

STATISTICAL RESEARCH REPORT  
Institute of Mathematics  
University of Oslo

No 1  
1971

UNBIASEDNESS OF LIKELIHOOD RATIO CONFIDENCE  
SETS IN CASES WITHOUT NUISANCE PARAMETERS

by

Emil Spjøtvoll

This is a revised version of the article "Likelihood Ratio Confidence Sets", Technical Report No 219, Department of Statistics, University of Wisconsin.

## S U M M A R Y

Confidence sets based upon the likelihood function is defined in the case without nuisance parameters. It is shown under certain assumptions on the density function that the confidence sets are unbiased. Various examples are given.

## 1. INTRODUCTION

Let  $X$  be a random variable with probability density  $f(x, \theta)$  with respect to some  $\sigma$ -finite measure  $\mu$ . The parameter  $\theta$  is unknown and may have values in some set  $\Omega$ . The problem we shall study is to find confidence sets for  $\theta$ , based upon the likelihood function.

Let

$$L(x, \theta) = \frac{f(x, \theta)}{\sup_{\theta} f(x, \theta)} \quad (1)$$

and define

$$S(x) = \{\theta : L(x, \theta) \geq c(\theta)\}, \quad (2)$$

where  $c$  is some function of  $\theta$ . We will call the family of subsets  $S(x)$  a family of likelihood ratio confidence sets (LRCS) for  $\theta$  at confidence level  $1-\alpha$  if  $c(\theta)$  is chosen such that

$$P_{\theta}\{\theta \in S(X)\} \geq 1-\alpha \quad \text{for all } \theta \in \Omega. \quad (3)$$

Clearly there always exist such a family of LRCS, we only need to choose  $c(\theta)$  less or equal to the lower  $\alpha$ -point of the distribution of  $L(X, \theta)$ . Furthermore the family LRCS is uniquely determined if we for each  $\theta$  choose  $c(\theta)$  as the largest lower  $\alpha$ -point of the distribution of  $L(X, \theta)$ .

The intuitive interpretation of the confidence set (2) should be clear. When  $x$  is observed we take a point  $\theta$  into our confidence set if the ratio of the density at  $\theta$  to the maximum obtainable density for the given  $x$  is greater than some constant. That this constant is allowed

to depend upon  $\theta$  might seem somewhat confusing. A technical reason for this is that it might be necessary if we want (3) to hold. But we also have a deeper reason. For some values of  $\theta$  the density  $f(x, \theta)$  may be quite flat, while for others the density has a peak around the true value. (an example could be the Poisson distribution.) It seems that it is easier for the former to get large values for  $L(x, \theta)$ , and therefore such values would have an advantage over the latter, if we used a  $c(\theta)$  not depending upon  $\theta$ . It will, however, be shown that for a number of problems  $c(\theta)$  is actually a constant independent of  $\theta$ .

It is easily seen that the maximum likelihood estimate of  $\theta$  is always in the confidence set, and that if  $\gamma(\theta)$  is a one-to-one transformation of  $\theta$ , then the family of LRCS for  $\gamma$  is  $\gamma(S(x))$ .

To use the function  $L(x, \theta)$  to derive confidence sets is not a new idea. In a recent paper Kalbfleisch and Sprott (1970) use  $L(x, \theta)$  and other forms of the likelihood function as a basis for statistical inference. Box and Cox (1964) used an asymptotic version of the sets (2). Barnard (1965) advocated the use of (2) (Barnard wanted the function  $c(\theta)$  in (2) to be a function of  $x$  but that may be a misprint). In an unpublished paper Hudson (1968) used a version of (2) for the binomial distribution. For other references see Hudson (1968) and Kalbfleisch and Sprott (1970).

The purpose of this paper is to show that in many cases the LRCS are unbiased confidence sets.

## 2. UNBIASEDNESS OF LIKELIHOOD RATIO CONFIDENCE SETS

A family  $S(x)$  of confidence sets is unbiased if (3) is satisfied and

$$P_{\theta}\{\theta' \in S(X)\} \leq 1-\alpha$$

when  $\theta \neq \theta'$ . We will now show that under certain conditions the LRCS are unbiased.

We will make the following assumptions:

A 1.  $\Omega$  is separable.

A 2.  $f(x, \theta)$  is continuous in  $\theta$  for all  $x$ .

