On After-Trial Properties of Best Neyman-Pearson Confidence Intervals

Teddy Seidenfeld

Philosophy of Science, Vol. 48, No. 2. (Jun., 1981), pp. 281-291.

Stable URL:
http://links.jstor.org/sici?sici=0031-8248%28198106%2948%3A2%3C281%3AOAPOBN%3E2.0.CO%3B2-A

Philosophy of Science is currently published by The University of Chicago Press.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ucpress.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Tue Mar 4 10:36:37 2008

ON AFTER-TRIAL PROPERTIES OF BEST NEYMAN-PEARSON CONFIDENCE INTERVALS*

TEDDY SEIDENFELD†

Department of Philosophy
Washington University, St. Louis

On pp. 55–58 of *Philosophical Problems of Statistical Inference* (Seidenfeld 1979), I argue that in light of unsatisfactory after-trial properties of "best" Neyman-Pearson confidence intervals, we can strengthen a traditional criticism of the orthodox N-P theory. The criticism is that, once particular data become available, we see that the pre-trial concern for tests of maximum power (and for their derivative confidence intervals of shortest expected length) may then misrepresent the conclusion of such a test (or interval estimate). Specifically, I offer a statistical example where there exists a Uniformly Most Powerful test (a UMP-test), a test of highest N-P credentials, which generates a system of "best" confidence intervals (the $[CI_\lambda]$ interval system) with exact confidence coefficients. But the $[CI_\lambda]$ intervals have the unsatisfactory feature that, for a recognizable set of outcomes, the interval estimates cover *all* parameter values consistent with the data, at strictly less than 100% confidence.

Even by Neyman's standards, there is a probability for such a trivial interval estimate given the data and statistical model. To wit, when the interval estimate covers all parameter values consistent with the data and model, the probability is 1 that the unknown parameter (perhaps a constant, perhaps a random variable with unknown "prior" probability) falls within the interval. To quote Neyman on this point,

> If θ is a constant, then whatever $a < b$, and $\boldsymbol{B}$, the probability $\boldsymbol{P}\{a \leq \theta \leq b/\boldsymbol{B}\}$ may have only values unity or zero according to whether θ falls in between a and b or not. (Neyman 1937, p. 256)

Thus, the system of "best" confidence intervals (best according to Neyman's standards) generates particular interval estimates which, though *known* with probability 1, carry a confidence coefficient of less than

*Received January 1981.

†I thank Isaac Levi and Carl Posy for helpful comments on an earlier draft of this paper. Also, I appreciate the opportunity to read and discuss Professor Mayo's paper (Mayo 1981) with her and Ronald Giere in advance of its publication.

Philosophy of Science, 48 (1981) pp. 281–291.
Copyright © 1981 by the Philosophy of Science Association.

100%. My concern, then, over trivial interval estimates reflects the tension between the confidence level (less than 100%) and a known probability (of exactly 100%).

Neyman's theory of confidence interval estimation was designed, as he reports, to avoid the variety of perceived deficiencies in solutions to problems of estimation advanced by others. [Neyman 1937, §1] In particular, with regard to the Bayesian strategy for solving interval estimation problems by Bayes' theorem, i.e., for calculating probabilities of hypotheses about the unknown parameters *h*, given statistical data *d*, using the theorem: $\boldsymbol{P}(h/d) \propto \boldsymbol{P}(d/h) \cdot \boldsymbol{P}(h)$, Neyman's objections focus on the "prior" probability component "$\boldsymbol{P}(h)$." He summarizes his dissatisfaction with this approach as follows:

> It is known that, as far as we work with the conception of probability as adopted in this paper, the above [Bayesian] theoretically perfect solution may be applied in practice only in quite exceptional cases, and this for two reasons:
>
> (*a*) It is only very rarely that the parameters . . . are random variables. They are generally unknown constants and therefore their probability law *a priori* has no meaning.
>
> (*b*) Even if the parameters to be estimated . . . could be considered as random variables, the elementary probability law *a priori* . . . is usually unknown, and hence the formula [Bayes' Theorem] cannot be used because of the lack of the necessary data. (Neyman 1937, p. 258)

Unfortunately, the two reasons quickly collapse into one, as Neyman notes that even "constants" have degenerate prior probability distributions concentrated on the two extreme probability values 0 and 1.

