E. T. JAYNES

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[1]

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.[2] So it seemed time to widen the front, and enter the arena on issue (2).

Harper and Hooker (eds.), Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II, 175–257. All Rights Reserved. Copyright © 1976 by D. Reidel Publishing Company Dordrecht-Holland.

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 are: (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 of 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 attention now to statistics, is more of a homecoming 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 statisfactory 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 common sense, 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

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

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:

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

where $\bar{t}_A$, $s_A^2=\overline{t_A^2}-\bar{t}_A^2$ 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

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

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 $(\bar{t}_B-\bar{t}_A)$ appear, unjustifiedly, 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 one-*

sided 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 one-sided 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_0?

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 $r\leqslant C$). On the hypothesis that we have the maximum tolerable failure rate, $\lambda_0=\theta_0^{-1}$, the probability that we shall see r or fewer failures is the binomial sum

$$(4) \qquad W(n, r)=\sum_{k=0}^{r}\binom{n}{k} e^{-(n-k)\lambda_0 t}(1-e^{-\lambda_0 t})^k$$

and so, setting $W(n, C) \leqslant 1-P$ gives us the sample size n required in order that this test will assure $\theta \geqslant \theta_0$ 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 $t=0.01\ \theta_0$ 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 $t=\theta_0$ with $C=1$, we need test only 5 units.

The amount of testing called for is appalling if $t \ll \theta_0$; 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 $\theta_0 \geqslant 10$ 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 informaion 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 $\lambda^{-1}=10$ 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

$$(5) \qquad p(r \mid n, \lambda, t) = \binom{n}{r} e^{-(n-r)\lambda t}(1-e^{-\lambda t})^r .$$

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

$$(6) \qquad p(d\lambda \mid n, r, t) = \frac{n!}{(n-r-1)!\, r!} e^{-(n-r)\lambda t}(1-e^{-\lambda t})^r \, d(\lambda t) .$$

The Bayesian acceptance criterion, which ensures $\theta \geqslant \lambda_0^{-1}$ with 100 P

percent probability, is then

$$(7) \qquad \int_{\lambda_0}^{\infty} p(d\lambda \mid n, r, t) \leqslant 1 - P.$$

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 $\{dt_1 \dots dt_r\}$ respectively, and the remaining $(n-r)$ units do not fail in time t, is

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

where $\bar{t} \equiv r^{-1} \sum t_i$ is the 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 $p(r, d\bar{t} \mid n, \lambda, t)$ 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 $\bar{t}$, 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

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

where

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

is the total unit-hours of failure-free operation observed. The posterior probability that $\lambda \geqslant \theta_0$ is now

$$(11) \qquad B(n, r) = \frac{1}{r!} \int_{\lambda_0 T}^{\infty} x^r e^{-x} dx = e^{-\lambda_0 T} \sum_{k=0}^{r} \frac{(\lambda_0 T)^k}{k!}$$

and so, $B(n, r) \leqslant 1-P$ 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_0 and observe only a very few failures. For, if $\lambda_0 t \ll 1$, 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_0. 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 $\bar{t} \geqslant 1.58\, \theta_0$. If we test 10 units and 9 fail, the ST test says we can assert with 90 percent confidence that $\theta \geqslant 0.22\, t$; the Bayesian test (11) says there is 90 percent probability that $\theta \geqslant 0.63\, \bar{t} + 0.07\, t$. 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_0, is

$$(12) \qquad p(\theta \geqslant \theta_0 \mid \lambda) = e^{-\lambda\theta_0}.$$

The probability that $\theta \geqslant \theta_0$, conditional on the evidence of the test, is therefore

$$(13) \qquad p(\theta \geqslant \theta_0 \mid n, t_1 \ldots t_r) =$$
$$= \int_0^\infty e^{-\lambda\theta_0}\, p(d\lambda \mid n, t_1 \ldots t_r) = \left(\frac{T}{T+\theta_0}\right)^{r+1}.$$

Thus, the Bayesian test which ensures, with 100 P percent probability, that the life of an *individual unit* is at least θ_0, has an acceptance criterion that the expression (13) is $\geqslant P$; 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 $\lambda_i \simeq t_i^{-1}$. 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

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

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 λ 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 $t_i \gtrsim \theta_0$, 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 $t_i \ll \theta_0$, 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 $\lambda = \lambda_0$ against $\lambda > \lambda_0$ 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

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

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 $(n-\mathrm{r}) \gg 1$, 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

$$(15) \qquad p(\mathrm{d}x \mid \theta) = \begin{Bmatrix} \exp(\theta - x)\,\mathrm{d}x\,, & x > \theta \\ 0 \qquad , & x < \theta \end{Bmatrix}.$$

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

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

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 $y \equiv (\theta^* - \theta + 1)$. 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 $F'(y_1)=F'(y_2)$. By computer, this yields the interval

$$(17) \qquad \theta^* - 0.8529 < \theta < \theta^* + 0.8264$$

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

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

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

$$(19) \qquad p(\theta \mid x_1 \dots x_N) = \begin{Bmatrix} N \exp N(\theta - x_1), & \theta < x_1 \\ 0 \quad , & \theta > x_1 \end{Bmatrix}$$

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

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

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 $0^* > x_1 + 0.85$; 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 on 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

$$(21) \qquad p(\mathrm{d}x \mid \theta) = \frac{1}{\pi} \frac{\mathrm{d}x}{1 + (x - \theta)^2}$$

