Statistical Methodology 9 (2012) 555–561

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Short communication

Revisiting the Berger location model: Fallacious confidence interval or a rigged example?

Aris Spanos *

Department of Economics, Virginia Tech, Blacksburg, VA 24061, United States

ARTICLE INFO

Article history:
Received 13 March 2011
Received in revised form
27 February 2012
Accepted 27 February 2012

Keywords:
Non-regular probability model
Support set
Error probability
Entrapment probability
Pre-data versus post-data information

ABSTRACT

Since the 1960s the Bayesian case against frequentist inference has been partly built on several "classic" examples which are devised to show how frequentist inference procedures can give rise to fallacious results; see Berger and Wolpert (1988) [2]. The primary aim of this note is to revisit one of these examples, the Berger location model, that is supposed to demonstrate the fallaciousness of frequentist Confidence Interval (CI) estimation. A closer look at the example, however, reveals that the fallacious results stem primarily from the problematic nature of the example itself, since it is based on a non-regular probability model that enables one to (indirectly) assign probabilities to the unknown parameter. Moreover, the proposed confidence set is not a proper frequentist CI in the sense that it is not defined in terms of legitimate error probabilities.

1. Introduction

The Bayesian case against frequentist inference is partly based on a number of carefully selected examples intended to call into question the soundness of frequentist reasoning by demonstrating the fallacious results that it gives rise to; see [4,5,12] inter alia.

The aim of this paper is to revisit one of these examples in [2], that was initially introduced by Berger [1] as a simplified version of the simple uniform model proposed by Welch [11]. The primary argument of this paper is that whatever the merit of the Bayesian case against frequentist inference, it is not enhanced by such examples. This is because what is *not* mentioned in these discussions is the *problematic nature* of the examples themselves.

It is argued that the Berger location example is 'rigged' by inducing a non-regularity in the form of a dependence of the probabilistic model's support on the unknown parameter θ. This creates a situation

* Tel.: +1 540 2317981.
E-mail address: aris@vt.edu.

1572-3127/$ – see front matter
doi:10.1016/j.stamet.2012.02.004

Table 1
Distribution of (Z_1, Z_2).

$z_1 \setminus z_2$	1	-1	$f(z_1)$	
1	0.25	0.25	0.5	
-1	0.25	0.25	0.5	(3)
$f(z_2)$	0.5	0.5	1	

where one can indirectly attach probabilities to the unknown parameter θ by blurring the distinction between the random variables (X_1, X_2) and their observed values (x_1, x_2). This induced non-regularity and the misinterpretation of relevant frequentist error probabilities are the real culprits behind the nonsensical confidence set. Viewed from a frequentist perspective, the latter makes no sense as a proper confidence interval.

2. A 'rather silly' confidence set

The centerpiece of the discussion that follows is the example from [2, pp. 5–6]:

Example 1. Suppose that X_1 and X_2 are independent and

$$P_\theta(X_i = \theta - 1) = P_\theta(X_i = \theta + 1) = \frac{1}{2}, \quad i = 1, 2. \tag{1}$$

Here $-\infty < \theta < \infty$ is an unknown parameter to be estimated from X_1 and X_2. It is easy to see that a 75% confidence set of smallest size for θ is

$$C(X_1, X_2) = \begin{cases} \text{the point } \dfrac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ \text{the point } (X_1 - 1) & \text{if } X_1 = X_2. \end{cases} \tag{2}$$

Thus, if repeatedly used in this problem, $C(X_1, X_2)$ would contain θ with probability 0.75.

Notice, however, that when $x_1 \neq x_2$ it is *absolutely certain* that $\theta = \frac{1}{2}(x_1 + x_2)$, while when $x_1 = x_2$ it is equally uncertain whether $\theta = \frac{1}{2}(x_1 - 1)$ or $\theta = \frac{1}{2}(x_1 + 1)$ (assuming no prior knowledge about θ). Thus, from a post-experimental viewpoint, one would say that $C(X_1, X_2)$ contains θ with "confidence" 100% when $x_1 \neq x_2$, but only with "confidence" 50% when $x_1 = x_2$. Common sense certainly supports the post-experimental view here. It is technically correct to call $C(X_1, X_2)$ a 75% confidence set, but if after seeing the data we know whether it is really a 100% or a 50% set, reporting 75% seems rather silly.