A 3. The family of densities  $\{f(x, \theta) : \theta \in \Omega\}$  is invariant under a group  $G$  of measurable transformations, and  $\mu$  is absolutely continuous with respect to  $\mu g^{-1}$  for all  $g \in G$ . Furthermore, the induced group  $\bar{G}$  of transformations of  $\Omega$  is transitive over  $\Omega$ .

For the concepts used in the assumptions see for example Lehmann (1959), Chapter 6.

We have

Lemma If Assumptions 1-3 hold, then  $c(\theta)$  in (2) is a constant not depending upon  $\theta$ .

Proof. Let  $\bar{G}$  be the induced group of transformations of  $\Omega$ . When A 3 holds we have (See Lehmann (1959), p. 252, Problem 7)

$$f(x, \theta) = a(gx) f(gx, \bar{g}\theta) \quad \text{a.e.}\mu, \quad (4)$$

where

$$a(gx) = \frac{d\mu}{d\mu g^{-1}}(gx) .$$

The null set for which the equality in (4) does not hold may depend upon  $\theta$  and  $g$ . Under A 1-2 we find (see the above

reference to Lehmann, 1959)

$$\sup_{\theta} f(x, \theta) = a(gx) \sup_{\theta} f(gx, \bar{g}\theta) \quad \text{a.e.}\mu. \quad (5)$$

Combining (4) and (5) we get

$$L(x, \theta) = L(gx, \bar{g}\theta) \quad \text{a.e.}\mu \quad (6)$$

Using this and the fact that if  $X$  has distribution  $P_{\theta}$  than  $gX$  has distribution  $P_{\bar{g}\theta}$  we get

$$\begin{aligned} P_{\theta}\{L(X, \theta) \geq c\} &= P_{\theta}\{L(gX, \bar{g}\theta) \geq c\} \\ &= P_{\bar{g}\theta}\{L(X, \bar{g}\theta) \geq c\} . \end{aligned}$$

Since this holds for all  $\bar{g}$  and  $c$  the lemma is proved.

We have the following theorem

Theorem If Assumptions 1-3 hold, then the family of confidence sets (2) is unbiased.

Proof. Since the densities  $f(x, \theta)$  are invariant under  $G$ , it follows from the lemma that  $c(\theta)$  is a constant  $c$  not depending upon  $\theta$ . From (5) we find by replacing  $x$  by  $g^{-1}x$

$$\begin{aligned} \sup_{\theta} f(g^{-1}x, \theta) &= a(x) \sup_{\theta} f(x, \bar{g}\theta) \\ &= a(x) \sup_{\theta} f(x, \theta) \quad \text{a.e.}\mu \end{aligned} \quad (7)$$

Furthermore by substituting  $g^{-1}x$  for  $x$  in (4), we find

$$f(g^{-1}x, \theta) = a(x) f(x, \bar{g}\theta) \quad \text{a.e.}\mu. \quad (8)$$

Define

$$A(\theta) = \{x: f(x, \theta) \geq c \sup_{\theta} f(x, \theta)\}.$$

We have

$$\begin{aligned}
 A(\bar{g}\theta) &= \{x: f(x, \bar{g}\theta) \geq c \sup_{\theta} f(x, \bar{g}\theta)\} \\
 &= \{x: a(x) f(x, \bar{g}\theta) \geq c a(x) \sup_{\theta} f(x, \bar{g}\theta)\} \\
 &= \{x: f(g^{-1}x, \theta) \geq c \sup_{\theta} f(g^{-1}x, \theta)\} = gA(\theta),
 \end{aligned} \tag{9}$$

where the third equality follows from (7) and (8).

Now consider the integral

$$\int_{A(\theta)} \sup_{\theta} f(x, \theta) d\mu(x) = \int_{A(\theta)} L^{-1}(x, \theta) f(x, \theta) d\mu(x). \tag{10}$$

The integral exists since  $L^{-1}(x, \theta) \leq c$  on  $A(\theta)$ . We also have

$$\begin{aligned}
 \int_{A(\theta)} L^{-1}(x, \theta) f(x, \theta) d\mu(x) &= \int_{gA(\theta)} L^{-1}(g^{-1}x, \theta) f(x, \bar{g}\theta) d\mu(x) \\
 &= \int_{gA(\theta)} \frac{\sup_{\theta} f(g^{-1}x, \theta)}{f(g^{-1}x, \theta)} f(x, \bar{g}\theta) d\mu(x) \\
 &= \int_{A(\bar{g}\theta)} \sup_{\theta} f(x, \theta) d\mu(x),
 \end{aligned} \tag{11}$$

where the first equality is obtained from the fact that  $f(x, \theta)$  is invariant under the transformation  $g$  (see Lehmann, 1959) p. 252, Problem 16), and the third equality is obtained by using (7), (8) and (9).