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

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

which has a sampling distribution $p(\mathrm{d}\theta^* \mid \theta)$ 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

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

and so the shortest 100 P percent confidence interval is

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

where

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

At the 90% level, $P=0.9$, we find $q=\tan(81°)=6.31$. 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

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

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

Next, we transform the distribution of samples, $p(\mathrm{d}x_1, \mathrm{d}x_2 \mid \theta) = p(\mathrm{d}x_1 \mid \theta)\, p(\mathrm{d}x_2 \mid \theta)$ to the new variables (θ^*, y). The jacobian of the transformation is just 2, and so the joint distribution is

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

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

$$p(\theta^* < \theta' \mid \theta, y) = \tfrac{1}{2} + \frac{1}{2\pi}[\tan^{-1}(\theta' - \theta + y) + \tan^{-1} \times (\theta' - \theta - y)] + \frac{1}{4\pi y} \log\left[\frac{1 + (\theta' - \theta + y)^2}{1 + (\theta' - \theta - y)^2}\right] \tag{28}$$

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

$$\begin{aligned} w(y, q) = {} & \frac{1}{\pi}[\tan^{-1}(q + y) + \tan^{-1}(q - y)] + \\ & + \frac{1}{2\pi y} \log\left[\frac{1 + (q + y)^2}{1 + (q - y)^2}\right]. \end{aligned} \tag{29}$$

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	$\frac{4q}{\pi(1 + y^2)}$	$\frac{2}{\pi y}$

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,

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

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

$$(31)\qquad p(\theta < \theta' \mid x_1, x_2) = \tfrac{1}{2} + \frac{1}{2\pi}\left[\tan^{-1}(\theta' - x_1) + \tan^{-1} \times \right.$$
$$\left. \times\ (\theta' - x_2)\right] + \frac{1}{4\pi y} \log\left[\frac{1+(\theta' - x_2)^2}{1+(\theta' - x_1)^2}\right]$$

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

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

The statement that θ is a location parameter means that

$$(33)\qquad f(x_1 + a, x_2 + a, \dots x_n + a; \theta + a) = f(x_1 \dots x_n; \theta),$$
$$-\infty < a < \infty .$$

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

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

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

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:

$$\begin{aligned} &p(\mathrm{d}\theta \mid x_1 \ldots x_n) = p(\mathrm{d}\theta \mid y_1 \ldots y_n) = \\ &= Kg(y_1 - \theta; y_2 \ldots y_n)\,\mathrm{d}\theta \end{aligned} \tag{38}$$

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 $\mathrm{d}\sigma/\sigma$.

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 extoll 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 $\mathrm{d}\theta$ are indicated by the sample to an extent precisely proportional to $L(\theta)\,\mathrm{d}\theta$. 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 are 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., $p(AB \mid C)$ is symmetric in the propositions A and B, and so it can be expanded two different ways: $p(AB \mid C)= =p(A \mid BC)\, p(B \mid C)=p(B \mid AC)\, p(A \mid C)$. If $p(B \mid C)\neq 0$, the last equality is just Bayes' theorem:

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

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 are 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, 1968), 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 are 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.

DISCUSSION

INTERFACES BETWEEN STATISTICS AND CONTENT

(Remarks on the paper 'Confidence Intervals vs Bayesian Intervals' by E. T. Jaynes by Margaret W. Maxfield)

Professor Jaynes recommends common sense as a 'Court of Last Resort' for statistics. He calls for applying competing statistical methods and choosing the one whose result conforms best with our common sense.

However, one of the main reasons we apply statistical methods at all is to inform our sense in an area where we find it hazy. We want to know 'how big is big'.

Under Part II, Significance Tests, Example 1, Professor Jaynes explains an example from Roberts (1964; references are to the bibliography of Jaynes' paper):

> 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.

Roberts concludes from an F test that the two variances are not significantly different, whereupon he pools the estimates of variance, an error in Jaynes' judgment. In any case, Roberts pools the estimates incorrectly, using a formula

$$s^2_{\text{pooled}} = \frac{n_A s_A^2 + n_B s_B^2}{n_A + n_B - 2},$$

instead of the correct formula

$$s^2_{\text{pooled}} = \frac{(n_A - 1)\, s_A^2 + (n_B - 1)\, s_B^2}{n_A + n_B - 2},$$

(See Crow, page 68, for instance), thus sufficiently overestimating the pooled variance to yield a conclusion of no significant difference in means at the 90% level, with an equal-tails t-test. With the correct estimate, the difference in sample means proves significant.

(a) POOLING VARIANCES AS AN INTERFACE

Neither Roberts nor Jaynes mentions the use of the problem context in the decision whether to pool variance estimates. The reason for pooling estimates is that we consider that there is a common variance to be estimated. Ideally, we base this model, not on inspection of the data, but on the content of the application. For instance, there may be a 'state of the art' limitation in production of components, and experience may suggest that both suppliers are near that limit. Or, the tolerances on components that fulfill other specifications may be quite tight, so that unless the variance is 'in control', the components will fail grosser tests.

In Roberts' problem we might suppose that the buyer expects the variance to be the same, but does not trust his common sense as to whether the ratio of the observed estimates is improbably large. This use of statistics and their distributions to guide common sense is a good example of an interface between statistics and content.

A nonstatistical data analysis is quite appropriate, also. If the buyer must make his decision on the basis of the submitted samples alone, he may observe that the components from manufacturer B have both the better (longer) mean component life and the smaller estimate of variance, which would fit a rationale of better quality control by manufacturer B – better and less variable product.

