9. CONFIDENCE INTERVALS vs BAYESIAN INTERVALS (1976)

We come now to the most polemical of all my articles. There are several reasons for this heated style. Firstly, most of it was written in 1963, as a reply to the 'astonishing article' of Bross referred to, whose polemics make mine seem unimaginative. Indeed, his anti-Bayesian tirade contained nothing but polemics, unsupported by a single technical fact. But had he taken the trouble to read Jeffreys, he would have found demonstrations, on the level of technical fact with no polemics, of the falsity of his charges.

But this distortion of known facts – naturally infuriating to a Bayesian – was hardly limited to Bross. Almost every 'orthodox' textbook written for decades had charged Bayesian methods with nonexistent defects, while ignoring the demonstrable defects in orthodox methods.

It seemed appropriate that a Bayesian point out these things, and so I collected a number of case histories, with mathematical details, from the same areas that Bross had alluded to and showed that, contrary to his assertions, Bayesian methods correct the shortcomings of orthodox methods.

My attempts to get the work published met with rebuff twice from those who had so quickly accepted the Bross article; clearly, different standards of acceptance existed for works differently slanted. After ten years of waiting, the opportunity arrived in an invitation to present a paper at the 1973 London Symposium. But three more years passed before the Proceedings Volume appeared, and so it required thirteen years to get this reply before the public. Still, none of it was obsolete - an interesting commentary on the rate of progress of orthodox thinking.

In the Proceedings Volume, there appear also comments on my presentation by Margaret Maxfield and Oscar Kempthorne, and my replies. Portions of the latter, which extend the main message, are included here.

The reply to Kempthorne also becomes polemical in places, but for an entirely different reason. Just as it was being written, a student of mine went forth into the world with a fresh Ph.D. degree, seeking a teaching position. Since he knew some statistical theory as well as some theoretical physics, one well-known institution suggested that he teach a statistics course, which he was well qualified to do. But on learning that he had been exposed to Bayesian

thinking, he was taken aside and told that it was a condition of his employment that he agree to teach straight out of Hoel, expounding no Bayesian ideas.

My revulsion at such thought control - by persons who would rightly denounce it anywhere else - resulted in my coming down on poor old Oscar a bit harder than I would have otherwise (still, he will be the first to admit that he delivers fully as much as he receives, and neither of us takes it personally).

The manuscript was sent to the publisher with some trepidations over the polemical style, but those fears were groundless; in fact, I have received more favorable fan mail over this article than on any two others. Many quite well-known figures have told me in confidence: 'Bravo! These things needed to be said, but I cannot say them because my position requires me to maintain diplomatic relations with both sides in the Great Debate. You are just enough of an outsider so you can get away with it.'

CONFIDENCE INTERVALS VS BAYESIAN INTERVALS

ABSTRACT. For many years, statistics textbooks have followed this 'canonical' procedure: (1) the reader is warned not to use the discredited methods of Bayes and Laplace, (2) an orthodox method is extolled as superior and applied to a few simple problems, (3) the corresponding Bayesian solutions are *not* worked out or described in any way. The net result is that no evidence whatsoever is offered to substantiate the claim of superiority of the orthodox method.

To correct this situation we exhibit the Bayesian and orthodox solutions to six common statistical problems involving confidence intervals (including significance tests based on the same reasoning). In every case, we find that the situation is exactly the opposite; i.e., the Bayesian method is easier to apply and yields the same or better results. Indeed, the orthodox results are satisfactory only when they agree closely (or exactly) with the Bayesian results. No contrary example has yet been produced.

By a refinement of the orthodox statistician's own criterion of performance, the best confidence interval for any location or scale parameter is proved to be the Bayesian posterior probability interval. In the cases of point estimation and hypothesis testing, similar proofs have long been known. We conclude that orthodox claims of superiority are totally unjustified; today, the original statistical methods of Bayes and Laplace stand in a position of proven superiority in actual performance, that places them beyond the reach of mere ideological or philosophical attacks. It is the continued teaching and use of orthodox methods that is in need of justification and defense.

I. INTRODUCTION

The theme of our meeting has been stated in rather innocuous terms: how should probability theory be (1) formulated, (2) applied to statistical inference; and (3) to statistical physics? Lurking behind these bland generalities, many of us will see more specific controversial issues: (1) frequency vs. nonfrequency definitions of probability, (2) 'orthodox' vs. Bayesian methods of inference, and (3) ergodic theorems vs. the principle of maximum entropy as the basis for statistical mechanics.

When invited to participate here, I reflected, that I have already held forth on issue (3) at many places, for many years, and at great length. At the moment, the maximum entropy cause seems to be in good hands and advancing well, with no need for any more benedictions from me; in any event, I have little more to say beyond what is already in print.' So it seemed time to widen the front, and enter the arena on issue (2).

Why a physicist should have the temerity to do this, when no statistician has been guilty of invading physics to tell us how we ought to do our jobs, will become clear only gradually; but the main points arc: (A) we were here first, and (B) because of our past experiences, physicists may be in a position to help statistics in its present troubles, well described by Kempthorne (1971). More specifically:

(A) Historically, the development of probability theory in the 18'th and early 19'th centuries from a gambler's amusement to a powerful research tool in science and many other areas, was the work of people - Daniel Bernoulli, Laplace, Poisson, Legendre, Gauss, and several others - whom we would describe today as mathematical physicists. In the 19'th century, a knowledge of their work was considered an essential part of the training of any scientist, and it was taught largely as a part of physics.

A radical change took place early in this century when a new group of workers, not physicists, entered the field. They proceeded to reject virtually everything done by Laplace and sought to develop statistics anew, based on entirely different principles. Simultaneously with this development, the physicists - with Sir Harold Jeffreys as almost the sole exception - quietly retired from the field, and statistics disappeared from the physics curriculum.

This departure of physicists from the field they had created was not, of course, due to the new

competition; rather, it was just at this time that relativity theory burst upon us, X-rays and radioactivity were discovered, and quantum theory started to develop. The result was that for fifty years physicists had more than enough to do unravelling a host of new experimental facts, digesting these new revolutions of thought, and putting our house back into some kind of order. But the result 0 our departure was that this extremely aggressive new school in statistics soon dominated the field so completely that its methods are now known as "orthodox statistics". For these historical reasons, I ask you to think with me, that for a physicist to turn his attrition now to statistics, is more of a home-coming than an invasion.

(B) Today, a physicist revisiting statistics to see how it has fared in our absence, sees quickly that something has gone wrong. For over fifteen years now, statistics has been in a state of growing ideological crisis - literally a crisis of conflicting ideas - that shows no signs of resolving itself, but yearly grows more acute; but it is one that physicists can recognize as basically the same thing that physics has been through several times (Jaynes, 1967). Having seen how these crises work themselves out, I think physicists may be in a position to prescribe a physic that will speed up the process in statistics.

The point we have to recognize is that issues of the kind facing us are never resolved by mere philosophical or ideological debate. At that level of discussion, people will persist in disagreeing, and nobody will be able to prove his case. In physics, we have our own ideological disputes, just as deeply felt by the protagonists as any in statistics; and at the moment I happen to be involved in one that strikes at the foundations of quantum theory (Jaynes, 1973). But in physics we have been perhaps more fortunate in that we have a universally recognized Supreme Court, to which all disputes are taken eventually, and from whose verdict there is no appeal. I refer, of course, to direct experimental observation of the facts.

This is an exciting time in physics, because recent advances in technology (lasers, fast computers, etc.) have brought us to the point where issues which have been debated fruitlessly on the philosophical level for 45 years, are at last reduced to issues of fact, and experiments are now underway testing controversial aspects of quantum theory that have never before been accessible to direct check. We have the feeling that, very soon now, we are going to know the real truth, the long debate can end at last, one way or the other; and we will be able to turn a great deal of energy to more constructive things. Is there any hope that the same can be done for statistics?

I think there is, and history points the way. It is to Galileo that we owe the first demonstration that ideological conflicts are resolved, not by debate, but by observation of fact. But we also recall that he ran into some difficulties in selling this idea to his contemporaries. Perhaps the most striking thing about his troubles was not his eventual physical persecution, which was hardly uncommon in those days; but rather the quality of logic that was used by his adversaries. For example, having turned his new telescope to the skies, Galileo announced discovery of the moons of Jupiter. A contemporary scholar ridiculed the idea, asserted that his theology had proved there could be *no* moons about Jupiter; and steadfastly refused to look through Galileo's telescope. But to everyone who did take a look, the evidence of his own eyes somehow carried more convincing power than did any amount of theology.

Galileo's telescope was able to reveal the truth, in a way that transcended all theology, because it could *magnify* what was too small to be perceived by our unaided senses, up into the range where it could be seen directly by all. And that, I suggest, is exactly what we need in statistics if this conflict is ever to be resolved. Statistics cannot take its dispute to the Supreme Court of the physicist; but there is another. It was recognized by Laplace in that famous remark, "Probability theory is nothing but common sense reduced to calculation".

Let me make what, I fear, will seem to some a radical, shocking suggestion: *the merits of any statistical method are not determined by the ideology which led to it.* For, many different, violently opposed

ideologies may all lead to the same final 'working equations' for dealing with real problems. Apparently, this phenomenon is something new in statistics; but it is so commonplace in physics that we have long since learned how to live with it. Today, when a physicist says, "Theory A is better than theory B", he does not have in mind any ideological considerations; he means simply, "There is at least one specific application where theory A leads to a better result than theory B".

I suggest that we apply the same criterion in statistics: *the merits of any statistical method are determined by the results it gives when applied to specific problems*. The Court of Last Resort in statistics is simply our commonsense judgment of those results. But our common sense, like our unaided vision, has a limited resolving power. Given two different statistical methods (e.g., an orthodox and a Bayesian one), in many cases they lead to final numerical results which are so nearly alike that our common sense is unable to make a clear decision between them. What we need, then, is a kind of Galileo telescope for statistics; let us try to invent an extreme case where a small difference is magnified to a large one, or if possible to a qualitative difference in the conclusions. Our common sense will then tell us which method is preferable, in a way that transcends all ideological quibbling over 'subjectivity', 'objectivity', the 'true meaning of probability', etc.

I have been carrying out just this program, as a hobby, for many years, and have quite a mass of results covering most areas of statistical practice. They all lead to the same conclusion, and I have yet to find one exception to it. So let me give you just a few samples from my collection.