> It is true that any constant, ξ, might be formally considered as a random variable with the integral probability law $\boldsymbol{P}\{a \leq \xi < b\}$ having only values unity or zero according to whether ξ falls between *a* and *b* or not. (Neyman 1937, p. 257)

Thus, I have interpreted Neyman's objection to the Bayesian solution to be based on the accurate observation that typically the investigator is *ignorant* of any precise, frequency based (chance based) prior probability for the unknown quantities (parameters). (Seidenfeld, pp. 29–36) That is, I understand Neyman to reject the Bayesian solution since, in most cases, the investigator has merely *indeterminate* knowledge of "prior" chances for the unknown quantities.[1] However, this criticism of Bayesian

[1] On this point, I wish to correct a misprint on page 35 of my book. In the first sentence of the last paragraph, '*intermediate*' should read '*indeterminate*'.

inference does *not* excuse Neyman from paying attention to consequences of the "theoretically perfect solution," where those consequences are independent of any specific "prior". That is, the tension between confidence intervals which cover the full parameter space at less than 100% confidence and the *known* probability 1 for such interval estimates reflects a conflict between Neyman's recommended solution and a consequence of the Bayesian solution that is independent of the controversial "prior" probability.

In my book, I conclude the discussion of the statistical example (where the N-P "best" intervals are recognizably trivial) by pointing out there exists an alternative system of confidence sets, denoted $[\mathrm{CI}_{\mathrm{alt.}}]$, also with exact confidence levels, which dominates the "best" intervals for the undesirable property of covering the full parameter space at less than 100% confidence. However, the alternative system $[\mathrm{CI}_{\mathrm{alt.}}]$ is generated from a severely biased N-P family of tests, tests of lowest N-P standards.[2] That is, I present a *reductio* argument against the thesis that $[\mathrm{CI}_{\lambda}]$ is the best system of estimates (which it is by Neyman's standards), since one may improve on the N-P "best" intervals for the purpose of minimizing the set of observations leading to trivial intervals. In other words, by shifting from the N-P "best" estimates to one deemed "worse than useless", we can reduce the instances of conflict between a known probability and the confidence coefficient.

In "In Defense of the Neyman-Pearson Theory of Confidence Intervals", D. Mayo expresses several points of dissent with the analysis I have provided on this matter. Most general, and to my mind most central, is her claim that my concern with interval estimates that cover the full parameter space at less than 100% confidence, "trivial" intervals, reflects an illegitimate interpretation of confidence intervals based on an inappropriate concern for "measures of final precision". For example, she says,

> It must be stressed, however, that having seen the value x, NP theory *never* permits one to conclude that the specific confidence interval formed covers the true value of θ with either $(1 - \alpha)$ 100% probability or $(1 - \alpha)$ 100% degree of confidence. (Mayo 1981, p. 272)

But, as I have argued (above), even on Neyman's conception of probability there is an acceptable probability for the trivial intervals. They carry a known probability 1. Thus, I dispute Mayo's assertion that, in focusing

[2] Confidence intervals at the $(1 - \alpha)$ level can be generated from families of hypothesis tests with size α. The estimate is formed, for particular data, as the union of (null) values, corresponding to null hypotheses, left unrejected by those observations. The reader should note that, in generating estimates from families of hypothesis tests, tests biased to one side of the null hypothesis yield estimates biased on the other side of the true parameter value.

on the triviality of certain N-P estimates, I rely on an illegitimate interpretation of Neyman's theory. In fact, I chose to attend to those cases exactly because they admit known probabilities (in conflict with their confidence level), where the probabilities satisfy Neyman's constraints and avoid his objections to "priors".

Mayo adds to this general criticism a number of objections to my analysis of the specific statistical example I construct. In particular, she alleges that: (*A*) I misidentify $[\mathrm{CI}_\lambda]$ as the N-P "best" system of interval estimates; (*B*) a different system, her $[\mathrm{CI}_0]$, is the N-P "best" one for the problem; and (*C*) since $[\mathrm{CI}_0]$ never provides trivial intervals, I have no ground on which to object to N-P theory. I reject each of these claims, and in what follows I offer reasons for my judgment that Mayo has failed to respond to my argument against confidence interval theory. Let me begin with a brief rehearsal of the example.