The issue to be discussed in what follows is not whether this is a silly confidence set (it is), but whether:

(a) the silliness stems from an inherent problem with frequentist CIs or from other sources, and
(b) the above confidence set constitutes a legitimate frequentist CI.

3. Induced non-regularity in a probability model

To shed some light on the above example, let us begin the discussion with two random variables (r.v.'s) Z_1 and Z_2 assumed to be Independent and Identically Distributed (IID) with a joint distribution given in Table 1.

Note that $E(Z_1) = E(Z_2) = 0, \mathrm{Var}(Z_1) = \mathrm{Var}(Z_2) = 1$ and the *sample space* for (Z_1, Z_2) is $\mathcal{Z} = \{-1, 1\} \times \{-1, 1\}$.

The latent r.v.'s (X_1, X_2) in the Berger location model can be viewed as arising from a one-to-one transformation of the r.v.'s (Z_1, Z_2) of the form

$$X_1 = Z_1 + \theta, \qquad X_2 = Z_2 + \theta. \tag{4}$$

Hence, the joint distribution of (X_1, X_2) takes the form given in Table 2.

Note that $E(X_1) = E(X_2) = \theta$, $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1$. The crucial difference between the distributions of Tables 1 and 2 is that the former is regular but the latter is *non-regular* in the sense

Table 2
Distribution of (X_1, X_2).

$x_1 \setminus x_2$	$\theta+1$	$\theta-1$	$f(x_1)$	
$\theta+1$	0.25	0.25	0.5	
$\theta-1$	0.25	0.25	0.5	(5)
$f(x_2)$	0.5	0.5	1	

that the support of $f(x_1, x_2; \theta)$:

$$\mathbb{R}_{\mathbf{X}}(\theta) = \{(x_1, x_2) : f(x_1, x_2; \theta) > 0\},$$
$$\text{for } (x_1, x_2) \in \mathcal{X} := \{\theta - 1, \theta + 1\} \times \{\theta - 1, \theta + 1\}, \tag{6}$$

depends on θ; see [3, p. 112]. Note that non-regularity was introduced via the transformation (4) that has created an overlap between the sample ($\mathcal{X}$) and parameter ($\theta \in \Theta := \mathbb{R}$) spaces.

Such non-regularity creates a variety of problems in frequentist inference including the non-applicability of Maximum Likelihood procedures in both estimation and testing; see [8]. An equally problematic, but somewhat different, problem arising in the Berger location example is that the induced *overlap* gives rise to the opportunity to assign probabilities to θ using a subtle sleight of hand. It is shown below that this overlap plays a key role in producing the "rather silly" confidence set mentioned in the quotation above.

Due to the non-regularity, the *post-data* support of $f(x_1, x_2; \theta)$ can be used to convey information pertaining to θ. In particular, for the observed values x_1 and x_2 of the r.v.'s X_1 and X_2, one of two situations holds:

$$\text{when } x_1 \neq x_2, \quad \text{either } x_1 = (\theta - 1), x_2 = (\theta + 1) \quad \text{or} \quad x_1 = (\theta + 1), x_2 = (\theta - 1).$$

This means that *post-data* (after the data come in) it is known with certainty that

$$\text{when } x_1 \neq x_2, \qquad (x_1 - x_2) = \pm 2 \quad \text{and} \quad (x_1 + x_2) = 2\theta \Rightarrow \theta = \frac{1}{2}(x_1 + x_2). \tag{7}$$