(a) INTERVAL ESTIMATION

Time not permitting even a hurried glimpse at the entire field of statistical inference, it is better to pick out a small piece of it for close examination. Now we have already a considerable Underground Literature on the relation of orthodox and Bayesian methods in the areas of point estimation and hypothesis testing, the topics most readily subsumed under the general heading of Decision Theory. [I say underground, because the orthodox literature makes almost no mention of it. Not only in textbooks, but even in such a comprehensive treatise as that of Kendall and Stuart (1961), the reader can find no hint of the existence of the books of Good (1950), Savage (1954), Jeffreys (1957), or Schlaifer (1959), all of which are landmarks in the modern development of Bayesian statistics].

It appears that much less has been written about this comparison in the case of interval estimation; so I would like to examine here the orthodox principle of confidence intervals (including significance tests based on the same kind of reasoning), as well as the orthodox criteria of performance and method of reporting results; and to compare these with the corresponding Bayesian reasoning and results, with magnification.

The basic ideas of interval estimation must be ancient, since they occur inevitably to anyone involved in making measurements, as soon as he ponders how he can most honestly communicate what he has learned to others, short of giving the entire mass of raw data. For, if you merely give your final best number, some troublesome fellow will demand to know how accurate the number is. And you will not appease him merely by answering his question; for if you reply, "It is within a tenth of a percent", he will only ask, "How sure are you of that? Will you make a 10:1 bet on it?"

It is not enough, then, to give a number or even an interval of possible error; at the very minimum, one must give both an interval and some indication of the reliability with which one can assert that the true value lies within it. But even this is not really enough; ideally (although this goes beyond current practice) one ought to give many different intervals - or even a continuum of all possible intervals - with some kind of statement about the reliability of each, before he has fully described his state of knowledge. This was noted by D. R. Cox (1958), in producing a nested sequence of confidence intervals; evidently, a Bayesian posterior probability accomplishes the same thing in a simpler way.

Perhaps the earliest formal quantitative treatment of interval estimation was Laplace's analysis of the accuracy with which the mass of Saturn was known at the end of the 18'th century. His method was to apply Bayes' theorem with uniform prior density; relevant data consist of the mutual perturbations of Jupiter and Saturn, and the motion of their moons, but the data are imperfect because of the finite accuracy with which angles and time intervals can be measured. From the posterior distribution $P(M)dM$ conditional on the available data, one can determine the shortest interval which contains a specified amount of posterior probability, or equally well the amount of posterior probability contained in a specified interval. Laplace chose the latter course, and announced his result as follows: "... it is a bet of 11000 against 1 that the error of this result is not 1/100 of its value". In the light of present knowledge, Laplace would have won his bet; another 150 years' accumulation of data has increased the estimate by 0.63 percent.

Today, orthodox teaching holds that Laplace's method was, in Fisher's words, "founded upon an error". While there are some differences of opinion within the orthodox school, most would hold that the proper method for this problem is the confidence interval. It would seem to me that, in order to substantiate this claim, the orthodox writers would have to (1) produce the confidence interval for Laplace's problem, (2) show that it leads us to numerically different conclusions, and (3) demonstrate that the confidence interval conclusions are more satisfactory than Laplace's. But, in some twenty years of searching the orthodox literature, I have yet to find one case where such a program is carried out, on any statistical problem.

Invariably, the superiority of the orthodox method is asserted, not by presenting evidence of superior performance, but by a kind of ideological invective about 'objectivity' which perhaps reached its purple climax in an astonishing article of Bross (1963), whose logic recalls that of Galileo's colleague. In his denunciation of everything Bayesian, Bross specifically brings up the matter of confidence intervals and orthodox significance tests (which are based on essentially the same reasoning, and often amount to one-sided confidence intervals). So we will do likewise; in the following, we will examine these same methods and try to supply what Bross omitted; the demonstrable facts concerning them.

We first consider three significance tests appearing in the recent literature of reliability theory. The first two, which turn out to be so clear that no magnification is needed, will also bring out an important point concerning orthodox methods of reporting results.

II. SIGNIFICANCE TESTS

> Significance tests, in their usual form, are not compatible with a Bayesian attitude.
> C. A. B. Smith (1962)

> At any rate, what I feel quite sure at the moment to be needed is simple illustration of the new [i.e., Bayesian] notions on real, everyday statistical problems.
> E. S. Pearson (1962)

(a) EXAMPLE 1. DIFFERENCE OF MEANS

One of the most common of the 'everyday statistical problems' concerns the difference of the means of two normal distributions. A good example, with a detailed account of how current orthodox practice deals with such problems, appears in a recent book on reliability engineering (Roberts, 1964).

Two manufacturers, *A* and *B,* are suppliers for a certain component, and we want to choose the one which affords the longer mean life. Manufacturer *A* supplies 9 units for test, which turn out to have a (mean ± standard deviation) lifetime of (42±7.48) hours. *B* supplies 4 units, which yield (50±6.48) hours.

I think our common sense tells us immediately, without any calculation, that this constitutes fairly substantial (but not overwhelming) evidence in favor of *B*. While we should certainly prefer a larger sample, *B*'s units did give a longer mean life, the difference being appreciably greater than the sample standard deviation; and so if a decision between them must be made on this basis, we should have no hesitation in choosing *B*. However, the author warns against drawing any such conclusion, and says that, if you are tempted to reason this way, then "perhaps statistics is not for you!" In any event, when we have so little evidence, it is imperative that we analyze the data in a way that does not throw any of it away.

The author then offers us the following analysis of the problem. He first asks whether the two variances are the same. Applying the *F*-test, the hypothesis that they are equal is not rejected at the 95 percent significance level, so without further ado he assumes that they are equal, and pools the data for an estimate of the variance. Applying the *t*-test, he then finds that, at the 90 percent level, the sample affords no significant evidence in favor of either manufacturer over the other.

Now, any statistical procedure which fails to extract evidence that is already clear to our unaided commonsense, is certainly *not* for me! So, I carried out a Bayesian analysis. Let the unknown mean lifetimes of *A*'s and *B*'s components be *a*, *b* respectively. If the question at issue is whether *b*> *a*, the way to answer it is to calculate the probability that *b* > *a*, conditional on all the available data. This is

$$(1) \qquad Prob(b>a)=\int_{-\infty}^{\infty} da \int_{a}^{\infty} db\, P_n(a) P_m(b)$$

where $P_n(a)$ is the posterior distribution of *a,* based on the sample of n=9 items supplied by *A*, etc. When the variance is unknown, we find that these are of the form of the 'Student' *t*-distribution:

$$(2) \qquad P_n(a) \sim \left[s_A^2 + (a - \bar{t}_A)^2 \right]^{-n/2}$$

where $\bar{t}_A$, $s_A^2 = \bar{t}_A^2 - \bar{t}_A^2$ are the mean and variance of sample *A*. Carrying out the integration (1), I find that the given data yield a probability of 0.920, or odds of 11.5 to 1, that *B's* components *do* have a greater mean life - a conclusion which, I submit, conforms nicely to the indications of common-sense. [3]

But this is far from the end of the story; for one feels intuitively that if the variances are assumed equal, this ought to result in a more selective test than one in which this is not assumed; yet we find the Bayesian test without assumption of equal variance yielding an apparently sharper result than the orthodox one with that assumption. This suggests that we repeat the Bayesian calculation, using the author's assumption of equal variances. We have again an integral like (1), but *a* and *b* are no longer independent, their joint posterior distribution being proportional to

$$(3) \qquad P(a,b) \sim \left\{ n \left[s_A^2 + (a - \bar{t}_A)^2 \right] + m \left[s_B^2 + (b - \bar{t}_B)^2 \right] \right\}^{-1/2(n+m)}$$

Integrating this over the same range as in (1) - which can be done simply by consulting the *t*-tables after carrying out one integration analytically - I find that the Bayesian analysis now yields a probability of 0.948, or odds of 18:1, in favor of *B*.

How, then, could the author have failed to find significance at the 90 percent level? Checking the tables used we discover that, without having stated so, he has applied the *equal tails t*-test at the 90 percent level. But this is surely absurd; it was clear from the start that there is no question of the data supporting *A*; the only purpose which can be served by a statistical analysis is to tell us *how strongly* it supports *B*.

The way to answer this is to test the null hypothesis $b=a$ against the one-sided alternative $b>a$ already indicated by inspection of the data; using the 90 percent equal-tails test throws away half the 'resolution' and gives what amounts to a one-sided test at the 95 percent level, where it just barely fails to achieve significance.

In summary, the data yield clear significance at the 90 percent level; but the above orthodox procedure (which is presumably now being taught to many students) is a compounding of two errors. Assuming the variances equal makes the difference $\bar{t}_A - \bar{t}_B$ appear, unjustifiably, even more significant; but then use of the equal tails criterion throws away more than was thus gained, and we still fail to find significance at the 90 percent level.

Of course, the fact that orthodox methods are capable of being misused in this way does not invalidate them; and Bayesian methods can also be misused, as we know only too well. However, there must be something in orthodox teaching which predisposes one toward this particular kind of misuse, since it is very common in the literature and in everyday practice. It would be interesting to know why most orthodox writers will not use - or even mention - the Behrens-Fisher distribution, which is clearly the correct solution to the problem, has been available for over forty years (Fisher, 1956; p. 95), and follows immediately from Bayes' theorem with the Jeffreys prior (Jeffreys, 1939; p. 115).

(b) EXAMPLE 2. SCALE PARAMETERS

A recent Statistics Manual (Crow *et al.,* 1960) proposes the following problem: 31 rockets of type 1 yield a dispersion in angle of 2237 mils^2, and 61 of type 2 give instead 1347 mils^2. Does this constitute significant evidence for a difference in standard deviation of the two types?

I think our common sense now tells us even more forcefully that, in view of the large samples and the large observed difference in dispersion, this constitutes absolutely unmistakable evidence for the superiority of type 2 rockets. Yet the authors, applying the equal-tails *F*-test at the 95 percent level, find it not significant, and conclude: "We need not, as far as this experiment indicates, differentiate between the two rockets with respect to their dispersion".

Suppose you were a military commander faced with the problem of deciding which type of rocket to adopt. You provide your statistician with the above data, obtained at great trouble and expense, and receive the quoted report. What would be your reaction? I think that you would fire the statistician on the spot; and henceforth make decisions on the basis of your own common sense, which is evidently a more powerful tool than the equal-tails *F*-test.