(b) ALTERNATIVE HYPOTHESES AS AN INTERFACE

Jaynes' second objection to Roberts' solution is to his use of an equal-tails test, of which Jaynes says: "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". Of course, there is nothing immoral about abandoning all statistical procedures and awarding the contract to the winner, whether he won by a nose or a mile. Undoubtedly, most decisions are made in this commonsense way, a comparison of means with no examination of dispersions. Dispersions, and their effect on reliability of estimation, are demonstrably beyond sense that is common, as any introductory statistics class will reveal.

If a buyer does realize, however, that fluctuation affects the reliability of his estimate of the difference in means, he will want to consult t tables, and in entering those tables, use an equal-tails model.

Full decision-theoretic analyses are rare, partly because of buyers' inability or reluctance to quantify their loss estimates. Either the losses are very one-sided (the buyer awards the contract to his brother-in-law), or they are as hard to estimate as the difference in means. In the absence of any utility or loss functions, the Bayesian analysis is perfectly suitable. However, if there are any lower-order criteria than difference in 'mean component life available for the choice between manufacturers – transportation costs, tie-in sales, etc., – the buyer needs to know not only who won, but by how much.

Jaynes next attacks a problem from Crow *et al.* (1960) about comparison of variances. The problem in this 'Manual' is introduced explicitly as a mere exercise in calculating an F statistic and entering the F table. There is no surrounding information offered. The F statistic is calculated as 1.66, for a ratio of standard deviations of 1.3 to 1. The finding that the F statistic in this case is not significant at the 5% level violates Jaynes' common sense.

(c) SIGNIFICANCE AS AN INTERFACE

Upon learning that an F statistic is not especially improbable, we are surprised at what different variance estimates can arise from the same variance, perhaps. We may recheck our calculations. Then we revise our common sense.

Since in both the problems quoted, the experimenter originally does not know which competitor to choose, and, in fact, that is the point of the problem, a one-tailed model is inappropriate.

The policy Jaynes recommends, of reporting the significance level at which the result would be just significant, may seem to take the binary curse from significance testing for us. However, it must be emphasized that results in a critical region are improbable *in the aggregate,* not as individuals, every single one of which has zero probability, wherever it lies.

(d) PRIMITIVE NOTIONS

People use common nontechnical connotative understanding to draw commonsense conclusions about points, lines, sets, and so on – the primitive undefined terms in mathematics. Until ‘random’ and ‘probable’, or their successors as basic terms in statistics, are understood and used in nontechnical language, we must slowly develop common sense about statistics from experience.

COMMON SENSE AS AN INTERFACE

(Reply of E. T. Jaynes to comments of Margaret Maxfield)

It has been recognized by all, beginning with Laplace, that the purpose of a statistical analysis is to aid our common sense by giving a quantitative measure to what we feel intuitively. If it is thought that there is an inconsistency between this and my program, please note the distinction between (1) using a statistical method to help our common sense; and (2) judging the relative merits of two statistical methods by magnifying their differences up to the range where common sense needs no help.

Whether we use the coefficient n or $(n-1)$ in the pooled variance estimate depends on whether we define the symbol s^2 as the sample variance or the unbiased estimate of the population variance. Since Roberts uses n as the weighting coefficient and (p. 86) explicitly calls s the sample standard deviation, I assumed that he was using the former convention. If we reinterpret s^2 as does Ms Maxfield, then all the numerical results – both Roberts' and mine – will be changed slightly, but in the same direction and by nearly the same amount. This will not affect the comparison of our methods.

The remarks about one-sided and equal-tails tests, and about reporting critical significance levels, ignore some elementary facts that I tried to point out. That in an equal-tails test "the F statistic in this case is not significant at the 5% [my 95%] level," is just a mathematical fact, and in no way violates my common sense. On the contrary, it confirms my common sense by demonstrating the folly of using an equal-tails test. Ponder the proper fate of a Public Health official who obtains evidence for a difference in side effects of two polio vaccines that is significant at the 95.7% level by a one-sided test, and concludes: "We need not differentiate between the vaccines."

The purpose of these tests is to give an indication whether our data are consistent with some nominal value θ_0 for a parameter, or whether there is statistically significant evidence for a departure from θ_0. In deciding which test will best serve this purpose, then, we need to ask, "How much information bearing on this question do you convey when you report the result of the test?" This establishes an ordering much like the notion of admissibility.

Thus, for the t-test, denote the cumulative distribution by $\text{Prob}[t<t(P)]=P$, and consider what we learn from the results of one-sided and equal-tails tests. If Mr A tells us that the null hypothesis $[H_0:\theta=\theta_0]$ was not rejected by the equal-tails test at the 90% level, then we know t was somewhere in $|t|<t\,(0.95)$.We can't tell from this whether H_0 would be rejected at the 80% level. If he tells us that H_0 was rejected at the 90% level, then we know t was in one of the tails, $|t|\geqslant t\,(0.95)$; but we don't know whether H_0 would have been rejected at the 92% level, or which alternative $[H_1:\theta>\theta_0]$ or $[H_2:\theta<\theta_0]$ is favored by the data.

If Mr A would report instead the critical level P for the equal-tails test, we would know far more. This determines that $t=\pm t\,(P')$, where $2P'=1+P$, and we know what the verdict would be at any level, for the equal-tails test. The critical level P_1 for the one-sided test $(H_0\text{ vs. }H_1)$ is either P' or $(1-P')$, but we can't tell which.

