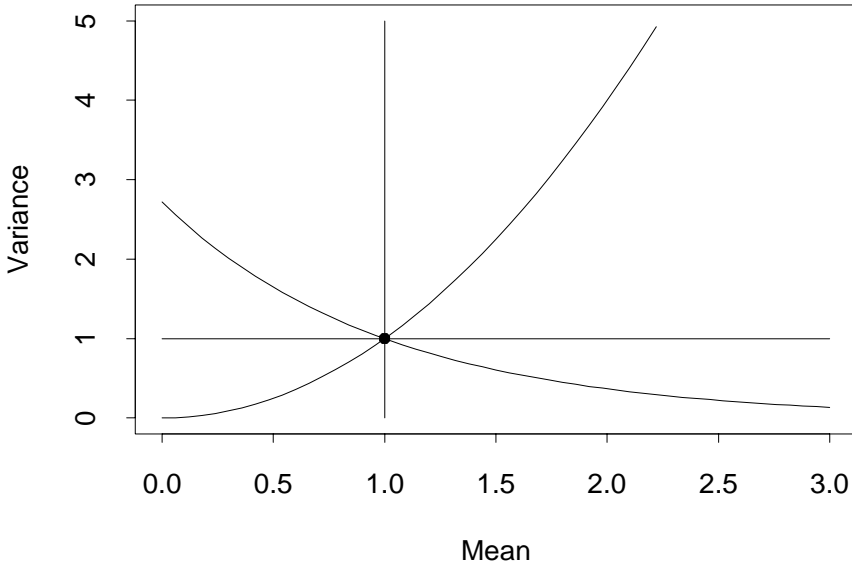


Figure 2.6 The curves depict four single parameter families through the distribution $N(1, 1)$, shown as a point. For each family, the variance is plotted versus the mean. The families are given in the text. Only the portions for which $0 \leq \mu \leq 3$ and $0 \leq \sigma^2 \leq 5$ are shown.

2.11 Bibliographic notes

Nonparametric maximum likelihood

The fact that the ECDF is an NPMLE was first noticed by Kiefer & Wolfowitz (1956). The NPMLE idea was used by Kaplan & Meier (1958) to derive the product-limit estimator for the CDF from censored data. Johansen (1978) shows that product-limit estimators are NPMLE's for transition probabilities of continuous time Markov chains. Grenander (1956) constructs an NPMLE for a distribution known to have a monotone decreasing density over $[0, \infty)$. Hartley & Rao (1968) show how to construct the NPMLE from a simple random sample of a finite population.

The star velocities are from the bright star catalogue of Hoffleit & Warren (1991). These data are also available on the Internet. Only those stars for which both velocities are recorded were used. These tend to be the stars nearest the sun. Binney & Merrifield (1998) provide information on stars and their movements.

Likelihood ratios

Theorem 2.2 on empirical likelihood was proved by Owen (1988b), who shows how ties may be ignored, and also presents an ELT for (Fréchet) differentiable statistical functions. An ELT for the mean of bounded scalar random variable appears in Owen (1985). The discussion of computing is based on Owen (1988b), as is the noncentral χ^2 result at (2.7). Brent's method and other safeguarded searches are described in Press, Flannery, Teukolsky & Vetterling (1993).

Neyman & Scott (1948) provide a famous example of the adverse consequences to likelihood methods when the number of parameters goes to infinity with the sample size. They have $X_{ij} \sim N(\mu_j, \sigma^2)$ for $j = 1, 2$ and $i = 1, \dots, n$. Then as $n \rightarrow \infty$ the MLE $\hat{\sigma}^2 \rightarrow \sigma^2/2$ and so is not consistent.

The earliest known use of an empirical likelihood ratio function is Thomas & Grunkemeier (1975). They consider the problem of forming a confidence interval for the survival function based on censored data. The Kaplan-Meier estimate gives the NPML, and nonparametric likelihood arguments can be used to form a confidence interval.

Cavendish's data appear in Stigler (1977). That article compares methods of estimating a parameter from data in cases where we now know the true value. Stigler's motivation was to assess whether robust estimators would have helped. One particularly delicate issue is that each scientist's instruments had built-in biases.

Coverage levels

Bahadur & Savage (1956, Corollary 2) show that no nontrivial confidence interval can be computed for the mean if the family \mathcal{F} of possible distributions is sufficiently rich. Their results rule out the existence of exact or even of conservative nonparametric confidence intervals for the mean. They let \mathcal{F} be any set of distributions for $X \in \mathbb{R}$, satisfying three conditions:

1. $\int |x|dF(x) < \infty$ for all $F \in \mathcal{F}$,
2. for each $m \in \mathbb{R}$ there is an $F \in \mathcal{F}$ with $\int xdF(x) = m$, and
3. if $F, G \in \mathcal{F}$, then $\lambda F + (1 - \lambda)G \in \mathcal{F}$ for $0 \leq \lambda \leq 1$.