However, if your statistician happened to be a Bayesian, he would report[4] instead: "These data yield a probability of 0.9574 or odds of 22.47:1, in favor of type 2 rockets". I think you would decide to keep this fellow on your staff, because his report not only agrees with common sense; it is stated in a far more useful form. For, you have little interest in being told merely whether the data constitute 'significant evidence for a difference'. It is already obvious without any calculation that they *do* constitute highly significant evidence in favor of type 2; the only purpose that can be served by a statistical analysis is, again, to tell us quantitatively *how significant* that evidence is. Traditional orthodox practice fails utterly to do this, although the point has been noted recently by some.

What we have found in these two examples is true more generally. The orthodox statistician conveys little useful information when he merely reports that the null hypothesis is or is not rejected

at some arbitrary preassigned significance level. If he reports that it is rejected at the 90 percent level, we cannot tell from this whether it would have been rejected at the 92 percent, or 95 percent level. If he reports that it is not rejected at the 95 percent level, we cannot tell whether it would have been rejected at the 50 percent, or 90 percent level. If he uses an equal-tails test, he in effect throws away half the 'resolving power' of the test, and we are faced with still more uncertainty as to the real import of the data.

Evidently, the orthodox statistician would tell us far more about what the sample really indicates if he would report instead *the critical significance level at which the null hypothesis is just rejected in favor of the onesided alternative indicated by the data;* for we then know what the verdict would be at all levels, and no resolution has been lost to a superfluous tail. Now two possible cases can arise: (I) the number thus reported is identical with the Bayesian posterior probability that the alternative is true; (II) these numbers are different.

If case (I) arises (and it does more often than is generally realized), the Bayesian and orthodox tests are going to lead us to exactly the same numerical results and the same conclusions, with only a verbal disagreement as to whether we should use the word 'probability' or 'significance' to describe them. In particular, the orthodox *t*-test and *F*-test against one-sided alternatives would, if their results were reported in the manner just advocated, be precisely equivalent to the Bayesian tests based on the Jeffreys prior $d\mu d\sigma/\sigma$. Thus, if we assume the variances equal in the above problem of two means, the observed difference is just significant by the one-sided *t*-test at the 94.8 percent level; and in the rocket problem a onesided *F*-test just achieves significance at the 95.74 percent level.

It is only when case (II) is found that one could possibly justify any 'objective' claim for superiority of either approach. Now it is just these cases where we have the opportunity to carry out our 'magnification' process; and if we can find a problem for which this difference is magnified sufficiently, the issue cannot really be in doubt. We find this situation, and a number of other interesting points of comparison, in one of the most common examples of acceptance tests.

(c) EXAMPLE 3. AN ACCEPTANCE TEST

The probability that a certain machine will operate without failure for a time t is, by hypothesis, $exp(-\lambda t)$, $0<t<\infty$. We test n units for a time t, and observe r failures; what assurance do we then have that the mean life $\theta=\lambda^{-1}$ exceeds a preassigned value θ_o?

Sobel and Tischendorf (1959) (hereafter denoted ST) give an orthodox solution with tables that are reproduced in Roberts (1964). The test is to have a critical number C (i.e., we accept only if $r\leq C$). On the hypothesis that we have the maximum tolerable failure rate, $\lambda_o = \theta_o^{-1}$, the probability that we shall see r or fewer failures is the binomial sum

$$W_t(n,r)=\sum_{k=0}^{r}\binom{n}{k}e^{-(n-k)\lambda_o t}\left(1-e^{-\lambda_o t}\right)^k \tag{4}$$

and so, setting $W(n,C) \leq 1-P$ gives us the sample size n required in order that this test will assure $\theta\geq\theta_o$, at the 100 P percent significance level. From the ST tables we find, for example, that if we wish to test only for a time $t=0.01\theta_o$ with $C=3$, then at the 90 percent significance level we shall require a test sample of $n=668$ units; while if we are willing to test for a time $t=\theta_o$ with $C=1$, we need test only 5 units.

The amount of testing called for is appalling if $t \ll \theta_o$; and out of the question if the units are complete systems. For example, if we want to have 95 percent confidence (synonymous with significance) that a space vehicle has $\theta_o \geq 10$ years, but the test must be made in six months, then

with C=1, the ST tables say that we must build and test 97 vehicles! Suppose that, nevertheless, it had been decreed on the highest policy level that this degree of confidence *must* be attained, and you were in charge of the testing program. If a more careful analysis of the statistical problem, requiring a few man-years of statisticians' time, could reduce the test sample by only one or two units, it would be well justified economically. Scrutinizing the test more closely, we note four points:

(1) We know from the experiment not only the total number r of failures, but also the particular times $\{t_1 \ldots t_r\}$ at which failure occurred. This information is clearly relevant to the question being asked; but the ST test makes no use of it.

(2) The test has a 'quasi-sequential' feature; if we adopt an acceptance number C=3, then as soon as the fourth failure occurs, we know that the units are going to be rejected. If no failures occur, the required degree of confidence will be built up long before the time t specified in the ST tables. In fact, t is the *maximum possible* testing time, which is actually required only in the marginal case where we observe exactly C failures. A test which is 'quasi-sequential' in the sense that it terminates when a clear rejection or the required confidence is attained, will have an expected length less than t; conversely, such a test with the expected length set at t will require fewer units tested.

(3) We have relevant prior information; after all, the engineers who designed the space vehicle knew in advance what degree of reliability was needed. They have chosen the quality of materials and components, and the construction methods, with this in mind. Each sub-unit has had its own tests. The vehicles would never have reached the final testing stage unless the engineers knew that they were operating satisfactorily. In other words, we are not testing a completely unknown entity. The ST test (like most orthodox procedures) ignores all prior information, except perhaps when deciding which hypotheses to consider, or which significance level to adopt.

(4) In practice, we are usually concerned with a different question than the one the ST test answers. An astronaut starting a five-year flight to Mars would not be particularly comforted to be told, "We are 95 percent confident that the average life of an imaginary population of space vehicles like yours, is at least ten years". He would much rather hear, "There is 95 percent probability that *this* vehicle will operate without breakdown for ten years". Such a statement might appear meaningless to an orthodox statistician who holds that (probability)≡(frequency). But such a statement would be very meaningful indeed to the astronaut.

This is hardly a trivial point; for if it were *known* that λ^{-1} =10 yr, the probability that a particular vehicle will actually run for 10 yrs would be only $1/e$=0.368*;* and the period for which we are 95 percent sure of success would be only -10 ln(0.95) years, or 6.2 months. Reports which concern only the 'mean life' can be rather misleading.

Let us first compare the ST test with a Bayesian test which makes use of exactly the same information; i.e., we are allowed to use only the total number of failures, not the actual failure times. On the hypothesis that the failure rate is λ, the probability that exactly r units fail in time t is

$$p(r|\ n,\lambda,t)=\binom{n}{r}e^{-(n-r)\lambda t}\left(1-e^{-\lambda t}\right)^r \tag{5}$$

I want to defer discussion of nonuniform priors; for the time being suppose we assign a uniform prior density to λ. This amounts to saying that, before the test, we consider it extremely unlikely that our space vehicles have a mean life as long as a microsecond; nevertheless it will be of interest to see the result of using this prior. The posterior distribution of λ is then

$$p(d\lambda|n,r,t)=\frac{n!}{(n-r-1)!\,r!}e^{-(n-r)\lambda t}\left(1-e^{-\lambda t}\right)^r d(\lambda t) \tag{6}$$

The Bayesian acceptance criterion, which ensures $\lambda \geq \theta_o^{-1}$ with 100 P percent probability, is then

(7) $$\int_{\lambda_o}^{\infty} p(d\lambda|n,r,t) \leq 1-P.$$

But the left-hand side of (7) is identical with $W(n,r)$ given by (4); this is just the well-known identity of the incomplete Beta function and the incomplete binomial sum, given already in the original memoir of Bayes (1763).

In this first comparison we therefore find that the ST test is mathematically identical with a Bayesian test in which (1) we are denied use of the actual failure times; (2) because of this it is not possible to take advantage of the quasi-sequential feature; (3) we assign a ridiculously pessimistic prior to λ; (4) we still are not answering the question of real interest for most applications.

Of these shortcomings, (2) is readily corrected, and (1) undoubtedly could be corrected, without departing from orthodox principles. On the hypothesis that the failure rate is λ, the probability that r specified units fail in the time intervals $\{dt_1 \ldots dt_r\}$ respectively, and the remaining $(n - r)$ units do not fail in time t, is

(8) $$p(dt_1 \ldots dt_r|n,r,t) = [\lambda^r e^{-r\lambda\bar{t}}\, dt_1 \ldots dt_r][e^{-(n-r)\lambda t}]$$

where $\bar{t} \equiv r^{-1}\sum t_i$ is mean life of the units which failed. There is no single 'statistic' which conveys all the relevant information; but r and $\bar{t}$ are jointly sufficient, and so an optimal orthodox test must somehow make use of both. When we seek their joint sampling distribution $p(r, d\bar{t}|n,\lambda,t)$ we find, to our dismay, that for given r the interval $0 < \bar{t} < t$ is broken up into r equal intervals, with a different analytical expression for each. Evidently a decrease in r, or an increase in, should incline us in the direction of acceptance; but at what rate should we trade off one against the other? To specify a definite critical region in both variables would seem to imply some postulate as to their relative importance. The problem does not appear simple, either mathematically or conceptually; and I would not presume to guess how an orthodox statistician would solve it.

The relative simplicity of the Bayesian analysis is particularly striking in this problem; for all four of the above shortcomings are corrected effortlessly. For the time being, we again assign the pessimistic uniform prior to λ; from (8), the posterior distribution of λ is then

(9) $$p(d\lambda|n,t,t_1 \ldots t_r) = \frac{(\lambda T)^r}{r!} e^{-\lambda T}\, d(\lambda T)$$

where

(10) $$T \equiv r\bar{t} + (n-r)t$$

is the total unit-hours of failure-free operation observed. The posterior probability that $\lambda_o \geq \theta_o$ is now

(11) $$B(n,r) = \frac{1}{r!}\int_{\lambda_o T}^{\infty} x^r e^{-x}\, dx = e^{-\lambda_o T}\sum_{k=0}^{r}\frac{(\lambda_o T)^k}{k!}$$

and so, $B(n, r) \leq 1-P$ is the new Bayesian acceptance criterion at the 100 P percent level; the test can terminate with acceptance as soon as this inequality is satisfied.

Numerical analysis shows little difference between this test and the ST test in the usual range of practical interest where we test for a time short compared to θ_o and observe only a very few failures. For, if $\lambda_o t \ll 1$, and $r \ll n$, then the Poisson approximation to (4) will be valid; but this is just the expression (11) except for the replacement of T by nt, which is itself a good approximation. In this region the Bayesian test (11) with maximum possible duration t generally calls for a test sample one

or two units smaller than the ST test. Our common sense readily assents to this; for if we see only a few failures, then information about the actual failure time adds little to our state of knowledge.

