Biometrika (1975), **62**, 1, *p.* 155
With 2 text-figures
Printed in Great Britain

Some counterexamples to the theory of confidence intervals

BY G. K. ROBINSON

Department of Probability and Statistics, University of Sheffield

SUMMARY

Some families of distributions are presented for which certain Neyman confidence intervals have very poor conditional properties. In each case there is a 50% Neyman confidence interval $I(x)$ for a parameter θ and a subset A of the sample space such that the conditional probability that $I(X)$ covers θ given that X belongs to A is less than 0·2 for all θ and the conditional probability that $I(X)$ covers θ given that X is not in A is at least 0·8 for all θ. The families of distributions are somewhat alike. More than one example is presented in order to show that the theory of confidence intervals cannot easily step around the difficulties presented to it.

Some key words: Conditional probability; Fiducial argument; Neyman confidence intervals; Relevant subsets.

1. INTRODUCTION

When claiming that the theory of confidence intervals leads to intuitively unreasonable statistical procedures, a problem which immediately asserts itself is that there is no definitive version of the theory. Neyman (1941) says that all confidence interval statements have the same justification through the fact that confidence intervals at confidence level α cover the true parameter value with probability α for all possible parameter values. However, largely as a result of attacks on confidence interval theory, it has become part of the theory that if a reason can be found for preferring one set of confidence limits to a second, then the second set is deemed not to be a proper confidence interval. For instance, although Neyman's (1937) concept of a 'shortest' confidence interval would only have been used by Neyman (1941) to select a single confidence interval which most people would prefer to use from the many available confidence intervals, all of which are equally supported by confidence interval theory, many statisticians now regard the fact that a confidence interval is not shortest as a reason for not considering it to be supported by the theory.

Possibly the best known counterexample for Neyman's version of confidence interval theory is the situation where a random variable X has a uniform distribution on the interval $(\theta, \theta+1)$ for a real parameter θ. If more than one observation is made it seems to be necessary to condition on the value of an ancillary statistic, largest observed value minus smallest observed value, in order to obtain sensible confidence intervals for θ. Pitman (1938), Welch (1939), Basu (1964) and Pierce (1973) have all discussed this example. Today it is widely accepted by adherents of confidence interval theory that they should perform their analyses conditional on the value of ancillary statistics. Basu's work makes it appear that there is no answer to the question of which ancillaries to condition upon, but this is regarded as a different question.

The concept of a relevant subset is very useful for discussing confidence intervals. When we make an inference which asserts that for a statement S, $\mathrm{pr}(S) = \alpha$, Buehler (1959) has called a subset C of the sample space a relevant subset if, for some $\epsilon > 0$, either

(i) $\mathrm{pr}(S|C) \geqslant \alpha+\epsilon$ for all parameter values, when C is called a positively biased relevant subset, or

(ii) $\mathrm{pr}(S|C) \leqslant \alpha-\epsilon$ for all parameter values, when C is called a negatively biased relevant subset.

Buehler & Fedderson (1963) presented a positively biased relevant subset for a confidence interval based on the t distribution. Since that confidence interval is generally regarded as being the correct one to use the net effect of this example seems to have been to make people believe that the existence of positively biased relevant subsets is not a severe criticism.

In Buehler & Fedderson's example the complement of their positively biased relevant subset is not a negatively biased relevant subset. However, for the examples in the present paper there are complementary sets one of which is a negatively biased relevant subset and the other of which is a positively biased relevant subset. Furthermore, for the third example, this difficulty cannot be overcome by conditioning on an ancillary, or deploying concepts like shortest.

2. The first example

Consider the family of distributions shown in Fig. 1 where it is intended that the pattern repeats itself indefinitely in both directions parallel to the line $\theta = x$. Diagrams of this sort have been used by Kendall & Stuart (1961, Chapter 20). Shown are the contours of $F(x|\theta) = \mathrm{pr}(X \leqslant x|\theta)$, where X is a random variable whose distribution depends upon a real parameter θ. For instance, with $x = 2$, $\theta = 4{\cdot}2$ we read $\mathrm{pr}(X \leqslant 2|\theta = 4{\cdot}2) = 0{\cdot}4$.