Evidently, we would know still more if Mr A would report the critical level P_1 for the one-sided test, instead of the equal-tails test. From P_1 we know what the verdict would be, at any level, for the equal-tails test *and* for both of the one-sided tests. But it is straightforward mathematics to show that P_1 is identical with the Bayesian posterior probability that H_1 is true.

All these considerations apply equally well to the F-test, and many others. A one-sided test tells us everything an equal-tails test does; and more. Where, then, is the justification for ever using an equal-tails test, or for claiming that "a one-tailed model is inappropriate?"

More importantly, it is not a matter of personal opinion, but a mathematically demonstrable fact, that the Bayesian method of significance testing, originated by Laplace, leads us at once to the maximum information given by the optimal orthodox test. Obviously, then, orthodox rejection of Bayesian tests cannot be justified on grounds of their actual performance.

Finally, I call the readers' attention to the devastating criticisms of orthodox hypothesis testing theory by Pratt (1961) and L. J. Savage (1962) which, to the best of my knowledge, remain unanswered to this day.

COMMENTS ON PAPER BY DR E. T. JAYNES

Oscar Kempthorne

(1) The paper by Jaynes is clearly a very seriously developed discussion of some of the problems and obscurities of statistical inference. My own views are given briefly in my presentation at this confernce. I shall not give these here but shall attempt a critique, not necessarily critical, of Jaynes' paper.

(2) I am very concerned that the picture of 'orthodox' statistics presented by Jaynes will lead philosophers of science and physicists to the view that statistics *as it is often practised* is stupid. It is a hard, bare fact that workers in noisy sciences use statistical methods, as presented, for example, by Snedecor and Cochran, very widely. Part of the thesis of my own presentation is that there has been very little attention to unavoidable noise in philosophy and physics. I refuse to discuss the matter with anyone who does not admit the problem of noise (or error, or wandering, or variability, whatever term appeals).

(3) At the beginning we are presented with a polarity, *the* orthodox solution versus *the* Bayesian solution. This needs clarification.

(a) What is the orthodox solution? There are in fact at least two streams of thought and statistical practice arising from a common origin. Fisher in his first paper used a Bayesian argument, but then by 1922 ('The Mathematical Foundations of Theoretical Statistics') had rejected Bayesian ideas. Instead, he produced (i) a large number of significance tests and (ii) a theory of statistical estimation. A natural step was then to construct some sort of statistical interval of uncertainty by inverting a significance test. This led later to Fisher's fiducial inference, which is a mystery to all but a very few, and has, I think, been rejected by almost all statisticians (including, perhaps, myself). In 1928, Neyman and Pearson tried to give more exact and more mathematical structure to the idea of tests of singificance. By 1933 they had replaced, in my opinion, significance tests by accept-reject rules and had cast the whole matter into a simple decision theoretic-structure. This led to some of the procedures which Jaynes justly castigates. The decision-theoretic approach with emphasis on frequency of errors of the two types was seen by Fisher to be in conflict with a need for quantifying in a reasonable way what may be termed the evidential content of data. There has been much ferment on the basic issue. The literature is now huge. The upshot has been a rejection by very many of the idea that Neyman-Pearson accept-reject rule theory necessarily leads to a valid quantification of evidence from data. I reviewed (*Biometrics* **25**, 647–654, 1969, discussion of paper by Cornfield) in an elementary way some of the problems and gave an example very much like Jaynes' Example 5. A serious attempt was made in the book *Probability, Statistics and Data Analysis,* by O. Kempthorne and J. L. Folks, Iowa State University Press, 1971 to present the problems. I think it is described there in what ways the Neyman-Pearson processes break down, for example, how the idea of unbiased *tests of significance* breaks down for the simplest case of independent Bernouilli trials. The story is a very long one. If, however, one wishes to enter deep discussion of the controversies a huge amount of literature, due to Fisher, Barnard, Cox, Birnbaum and others (perhaps including myself), must be read with a critical mind. At the same time one must read with a critical mind the writing of L. J. Savage, Lindley, I. R. Savage, Box and others. And also one must read with a critical mind the theory of games (especial Section 4.8.2 of the von Neumann-Morgenstern classic) and the theory of decision which is closely related to the theory of utility (whence the theory of preferences). I have found *Games and Decisions* by Luce and Raiffa very informative and I would like readers to pay especial attention to pages 33 to 37. The overwhelmingly strong message for me from the London conference was that a large portion of the ideas of outstanding workers can reasonably be subjected to severe criticism. *No one's work is sacrosanct.* We heard criticisms of the basic work in modern physics, and we know of the extreme doubts of Einstein about quantum physics and also of the extreme doubts of the validity of Einstein's criticisms. So I put forward a guiding principle: *It is complete naiveté to assume that a presentation by worker X (very good though he may be) of a theory of physics or a theory of statistics, or a theory of decision, or whatever can be taken as definitive and totally forcing.*

Any writer presents, of course, as convincing a case as possible, and almost every writer has contributed to understanding.

The basic point about 'orthodox' statistics is that there is an orthodoxy in the books on mathematical statistics, but this orthodoxy is present only mildly, and almost tangentially, in the orthodoxy (if there is one) of practising statisticians, as represented by a number of texts on statistical methods. To attack Neyman-Pearson orthodoxy is one thing (which may be accomplished with some success, I think) but to assume that this orthodoxy has had any deep influence on *working* statistical practice is, I think, largely fallacious.