Now let us magnify. The big differences between (4) and (11) will occur when we find many failures; if all n units fail, the ST test tells us to reject at all confidence levels, even though the observed mean life may have been thousands of times our preassigned θ_o. The Bayesian test (11) does not break down in this way; thus if we test 9 units and all fail, it tells us to accept at the 90 percent level if the observed mean life $\bar{t} \leq 1.58\,\theta_o$. If we test 10 units and 9 fail, the ST test says we can assert with 90 percent confidence that $\theta \geq 0.22\ t$; the Bayesian test (11) says there is 90 percent probability that $\theta \geq 0.63\,\bar{t} + 0.07\,t$. Our common sense has no difficulty in deciding which result we should prefer; thus taking the actual failure times into account leads to a clear, although usually not spectacular, improvement in the test. The person who rejects the use of Bayes' theorem in the manner of Equation (9) will be able to obtain a comparable improvement only with far greater difficulty.

But the Bayesian test (11) can be further improved in two respects. To correct shortcoming (4), and give a test which refers to the reliability of the individual unit instead of the mean life of an imaginary 'population' of them, we note that if λ were known, then by our original hypothesis the probability that the lifetime θ of a given unit is at least θ_o, is

$$p(\theta \geq \theta_o | \lambda) = e^{-\lambda \theta_o} \quad . \tag{12}$$

The probability that $\theta \geq \theta_o$, conditional on the evidence of the test, is therefore

$$p(\theta \geq \theta_o | n, t_1 \ldots t_r) = \int_0^\infty e^{-\lambda \theta_o}\, p(d\lambda | n, t_1 \ldots t_r) = \left(\frac{T}{T + \theta_o} \right)^{r+1} \quad . \tag{13}$$

Thus, the Bayesian test which ensures, with 100 P percent probability, that the life of an *individual unit* is at least θ_o, has an acceptance criterion that the expression (13) is $\geq P$; a result which is simple, sensible, and as far as I can see, utterly beyond the reach of orthodox statistics.

The Bayesian tests (11) and (13) are, however, still based on a ridiculous prior for λ; another improvement, even further beyond the reach of orthodox statistics, is found as a result of using a reasonable prior. In 'real life' we usually have excellent grounds based on previous experience and theoretical analyses, for predicting the general order of magnitude of the lifetime in advance of the test. It would be inconsistent from the standpoint of inductive logic, and wasteful economically, for us to fail to take this prior knowledge into account.

Suppose that initially, we have grounds for expecting a mean life of the order of t_i; or a failure rate of about $\lambda_i \approx t_i^{-1}$. However, the prior information does not justify our being too dogmatic about it; to assign a prior centered sharply about λ_i; would be to assert so much prior knowledge that we scarcely need any test. Thus, we should assign a prior that, while incorporating the number t_i, is still as 'spread out' as possible, in some sense.

Using the criterion of maximum entropy, we choose that prior density $p_i(\lambda)$ which, while yielding an expectation equal to λ_i, maximizes the 'measure of ignorance' $H = -\int p_i(\lambda) \log p_i(\lambda)\, d\lambda$ The solution is: $p_i(\lambda) = t_i \exp(-\lambda t_i)$. Repeating the above derivation with this prior, we find that the posterior distribution (9) and its consequences (11)-(13) still hold, but that Equation (11) is now to be replaced by

$$T \equiv r\,\bar{t}\,(n - r)\,t + t_i \quad . \tag{14}$$

Subjecting the resulting solution to various extreme conditions now shows an excellent

correspondence with the indications of common sense. For example, if the total unit-hours of the test is small compared to t_i, then our state of knowledge about A can hardly be changed by the test, unless an unexpectedly large number of failures occurs. But if the total unit-hours of the test is large compared to t_i, then for all practical purposes our final conclusions depend only on what we observed in the test, and are almost independent of what we thought previously. In intermediate cases, our prior knowledge has a weight comparable to that of the test; and if $t_i \geq \theta_o$, the amount of testing required is appreciably reduced. For, if we were already quite sure the units *are* satisfactory, then we require less additional evidence before accepting them. On the other hand, if $t_i \ll \theta_o$, the test approaches the one based on a uniform prior; if we are initially very doubtful about the units, then we demand that the test itself provide compelling evidence in favor of them.

These common-sense conclusions have, of course, been recognized qualitatively by orthodox statisticians; but only the Bayesian approach leads automatically to a means of expressing all of them explicitly and quantitatively in our equations. As noted by Lehmann (1959), the orthodox statistician can and does take his prior information into account, in some degree, by moving his significance level up and down in a way suggested by the prior information. But, having no formal principle like maximum entropy that tells him how much to move it, the resulting procedure is far more 'subjective' (in the sense of varying with the taste of the individual) than anything in the Bayesian approach which recognizes the role of maximum entropy and transformation groups in determining priors.

No doubt, the completely indoctrinated orthodoxian will continue to reject priors based even on the completely impersonal (and parameter - independent) principles of maximum entropy and transformation groups, on the grounds that they are still 'subjective' because they are not frequencies [although I believe I have shown (Jaynes, 1968, 1971) that if a random experiment is involved, the probabilities calculated from maximum entropy and transformation groups have just as definite a connection with frequencies as probabilities calculated from any other principle of probability theory]. In particular, he would claim that the prior just introduced into the ST test represents a dangerous loss of 'objectivity' of that test.

To this I would reply that the judgment of a competent engineer, based on data of past experience in the field, represents information fully as 'objective' and reliable as anything we can possibly learn from a random experiment. Indeed, most engineers would make a stronger statement; since a random experiment is, by definition, one in which the outcome - and therefore the conclusion we draw from it - is subject to uncontrollable variations, it follows that the only fully 'objective' means of judging the reliability of a system is through analysis of stresses, rate of wear, etc., which avoids random experiments altogether.

In practice, the real function of a reliability test is to check against the possibility of completely unexpected modes of failure; once a given failure mode is recognized and its mechanism understood, no sane engineer would dream of judging its chances of occurring merely from a random experiment.

(d) SUMMARY

In the article of Bross (1963)-and in other places throughout the orthodox literature - one finds the claim that orthodox significance tests are 'objective' and 'scientific', while the Bayesian approach to these problems is erroneous and/or incapable of being applied in practice. The above comparisons have covered some important types of tests arising in everyday practice, and in no case have we found any evidence for the alleged superiority, or greater applicability, of orthodox tests. In every case, we have found clear evidence of the opposite.

The mathematical situation, as found in these comparisons and in many others, is just this: some

orthodox tests are equivalent to the Bayesian ones based on non-informative priors, and some others, when sufficiently improved both in procedure and in manner of reporting the results, can be made Bayes-equivalent. We have found this situation when the orthodox test was (A) based on a sufficient statistic, and (B) free of nuisance parameters. In this case, we always have asymptotic equivalence for tests of a simple hypothesis against a one-sided alternative. But we often find exact equivalence for all sample sizes, for simple mathematical reasons; and this is true of almost all tests which the orthodox statistician himself considers fully satisfactory.

The orthodox *t*-test of the hypothesis $\mu=\mu_0$, against the alternative $\mu>\mu_0$ is exactly equivalent to the Bayesian test for reasons of symmetry; and there are several cases of exact equivalence even when the distribution is not symmetrical in parameter and estimator. Thus, for the Poisson distribution the orthodox test for $\lambda=\lambda_o$ against $\lambda>\lambda_o$, is exactly equivalent to the Bayesian test because of the identity

$$\frac{1}{n!}\int_{\lambda}^{\infty} x^n e^{-x}\,dx = \sum_{k=0}^{n}\frac{e^{-\lambda}\lambda^k}{k!}$$

and the orthodox *F*-test for $\sigma_1=\sigma_2$ against $\sigma_1>\sigma_2$ is exactly Bayes equivalent because of the identity

$$\frac{(n+m+1)!}{n!\,m!}\int_{0}^{P} x^n(1-x)^m\,dx = \sum_{k=0}^{m}\frac{(n+k)!}{n!\,k!}P^{n+1}(1-P)^k \quad .$$

In these cases, two opposed ideologies lead to just the same final working equations.

If there is no single sufficient statistic (as in the ST test) the orthodox approach can become extremely complicated. If there are nuisance parameters (as in the problem of two means), the orthodox approach is faced with serious difficulties of principle; it has not yet produced any unambiguous and fully satisfactory way of dealing with such problems.

In the Bayesian approach, neither of these circumstances caused any difficulty; we proceeded in a few lines to a definite and useful solution. Furthermore, Bayesian significance tests are readily extended to permit us to draw inferences about the specific case at hand, rather than about some purely imaginary 'population' of cases. In most real applications, it is just the specific case at hand that is of concern to us; and it is hard to see how frequency statements about a mythical population or an imaginary experiment can be considered any more 'objective' than the Bayesian statements. Finally, no statistical method which fails to provide any way of taking prior information into account can be considered a full treatment of the problem; it will be evident from our previous work (Jaynes, 1968) and the above example, that Bayesian significance tests are extended just as readily to incorporate any testable prior information.

III. TWO-SIDED CONFIDENCE INTERVALS

> The merit of the estimator is judged by the distribution of estimates to which it gives rise, i.e., by the properties of its sampling distribution.
>
> We must content ourselves with formulating a rule which will give good results 'in the long run' or 'on the average' ....
>
> Kendall and Stuart (1961)

The above examples involved some one-sided confidence intervals, and they revealed some cogent evidence concerning the role of sufficiency and nuisance parameters; but they were not well

adapted to studying the principle of reasoning behind them. When we turn to the general principle of two-sided confidence intervals some interesting new features appear.

(a) EXAMPLE 4. BINOMIAL DISTRIBUTION

Consider Bernoulli trials B_2 (i.e., two possible outcomes at each trial, independence of different trials). We observe r successes in n trials, and asked to estimate the limiting frequency of success f, and give a statement about the accuracy of the estimate. In the Bayesian approach, this is a very elementary problem; in the case of a uniform prior density for f [the basis of which we have indicated elsewhere (Jaynes, 1968) in terms of transformation groups; it corresponds to prior knowledge that it is possible for the experiment to yield either success or failure], the posterior distribution is proportional to $f^r(1-f)^{n-r}$ as found in Bayes' original memoir, with mean value $\bar{f} = (r+1)/(n+2)$ as given by Laplace (1774), and variance $\sigma^2 = \bar{f}(1-\bar{f})/(N+3)$.