It is important, however, to emphasize that this information pertains to the observed values (x_1, x_2) and not the underlying r.v.'s (X_1, X_2) . This is because it concerns the *post-data* support of $f(x_1, x_2; \theta)$, and holds *irrespective of its probabilistic assignments*. That is, one can change the probabilities in (5) to any non-negative numbers that add up to 1 without affecting this result. In this sense, $\theta = \frac{1}{2}(x_1 + x_2)$ is essentially *deterministic information*. Hence, the move from the post-data relationship $\theta = \frac{x_1+x_2}{2}$ to viewing $\theta = \frac{1}{2}(X_1 + X_2)$ as a legitimate "confidence" set is highly questionable. It amounts to transforming legitimate post-data deterministic information pertaining to the support of $f(x_1, x_2; \theta)$ into illegitimate probabilistic assignments pertaining to θ. This can be seen more clearly in the case of the [11] model which is more realistic; see [9]. Hence, the claim by Berger and Wolpert that "It is technically correct to call $C(X_1, X_2)$ a 75% confidence set" is called seriously into question.

Returning to the above quotation, the sleight of hand can be seen by comparing the use of capital letters to denote the r.v.'s (X_1, X_2) in stating the apparent CI in (2), and the switch to the small letters (x_1, x_2), denoting the particular observed values.

4. What constitutes a proper frequentist CI?

In frequentist inference, θ is an unknown parameter, and for regular probability models it is unrelated to the support of the density function $f(x; \theta)$. To bring out the differences between a proper and an improper CI consider the following regular model.

Example 2. Consider *the simple Normal model*:

$$X_k \backsim \mathsf{NIID}(\theta, \sigma^2), \quad k = 1, 2, \ldots, n, \tag{8}$$

where 'NIID (θ, σ^2)' stands for 'Normal, IID with mean θ and variance σ^2'. This defines a *regular* statistical model since the support $\mathbb{R}_X = \{x: f(x) > 0\} = \mathbb{R} := (-\infty, \infty)$ is free of θ.

Let us consider the question of constructing an optimal CI for θ. Assuming that σ^2 is *known* for simplicity, it can be shown that an optimal CI can be based on the pivotal quantity $d(\mathbf{X};\theta)=\frac{\sqrt{n}(\overline{X}_n-\theta)}{\sigma}$, whose sampling distribution under the 'true state of nature' ($\theta=\theta^*$) is

$$d(\mathbf{X};\theta)=\frac{\sqrt{n}(\overline{X}_n-\theta)}{\sigma}\overset{\theta=\theta^*}{\backsim}\mathsf{N}\left(0,1\right), \tag{9}$$

where θ^* denotes the true value of θ—whatever that happens to be. This result follows from the sampling distribution of the estimator $\overline{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k$ taking the form

$$\overline{X}_n\backsim\mathsf{N}\left(\theta,\frac{\sigma^2}{n}\right), \tag{10}$$

since $n\overline{X}_n$ is the sum of n NIID random variables. The result in (9) can be used to define a $(1-\alpha)$ confidence interval (CI) for θ:

$$\mathbb{P}\left(\overline{X}_n-c_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\leq\theta\leq\overline{X}_n+c_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}};\theta=\theta^*\right)=1-\alpha, \tag{11}$$

where $-c_{\frac{\alpha}{2}}$ denotes the $\frac{\alpha}{2}$ quantile of $\mathsf{N}(0,1)$; see [8].

The main features of a proper frequentist CI

(i) In principle one can define an infinity of $(1-\alpha)$ CIs but the overwhelming majority are likely to be improper in the sense that they do not satisfy even minimal optimal properties. It turns out that the optimality of CIs is closely related to that of Neyman–Pearson (N–P) hypothesis testing in the sense that there is a duality between their respective properties. For example, a uniformly most powerful test can be shown to be the dual to a Uniformly Most Accurate CI; see [6]. Hence, the choice of the interval in (11) is made on *optimality* grounds, which minimally includes the property of *consistency*, in the sense that the coverage probability is inherently a function of the sample size n with the width of the CI decreasing to zero as $n\to\infty$. In the case of an N–P test, consistency refers to its power at any nonzero discrepancy from the null increasing to one $n\to\infty$; see Rao [8].