Now suppose that a confidence interval method has probability at least $1 - \alpha$ of covering the mean of F , with this holding for all $F \in \mathcal{F}$. Then for every $F \in \mathcal{F}$ and every $m \in \mathbb{R}$, that method has probability at least $1 - \alpha$ of covering m when sampling from F . This failure cannot be escaped by taking random endpoints for the interval, or by taking a random (but always finite) number of observations. The problem is that F can place a very small probability p on a very large value X_0 . The probability p can be small enough that X_0 is unlikely to be seen in a sample of size n . But X_0 can be large enough that pX_0 is not small compared to $E(X)$.

Nonparametric confidence intervals typically approach their asymptotic coverage levels from below. For finite n , the true coverage level is usually, though

not always, below the nominal level. This has been observed for empirical likelihood by Owen (1988*b*), Owen (1990*a*), and others, for generalized method of moments by Imbens, Spady & Johnson (1998) and others cited therein, and for the bootstrap by Schenker (1985) and others. Kauermann & Carroll (2000) give explicit undercoverage formulas for some confidence intervals based on sandwich estimators of variance. Undercoverage can also arise for asymptotically justified confidence intervals in parametric problems.

As Efron (1988) shows, practically significant errors in coverage levels can correspond to very minor-looking errors in the endpoints of confidence intervals; the noncoverage events are typically near misses. Asymptotic confidence regions give a realistic separation between more and less plausible parameter values, but if we see a special value θ^* just barely outside an asymptotic confidence region, we cannot be sure of the p -value for rejecting θ^* . It is a good practice to plot the confidence interval or region, and then annotate the plot with any values that are special in the context of the data. Still, undercoverage is to be avoided where possible, and it can be greatly alleviated by using bootstrap calibration as described in Chapter 3.3.

Hall (1986) establishes formulas for the coverage error of nonparametric confidence interval methods for the mean. Owen (1990*a*) provides extensive simulation of various sample sizes and distributions, and finds that by $n = 20$, Hall's formulas are within roughly 1% of the true coverage errors, except for very heavy tailed distributions.

The bibliographic notes for Bartlett correction, signed root corrections and bias shifting of empirical likelihood appear on page 257 at the end of Chapter 13.

Very general power optimality was shown for likelihood ratio tests by Hoeffding (1965). A version for empirical likelihood, discussed in Chapter 13.5, is due to Kitamura (2001), who also presents some simulations. The simulations in Owen (1990*a*) showed that the bootstrap- t produced the best confidence intervals for the univariate mean. The hardest problem turned out to be covering the mean of the lognormal distribution. In each simulated data set, every competing method constructed a confidence interval. The lengths of these intervals were recorded. Then each method constructed a set of intervals, using the interval lengths chosen by every other method. Empirical likelihood intervals often achieved better coverage when using an interval of a given length than did the method whose nominal 95% interval was of that length.

2.12 Exercises

Exercise 2.1 Another approach to breaking ties is to perturb $X_i \in \mathbb{R}^d$ into $X_i^\epsilon = X_i + \epsilon Z_i$, where $Z_i \sim N(0, I_d)$ are independent of each other and the X_i . Let T be a function of the distribution of X_i . Let $\mathcal{R}(\theta, \epsilon) = \max\{\prod_{i=1}^n n w_i \mid T(\sum_{i=1}^n w_i \delta_{X_i}) = \theta, w_i \geq 0, \sum_{i=1}^n w_i = 1\}$. Does $\lim_{\epsilon \rightarrow 0} \mathcal{R}(\theta, \epsilon)$ always equal the unperturbed empirical likelihood ratio $\mathcal{R}(\theta, 0)$?

Exercise 2.2 The empirical likelihood ratio is $\prod_{i=1}^n nw_i = n^n \times \prod_{i=1}^n w_i$. The disadvantage of the second expression is that it is a product of one very large factor and one very small factor. A computer might end up trying to multiply an overflowing number by an underflowing one. For one specific computer, find out how large n has to be for n^n to overflow its floating point representation. Find out how large n must be for $(1/n)^n$ to underflow. In practice accuracy can be lost at values of n smaller than those causing underflow or overflow, and it is better to work with the log likelihood ratio.

Exercise 2.3 Exact confidence intervals are possible for the mean μ in the family $N(\mu, 1)$, and for μ in the family $N(\mu, \sigma^2)$, assuming $n \geq 1$ in the first case and $n \geq 2$ in the second. Explain why this does not contradict the result of Bahadur & Savage (1956) quoted in Chapter 2.11.

Exercise 2.4 This exercise and the next provide a nonrigorous sketch of the proof of [Theorem 2.2](#). Expand equation [\(2.8\)](#) in a Taylor series about $\lambda = 0$. Using the leading terms show that $\lambda \doteq (\bar{X} - \mu)/S(\mu)$ where $S(\mu) = (1/n) \sum_{i=1}^n (X_i - \mu)^2$.

Exercise 2.5 Substitute $\lambda = (\bar{X} - \mu)/S(\mu)$ from [Exercise 2.4](#) into an expression for $-2 \log \mathcal{R}(\mu_0)$ and show, after a Taylor approximation, that the result is nearly equal to $n(\bar{X} - \mu_0)^2/S(\mu_0)$. A $\chi^2_{(1)}$ limit is then reasonable when $\sqrt{n}(\bar{X} - \mu_0)$ is asymptotically normal with a variance estimated by $S(\mu_0)$.