It is a bad thing, I believe, to suggest, for example, that working statisticians will estimate a probability by a formula that can give an answer of (2), as a leading text suggests.

Or to suggest that working statisticians will estimate a probability to be a negative number, as a Nobel laureate in physics, P. A. M. Dirac, suggested in his Baker lecture to the Royal Society of London in 1942.

Dr. Jaynes and the neo-Bayesians have a very difficult problem arising from the fact that there are few (perhaps no) books which describe *orthodox statistical practice* from a theoretical viewpoint (though I suggest, perhaps naturally, that the book by Kempthorne and Folks attempts this).

I repeat that the whole field is very, very difficult, and give my view that the difficulties are not mitigated by the existence of many books on mathematical statistics that give a picture, which is ludicrous, as our neo-Bayesian friends are asserting, and correctly so, I think. I now turn to a detailed reaction to the Jaynes paper.

(b) Jaynes talks about *the* Bayesian solution. I state emphatically with no fear of being proved wrong, that there is *not* a definite Bayes solution. There is a Bayes 'solution' associated with each choice for a prior distribution. This ambiguity would be resolved if there were a completely compelling way to produce a particular prior distribution for each problem by purely logical analysis. The attempts to do this have failed, I believe. Professor Lindley himself has so admitted, I believe.

If this is accepted, then it is incorrect to talk about *the* Bayesian result. If the attitude to the whole problem is changed, so that the prior distribution is a summarization of the *beliefs* of the individual investigator (a viewpoint which L. J. Savage proposed, I believe), then the result must be labelled as the result of the individual's prior. So if Jaynes is using his personal prior, the result of the Bayes algorithm should be called 'Jaynes' probability' or 'Jaynes' interval'. I do not find such statements offensive or misleading. In a real sense, all probabilities of future contingencies are personal. I have no objection at all to stating 'Kempthorne's probability that it will rain tomorrow is 0.3'. Whether anyone else should accept this as a reasonable probability *for him* is another problem. I do accept probabilities of authorities, e.g. in the genetic counselling area. If Professor Lindley would state 'My probability of hypothesis H is 0.7', I cannot possibly quarrel with him. He is making an assertion about himself and I regard him as the best available authority on himself and his beliefs.

(4) When Jaynes talks about the 'orthodox' solution I am of the opinion that he is talking about the Neyman-Pearson decision-theory based solution. That this became orthodoxy was a frequently and strongly voiced objection by R. A. Fisher. References are very well known and are manifold. Hence to use a phrase such as the Fisher-Neyman-Pearson approach is an obvious calumny of Fisher's views and situation. In his name I protest most strongly. For the benefit of readers who were not present at the conference, it may be stated that *precisely* this phrase was used.

(5) That the obscurities of modern physics will be resolved shortly may well be hoped, but a message I obtained from the London conference was that the foundations of modern physics are in a shambles. We have the appeal to the two-hole experiment, which we are then told is a 'Gedanken' or 'thought' experiment which has never been done. [Though again, there is a suggestion that this basic experiment was done about a year ago by someone, this 'year ago' being some fifty years after the theory was promulgated as *the final answer*. (See the quotations in my own essay.)] So I regret that I do not share Jaynes' optimism, though I do share a wonderment at the advances in technology that have grown out of modern physics, and I support strongly research in physics.

(6) Jaynes appeals to 'common sense'. I suggest most strongly that our problem is to decide what 'common sense' is. We have seen 'common sense' to be totally fallacious a huge number of times in the history of human thought. And to attempt to push the point home, I suggest that Jaynes would have very great, and perhaps insuperable, difficulties in reconciling quantum physics with common sense. It is true that we all have a vague idea of something which we call 'common sense', but our problem is to make this vague feeling sufficiently precise to use in scientific discourse. One cannot help recalling that appeals to common sense, to motherhood, to land and order and so on have led to some of the worst atrocities that humanity has perpetrated.

So I assert my opinion that Jaynes is being strongly misleading in talking about what common sense tells us. I am sure that Jaynes is highly proficient in branches of physics and that he would reject my 'common sense' and justifiably so, perhaps.

A 'publicly agreed verdict' may well be terribly wrong, and history is replete with examples.

(7) Jaynes and I are agreed that a statistical method should be judged 'by the results which it gives in practice'. It was precisely on this basis that the Bayesian idea was rejected by Boole, Venn, Fisher, Neyman and so on. I hope that this list of names gives Jaynes pause. The simple fact is that a Bayesian interval has *no* predictive verifiability. The neo-Bayesian cannot successfully back his probability assertions by accepting money challenges. On the other hand, the Neyman-Pearson interval assertion is straightforward: he asserts: I will bet $ 19 to $ 1 that my interval contains the true unknown parameter value. It is a matter of mathematics that this claim can be sustained. Neyman can issue this assertion and will apart from random fluctuations remain solvent. The neo-Bayesian will, on the other hand, 'lose his shirt'. He will be coherent in his whole battery of probability assertions, but he will be *coherently wrong,* in a situation in which an individual A chooses a probability structure and valid data, individual B, a neo-Bayesian, makes his probability assertions, and individual C (say Neyman) challenges B's assertions, with individual A then verifying correctness of assertion.

(8) The reader may check whether the book by Kempthorne and Folks mentions the works of Good, Savage and Jeffreys.

(9) Jaynes says "The basic ideas of interval estimation must be ancient". We will agree with him, but the problem is surely to give some logically satisfying structure for this basic idea. He gives a view, the need for an interval of uncertainty or a final best number, with which I agree but a view with which L. J. Savage who is surely one of the originators of neo-Bayesianism, disagrees in his foundations. Savage's book should be read, and the assertion of this view will be found.