The $(\bar{f} \pm \sigma)$ thus found provide a good statement of the 'best' estimate of f, and if $\bar{f}$ is not too close to 0 or 1, an interval within which the true value is reasonably likely to be. The full posterior distribution of f yields more detailed statements; if $r \gg 1$ and $(n-r) \gg 1$, it goes into a normal distribution $(\bar{f}, \sigma)$. The 100 P percent interval (i.e., the interval which contains 100 P percent of the posterior probability) is then simply $(\bar{f} \pm q\sigma)$, where q is the $(1+P)/2$ percentile of the normal distribution; for the 90, 95, and 99% levels, q= 1.645, 1.960, 2.576 respectively.

When we treat this same problem by confidence intervals, we find that it is no longer an undergraduate-level homework problem, but a research project. The final results are so complicated that they can hardly be expressed analytically at all, and we require a new series of tables and charts.

In all of probability theory there is no calculation which has been subjected to more sneering abuse from orthodox writers than the Bayesian one just described, which contains Laplace's rule of succession. But suppose we take a glimpse at the final numerical results, comparing, say, the 90% confidence belts with the Bayesian 90% posterior probability belts.

This must be done with caution, because published confidence intervals all appear to have been calculated from approximate formulas which yield wider intervals than is needed for the stated confidence level. We use a recently published (Crow *et al.* 1960) recalculated table which, for the case n=10, gives intervals about 0.06 units smaller than the older Pearson-Clopper values.

If we have observed 10 successes in 20 trials, the upper 90% confidence limit is given as 0.675; the above Bayesian formula gives 0.671. For 13 successes in 26 trials, the tabulated upper confidence limit is 0.658; the Bayesian result is 0.652.

Continued spot-checking of this kind leads one to conclude that, quite generally, the Bayesian belts lie just inside the confidence belts; the difference is visible graphically only for wide belts for which, in any event, no accurate statement about f was possible. The inaccuracy of published tables and charts is often greater than the difference between the Bayesian interval and the correct confidence interval. Evidently, then, claims for the superiority of the confidence interval must be based on something other than actual performance. The differences are so small that I could not magnify them into the region where common sense is able to judge the issue.

Once aware of these things the orthodox statistician might well decide to throw away his tables and charts, and obtain his confidence intervals from the Bayesian solution. Of course, if one demands very accurate intervals for very small samples, it would be necessary to go to the incomplete Beta-function tables; but it is hard to imagine any real problem where one would care about the exact width of a very wide belt. When $r \gg 1$ and $(n-r) \gg 1$, then to all the accuracy one can ordinarily use, the required interval is simply the above $(\bar{f} \pm q\sigma)$. Since, as noted, published confidence intervals are 'conservative' - a common euphemism - he can even improve his results by this procedure.

Let us now seek another problem, where differences can be magnified to the point where the

equations speak very clearly to our common sense.

(b) EXAMPLE 5. TRUNCATED EXPONENTIAL DISTRIBUTION

The following problem has occurred in several industrial quality control situations. A device will operate without failure for a time θ because of a protective chemical inhibitor injected into it; but at time θ the supply of this chemical is exhausted, and failures then commence, following the exponential failure law. It is not feasible to observe the depletion of this inhibitor directly; one can observe only the resulting failures. From data on actual failure times, estimate the time θ of guaranteed safe operation by a confidence interval. Here we have a continuous sample space, and we are to estimate a location parameter θ, from the sample values $\{x_1 \ldots x_N\}$, distributed according to the law

$$p(dx|\theta)=\begin{cases}\exp(\theta-x)\,dx\,, & x>\theta\\ 0\,, & x<\theta\end{cases} \tag{15}$$

Let us compare the confidence intervals obtained from two different estimators with the Bayesian intervals. The population mean is $E(x) = \theta + 1$, and so

$$\theta^*(x_1,\ldots x_N)\equiv\frac{1}{N}\sum_{i=1}^{N}(x_i-1)=\bar{x}-1 \tag{16}$$

is an unbiased estimator of θ. By a well-known theorem, it has variance $\sigma^2=N^{-1}$, as we are accustomed to find. We must first find the sampling distribution of θ^*; by the method of characteristic functions we find that it is proportional to $y^{N-1}\exp(-Ny)$ for $y > 0$, where $y\equiv(\theta^*-\theta+1)$.

Proof added. Using the RVT one has:

$$\int_{-\infty}^{\infty} e^{-N(\bar{x}-\theta)}\,\delta(\bar{x}-y)\prod_{i=1}^{N}[\theta(x_i-\theta)d\,x_i]\rightarrow\{x_i\Rightarrow x_i-\theta\}\rightarrow e^{-N(\bar{x}-\theta)}\int_{0}^{\infty}\delta(\bar{x}+\theta-y)\prod_{i=1}^{N}[dx_i]$$

this last integral evaluates to

$$\int_{-\infty}^{\infty}d\lambda\int_{0}^{\infty}d\,x_i\,e^{i\lambda(\bar{x}+\theta-y)}\rightarrow\int_{-\infty}^{\infty}d\,\lambda\,e^{i\lambda(\theta-y)}\left[\frac{1}{\lambda}\int_{0}^{\infty}e^{ix}\right]^N\approx\int_{-\infty}^{\infty}\frac{d\lambda}{\lambda^N}e^{i\lambda(\theta-y)}=(\theta-y)^{N-1}\int_{-\infty}^{\infty}\frac{dz}{z^N}e^{iz}$$

and so... $p(x_1,..x_N)=p(\bar{x})\approx(\bar{x}-\theta)^{N-1}\exp-N(\bar{x}-\theta)$

Evidently, it will not be feasible to find the shortest confidence interval in closed analytical form, so in order to prevent this example from growing into another research project, we specialize to the case N=3, suppose that the observed sample values were $\{x_1, x_2, x_3\}$= {12, 14, 16}; and ask only for the shortest 90% confidence interval.

A further integration then yields the cumulative distribution function $F(y)$= [1-(1+3y+9y^2/2) exp(-3y)], y>0. Any numbers y_1, y_2 satisfying $F(y_2)$–$F(y_1)$=0.9 determine a 90% confidence interval. To find the shortest one, we impose in addition the constraint $F'(y_2)=F'(y_1)$. By computer, this yields the interval

$$\theta^* - 0.8529 < \theta < \theta^* + 0.8264 \tag{17}$$

or, with the above sample values, the shortest 90% confidence interval is

(18) $12.1471 < \theta < 13.8264.$

The Bayesian solution is obtained from inspection of (15); with a constant prior density [which, as we have argued elsewhere (Jaynes, 1968) is the proper way to express complete ignorance of location parameter], the posterior density of θ will be

(19) $$p(\theta|x_1 \ldots x_N) = \begin{cases} N \exp N(\theta - x_1), & \theta < x_1 \\ 0, & \theta > x_1 \end{cases}$$

where we have ordered the sample values so that x_1 denotes the least one observed. The shortest posterior probability belt that contains 100 P percent of the posterior probability is thus $(x_1 - q) < \theta < x_1$ where $q = -N^{-1}\log(1-P)$. For the above sample values we conclude (by slide-rule) that, with 90% probability, the true value of θ is contained in the interval

(20) $11.23 < \theta < 12.0$.

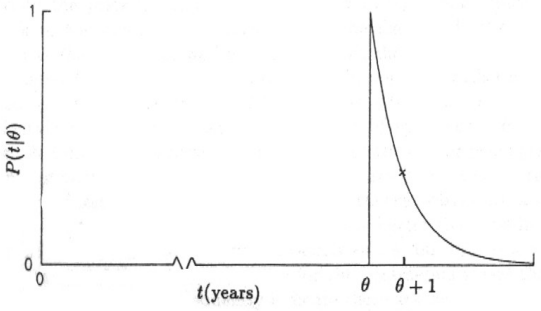

Fig. 6.3. The sampling density $P(t|\theta)$.

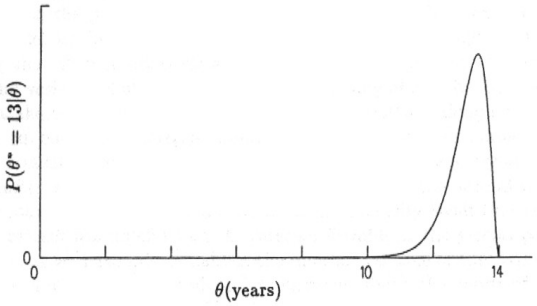

Fig. 6.4. The unbiased estimator density for θ.

Now what is the verdict of our common sense? The Bayesian interval corresponds quite nicely to our common sense; the confidence interval (18) is over twice as wide, and *it lies entirely in the region* $\theta < x_1$ *where it is obviously impossible for* θ *to be!*

I first presented this result to a recent convention of reliability and quality control statisticians working in the computer and aerospace industries; and at this point the meeting was thrown into an uproar, about a dozen people trying to shout me down at once. They told me, "This is complete nonsense. A method as firmly established and thoroughly worked over as confidence intervals couldn't possibly do such a thing. You are maligning a very great man; Neyman would never have advocated a method that breaks down on such a simple problem. If you can't do your arithmetic right, you have no business running around giving talks like this".

After partial calm was restored, I went a second time, very slowly and carefully, through the numerical work leading to (18), with all of them leering at me, eager to see who would be the first to catch my mistake [it is easy to show the correctness of (18), at least to two figures, merely by applying parallel rulers to a graph of $F(y)$]. In the end they had to concede that my result was correct after all.

To make a long story short, my talk was extended to four hours (all afternoon), and their reaction finally changed to: "My God - why didn't somebody tell me about these things before? My professors and textbooks never said anything about this. Now I have to go back home and recheck everything I've done for years".

This incident makes an interesting commentary on the kind of indoctrination that teachers of orthodox statistics have been giving their students for two generations now.

(c) WHAT WENT WRONG?

Let us try to understand what is happening here. It is perfectly true that, *if* the distribution (15) is indeed identical with the limiting frequencies of various sample values, and *if* we could repeat all this an indefinitely large number of times, then use of the confidence interval (17) *would* lead us, in the

long run, to a correct statement 90% of the time. But it would lead us to a wrong answer 100% of the time in the subclass of cases where $\theta^* > x_1+0.85$; and *we know from the sample whether we are in that subclass.*

That there must be a very basic fallacy in the reasoning underlying the principle of confidence intervals, is obvious from this example. The difficulty just exhibited is generally present in a weaker form, where it escapes detection. The trouble can be traced to two different causes.