One way of looking at the diagram is for a fixed value of θ; we can see the distribution of X given θ. For instance, given $\theta = 2{\cdot}1$ we see that $\mathrm{pr}(X \leqslant -0{\cdot}5) = 0$, $\mathrm{pr}(X \leqslant 0) = 0{\cdot}45$, $\mathrm{pr}(X \leqslant 0{\cdot}5) = 0{\cdot}5 = \mathrm{pr}(X \leqslant 2{\cdot}5)$, $\mathrm{pr}(X \leqslant 3) = 0{\cdot}95$ and $\mathrm{pr}(X \leqslant 3{\cdot}5) = 1$. The lines $F(x|\theta) = 0$ and $F(x|\theta) = 1$ give upper and lower limits to the value of X for various θ values. We can also see that there is zero probability that X lies between the two jagged lines $\ldots B_{-1}B_0B_1B_2\ldots$ and $\ldots C_{-1}C_0C_1C_2\ldots$, since $F(x|\theta) = 0{\cdot}5$ for both of these lines. Throughout the region $A_0A_1B_1B_0$ and throughout other regions of the same shape the density of X given θ, $f(x|\theta)$, is $0{\cdot}1$. Throughout $A_1A_2B_2B_1$ and other regions of the same shape the density of X given θ is $0{\cdot}9$. The density function takes only the values 0, $0{\cdot}1$ and $0{\cdot}9$, but the pattern of regions where it assumes its three values is rather complicated.

Another way of looking at the diagram is to consider the likelihood function $L(\theta|x) = f(x|\theta)$ for a fixed value of x. For instance, given $x = 1{\cdot}5$ the likelihood of θ given x is $0{\cdot}9$ for $3{\cdot}1 \leqslant \theta \leqslant 4{\cdot}1$, $0{\cdot}1$ for $0{\cdot}1 \leqslant \theta \leqslant 1{\cdot}1$ and zero otherwise.

We define I and J to be set functions such that $I(x) = \{\theta\colon (x,\theta)$ lies between the lines $\ldots A_{-1}A_0A_1A_2\ldots$ and $\ldots B_{-1}B_0B_1B_2\ldots\}$, and $J(x) = \{\theta\colon (x,\theta)$ lies between the lines $\ldots C_{-1}C_0C_1C_2\ldots$ and $\ldots D_{-1}D_0D_1D_2\ldots\}$. Note that for every θ only values x of X such that $\theta \in I(x)$ or $\theta \in J(x)$ are possible. Thus the events $\theta \in I(X)$ and $\theta \in J(X)$ are complementary events.

Now the interval $I(x)$ is a 50% Neyman confidence interval for θ since it satisfies for all θ the equation

$$\mathrm{pr}\{\theta \in I(X)|\theta\} = 0{\cdot}5. \tag{1}$$

If we state that θ lies in $I(x)$ whenever we observe the value x of a random variable X having a distribution from amongst the family specified, then we will be right in 50 % of cases in the long run.

However, taking A to be the set of real numbers whose integer part is even, we can see from Fig. 1 that for all θ

$$\mathrm{pr}\{X \in A \text{ and } \theta \in I(X) | \theta\} \leqslant 0{\cdot}1, \tag{2}$$

$$\mathrm{pr}\{X \notin A \text{ and } \theta \in J(X) | \theta\} \leqslant 0{\cdot}1. \tag{3}$$