(10) While Neyman-Pearson orthodoxy states that the "proper method for this problem is the confidence interval", Fisher objected strongly and consistently that the proper method was *not* the confidence interval. A view is expressed in the Kempthorne-Folks book which is along the same lines. It seems to be agreed by a sizeable group of practising statisticians that one cannot *necessarily* have confidence in confidence

intervals. It was for this reason that Folks and I introduced the term 'consonance interval' and this suggestion has met with approval by *some* workers.

(11) The words subjectivity and objectivity should be banned from the literature. But the situation is not simple. We cannot take Jaynes' common sense to be a substitute for a requirement of interpersonal validation of subjectivety formed assertions which is what 'objectivity' usually means. The comments of Bross cannot, I believe, be dismissed by a few words and a wave of the hand. I do not thereby imply that I agree *totally* with him.

(12) A question that recurs again and again is "What is a significance test?" There has been in the opinion of some statisticians, a vast confusion on the matter. A distinction between a significance test and an accept-reject rule has been made by myself, by Kempthorne and Folks in their book, by Kalbfleisch in his book, and by Kalbfleisch and Sprott in a paper presented at this conference. The confusion has been generated by a mixing of phrases which should be kept separate. One meets an infelicity of phrase of an accept-reject rule at 5% significance level. This mixes up significance tests and accept-reject rules to the confusion of most readers. A significance test is a quantification of 'strength of evidence against' expressed on the probability (ie. the [0, 1]) scale. The confusion is not confined to Jaynes' paper but permeates the mathematical statistics area.

(13) On Example 1, Jaynes' common sense tells him what he says. I suggest, however, that the standard error of the difference of means is $\sqrt{7.48^2+6.48^2} \doteq \sqrt{98} \doteq 10$. Hence the difference of means is 8 with a standard error of 10. There is evidence that B is better than A, but not 'fairly substantial' in my opinion. No one would quarrel, I believe, with a *decision* to choose B, but the situation evidentially is by no means as clear as Jaynes with his 'common sense' approach suggests.

(14) The statement "any statistical procedure which fails to extract evidence that is already clear to our unaided common sense, is certainly *not* for me", is excellent polemics, but, I suggest, no more than that. If Jaynes' circulated document is his 'common sense', then I believe we must reject a sizeable portion of this polemic.

(15) In Example 1 and elsewhere, Jaynes uses improper priors. Hacking in 1962, or thereabouts, raised the question of how a quantity

which is not a probability could be convoluted with a probability to yield a probability. Many of us have been utterly queasy about improper priors. In discussion of Fraser's presentation, Lindley mentioned recent work by Dawid, Stone and Zidek which appears to show that Fraser's structural inference has consistency problems. I am very sorry that the fact that *the same paper* appears to show irremovable defects in the neo-Bayesian process using improper priors was not mentioned. If the force of the Dawid-Stone-Zidek paper is accepted, as Lindley does, it seems, because he used it as an argument against Fraser's structural inference, then it would seem that the 1967 book of Lindley and the bulk of the Jaynes paper should be rejected, as well as other recent Bayesian books. At the conference, I attempted to obtain Lindley's view of the effect of the Dawid-Stone-Zidek paper on his own earlier work, and hence on Jaynes' work reported at this conference, but was *unsuccessful.* I hope that Lindley will contribute a definitive statement to the proceedings of this conference.

(16) We see the sentence 'How then, could the author have failed to find significance at the 90% level?'. This raises the question of what a significance test is. Some obliquely directed views of Jaynes seem to be aimed in the direction put forward in the Kempthorne-Folks book. But most of the mathematical-statistics literature is unclear on the matter. I recall Wolfowitz saying twenty years ago that Fisher had never defined a test of significance. He was correct, I believe. I would also add that I am not entirely clear in my own mind on what a significance test is, though what is given in the Kempthorne-Folks book and in my Snedecor essay represents an initial stab.

(17) In Example 2 we are told there is "absolutely unmistakable evidence for the superiority of type 2 rockets". I believe the critical reader should jib at this. I certainly do. I would give a statement of "quantitatively *how significant* that evidence is". I would report "*the* critical significance level at which etc". I am on record with Folks as advocating precisely this mode of statement. But, then, perhaps I and Folks are not 'orthodox' statisticians, or perhaps we are not statisticians at all!

(18) Throughout the examples we are given in Jaynes' words *the* Bayesian solution. But Dawid-Stone-Zidek tell us that *the* Bayesian solution (as given by Jaynes) has defects, and we can only surmise, in the absence of clear statement, that Lindley agrees on the presence of defects.

(19) We see that certain results are based on a 'ridiculous prior for λ'. This raises the question of what is a poor prior. Clearly Jaynes admits that possibility. On the other hand, Lindley states that the quesion of accuracy of a prior does not arise. And, finally, we heard a lecture by Patrick Suppes which cast extreme (for me) doubts on the whole Savage prescription, which Lindley has stated *at several times* to be his intellectual basis. To place the matter in focus, a prior that Jaynes uses may be 'ridiculous' to me and vice versa, and that is *precisely* why the Bayesian process was rejected by Boole, Venn, etc., etc.

(20) It is surely the case that a Bayesian interval makes predictions about an *imaginary* population of repetitions. The neo-Bayesians do not like such a phrase, but their opponents state something which is as close to this as the looseness of language permits.