Firstly, it has never been a part of 'official' doctrine that confidence intervals must be based on sufficient statistics; indeed, it is usually held to be a particular advantage of the confidence interval method that it leads to exact frequency-interpretable intervals without the need for this. Kendall and Stuart (1961), however, noting some of the difficulties that may arise, adopt a more cautious attitude and conclude (loc. cit., p. 153): "... confidence interval theory is possibly not so free from the need for sufficiency as might appear".

We suggest that the general situation, illustrated by the above example, is the following: whenever the confidence interval is not based on a sufficient statistic, it is possible to find a 'bad' subclass of samples, *recognizable from the sample,* in which use of the confidence interval would lead us to an incorrect statement more frequently than is indicated by the confidence level; and also a recognizable 'good' subclass in which the confidence interval is wider than it needs to be for the stated confidence level. The point is not that confidence intervals fail to do what is claimed for them; the point is that, if the confidence interval is not based on a sufficient statistic, it is possible to do better in the individual case by taking into account evidence from the sample that the confidence interval method throws away.

The Bayesian literature contains a multitude of arguments showing that it is precisely the original method of Bayes and Laplace which does take into account all the relevant information in the sample; and which will therefore always yield a superior result to any orthodox method not based on sufficient statistics. That the Bayesian method does have this property (i.e., the 'likelihood principle') is, in my opinion, now as firmly established as any proposition in statistics. Unfortunately, many orthodox textbook writers and teachers continue to ignore these arguments; for over a decade hardly a month has gone by without the appearance of some new textbook which carries on the indoctrination by failing to present both sides of the story.

If the confidence interval is based on a sufficient statistic, then as we saw in Example 4, it turns out to be so nearly equal to the Bayesian interval that it is difficult to produce any appreciable difference in the numerical results; in an astonishing number of cases, they are identical. That is the case in the example just given, where x_1 is a sufficient statistic, and it yields a confidence interval identical with the Bayesian one (20).

Similarly, the shortest confidence interval for the mean of a normal distribution, whether the variance is known or unknown; and for the variance of a normal distribution, whether the mean is known or unknown; and for the width of a rectangular distribution, all turn out to be identical with the shortest Bayesian intervals at the same level (based on a uniform prior density for location parameters and the Jeffreys prior $d\sigma/\sigma$ for scale parameters). Curiously, these are just the cases cited most often by textbook writers, after warning us not to use those erroneous Bayesian methods, as an illustration of their more 'objective' orthodox methods.

The second difficulty in the reasoning underlying confidence intervals concerns their criteria of performance. In both point and interval estimation, orthodox teaching holds that the reliability of an estimator is measured by its performance 'in the long run', i.e., by its sampling distribution. Now there are some cases (e.g., fixing insurance rates) in which long run performance is the sole, all-important consideration; and in such cases one can have no real quarrel with the orthodox reasoning (although the same conclusions are found just as readily by Bayesian methods). However, in the great majority of real applications, long-run performance is of no concern to us, because it will never be

realized.

Our job is not to follow blindly a rule which would prove correct 90% of the time in the long run; there are an infinite number of radically different rules, all with this property. Our job is to draw the conclusions that are most likely to be right in the specific case at hand; indeed, the problems in which it is most important that we get this theory right are just the ones (such as arise in geophysics, econometrics, or antimissile defense) where we know from the start that the experiment can *never* be repeated.

To put it differently, the sampling distribution of an estimator is not a measure of its reliability in the individual case, because considerations about samples that have *not* been observed, are simply not relevant to the problem of how we should reason from the one that *has* been observed. A doctor trying to diagnose the cause of Mr. Smith's stomachache would not be helped by statistics about the number of patients who complain instead of a sore arm or stiff neck.

This does not mean that there are no connections at all between individual case and long-run performance; for if we have found the procedure which is 'best' in each individual case, it is hard to see how it could fail to be 'best' also in the long run.

The point is that the converse does not hold; having found a rule whose long-run performance is proved to be as good as can be obtained, it does not follow that this rule is necessarily the best in any particular individual case. One can trade off increased reliability for one class of samples against decreased reliability for another, in a way that has no effect oil long-run performance; but has a very large effect on performance in the individual case.

Now, if I closed the discussion of confidence intervals at this point, I know what would happen; because I have seen it happen several times. Many persons, victims of the aforementioned indoctrination, would deny and ridicule what was stated in the last five paragraphs, claim that I am making wild, irresponsible statements; and make some reference like that of Bross (1963) to the 'first-rate mathematicians' who have already looked into these matters.

So, let us turn to another example, in which the above assertions are demonstrated explicitly, and so simple that all calculations can be carried through analytically.

(d) EXAMPLE 6. THE CAUCHY DISTRIBUTION

We sample two members $\{x_1, x_2\}$ from the Cauchy population

$$p(dx|\theta) = \frac{1}{\pi}\frac{dx}{1+(x-\theta)^2} \tag{21}$$

and from them we are to estimate the location parameter θ. The translational and permutation symmetry of this problem suggests that we use the estimator

$$\theta^*(x_1, x_2) = \tfrac{1}{2}(x_1 + x_2) \tag{22}$$

which has a sampling distribution $p(d\theta^*|\theta)$ identical with the original distribution (21); an interesting feature of the Cauchy law.

It is just this feature which betrays a slight difficulty with orthodox criteria of performance. For x_1, x_2, and θ^* have identical sampling distributions; and so according to orthodox teaching it cannot make any difference which we choose as our estimator, for either point or interval estimation. They will all give confidence intervals of the same length, and in the long run they will all yield correct statements equally often.

But now, suppose you are confronted with *a specific* problem; the first measurement gave $x_1=3$, the second $x_2=5$. You are not concerned in the slightest with the 'long run', because you know that, if

your estimate of θ *in this specific case* is in error by more than one unit, the missile will be upon you, and you will not live to repeat the measurement. Are you now going to choose $x_1=3$ as your estimate when the evidence of that $x_2=5$ stares you in the face? I hardly think so! Our common sense thus forces us to recognize that, contrary to orthodox teaching, the reliability of an estimator is not determined merely by its sampling distribution.

The Bayesian analysis tells, us, in agreement with common sense, that for this sample, by the criterion of any loss function which is a monotonic increasing function of $|\theta^*-\theta|$ (and, of course, for which the expected loss converges), the estimator (22) is uniquely determined as the optimal one. By the quadratic loss criterion, $L(\theta^*, \theta)=(\theta^*-\theta)^2$, it is the unique optimal estimator whatever the sample values.

The confidence interval for this problem is easily found. The cumulative distribution of the estimator (22) is

$$\text{(23)} \qquad p(\theta^* < \theta' \mid \theta) = \tfrac{1}{2} + (1/\pi)\tan^{-1}(\theta'-\theta)$$

and so the shortest 100 P percent confidence interval is

$$\text{(24)} \qquad (\theta^* - q) < \theta < (\theta^* + q)$$

where

$$\text{(25)} \qquad q = \tan(\pi P/2).$$

At the 90% level, $P=0.9$, we find $q=\tan(81^\circ)=6.31$. Let us call this the 90% CI.

Now, does the CI make use of all the information in the sample that is relevant to the question being asked? Well, we have made use of (x_1+x_2); but we also know (x_1-x_2). Let us see whether this extra information from the individual sample can help us. Denote the sample half-range by

$$\text{(26)} \qquad y = \tfrac{1}{2}(x_1-x_2).$$

The sampling distribution $p(dy\mid\theta)$ is again a Cauchy distribution with the same width as (21) but with zero median.

Next, we transform the distribution of samples, $p(dx_1,dx_2\mid\theta)=p(dx_1\mid\theta)p(dx_2\mid\theta)$ to the new variables (θ^*, y). The jacobian of the transformation is just 2, and so the joint distribution is

$$\text{(27)} \qquad p(d\theta^*, dy\mid\theta)=\frac{2}{\pi^2}\frac{d\theta^* dy}{[1+(\theta^*-\theta+y)^2][1+(\theta^*-\theta-y)^2]}$$

While (x_1+x_2) are independent, (θ^*, y) are not. The conditional cumulative distribution of θ^*, when y is known, is therefore not (23), but

$$\text{(28)} \qquad \begin{aligned} p(\theta^*<\theta'\mid\theta, y)=&\frac{1}{2}+\frac{1}{2\pi}[\tan^{-1}(\theta'-\theta+y)+\tan^{-1}(\theta'-\theta-y)] \\ &+\frac{1}{4\pi y}\log\left[\frac{1+(\theta'-\theta+y)^2}{1+(\theta'-\theta-y)^2}\right] \end{aligned}$$

and so, in the subclass of samples with given (x_1-x_2), the probability that the confidence interval (24) will yield a correct statement is not $P=(2/\pi)\tan^{-1}q$, but

$$\text{(29)} \qquad w(y,q)=\frac{1}{\pi}[\tan^{-1}(q+y)+\tan^{-1}(q-y)]+\frac{1}{2\pi y}\log\left[\frac{1+(q+y)^2}{1+(q-y)^2}\right] \quad .$$

Numerical values computed from this equation are given in Table I,

TABLE I
Performance of the 90 % confidence interval for various sample half-ranges y

y	w(y, 6.31)	$F(y)$
0	0.998	1.000
2	0.991	0.296
4	0.952	0.156
6	0.702	0.105
8	0.227	0.079
10	0.111	0.064
12	0.069	0.053
14	0.047	0.046
>14	$4q/\pi(1+y^2)$	$2/\pi y$

in which we give the actual frequency $w(y, 6.31)$ of correct statements obtained by use of the 90% confidence interval, for various half-ranges y. In the third column we give the fraction of all samples, $F(y)=(2/\pi)\tan^{-1}(1/y)$ which have half-range greater than y.

It appears that information about $(x_1 - x_2)$ was indeed relevant to the question being asked. In the long run, the 90% CI will deliver a right answer 90% of the time; however, its merits appear very different in the individual case. In the subclass of samples with reasonably small range, the 90% CI is too conservative; we can choose a considerably smaller interval and still make a correct statement 90% of the time. If we are so unfortunate as to get a sample with very wide range, then it is just too bad; but the above confidence interval would have given us a totally false idea of the reliability of our result. In the 6% of samples of widest range, the supposedly '90%' confidence interval actually yields a correct statement less than 10% of the time - a situation that ought to alarm us if confidence intervals are being used to help make important decisions.