The statistical problem I develop is a variant of one presented by Neyman in his classic paper (Neyman 1937) on the theory of confidence intervals.[3] In Neyman's version we observe a continuous random variable, X, uniformly distributed on the closed interval $[0,\theta]$, with $\theta > 0$. As Neyman points out, there is a family of UMP-tests for a simple (null) hypothesis h_0: $\theta = \theta_0$, against the composite alternative $\theta \neq \theta_0$. The family of UMP-tests generates the $[\mathrm{CI}_\lambda]$ system of confidence intervals, which are best according to Neyman's standard for minimizing the probability of including false values of the parameter.[4] Equivalently, the $[\mathrm{CI}_\lambda]$ intervals have uniformly shortest expected length.[5]

If we truncate the parameter space by setting an upper bound, $0 < \theta \leq \bar{\theta}$, we arrive at the desired variant of Neyman's original problem. Since the $[\mathrm{CI}_\lambda]$ intervals are based on a family of UMP-tests, and since such tests retain their optimum properties even when the space of alternative parameter values is truncated, the truncated $[\mathrm{CI}_\lambda]$ interval system (see figure 1) remains (uniquely) "best" in Neyman's sense. That is, the truncated $[\mathrm{CI}_\lambda]$ system has minimum probability of covering false parameter values and has uniformly shortest expected length. However, for sample points $x \geq X$ (see figure 1, p. 288), the $[\mathrm{CI}_\lambda]$ interval is trivial, i.e., it covers all parameter values consistent with the data and model at *less* than 100% confidence. For instance, if we set a .95 confidence level and upper bound $\bar{\theta} = 15$, then for all $x \geq 3/4$ the truncated $[\mathrm{CI}_\lambda]$ interval estimate is $[x,15]$, which is known to cover the true value of θ with probability 1.

[3] (Neyman 1937, pp. 269–74) The reader is alerted to inaccuracies in Neyman's formulas on p. 271, as shown to me by H. Kyburg. Corrections are given on p. 53 of my book.

[4] Neyman defends this criterion on p. 282 of (Neyman 1937).

[5] The equivalence is demonstrated in (Ghosh 1961) and (independently) (Pratt 1961).

In section 3 of her paper, Mayo finds that the [CI$_\lambda$] system is not "best" when the parameter space contains the upper bound. She says,

> In the case were θ is truncated from above, however, it seems that a one-sided test would generate a more appropriate confidence interval; namely, one which is one-sided. (Mayo, p. 274)

and,

> I suggest that in the situation where θ is truncated from above, a one-sided (lower) confidence interval is called for. (Mayo, p. 274)

It is on the basis of this suggestion that Mayo forwards her [CI$_0$] system as the "best" N-P candidate solution.

Though yielding a "one-sided" estimate (an upper bound) for θ, the [CI$_\lambda$] system is based on a "two-sided" test in the sense that one attempts to minimize the probability of including false values of the parameter in the estimate, be those values above or below the true parameter value. That is, one pays equal attention to errors in estimation that arise from extending the estimate unnecessarily far above or below the true value. In a "one-sided" test, and derivative interval estimates, one discounts errors in one direction, i.e., the demands of the problem are such that one may ignore errors to one side of the true value. As Neyman points out (Neyman 1937, pp. 284–86) and as Mayo reminds us (Mayo, p. 275) such might be the constraints in an inquiry as to the minimum gain in yield of a new grain over the established one, or an inquiry as to the upper limit of the percentage of defective items in a manufactured batch. But I fail to understand why Mayo thinks that, with the introduction of truncation of the parameter space, the demands for information *must* change so that we no longer care about errors on one side of the true value. Specifically, Mayo's suggestion is that, with the truncation from above, $\theta \leq \bar{\theta}$, we *ought* to discount errors in estimation due to including unnecessarily many false values θ' greater than the true value of θ.