(21) On Example 5, the points that Jaynes makes are extremely close to those that Fisher made in 1934 (39 *years ago*) in rejecting the Neyman-Pearson prescription. The notion of *recognizable* subsets was put forward very early, if not first, by Fisher. I *infer* that the literature has not been read.

(22) I have already indicated implicitly that there are difficulties with the Savage axiomatic system. I call Patrick Suppes to witness.

(23) I am of the opinion that the consideration of axiomatic structures is a very important part of logical thinking. But I believe it is exceedingly dangerous to accept any set of axioms, no matter how 'true' and 'correct' they appear to be. I have seen writings recently that indicate that certain workers are very doubtful of 'the sure-thing principle'. I suggest that we listen a bit to these workers.

(24) The neo-Bayesians quote F. P. Ramsey as the originator. They fail, however, to record that Ramsey recommended to refer to Fisher on applicational matters. The last sentence of Ramsey goes like "For all this see Fisher".

(25) We await with eagerness the Fisher lecture of L. J. Savage, which was not as assertive as his followers might well expect. (I was chairman of the session.)

(26) Jaynes says that the neo-Bayesian prescription 'works'. Just what does he mean and what is the evidence, apart from *his* common sense.

(27) That multiple comparison procedures are thought to be defective by many practitioners (and non-Bayesians) is well-known.

(28) L. J. Savage was crystal clear in his presentation that he was giving a *theory of decision for one person,* who, of course, has to be L. J. Savage. I do not, thereby, denigrate Savage. He was a fine individual with a very fine mind.

(29) Ultimately, the basic antithesis is between evidence and decision. To some everything is decision. To others and indeed everyone, decision is important. Every human and *animal* makes decisions. But there is another aspect, the accumulation and weighing of *evidence.* Just what this is, is not clear. But the lesson of all human thought is not to dismiss a vague but pervasive idea because one cannot formulate it tightly.

Statistical Laboratory, Iowa State University

REFERENCES

Cornfield, J., 1969, 'The Bayesian Outlook and its Applications', *Biometrics* **25**, 617–642, for a good listing of background material.

Dawid, A. P., M. Stone and J. V. Sidek, 1973, 'Marginalization Paradoxes in Bayesian and Structural Inference', *J. Roy. Stat. Soc. B.*, in press.

Dawid, A. P. and M. Stone, 1972, 'Expectation Consistency of Inverse Probability Distributions', *Biometrika* **59**, 486–489.

Easterling, R. G., 1972, 'A Personal View of the Bayesian Controversy in Reliability and Statistics', *IEEE Trans.* R-21, 186–194.

Fisher, R. A., 1959, *Statistical Methods and Scientific Inference*, 2nd ed., Hafner, New York.

Kalbfleisch, J. G., 1971, *Probability and Statistical Methods*, Dept. of Statistics, Waterloo.

Kalbfleisch, J. G., and D. A. Sprott, 1973, 'On Tests of Significance', this volume, p. 259.

Kempthorne, O., 1969, *Biometrics* **25**, 647–654. Discussion of paper by J. Cornfield.

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

Kempthorne, O., 1972, 'Theories of Inference and Data Analysis', in *Statistical Papers in Honor of George W. Snedecor*, Iowa State Univ. Press, Ames.

Kempthorne, O. and J. L. Folks, 1971, *Probability, Statistics, and Data Analysis*, Iowa State Univ. Press, Ames.

Luce, R. D. and H. Raiffa, 1957, *Games and Decisions*, Wiley, New York.

Ramsey, F. P., 1926, 'Truth and Probability', in *The Foundations of Mathematics and Other Logical Essays*, Kegan, London.

Savage, L. J., 1954, *The Foundations of Statistics*, Wiley, New York.

Snedecor, G. W. and W. G. Cochran, 1967, *Statistical Methods*, 6th ed., Iowa State Univ. Press, Ames.

Stone, M. and A. P. Dawid, 1972, 'Un-Bayesian Implications of Improper Bayes Inference in Routine Statistical Problems', *Biometrika* **59**, 369–375.

JAYNES' REPLY TO KEMPTHORNE'S COMMENTS

I am most grateful to Professor Kempthorne for this lengthy commentary. Such a magnificent confirmation of my main thesis could hardly have been hoped for; he has surely silenced those critics who thought that my account of the orthodox position was exaggerated.

Before venturing into areas where we presently differ I want to say that, during our five days acquaintance at this Conference, I have developed a warm personal affection for Oscar Kempthorne, and came to seek him out for many between-sessions and after-dinner discussions, all pleasant and valuable to me for reasons ranging from his interesting comments to the aroma of his cigars. Although it may not be apparent to the casual reader, there is a very wide area of agreement between us; on most of the issues discussed at this Conference, we would stand together.

For example, we both see at a glance the sterility of efforts to refine the mathematics without refining the concepts; or to axiomatize old ideas without any creative development of new ones. We are, I think, equally appalled at the prospect of changing the principles of logic to accommodate an illogical theory of physics.

We both tend to place more emphasis on the practical working rules and less on highflown mathematical and philosophical aspects of statistics than some of our younger colleagues, because we have seen enough ambitious but short-lived efforts with the generic title: 'A New Foundation for Statistics' to become a bit weary of them. And we have seen enough putative 'foundations' develop a fluid character unlike real foundations and adapt themselves to the unyielding practical realities, to become a bit wary of them.