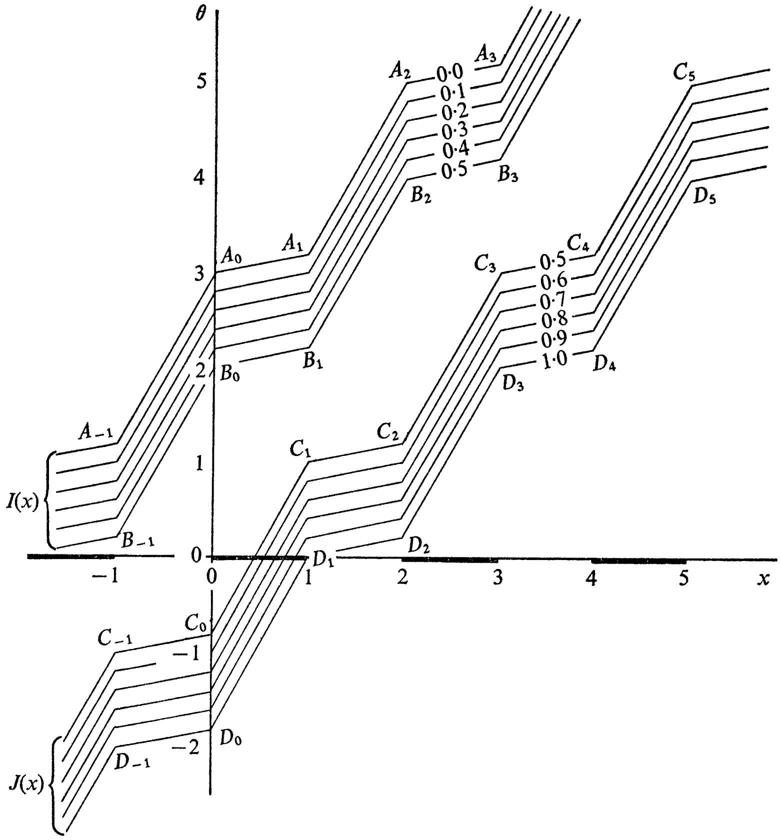

Fig. 1. Graph of contours of $F(x|\theta)$ for the family of distributions discussed in § 2. The label on a contour specifies pr $(X \leqslant x|\theta)$ for every point (x, θ) on that contour.

It follows that

$$\mathrm{pr}\{X \notin A \text{ and } \theta \in I(X) | \theta\} \geqslant 0{\cdot}4, \tag{4}$$

$$\mathrm{pr}\{X \in A \text{ and } \theta \in J(X) | \theta\} \geqslant 0{\cdot}4. \tag{5}$$

Hence

$$\mathrm{pr}\{\theta \in I(X) | X \in A, \theta\} = \left[1 + \frac{\mathrm{pr}\{X \in A \text{ and } \theta \notin I(X) | \theta\}}{\mathrm{pr}\{X \in A \text{ and } \theta \in I(X) | \theta\}}\right]^{-1},$$

$$\leqslant (1 + 0{\cdot}4/0{\cdot}1)^{-1} = 0{\cdot}2, \tag{6}$$

and

$$\begin{aligned}\operatorname{pr}\{\theta\in I(X)\,|\,X\notin A,\theta\} &= \left[1+\frac{\operatorname{pr}\{X\notin A \text{ and } \theta\notin I(X)\,|\,\theta\}}{\operatorname{pr}\{X\notin A \text{ and } \theta\in I(X)\,|\,\theta\}}\right]^{-1}, \\ &\geqslant (1+0{\cdot}1/0{\cdot}4)^{-1} = 0{\cdot}8. \end{aligned} \tag{7}$$

Thus A and its complement are both relevant subsets.

That equations (6) and (7) show the unreasonableness of using (1) as a basis for being 50% confident in the statement '$\theta\in I(x)$' can be seen as follows.

Suppose that a client comes several times to visit two consulting statisticians. Each time he has a single observation from the horrendous family shown in Fig. 1, and each time the first statistician tells him to quote $I(x)$ as his confidence interval with confidence coefficient 0·5. When the client visits the second statistician he asks him to verify the first statistician's conclusions. The second statistician disagrees with the conclusions of the first and, perhaps having noted (6) and (7), says that the client should really not be more than 20% confident that $\theta\in I(x)$ when $x\in A$, and be not less than 80% confident that $\theta\in I(x)$ when $x\notin A$.

By the nature of the client's work the true values of θ are available to him later in time than the values of X. He checks the first statistician's analysis and finds that he was correct in so far as θ did belong to $I(X)$ approximately 50% of the time. However, the second statistician was also correct in his statements and has been able to tell him roughly which times θ belonged to $I(X)$.

I believe that the correct explanation of what is happening here is in terms of the likelihood function. When $x\in A$ we have that $L(\theta|x) = 0{\cdot}1$ for $\theta\in I(x)$, and $L(\theta|x) = 0{\cdot}9$ for $\theta\in J(x)$. Our strength of belief that $\theta\in I(x)$ for a uniform improper prior distribution on θ would be 0·1. Similarly, when $x\notin A$ the likelihood function would lead us to think that $\theta\in I(x)$ is more likely than $\theta\in J(x)$.

Adherents of the classical theory of confidence intervals would try to avoid such a Bayesian interpretation. They might point out that, since $\operatorname{pr}(X\in A\,|\,\theta) = 0{\cdot}5$ for all θ, the indicator of A is an ancillary statistic, so that perhaps we should condition on its value. If we did this we would be led to consider the intuitively reasonable confidence interval

$$K(x) = \begin{cases} I(x) & (x\notin A), \\ J(x) & (x\in A), \end{cases} \tag{8}$$

at confidence level 80%. The set $K(x)$ is where the likelihood function takes the value 0·9. It covers θ with probability at least 0·8 for all θ.