The orthodox statistician can avoid this dangerous shortcoming of the confidence interval (24), without departing from his principles, by using instead a confidence interval based on the conditional distribution (28). For every sample he would choose a different interval located from (29) so as to be the shortest one which in *that subclass* will yield a correct statement 90% of the time. For small-range samples this will give a narrower interval, and for wide-range samples a correct statement more often, than will the confidence interval (24). Let us call this the 90% 'uniformly reliable' (UR) estimation rule.

Now let us see some numerical analysis of (29), showing how much improvement has been found. The 90% UR rule will also yield a correct statement 90% of the time; but for 87% of all samples (those with range less than 9.7) the UR interval is shorter than the confidence interval (24). For samples of very small range, it is 4.5 times shorter, and for half of all samples, the UR interval is less than a third of the confidence interval (24). In the 13% of samples of widest range, the confidence interval (24) yields correct statements less than 90% of the time, and so in order actually to achieve the claimed reliability, the UR interval must be wider, if we demand that it be simply connected. But we can find a UR region of two disconnected parts, whose total length remains less than a third of the CI (24) as $y \to \infty$.

The situation, therefore, is the following. For the few 'bad' samples of very wide range, no accurate estimate of θ is possible, and the confidence interval (24), being of fixed width, cannot deliver the presumed 90% reliability. In order to make up for this and hold the average success for all samples at 90%, it is then forced to cheat us for the great majority of 'good' samples by giving us an interval far wider than is needed. The UR rule never misleads us as to its reliability, neither

underestimating it nor overestimating it for any sample; and for most samples it gives us a much shorter interval.

Finally, we note the Bayesian solution to this problem. The posterior distribution of θ is, from (21) in the case of a uniform prior density,

$$p(d\theta|x_1, x_2) = \frac{2}{\pi}\frac{(1+y^2)\,d\theta}{[1+(\theta-x_1)^2][1+(\theta-x_2)^2]} \tag{30}$$

and, to find the shortest 90% posterior probability interval, we compute the cumulative distribution:

$$\begin{aligned} p(\theta<\theta'|x_1, x_2) &= \frac{1}{2}+\frac{1}{2\pi}[\tan^{-1}(\theta'-x_1)+\tan^{-1}(\theta'-x_2)] \\ &\quad +\frac{1}{4\pi y}\log\left[\frac{1+(\theta'-x_2)^2}{1+(\theta'-x_1)^2}\right] \end{aligned} \tag{31}$$

and so, - but there is no need to go further. At this point, simply by comparing (31) with (28), the horrible truth appears: the uniformly reliable rule is precisely the Bayesian one! And yet, if I had simply introduced the Bayesian solution *ab initio,* the orthodox statistician would have rejected it instantly on grounds that have nothing to do with its performance.

(e) GENERAL PROOF

The phenomenon just illustrated is not peculiar to the Cauchy distribution or to small samples; it holds for any distribution with a location parameter. For, let the sampling distribution be

$$p(\mathrm{d}x_1 \ldots \mathrm{d}x_n \mid \theta) = f(x_1 \ldots x_n;\, \theta)\, dx_1 \ldots dx_n . \tag{32}$$

The statement that θ is a location parameter means that

$$f(x_1 + a,\, x_2 + a, \ldots x_n + a\,;\, \theta + a) = f(x_1 \ldots x_n;\, \theta), \quad -\infty < a < \infty. \tag{33}$$

Now transform the sample variables $\{x_1 \ldots x_n\}$ to a new set $\{y_1 \ldots y_n\}$:

$$y_1 \equiv = n^{-1}\sum x_i \tag{34}$$

$$y_i = x_i - x_1\,, \quad i = 2, 3, \ldots n. \tag{35}$$

From (33), (34), (35), the sampling distribution of the $\{y_1 \ldots y_n\}$ has the form

$$p(\mathrm{d}y_1 \ldots \mathrm{d}y_n \mid \theta) = g(y_1 - \theta\,;\, y_2 \ldots y_n)\, \mathrm{d}y_1 \ldots \mathrm{d}y_n . \tag{36}$$

If y_1 is not a sufficient statistic, a confidence interval based on the sampling distribution $p(\mathrm{d}\, y_1 \mid \theta)$ will be subject to the same objection as was (24); i.e., knowledge of $\{y_2 \ldots y_n\}$ will enable us to define 'good' and 'bad' subclasses of samples, in which the reliability of the confidence interval is better or worse than indicated by the stated confidence level. To obtain the Uniformly Reliable interval, we must use instead the distribution conditional on all the 'ancillary statistics' $\{y_2 \ldots y_n\}$. This is

$$p(\mathrm{d}y_1 \mid y_2 \ldots y_n\,;\, \theta) = Kg(y_1 - \theta\,;\, y_2 \ldots y_n)\, \mathrm{d}y_1 \tag{37}$$

where K is a normalizing constant. But the Bayesian posterior distribution of θ based on uniform

prior is:

$$(38) \qquad p(d\theta \mid x_1 \ldots x_n) = p(d\theta \mid y_1 \ldots y_n) = \\ = K g(y_1 - \theta\,;\; y_2 \ldots y_n)\, d\theta$$

which has exactly the same density function as (37). Therefore, by a refined orthodox criterion of performance, the 'best', (i.e., Uniformly Reliable) confidence interval for any location parameter is identical with the Bayesian posterior probability interval (based on a uniform prior) at the same level.

With a scale parameter σ, data $\{q_1 \ldots q_n\}$, set $\theta = \log\sigma$, $x_i = \log q_i$, and the above argument still holds; the UR confidence interval for any scale parameter is identical with the Bayesian interval based on the Jeffreys prior $d\sigma/\sigma$.

IV. POLEMICS

Seeing the above comparisons, one naturally asks: on what grounds was it ever supposed that confidence intervals represent an advance over the original treatment of Laplace? On this point the record is clear and abundant; orthodox arguments against Laplace's use of Bayes' theorem, and in favor of confidence intervals, have never considered such mundane things as demonstrable facts concerning performance. They consist of ideological slogans, such as "Probability statements can be made only about random variables. It is meaningless to speak of the probability that θ lies in a certain interval, because θ is not a random variable, but only an unknown constant".

On such grounds we are to be denied the derivation via Equations (1), (6), (9), (19), (30), (38) which in each case leads us in a few lines to a result that is either the same as the best orthodox result or demonstrably superior to it. On such grounds it is held to be very important that we use the words, "the probability that the interval covers the true value of θ" and we must *never, never* say, "the probability that the true value of θ lies in the interval". Whenever I hear someone belabor this distinction, I feel like the little boy in the fable of the Emperor's New Clothes.

Suppose someone proposes to you a new method for carrying out the operations of elementary arithmetic. He offers scathing denunciations of previous methods, in which he never examines the results they give, but attacks their underlying philosophy. But you discover that application of the new method leads to the conclusion that 2+2=5. I think all protestations to the effect that, "Well, the case of 2+2 is a peculiar pathological one, and I didn't intend the method to be used there", will fall on deaf ears. A method of reasoning which leads to an absurd result in *one* problem is thereby proved to contain a fallacy. At least, that is a rule of evidence universally accepted by scientists and mathematicians.

Orthodox statisticians appear to use different rules of evidence. It is clear from the foregoing that one can produce any number of examples, at first sight quite innocent-looking, in which use of confidence intervals or orthodox significance tests leads to absurd or dangerously misleading results. And, note that the above examples are not pathological freaks; every one of them is an important case that arises repeatedly in current practice. To the best of my knowledge, nobody has ever produced an example where the Bayesian method fails to yield a reasonable result; indeed, in the above examples, and in those noted by Kendall and Stuart (1961), the only cases where confidence intervals appear satisfactory at all are just the ones where they agree closely (or often exactly) with the Bayesian intervals. From our general proof, we understand why. And, year after year, the printing presses continue to pour out textbooks whose authors extol the virtues of confidence intervals and warn the student against the thoroughly discredited method of Bayes and Laplace.

A physicist viewing this situation finds it quite beyond human understanding. I don't think the history of science can offer any other example in which a method which has always succeeded was rejected on doctrinaire grounds in favor of one which often fails.

Proponents of the orthodox view often describe themselves, as did Bross (1963), as 'objective',

and 'fact-oriented', thereby implying that Bayesians are not. But the foundation-stone of the orthodox school of thought is this dogmatic insistence that the word 'probability' *must* be interpreted as 'frequency in some random experiment'; and that any other meaning is metaphysical nonsense. Now, assertions about the 'true meaning of probability', whether made by the orthodox or the Bayesian, are not statements of demonstrable fact. They are statements of ideological belief about a matter that cannot be settled by logical demonstration, or by taking votes. The only fully objective, fact-oriented criterion we have for deciding issues of this type, is just the one scientists use to test any theory: sweeping aside all philosophical clutter, which approach leads us to the more reasonable and useful results? I propose that we make some use of this criterion in future discussions.

Mathematically, or conceptually, there is absolutely nothing to prevent us from using probability theory in the broader Laplace interpretation, as the 'calculus of inductive reasoning'. Evidence of the type given above indicates that to do so greatly increases both the power and the simplicity of statistical methods; in almost every case, the Bayesian result required far less calculation. The main reason for this is that both the *ad hoc* step of 'choosing a statistic' and the ensuing mathematical problem of finding its sampling distribution, are eliminated. In particular, the F-test and the t-test, which require considerable mathematical demonstration in the orthodox theory, can each be derived from Bayesian principles in a few lines of the most elementary mathematics; the evidence of the sample is already fully displayed in the likelihood function, which can be written down immediately.

Now, I understand that there are some who are not only frightened to death by a prior probability, they do not even believe this last statement, the so-called 'likelihood principle', although a proof has been given (Birnbaum, 1962). However, I don't think we need a separate formal proof if we look at it this way. Nobody questions the validity of applying Bayes' theorem in the case where the parameter θ is itself a 'random variable'. But in this case the entire evidence provided by the sample *is* contained in the likelihood function; independently of the prior distribution, different intervals $d\theta$ are indicated by the sample to an extent precisely proportional to $L(\theta)d\theta$. It is already conceded by all that the likelihood function has this property when θ is a random variable with an arbitrary frequency distribution; is it then going to lose this property in the special case where θ is a constant? Indeed, isn't it a matter of the most elementary common sense to recognize that, in the specific problem at hand, θ is always just an unknown constant? Whether it would or would not be different in some other case that we arc not reasoning about, is just not relevant to our problem; to adopt different methods on such grounds is to commit the most obvious inconsistency.