By definition we have

$$P_{\theta}\{\theta \in S(X)\} = P_{\theta}\{X \in A(\theta)\} = \int_{A(\theta)} f(x, \theta) d\mu(x), \tag{12}$$

$$P_{\theta}\{\bar{g}\theta \in S(X)\} = P_{\theta}\{X \in A(\bar{g}\theta)\} = \int_{A(\bar{g}\theta)} f(x, \theta) d\mu(x). \quad (13)$$

Furthermore (10) and (11) imply

$$\int_{A(\theta)} \sup_{\theta} f(x, \theta) d\mu(x) = \int_{A(\bar{g}\theta)} \sup_{\theta} f(x, \theta) d\mu(x) = \beta, \quad (14)$$

where  $\beta$  is some constant. Now consider the problem to find a region  $B$  that will maximize

$$\int_B f(x, \theta) d\mu(x) \quad (15)$$

subject to

$$\int_B \sup_{\theta} f(x, \theta) d\mu(x) = \beta. \quad (16)$$

By the Neyman-Pearson fundamental lemma the maximizing region is  $A(\theta)$ .  $A(\bar{g}\theta)$  is another region satisfying (16). Hence

$$\int_{A(\bar{g}\theta)} f(x, \theta) d\mu(x) \leq \int_{A(\theta)} f(x, \theta) d\mu(x),$$

and therefore by (12) and (13) we have

$$1 - \alpha = P_{\theta}\{\theta \in S(X)\} \geq P_{\theta}\{\bar{g}\theta \in S(X)\} \quad \text{for all } \bar{g} \in \bar{G}.$$

The theorem is proved.

Remark. The conclusion of the Theorem holds if

$$S(x) = \{\theta: f(x, \theta) \geq h(x)\}$$

where  $h(x)$  is any function such that

$$h(g^{-1}x) = a(x) h(x).$$

Then (7) is satisfied, and that is all we need.



3. EXAMPLES

Example 1. Let  $X_1, \dots, X_n$  be independent  $N(\xi, \sigma^2)$ . If  $\sigma^2$  is known it is easily found that the LRCS for  $\xi$  is the standard confidence interval for  $\xi$ , and if  $\xi$  is known it is found after some calculations that the LRCS for  $\sigma^2$  is the best unbiased confidence interval as given in Lehmann (1959), pp. 129-130.

The joint LRCS for  $\xi$  and  $\sigma^2$  when both are unknown is found to be

$$\{(\xi, \sigma^2): \sum_{i=1}^n (x_i - \bar{x})^2 / \sigma^2 - n \log \left\{ \sum_{i=1}^n (x_i - \bar{x})^2 / \sigma^2 \right\} + n(\bar{x} - \xi)^2 / \sigma^2 \leq c_1\} \quad (20)$$

where  $c_1 = n-2 \log c$ .

Let  $Z$  and  $V$  denote chi-square distributed random variables with  $n-1$  and 1 degrees of freedom, respectively. Let also  $Z$  and  $V$  be independent. Then the left hand side of the inequality sign in (20) is distributed as  $W = Z - n \log Z + V$ . The confidence set (20) is a  $1-\alpha$  LRCS if we choose  $c_1$  equal to the upper  $\alpha$ -point of the distribution of  $W$ .

We will now examine the form of the confidence set (20). The inequality in (20) can be written

$$n(\bar{x} - \xi)^2 \leq g(\sigma^2) \quad (21)$$

where  $g(\sigma^2) = c_1 \sigma^2 + n\sigma^2 \log (S/\sigma)^2 - S^2$  and  $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$ .

The function  $g(\sigma^2)$  has a unique maximum at  $\hat{\sigma}^2 = S^2 \exp(c_1/n - S^2)$  and  $\lim_{\sigma^2 \rightarrow 0} g(\sigma^2) = -S^2$ ,  $\lim_{\sigma^2 \rightarrow \infty} g(\sigma^2) \rightarrow -\infty$ . From (21) and the form of  $g(\sigma^2)$  it follows that for each fixed  $\sigma^2$  the set

of possible values for  $\xi$  within the confidence set, is the interval  $[\bar{x} - \{g(\sigma^2)/n\}^{\frac{1}{2}}, \bar{x} + \{g(\sigma^2)/n\}^{\frac{1}{2}}]$ . The largest interval is obtained when  $\sigma^2 = \hat{\sigma}^2$ . Then  $g(\hat{\sigma}^2)/n = S^2\{S^2 \exp(c_1/n - S^2) - n^{-1}\}$ . When  $\sigma^2$  increases or decreases from  $\hat{\sigma}^2$  the lengths of the intervals decrease.