In the example under discussion, the upshot of this recommendation is quite serious from the standpoint of the "biased" status of the intervals Mayo's [CI$_0$] system produces. Just as in the alternative [CI$_{\text{alt.}}$] system I construct for the *reductio* argument (see figure 3), the [CI$_0$] estimates are severely biased for alternatives above the true value of θ (see figure 2, p. 289). Equivalently, both underlying families of hypothesis tests are biased with respect to alternatives below the null value, i.e., with respect to alternatives below h_0: $\theta = \theta_0$ the probability of *rejecting* h_0 is greater when true than when false! Thus, if we maintain the same demands for information in the case of truncation as is assumed in the original version (Neyman's formulation with no upper bound on θ), the [CI$_0$] system is ranked well below the [CI$_\lambda$] system according to the standards proposed by Neyman. Of course, I would insist that there is no regulation dictating that we *must* shift our concerns from "two-sided" to "one-sided" estimation when parameter spaces are truncated. Thus, I do *not* accept Mayo's suggestion that, when an upper bound $\bar{\theta}$ is fixed, [CI_0] is "better" than [CI_λ].[6]

In passing, I note that for many common statistical problems, e.g., estimation of a binomial parameter or estimation of the mean of a normal distribution, there is incentive to use one-sided procedures if the context allows. In such circumstances there are no UMP-tests against the two-sided alternative hypothesis, whereas there are UMP-tests for the one-sided alternatives. However, for the problem discussed here, there *is* a UMP-test for the two-sided alternative hypotheses. Thus, there can be no advantage gained, in terms of increasing the power of tests or decreasing the probability of covering false parameter values in estimates, by shifting from two-sided to one-sided procedures. For the example discussed, there is no improvement afforded by changing standards and discounting errors in estimation due to interval estimates that extend too far above the true value.[7]

Also, I wish to point out that Mayo's [CI_0] fails to be a confidence system, subject to Neyman's requirement that each possible observation

[6] The point is easily pressed. Invariably the investigator knows enough to adopt bounds for parameters, i.e., invariably the parameter spaces can be truncated on theoretical grounds (at least). Does such background information dictate that one-sided procedures are more appropriate than two-sided ones? I see no reason to believe so.

Also, just when one should agree truncation has taken place is open to dispute. Even in the original version of the statistical problem (θ unbounded above), one might argue that 0 (the lower bound for θ) represents truncation—the statistical model can be extended so that, for negative θ, X is uniformly distributed between 0 and θ. (The distribution of X for $\theta = 0$ is arbitrary, say then X is a point-mass with probability concentrated at the point $x = 0$.)

In short, on both practical and theoretical grounds, I find unwarranted Mayo's proposal to modify the concern with errors, i.e., to shift from two-sided to one-sided procedures, in the presence of a truncation in the parameter space. I see no reason to adopt the proposal as a general methodological rule.

[7] Let $\beta < 1$, so that $\beta\theta_0 < \theta_0 = \theta$ (by assumption). Fix the confidence level at $(1 - \alpha)$. Then the probability, given $\theta = \theta_0$, of including (covering) the false value $\beta\theta_0$ with the [CI_0] system of estimation is just the probability of an observation $x \leq (1 - \alpha)\beta\theta_0$, which is the value $(1 - \alpha)\beta$. Similarly, with [CI_λ], the probability of covering the false value $\beta\theta_0$ is just the probability of an observation x satisfying the inequalities: $\alpha\beta\theta_0 \leq x \leq \beta\theta_0$, which also is the value $(1 - \alpha)\beta$. Thus, with respect to false parameter values below the true one, the [CI_0] system is no more accurate than [CI_λ], whereas [CI_λ] is much more accurate with respect to alternative (false) values above the true one.

Moreover, because all (consistent) hypothesis tests are, for this problem, equally unbiased with respect to alternatives *above* the null value, all (consistent) systems of estimation are equally accurate with respect to false values below the true one. Hence, for this problem, shifting concern to one-sided procedures by dismissing errors in estimation for false values above the true one, leads to a situation in which *all* systems of estimation are judged equally accurate!

generate some estimate.[8] As can be seen from figure 2, the [CI_0] system provides no estimate of θ whenever $x > X^*$.[9] I note that the [$\mathrm{CI}_{\mathrm{alt.}}$] system satisfies Neyman's condition (that there always be an estimate) by including an (arbitrarily) narrow "strip of acceptance" along the line ($x = \theta$), see figure 3. In fact, my [$\mathrm{CI}_{\mathrm{alt.}}$] system and her [$\mathrm{CI}_0$] system diffeı only in this respect. Thus, the "arbitrary 'bite' " [Mayo, p. 278], which Mayo finds objectionable in the [$\mathrm{CI}_{\mathrm{alt.}}$] system is due to the satisfaction of a condition proposed by Neyman, a condition [CI_0] stands in violation of.[10]