This objection is easily answered by means of a second counterexample.

3. The second example

Suppose we moved the region spanned by $J(x)$ a small distance, say 0·1, in the negative θ direction on Fig. 1. This would entail moving each of the points C_0, C_1, C_2, C_{-1}, D_0, D_1, etc. 0·1 units downwards. The corresponding change in the family of distributions would mean that whether or not $X\in A$ was no longer an ancillary statistic, since $\operatorname{pr}(X\in A\,|\,\theta)$ would now vary with θ. Therefore classical arguments would not tell us to condition on its value. However, equations (1), (2) and (3) would remain true after this change so that (6) and (7) would also remain true, leaving the paradox intact.

The argument is not yet even nearly finished. The adherents of the theory of confidence intervals can now raise another objection to the use of the confidence region $I(x)$.

The confidence interval $K(x)$ defined by (8) is the same length as $I(x)$ but has a larger probability of covering θ, at least 0·8 as can be seen from inequalities (4) and (5). Therefore there must be confidence regions which are always shorter than $I(x)$ but which cover θ with probability more than 50% for all θ. It might be argued that the existence of confidence regions which are shorter than $I(x)$ in the sense of Neyman (1937) means that $I(x)$ is not supported by the theory of confidence intervals. Thus a criticism of $I(x)$ is not a criticism of that theory.

While answering this objection another objection can also be answered. This further objection was raised by a referee to an earlier version of this paper. We are, in a sense, accepting as possible those θ values in a confidence interval and rejecting the others. Consequently we should be thinking in terms of tests of hypotheses between the possible θ values. Thus a confidence interval should, for every x, contain θ values which have higher likelihoods than those θ values which it does not contain.

4. The third example

We now look at a confidence interval which appears to be completely supported by confidence interval theory but which is intuitively unreasonable despite this.

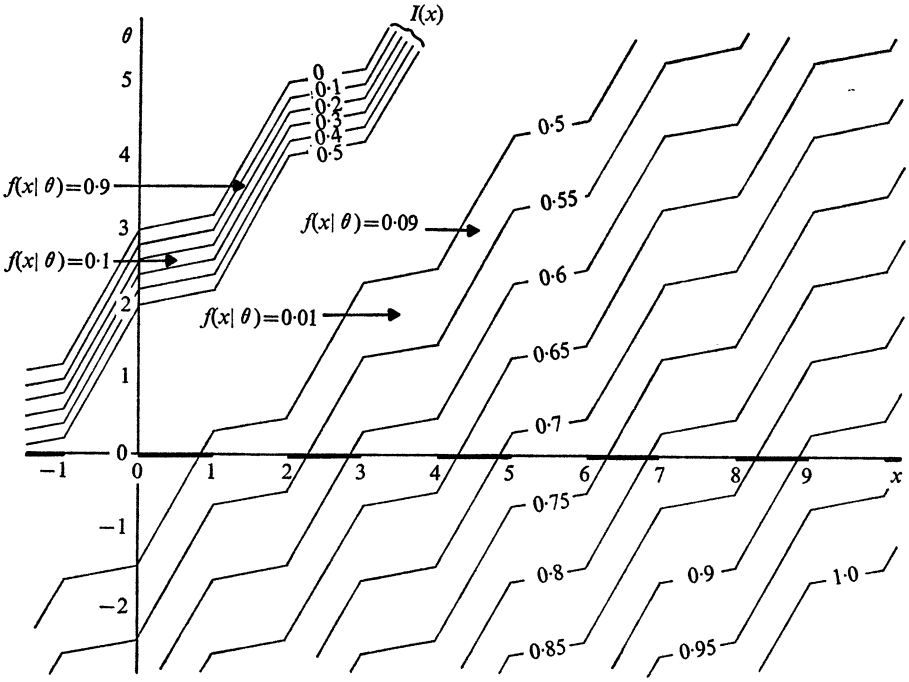

Fig. 2. Graph of contours of $F(x|\theta)$ for the family of distributions discussed in §4.