Similarly the set of possible values for  $\sigma^2$  for a fixed  $\xi$  is an interval with endpoints equal to the smallest and largest solution of  $n(\bar{x} - \xi)^2 = g(\sigma^2)$ , respectively. The largest interval is obtained when  $\xi = \bar{x}$ , and when  $\xi$  increases or decreases from  $\bar{x}$ , the lengths of the intervals decrease.

The assumption A 1-3 are easily seen to be satisfied in this example. For the confidence sets for  $\xi$  we use a group of translation, for  $\sigma^2$  we use a group of scale change while for the joint confidence sets for both  $\xi$  and  $\sigma^2$  we use both groups.

As other examples we could use other non-discrete exponential families. It turns out that we get the usual confidence sets. Instead we will now consider some more complicated examples..

Example 2. Let  $X_1, \dots, X_n$  be independent with the double exponential density, so that

$$f(x, \theta) = \left(\frac{1}{2}\right)^n \exp\left(-\sum_{i=1}^n |x_i - \theta|\right).$$

Let  $x_{(1)} \leq \dots \leq x_{(n)}$  be the ordered observations, and let  $y = x_{(m+1)}$  if  $n = 2m+1$  and  $y = x_{(m)}$  if  $n = 2m$ . Then

$$\sup_{\theta} f(x, \theta) = \left(\frac{1}{2}\right)^n \exp\left(-\sum_{i=1}^n |x_i - y|\right),$$

and hence

$$L(x, \theta) = \exp\{-\sum_{i=1}^n (|x_i - \theta| - |x_i - y|)\}.$$

The likelihood ratio confidence set for  $\theta$  is given by

$$\{\theta: \sum_{i=1}^n |x_i - \theta| - \sum_{i=1}^n |x_i - y| \leq c_1\} \quad (22)$$

where  $c_1 = \exp(-c)$ . The probability that the region (22) covers  $\theta$  is

$$P_{\theta} (\sum_{i=1}^n |x_i - \theta| - \sum_{i=1}^n |x_i - y| \leq c_1). \quad (23)$$

This does not depend upon  $\theta$ . When evaluating the probability (23) we may assume  $\theta = 0$ . Then we have to find  $P(T \leq c_1)$ , where  $T = \sum_{i=1}^n |x_i| - \sum_{i=1}^n |x_i - y|$ .

It is seen that

$$T = \begin{cases} -2(y + \dots + u) + y & \text{if } y < 0 \\ 2(v + \dots + y) - y & \text{if } y > 0, \end{cases}$$

where  $u$  is the largest negative  $x$  and  $v$  the smallest positive  $x$ , and where the summation is over all observation between  $y$  and  $u$ , and  $v$  and  $y$ , respectively. It is, of course, not easy to evaluate the distribution of  $T$ .

Example 3. Let  $X_1, \dots, X_n$  be independent with a Cauchy distribution, so that

$$f(x, \theta) = \pi^{-n} \prod_{i=1}^n \{1 + (x_i - \theta)^2\}^{-1}.$$

Then the LTCS for  $\theta$  is

$$\{\theta: \prod_{i=1}^n \{1 + (x_i - \hat{\theta})^2\} / \{1 + (x_i - \theta)^2\} \geq c\} \quad (24)$$

where  $\hat{\theta}$  is the maximum likelihood estimate of  $\theta$ .

The confidence coefficient of (24) is difficult to obtain when  $n$  is large. The case  $n=1$  is trivial. We will consider the case  $n=2$  in some detail, since that was used as an example of how to find a structural confidence set for  $\theta$  in Fraser, 1968. This gives us an opportunity to compare the two in a nontrivial case.