(ii) The coverage probability of a properly specified $(1-\alpha)$ CI relates to a hypothetical sequence of a reasonably large number of realizations of the sample $(X_1,X_2,\ldots,X_n)$, rendering illegitimate any assignment of probabilities to an *observed* $(1-\alpha)$ CI:

$$\left(\overline{x}_n-c_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{x}_n+c_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right). \tag{12}$$

This is because post-data, the 'true state of nature' scenario, underlying the frequentist confidence interval reasoning in (9), has already played out and the relevant error probability becomes degenerate in the sense that any observed CI either includes or excludes θ^*; see [7].

(iii) A $(1-\alpha)$ CI does *not* attach – directly or indirectly – any probabilities to θ, since it is treated as an unknown constant whose true value, θ^* (whatever that happens to be), is the focus of the inference. That is, probabilities are always calibrating the errors associated with the inference procedure in question and are firmly attached to the random bound(s) $\overline{X}_n\pm c_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$; never to θ.

5. Revisiting the Berger 'confidence set'

To bring out the nature of the 'questionable' confidence set in (2), let us compare it with the proper CI in (11) using the more precise formulation

$$\begin{aligned}\Pr(\theta\in C(X_1,X_2)) &= P(X_1=X_2)\Pr(\theta=x_1-1)\\ &\quad+P(X_1\neq X_2)\Pr\left(\theta=\frac{x_1+x_2}{2}\right)=0.75,\end{aligned} \tag{13}$$

where the random interval $C(X_1,X_2)$ takes the form

$$C(X_1,X_2)=\begin{cases}\text{the point }\dfrac{1}{2}(X_1+X_2) & \text{if } x_1\neq x_2\\ \text{the point }(X_1-1) & \text{if } x_1=x_2.\end{cases} \tag{14}$$

(i)* The choice of the Berger confidence set in (13) does not stem from any optimality criteria. Viewing $\frac{1}{2}(X_1+X_2)$ as the implicit estimator of θ, it is clear that it is a terrible estimator because it is *inconsistent*! By the same token one can claim that

$$\mathbb{P}(X_1-1\leq\theta\leq X_1+1)=1,$$

provides a sure 'confidence set' for θ, but this is practically useless. Analogously, in the case of the simple Normal model, nobody would seriously entertain specifying a $(1-\alpha)$ CI by replacing $\overline{X}_n$ in (10) with the estimator

$$\widetilde{X}_2=\frac{1}{2}(X_1+X_n)\backsim \mathsf{N}\left(\theta,\frac{\sigma^2}{2}\right),\tag{15}$$

primarily because $\widetilde{X}_2$ is an inconsistent estimator of θ, giving rise to an inconsistent CI. In contrast, $\overline{X}_n$ is not just a consistent estimator of θ, but also fully efficient and sufficient; see Rao [8]. Hence, just because one can standardize $\widetilde{X}_2$ to define the pivotal quantity

$$\frac{\widetilde{X}_2-E(\widetilde{X}_2)}{\sqrt{\mathrm{Var}(\widetilde{X}_2)}}=\frac{\sqrt{2}(\widetilde{X}_2-\theta)}{\sigma}\overset{\theta=\theta^*}{\backsim}\mathsf{N}(0,1),\tag{16}$$

this does *not* mean that the latter can be used to derive a 'sensible' frequentist CI. What is 'sensible' is defined in terms of its optimal properties which should at least include consistency as a minimal reliability property. The $(1-\alpha)$ confidence interval (CI) for θ stemming from (16):

$$\mathbb{P}\left(\widetilde{X}_2-c_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{2}}\leq\theta\leq\widetilde{X}_2+c_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{2}};\theta=\theta^*\right)=1-\alpha,\tag{17}$$

is clearly *inconsistent*. Moreover, for reasonably large sample sizes, say $n\geq 50$, the width of the interval (17) will be at least 5 times larger than that of (11), rendering the former practically useless for inference purposes. Hence, nobody would seriously propose (17) as a sensible confidence interval.