Consider the family of distributions illustrated in Fig. 2 which is of the same type as Fig. 1. In the region spanned by $I(x)$ the density of X given θ takes only the same two values as before: 0·1 and 0·9. The other region of nonzero probability is effectively ten copies of the region spanned by $J(x)$ in Fig. 1 placed one beneath the other with total probability 0·5 distributed between them. In this the density of X given θ takes only the values 0·01 and

0·09. Thus $I(x)$ does contain those θ values which have the highest likelihood. It is also shortest in the sense of Neyman (1937).

Equations (1) and (6) can be derived as for the first two examples. To derive (7) we must note that

$$\mathrm{pr}\{X \notin A \text{ and } \theta \notin I(X)|\theta\} \leqslant 0{\cdot}1$$

and that

$$\mathrm{pr}\{X \in A \text{ and } \theta \notin I(X)|\theta\} \geqslant 0{\cdot}4.$$

Now since (1), (6) and (7) are true for this family of distributions the criticism in § 2 of $I(x)$ as a 50 % confidence region for θ remains in force.

5. Further changes to the family of distributions

Four changes which are of some interest could be made to any of the three preceding examples.

(*a*) The shallow gradient sections, for example $I(x)$ between $x = 0$ and $x = 1$, could be made of even smaller gradient. Thus we could make the bounds on $\mathrm{pr}\{X \in A \text{ and } \theta \in I(X)|\theta\}$ and $\mathrm{pr}\{X \notin A \text{ and } \theta \notin I(X)|\theta\}$ as small as desired. Hence we could make the upper bound on $\mathrm{pr}\{\theta \in I(X)|X \in A, \theta\}$ arbitrarily small and make the lower bound on $\mathrm{pr}\{\theta \in I(X)|X \notin A, \theta\}$ arbitrarily near to unity.

(*b*) The probability content of $I(x)$ could be changed, making it a confidence interval at some other confidence level. The set A and its complement would remain relevant subsets.

(*c*) The density $f(x|\theta)$ could be changed to make it nonzero everywhere and infinitely differentiable with respect to both of its arguments with arbitrarily little effect on (1), (6) and (7).

(*d*) A transformation could be applied to x so that the contours of $F(x|\theta)$ within $I(x)$ became straight lines. Alternatively, the contours outside $I(x)$ could be made straight lines. This shows that the anomalous behaviour is not solely either inside or outside $I(x)$.

6. Applicability of these counterexamples to fiducial probability

The theory of confidence intervals and Fisher's fiducial argument perform similarly for one-dimensional problems like the examples in this paper. Thus my examples are also counterexamples for fiducial theory.

Fisher (1956*b*) criticized Welch's proposed solution of the two means problem on the grounds of its conditional behaviour. Buehler (1959) has shown that, in his terminology, the criticism is that a negatively biased relevant subset exists. It is apparent from Fisher's criticism and from some of his statements in Fisher (1956*a*, p. 55) that he believed that his own theory would not allow such a subset to exist. Yates (1964) argued that Buehler and Fedderson's positively biased relevant subset for a confidence/fiducial interval based on the t distribution does not contradict the fiducial argument because the complementary set is not a negatively biased relevant subset. His defence is not valid for the examples in this paper.

References

Basu, D. (1964). Recovery of ancillary information. *Sankhyā* **26**, 3–16.

Buehler, R. J. (1959). Some validity criteria for statistical inference. *Ann. Math. Statist.* **30**, 845–63.

Buehler, R. J. & Fedderson, A. P. (1963). Note on a conditional property of Student's *t*. *Ann. Math. Statist.* **34**, 1098–100.

Fisher, R. A. (1956*a*). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1956*b*). On a test of significance in Pearson's *Biometrika* Table No. 11. *J. R. Statist. Soc.* B **18**, 56–60.

Kendall, M. G. & Stuart, A. (1961). *The Advanced Theory of Statistics*, Vol. 2. London: Griffin.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. R. Soc.* A **236**, 333–80.

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika* **32**, 128–50.

Pierce, D. A. (1973). On some difficulties in a frequency theory of inference. *Ann. Statist.* **1**, 241–50.

Pitman, E. J. G. (1938). The estimation of location and scale parameters of a continuous population of any given form. *Biometrika* **30**, 391–421.

Welch, B. L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location. *Ann. Math. Statist.* **10**, 58–69.

Yates, F. (1964). Fiducial probability, recognisable subsets and Behrens' test. *Biometrics* **20**, 343–60.

[*Received February* 1974. *Revised October* 1974]