The density  $f(x, \theta)$  has a stationary point as a function of  $\theta$  for  $\theta_1 = \frac{1}{2}(x_1+x_2)$ ,  $\theta_2 = \frac{1}{2}[x_1+x_2 - \sqrt{\{(x_1-x_2)^2 - 4\}}]$  and  $\theta_3 = \frac{1}{2}[x_1+x_2 + \sqrt{\{(x_1-x_2)^2 - 4\}}]$ . The root  $\theta_1$  is the only real root if  $(x_1-x_2)^2 \leq 4$ . The function  $f(x, \theta)$  has a maximum at  $\theta_2$  and  $\theta_3$  if these are real, and a local minimum at  $\theta_1$ . If  $\theta_1$  is the only real root,  $f(x, \theta)$  has a maximum at  $\theta_1$ . It follows that

$$P_{\theta}\{L(X, \theta) \geq c\} = P_0[(X_1-X_2)^2 \leq 4 \text{ and } \{1 + (X_1-X_2)^2/4\}^2 / \{(1+X_1^2)(1+X_2^2)\} \geq c] + P_0[(X_1-X_2)^2 \geq 4 \text{ and } (X_1-X_2)^2 / \{(1+X_1^2)(1+X_2^2)\} \geq c].$$

The confidence set (24) can be written in the form

$$\{\theta: [\{\theta - (x_1+x_2)/2\}^2 + 1-d^2]^2 + 4d^2 \leq c^{-1} \{1 + (x_1-\hat{\theta})^2\} \{1 + (x_2-\hat{\theta})^2\}\}$$

where  $d = |x_1 - x_2|/2$ . If  $d \leq 1$  this reduces to the interval

$$(x_1+x_2)/2 - A_1 \leq \theta \leq (x_1+x_2)/2 + A_1,$$

where  $A_1^2 = d^2 - 1 + \{c^{-1}(1+d^2)^2 - 4d^2\}^{\frac{1}{2}}$ . If  $1 < d \leq (c^{-1}-1)^{\frac{1}{2}} + c^{-\frac{1}{2}}$

we get the interval

$$(x_1+x_2)/2 - A_2 \leq \theta \leq (x_1+x_2)/2 + A_2,$$

where  $A_2^2 = d^2 - 1 + 2d(c^{-1} - 1)^{\frac{1}{2}}$ . In the case  $d > (c^{-1} - 1)^{\frac{1}{2}} + c^{-\frac{1}{2}}$  we get a union of two intervals

$$\{(x_1+x_2)/2 - A_2 \leq \theta \leq (x_1+x_2)/2 - A_3\} \cup \{(x_1+x_2)/2 + A_3 \leq \theta \leq (x_1+x_2)/2 + A_2\},$$

where  $A_3^2 = d^2 - 1 - 2d(c^{-1} - 1)^{\frac{1}{2}}$ .

This is not the same as the structural confidence interval given in Fraser (1968), p. 16.

Example 4. Consider again Example 2. We will construct a confidence set by using the Remark in Section 2. In this example  $G$  is a group of translations, hence  $a(x) = 1$ . We choose  $h(x) = c$ , then we get

$$S(x) = \{\theta: \exp(-\sum_{i=1}^n |x_i - \theta|) > 2^n c\}$$

or

$$S(x) = \{\theta: 2\sum_{i=1}^n |x_i - \theta| \leq c_1\}$$

where  $c_1 = -2 \log 2^n c$ . Since  $2\sum_{i=1}^n |x_i - \theta|$  has a chi-square distribution with  $2n$  degrees of freedom, we get a  $1-\alpha$  confidence set for  $\theta$  by choosing  $c_1$  equal to the upper  $\alpha$ -point of that distribution. The confidence set is unbiased, and for a given set of data it is easily found. This confidence set has a peculiar behaviour, since with positive probability  $2\sum_{i=1}^n |x_i - y| > c_1$ , where  $y$  is the median, and for such values of  $x_1, \dots, x_n$  the confidence set will be empty.

## REFERENCES

- BARNARD, G.E. (1965) The use of the likelihood function in statistical practice. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1, 27-40.
- BOX, G.E.P. and COX, D.R. (1964). An analysis of transformations. J.R. Statist. Soc. B, 26, 211-252.
- FRASER, D.A.S. (1968). The Structure of Inference. Wiley, New York.
- HUDSON, D.J. (1968). Interval estimation from the likelihood function. Bell Telephone Lab. Tech. Rep.
- KALBFLEISCH, J.D. and SPOTT, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. J.R. Statist. Soc. B, 32, 175-208.
- LEHMANN, E.L. (1959). Testing Statistical Hypotheses. Wiley, New York.