Lastly, on pp. 58–63 of my book, I offer a rebuttal to the objection discussed here, the objection that estimates labeled "best" by N-P standards may be deficient with respect to the legitimate concern to avoid conflicts between confidence levels and known (precise) probabilities. I base the rebuttal on a novel criterion: confidence equivalence. Perhaps others will find that defense adequate to excuse the triviality of (some) N-P "best" procedures. I do not. Nor do I find Mayo's proposals sufficient for the question at hand.

[8] This is Neyman's condition (ii) (Neyman 1937, p. 267). He uses it to eliminate a candidate estimation system, his #(1), pp. 269–70.

[9] I have recently discovered that R. von Mises observed this same difficulty in the one-sided system [CI_0] and, to some extent, anticipated the discussion of trivial confidence intervals (von Mises 1941, pp. 202–03).

Mayo responds to this technical objection in her fourth footnote (Mayo, pp. 275), where she says,

> Hence whenever $x > (1 - \alpha)c_u$ [CI_0] collapses to the limiting case of the interval; namely, $\theta = c_u$.

This claim is false for the [CI_0] estimation system defined by Mayo (Mayo, p. 274–75). In order to modify the [CI_0] system so that it has this new feature, while retaining its status as a one-sided estimate, one must sacrifice the *exact* confidence level $(1 - \alpha)$, and report (merely) that estimates from the modified [CI_0] system carry a confidence level of *at least* $1 - \alpha$. Of course, intervals that cover all parameter values are not in conflict with known probabilities if they carry the "conservative" (at least) $1 - \alpha$ confidence level. Thus, if the only solution to the problem I raise is to revert to "conservative" estimates, then my objection stands since it would be admitted that the "best" N-P confidence intervals with *exact* confidence levels are deficient.

[10] Mayo states, in connection with her objection to the property (Seidenfeld, p. 57) that sometimes the [$\mathrm{CI}_{\mathrm{alt.}}$] estimates are not intervals, i.e., they might be two (disconnected) intervals,

> . . . it is counterintuitive to accept values of θ_0 both above and below a value of θ which is rejected. (Mayo, p. 278–79)

(This property of [$\mathrm{CI}_{\mathrm{alt.}}$] is equivalent to the existence of, what Mayo calls, a "bite" taken out of the rejection region.) However, I remind the reader that, both in practice and in theory, N-P procedures can recommend estimates with this property. A case in point (discussed in my book) is Fieller's solution to the problem of estimation of the ratio of means for two (independent) normally distributed random variables—a problem with considerable practical significance. Thus, I do not find the "bite" disturbing.