Returning to the Berger location example in (2), it is clear that, like for the CI in (17), the sample size plays no role in increasing the precision of inference as $n\to\infty$. Worse, unlike (17) the Berger location example in (2) does not even have legitimate error probabilities. Indeed, ignoring the notion of what constitutes a proper frequentist CI plays a key role in misconstruing illegitimately induced 'entrapment' probabilities as genuine error probabilities. Properly defined error probabilities provide the backbone of all frequentist inference, and dismissing them foregoes any possibility for an optimal theory of interval estimation or testing; see [10].

(ii)* The Berger confidence set in (13) *does* depend on the particular values taken by X_1 and X_2, which seems highly paradoxical. How is that achieved? By a sleight of hand! A closer look at the above quotation from [2] reveals that the two scenarios in (2) are supposed to be specified in terms of the random variables $X_1\neq X_2$ and $X_1=X_2$, but this makes no probabilistic sense without the observed values. The only way one can specify the two scenarios is in terms of different values $x_1\neq x_2$ and $x_1=x_2$; hence the restatement in (14). This key problem can be seen more clearly in the case of the Welch uniform example [9], being more realistic than the above oversimplified Berger location model.

The sleight of hand becomes less obscure by unveiling the way that one is supposed to derive the probability 0.75 in (2):

$$\begin{aligned}&\Pr(\theta\in C(X_1,X_2))\\&\quad=P((x_1,x_2):x_1=x_2)\Pr(\theta=x_1-1)+P((x_1,x_2):x_1\neq x_2)\Pr\left(\theta=\frac{x_1+x_2}{2}\right)\\&\quad=\frac{1}{2}\Pr(\theta=x_1-1)+\frac{1}{2}\Pr\left(\theta=\frac{x_1+x_2}{2}\right)=\frac{1}{2}\left(\frac{1}{2}\right)+\frac{1}{2}(1)=0.75.\end{aligned}\tag{18}$$

This brings us conveniently to the last feature.

(iii)* One can see from (18) that the Berger confidence set in (13) *does* in fact attach probabilities to θ. To bring that out more clearly, the probabilistic assignments in (18) are given two different symbols,

$P(\cdot)$ and $\Pr(\cdot)$. $P(\cdot)$ assigns probabilities to legitimate events pertaining to the r.v.'s (X_1, X_2), and $\Pr(\cdot)$ assigns (illegitimate) probabilities to θ by:

[a] abusing the non-regularity, and

[b] blurring the distinction between the r.v.'s (X_1, X_2) and their observed values (x_1, x_2). But where do the probabilities $\Pr(\theta = x_1 - 1)$ and $\Pr\left(\theta = \frac{x_1+x_2}{2}\right)$ come from?

In the case of Example 1, the non-regularity has created a situation where one seems to be able to 'legitimately' attach probabilities to θ since

x_1	$f(x_1)$		θ	$f(\theta)$	
$\theta+1$	.5	$\Leftrightarrow$	x_1-1	.5	(19)
$\theta-1$	.5		x_1+1	.5	

x_2	$f(x_2)$		θ	$f(\theta)$	
$\theta+1$	.5	$\Leftrightarrow$	x_2-1	.5	(20)
$\theta-1$	.5		x_2+1	.5	

Indeed, this indirect assignment of probabilities to θ lies behind the derivations in (2) alluded to in the above quotation from [2]. The claim "that when $x_1 \neq x_2$ it is *absolutely certain* that $\theta = \frac{1}{2}(x_1+x_2)$" is based on the evaluation

$$\Pr\left(\theta = \frac{x_1+x_2}{2}\right) = 1, \quad \text{for } x_1 \neq x_2. \tag{21}$$

Such an assignment, however, violates every aspect of a proper frequentist CI.