I am unable to see why 'objectivity' requires us to interpret every probability as a frequency in some random experiment; particularly when we note that in virtually every problem of real life, the direct probabilities are not determined by any real random experiment; they are calculated from a theoretical model whose choice involves 'subjective' judgment. The most 'objective' probabilities appearing in most problems are, therefore, frequencies only in an *ad hoc,* imaginary universe invented just for the purpose of allowing a frequency interpretation. The Bayesian could also, with equal ease and equal justification, conjure up an imaginary universe in which all his probabilities are frequencies; but it is idle to pretend that a mere act of the imagination can confer any greater objectivity on our methods.

According to Bayes' theorem, the posterior probability is found by multiplying the prior probability by a numerical factor, which is determined by the data and the model. The posterior probabilities therefore partake of whatever 'qualities' the priors have:

(A) If the prior probabilities are real frequencies, then the posterior probabilities are also real frequencies.

(B) If the prior probabilities are frequencies in an imaginary universe, then the posterior probabilities are frequencies in that same universe.

(C) If the prior probabilities represent what it is reasonable to believe before the experiment, by

any criterion of 'reasonable', then the posterior probabilities will represent what it is equally reasonable to believe after the experiment, by the same criterion.

In no case are there any grounds for questioning the use of Bayes' theorem, which after all is just the condition for consistency of the product rule of probability theory; i.e., $p(AB|C)$ is symmetric in the propositions A and B, and so it can be expanded two different ways: $p(AB|C) = = p(A|BC)p(B|C) = p(B|AC)p(A|C)$. If $p(B|C) \neq 0$, the last equality is just Bayes' theorem:

$$P(A|BC) = p(A|C)\frac{P(B|AC)}{P(B|C)} \quad .$$

To recognize these things in no way forces us to accept the 'personalistic' view of probability (Savage, 1954, 1962). 'Objectivity' clearly does demand at least this much: the results of a statistical analysis ought to be independent of the personality of the user. In particular, our prior probabilities should describe the prior information; and not anybody's vague personal feelings.

At present, this is an ideal that is fully achieved only in particularly simple cases where all the prior information is testable in the sense defined previously (Jaynes, 1968). In the case of the aforementioned 'competent engineer' the determination of the exact prior is, of course, not yet completely formalized. But, as stressed before, the measure of our success in achieving 'objectivity' is just the extent to which we are able to eliminate all personalistic elements, and approach a completely 'impersonalistic' theory of inference or decision; on this point I must agree whole-heartedly with orthodox statisticians.

The real issue facing us is not an absolute value judgment but a relative one; it is not whether Bayesian methods are 100% perfect, or whether their underlying philosophy is opprobrious; but simply whether, at the present time, they are better or worse than orthodox methods in the results they give in practice. Comparisons of the type given here and in the aforementioned Underground Literature - and the failure of orthodoxy to produce any counter-examples - show that the original statistical methods of Laplace stand today in a position of proven superiority, that places them beyond the reach of attacks on the philosophical level, and *a fortiori* beyond any need for defense on that level.

Presumably, the future will bring us still better statistical methods; I predict that these will be found through further refinement and generalization of our present Bayesian principles. After all, the unsolved problems of Bayesian statistics arc ones (such as treatment of nontestable prior information) that, for the most part, go so far beyond the domain of orthodox methods that they cannot even be formulated in orthodox terms.

It would seem to me, therefore, that instead of attacking Bayesian methods because we still have unsolved problems, a rational person would want to be constructive and recognize the unsolved problems as the areas where it is important that further research be done. My work on maximum entropy and transformation groups is an attempt to contribute to, and not to tear down, the beautiful and powerful intellectual achievement that the world owes to Bayes and Laplace.

Dept. of Physics, Washington University, St. Louis, Missouri 63130

REFERENCES

Note: Two recent objections to the principle of maximum entropy (Rowlinson, 1970; Friedman and Shimony, 1971) appear to be based on misunderstandings of work done seventeen years ago (Jaynes, 1957). In the meantime, these objections had been anticipated and answered in other articles (particularly Jaynes, 1965, 1967, 11968), of which these authors take no note. To help avoid further misunderstandings of this kind, the following references include a complete list of my publications in which maximum entropy is discussed, although not all arc relevant to the present topic of Bayesian interval estimation.

Bayes, Rev. Thomas, 'An Essay Toward Solving a Problem in the Doctrine of Chances', Phil. *Trans. Roy. Soc.* 330-418 (1763). Reprint, with biographical note by G. A. Barnard, in *Biometrika* **45***,* 293-315 (1958) and in *Studies in the History of Statistics and Probability,* E. S. Pearson and M. G. Kendall, (eds), C. Griffin and Co. Ltd., London, (1970). Also reprinted in *Two Papers by Bayes with Commentaries*, (W. E. Deming, ed.), Hafner Publishing Co., New York, (1963).

Birnbaum, Allen, 'On the Foundations of Statistical Inference', *J. Am. Stat. Ass'n* **57** 269 (1962).

Bross, Irwin D. J., 'Linguistic Analysis of a Statistical Controversy', *The Am. Statist*. **17***,* 18 (1963).

Cox, D. R., 'Some Problems Connected with Statistical Inference', *Ann. Math. Stat.* **29**, 357 (1958).

Crow, E. L., Davis, F. A., and Maxfield, M. W., *Statistics Manual,* Dover Publications, Inc., New York (1960).

Fisher, R. A., *Statistical Methods and Scientific Inference,* Hafner Publishing Co., New York (1956).

Friedman, K. and Shimony, A., 'Jaynes' Maximum Entropy Prescription and Probability Theory', *J. Stat. Phys. 3, 381-384* (1971).

Good, I. J.**,** *Probability and The Weighing of Evidence, C.* Griffin and Co. Ltd., London (1950).

Good, I. J., *The Estimation of Probabilities,* Research Monograph *#30,* The MIT Press, Cambridge, Mass. (1965); paperback edition, 1968.

Jaynes, E. T., 'Information Theory and Statistical Mechanics, I, II', *Phys. Rev.* **106***,* 620-630; **108,** 171-190 (1957).

Jaynes, E. T., *Probability Theory in Science and Engineering,* No. *4* of *Colloquium Lectures on Pure and Applied Science,* Socony-Mobil Oil Co., Dallas, Texas (1958).

Jaynes, E. T., 'Note on Unique Decipherability', IRE Trans. on Information Theory, p. 98 (September 1959).

Jaynes, E. T., 'New Engineering Applications of Information Theory', in *Engineering Uses of Random Function Theory and Probability,* J. L. Bogdanoff and F. Kozin, (eds.), J. Wiley & Sons, Inc., N.Y. (1963); pp. 163-203.

Jaynes, E. T., 'Information Theory and Statistical Mechanics', in *Statistical Physics,* K. W. Ford, (ed.), W. A. Benjamin, Inc., (1963); pp. 181-218.

Jaynes, E. T., 'Gibbs vs. Boltzmann Entropies', Am. J. Phys. 33, 391 (1965).

Jaynes, E. T., 'Foundations of Probability Theory and Statistical Mechanics', Chap. *6* in *Delaware Seminar in Foundations of Physics,* M. Bunge, (ed.), Springer-Verlag, Berlin (1967); Spanish translation in *Modern Physics,* David Webber, (ed.), Alianza Editorial s/a, Madrid 33 (1973).

Jaynes, E. T., 'Prior Probabilities', IEEE Trans. on System Science and Cybernetics, SSC-4, (September 1968), pp. 227-241.

Jaynes, E. T., 'The Well-Posed Problem', in *Foundations of Statistical Inference,* V. P. Godambe and D. A. Sprott, (eds.), Holt, Rinehart and Winston of Canada, Toronto (1971).

Jaynes, E. T., 'Survey of the Present Status of Neoclassical Radiation Theory', in *Coherence and Quantum Optics,* L. Mandel and E. Wolf, (eds.), Plenum Publishing Corp., New York (1973), pp. 35-81.

Jeffreys, *H., Theory of Probability,* Oxford University Press (1939).

Jeffreys, *H., Scientific Inference,* Cambridge University Press (1957).

Kempthorne, O., 'Probability, Statistics, and the Knowledge Business', in *Foundations of Statistical Inference, V*. P. Godambe and D. A. Sprott, (eds.), Holt, Rinehart and Winston of Canada, Toronto (1971).

Kendall, M. G. and Stuart, *A., The Advanced Theory of Statistics,* Volume *2, C.* Griffin and Co., Ltd., London (1961).

Lehmann, E. L., *Testing Statistical Hypotheses,* J. Wiley & Sons, Inc., New York (1959), p. 62.

Pearson, E. S., Discussion in Savage (1962); p. 57.

Roberts, Norman A., *Mathematical Methods in Reliability Engineering,* McGraw-Hill Book Co., Inc., New York (1964) pp. 86-88.

Rowlinson, J. S., 'Probability, Information and Entropy', *Nature* **225**, 1196-1198 (1970).

Savage, L. J., *The Foundations of Statistics,* John Wiley, & Sons, Inc., New York (1954).

Savage, L. J., *The Foundations of Statistical Inference,* John Wiley & Sons, Inc., New York (1962).

Schlaifer, *R., Probability and Statistics for Business Decisions,* McGraw-Hill Book Co., Inc., New York (1959).

Sobel, M. and Tischendorf, J. A., Proc. Fifth Nat'l Symposium on Reliability and Quality Control, I.R.E., pp. 108-118 (1959).

Smith, C. A. B., Discussion in Savage (1962); p. 60.

NOTES

[1] Supported by the Air Force Office of Scientific Research, Contract No. F44620-60-0121.

[2] For those who had hoped, or at least expected, to hear instead a summary of the present status of maximum entropy, see the Note at the beginning of the References.

[3] This analysis is mathematically equivalent to use of the Behrens-Fisher distribution; however, the numerical work was done directly from Equation (1) rather than relying on tables which have been so little used and which would require a risky kind of interpolation. The first integration can be done analytically, and the second is easily done numerically to all the accuracy needed. Tail areas for $a<0$ need not be truncated, since they contribute to (1) only in the sixth decimal place.

[4] IBM *7092* calculation by Mr. Robert Schainker. Using the Jeffreys prior, $d\sigma/\sigma$, the posterior distributions have the form $p(d\sigma \mid s) = -x^r e^{-x} dx/r!$, where $x \equiv ns^2/2\sigma^2$, $2r=n-3$, and $s_1^2=2{,}237$, etc. The required probability is then an integral like (1), which can be expressed as a finite sum for numerical work. Alternatively, it can be expressed in terms of the incomplete Beta function, so that in principle the *F*-tables could be used; however, these tables use too widely separated values of the significance level for accurate interpolation.