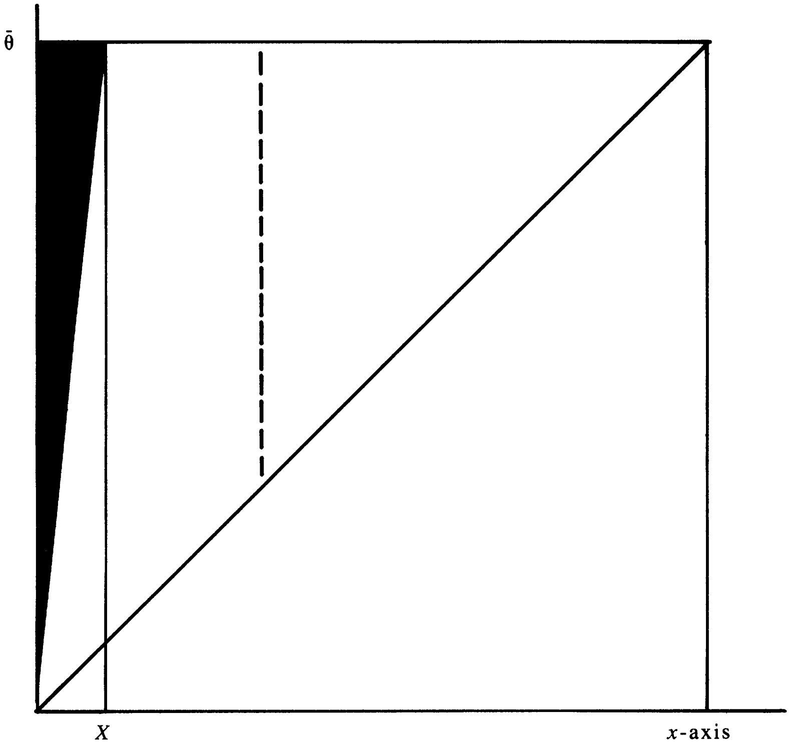

Figure 1

DIAGRAMS

Figures 1, 2, and 3 show the three interval systems [CI$_\lambda$], [CI$_0$], and [CI$_{\text{alt.}}$]. Diagrams are drawn with $\alpha = .1$, so that intervals have a 90% confidence level. In all three figures $\bar{\theta}$ is an upper bound for the parameter θ. The set of possible states (variable, parameter pairs) is the upper right triangle with coordinates (0,0), ($\bar{\theta}$,$\bar{\theta}$), and (0,$\bar{\theta}$). Rejection regions are blackened (for the set of possible states).

The truncated [CI$_\lambda$] "best" confidence intervals. The [CI$_\lambda$] interval (the dashed line) is: $x \leq \theta \leq \min[x/\alpha;\ \bar{\theta}]$. These intervals are trivial for all $x \geq X = \alpha\bar{\theta}$.

The interval system [CI$_0$] proposed by Mayo. There is no estimate of θ if $x > X^* = (1 - \alpha)\bar{\theta}$. If θ_0 is the true parameter value, the system is biased for all false values above θ_0. The bias is maximal for false param-

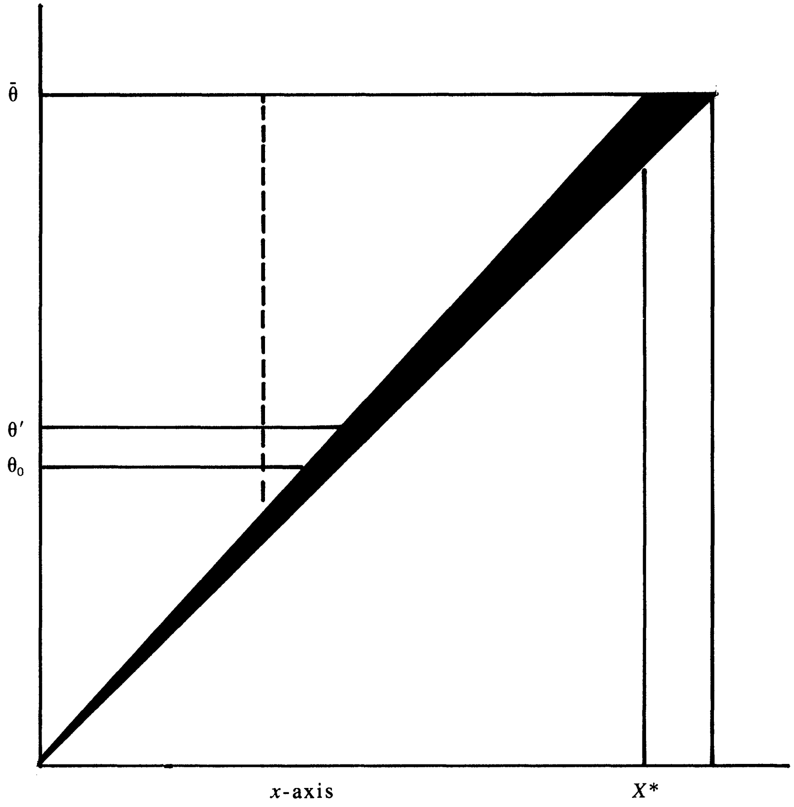

Figure 2

eter values at least as large as $\theta' = \theta_0/(1 - \alpha)$, when every interval estimate includes all such values. The $[\mathrm{CI}_0]$ interval (the dashed line) is:

$$x/(1 - \alpha) \leq \theta \leq \bar{\theta}.$$

The alternative $[\mathrm{CI}_{\mathrm{alt.}}]$ confidence intervals ($\alpha \leq .9$), used for the "reductio" argument. The reader will note that the sole difference between the $[\mathrm{CI}_0]$ and $[\mathrm{CI}_{\mathrm{alt.}}]$ systems is that the latter contain an (arbitrarily) narrow "strip of acceptance" along the diagonal line "$x = \theta$." This allows the $[\mathrm{CI}_{\mathrm{alt.}}]$ to be well defined for all possible observations, unlike the $[\mathrm{CI}_0]$ system.