The combination of [a] and [b] enables one to misinterpret the support information in (7) – which is deterministic in nature because it pertains exclusively to the post-data values (x_1, x_2) – as legitimate probabilistic information relating to the corresponding random variable:

$$\frac{X_1+X_2}{2} = \begin{cases} (\theta+1) & \text{if } x_1 = x_2 = (1+\theta) \\ (\theta-1) & \text{if } x_1 = x_2 = (\theta-1) \\ \theta & \text{if } x_1 = (\theta-1),\ x_2 = (1+\theta) \text{ or } x_1 = (\theta+1),\ x_2 = (\theta-1) \end{cases} \tag{22}$$

(21) appears to follow after blurring the distinction between the r.v.'s (X_1, X_2) and their observed values (x_1, x_2). Similarly, the random variable

$$X_1 - 1 = \begin{cases} \theta & \text{if } x_1 = (1+\theta) \\ (\theta-2) & \text{if } x_1 = (\theta-1) \end{cases} \tag{23}$$

is made to appear as a legitimate "confidence" set for θ by imposing the restriction $x_1 = x_2$ and invoking the indirect distribution of θ in (19). This gives rise to the equally questionable claim in (18) that

$$\Pr(\theta = x_1 - 1) = \frac{1}{2}, \quad \text{for } x_1 = x_2. \tag{24}$$

What made the post-data probabilistic assignment in (24) possible was the abuse of the indirect distribution of θ in (19). Hence, the claim that "the point $X_1 - 1$ contains θ with confidence 50% when $x_1 = x_2$" makes no sense in frequentist CI terms. Even if one were to ignore this, (24) could best be described as an 'entrapment' probability brought about by the 'induced' non-regularity by transforming (Z_1, Z_2) into (X_1, X_2).

6. Conclusion

In frequentist inference a probabilistic statement of the form

$$\Pr(\theta = \theta_0) = p, \tag{25}$$

where θ_0 is a known value is *meaningless*, irrespective of where the probabilities or the value θ_0 come from. A closer look at the Berger confidence set in (2) reveals that it is specified in terms of such illegitimate probabilistic assignments.

The problematic nature of (25) stems primarily from three interrelated sources:

(i) the *non-regularity* in (6) that renders the support of $f(x_1, x_2; \theta)$ a function of θ and creates an overlap between the sample and parameter spaces making it 'legitimate' to (indirectly) assign probabilities to θ,
(ii) the misuse of post-data information stemming from (i) as a pre-data assignment of probabilities to θ, and
(iii) the misinterpretation of 'entrapment' probabilities arising from (i) and (ii) as legitimate error probabilities.

Hence, the argument that the "silly" confidence set results from an inherent conceptual problem with frequentist CIs is rather misplaced and contributes nothing to the Bayesian case against frequentist inference.

References

[1] J.O. Berger, In defence of the likelihood principle: axiomatics and coherency, in: J.M. Bernardo, M.H. DeGroot, D. Lindley, A. Smith (Eds.), Bayesian Statistics II, North-Holland, Amsterdam, 1984.
[2] J.O. Berger, R.W. Wolpert, The likelihood principle, in: Institute of Mathematical Statistics, second ed., in: Lecture Notes—Monograph Series, vol. 6, Hayward, California, 1988.
[3] D.R. Cox, D.V. Hinkley, Theoretical Statistics, Chapman & Hall, London, 1974.
[4] M.H. DeGroot, M.J. Schervish, Probability and Statistics, third ed., Addison-Wesley, New York, 2002.
[5] J.K. Ghosh, M. Delampady, T. Samanta, An Introduction to Bayesian Analysis: Theory and Methods, Springer, New York, 2006.
[6] E.L. Lehmann, Testing Statistical Hypotheses, second ed., Wiley, New York, 1986.
[7] J. Neyman, Note on an article by sir Ronald Fisher, Journal of the Royal Statistical Society B 18 (1956) 288–294.
[8] C.R. Rao, Linear Statistical Inference and Its Applications, second ed., Wiley, New York, 1973.
[9] A. Spanos, Revisiting the Welch uniform model: a case for conditional inference? Advances and Applications in Statistical Sciences 5 (2011) 33–52.
[10] A. Spanos, Misplaced criticism of Neyman–Pearson (N–P) testing in the case of two simple hypotheses, Advances and Applications in Statistical Sciences 5 (2011) 229–242.
[11] B.L. Welch, On confidence limits and sufficiency, and particular reference to parameters of location, Annals of Mathematical Statistics 10 (1939) 58–69.
[12] G.A. Young, R.L. Smith, Essentials of Statistical Inference, Cambridge University Press, Cambridge, 2005.