The $[\mathrm{CI}_{\mathrm{alt.}}]$ interval (set) is: $x \leq \theta \leq \min[x/(1 - [.1 \cdot \alpha]); \bar{\theta}]$ & $x/(1 - [1.1 \cdot \alpha]) \leq \theta \leq \bar{\theta}$. The second interval may be empty. This system

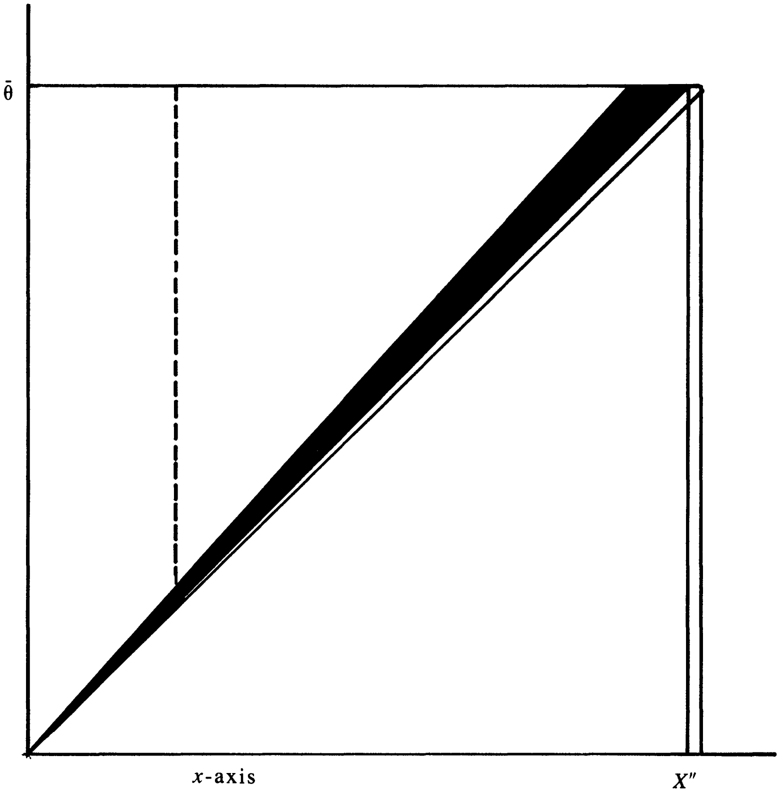

Figure 3

provides trivial intervals only for $x \geq X'' = (1 - [.1 \cdot \alpha])\bar{\theta}$. As required for the "reductio" argument, these interval estimates are severely biased for false values above the true one.

REFERENCES

Ghosh, J. K. (1961), "On the Relation Among Shortest Confidence Intervals of Different Types", *Calcutta Statistical Association Bulletin 10:* 147–152.

Mayo, D. G. (1981), "In Defense of the Neyman-Pearson Theory of Confidence Intervals", *Philosophy of Science 48:* pp. 269–80.

Neyman, J. (1937), "Outline of a theory of statistical estimation based on the classical theory of probability", *Philosophical Transactions of the Royal Society of London* Ser A, **236:** 333–380. Reprinted as paper 20 in *A Selection of Early Statistical Papers of J. Neyman* (1967), Berkeley and Los Angeles: University of California Press, pp. 250–290. Page references are to this reprinting.

Pratt, R. (1961), "Length of Confidence Intervals", *Journal of the American Statistical Association 56,* #295: 549–567.

Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference.* Dordrecht: Reidel.

von Mises, R. (1941), "On the Foundations of Probability and Statistics", *Annals of Mathematical Statistics 12:* 191–205.