It is clear to me that, on a much deeper level than the superficial differences being aired here, Oscar Kempthorne and I are kindred souls, with the same basic outlook and value judgments. On studying his comments, I am convinced that our differences arise almost entirely from misapprehensions concerning the nature of Bayesian methods *as they exist today,* which could have been cleared up if only we had more time to thresh matters out. Surely, there is no difference in our real aims to improve the power and scope of statistical methods at the practical, working level.

But granting all this, the differences between us do involve issues of crucial importance to statistics, and it would be a disservice to minimize them. This 120-year-old hangup over prior probabilities, started by Boole, must come to an end, because it is the direct cause of the troubles that today prevent orthodox statistics from giving any useful solutions to many important, real statistical problems.

Thus, linear regression with both variables subject to error is one of the most common statistical problems faced by experimenters; yet orthodox theory is helpless to deal with it because with n data points we have $(n+2)$ nuisance parameters. In irreversible statistical mechanics, and in some mathematically similar problems of communication theory and business decisions, the only probabilities involved are prior probabilities. The possibility of any useful solutions at all depends on principles such as maximum entropy, for translating prior information into prior distributions.

This debate has gone on for over 100 years, with the same old arguments and counter-arguments repeated back and forth for generations, without ever getting anywhere. Philosophical disputation may be great fun; but through recorded history its score for actually solving problems is, I believe, precisely zero. Anybody who genuinely wants to see these issues resolved must recognize the need for a better method.

Now the present condition of statistics is just the condition physics was in until the late 16th century, when Galileo showed us a better method – the direct cause of the advances that physics has made since. Instead of arguing about how objects 'ought' to move according to some philosophical or theological preconceptions, or by quoting ancient authorities such as Aristotle, why don't we just use the evidence of our own eyes? We are surrounded daily by moving objects; so any proposed theory about how they move can be tested by direct observation of the facts.

But, as this Conference showed very dramatically, 400 years of 'enlightenment' have not changed basic human nature. Today, statisticians regard themselves as the guardians of 'scientific objectivity' in drawing conclusions from data. Yet when I suggested that their own methods be judged, not by the philosophical preconceptions underlying them, but by examination of the facts of their actual performance, this appeared to many – as I knew it would – just as radical and shocking at as it did to Galileo's contemporaries. After my talk, a half-dozen people remonstrated with me, trying to inform me about the terrible defects of Bayesian methods by repeating the same tired old Boole-Venn clichés that we all learned as children. Not one of these individuals took the slightest note of the contrary facts (the mathematically demonstrable relations between actual performance of Bayesian and orthodox methods) that I had just pointed out. So we had an exact 20'th century repetition of Galileo's experience with the colleague who refused to look through his telescope.

To answer fully every point raised by Kempthorne would require a document much longer than my original presentation. Therefore, this reply must be confined to a brief summary of the situation, followed by specific comments only on those points of fact which are of general interest, and which would propagate confusion if they were allowed to go unanswered.

SUMMARY

My presentation was concerned with examining the relative merits of orthodox and Bayesian statistical methods by considering specific real problems, giving for each *an* orthodox solution which has been advocated in the recent literature, and adding what cannot be found in that literature, namely *the* Bayesian solution *which makes use of the same information* (i.e., is based on a noninformative prior). In Example 3, we also examined the further improvement obtainable when definite prior information is put in by maximum entropy. From these comparisons, several substantive conclusions emerge, which can be summed up as follows: Orthodox methods, when improved to the maximum possible extent (by using one-sided tests, reporting critical significance levels, using sufficient statistics or conditioning on all ancillary information, etc.) become mathematically equivalent to the Bayesian methods based on noninformative priors, provided that no nuisance parameters are present, and a sufficient statistic or complete set of ancillary statistics exists. Otherwise, mathematical equivalence cannot be achieved, and magnification then shows the Bayesian result to be superior.

This conclusion is supported in part by general theorems, in part by examination of specific cases. By now, we have a multitude of specific worked-out examples supporting it; and anyone who has understood my analysis can see that we are prepared to mass-produce any number of additional examples. Orthodox statistics has yet to produce *one* counter-example. The reason for this is clear to one who has studied the theorems of R. T. Cox (1946, 1961). He shows that any method of plausible reasoning in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to Laplace's, or inconsistent.

Even though an orthodox statistician may, in the words between his equations, vociferously denounce the use of Bayes' theorem, it is nevertheless a matter of straightforward mathematics to see if his actual conclusions can be derived from Bayes' theorem. Either they can or they cannot. If they can, then it is obvious that his rejection of the Bayesian method is not based on its actual performance. If the conclusions are different, then we have the opportunity to judge that difference by Galileo's method. If we can magnify the difference sufficiently, it will become quite obvious which method is giving sensible results, and which is not.

Let me stress this point. Doubtless, some readers will jump to the conclusion that I deliberately chose examples to support my prejudices; and that one can just as easily produce examples on the other side. In fact, I hope that every reader of the orthodox persuasion will come to exactly that conclusion, and set about immediately to produce six examples where an orthodox method yields a result that simple common sense can see is preferable to the Bayesian result. For it is not in the passive reading of my words, but in the active attempt to produce these counter-examples, that one's eyes will be opened.

Professor Kempthorne's appraisal of my efforts falls somewhat short of the warm approbation that I had naturally hoped for. As he notes (Item 3), any writer presents 'as convincing a case as possible'. Presumably, therefore, if he was in a position to refute any of my claims – whether by exhibiting an error in mathematics, a counter-example, or a documentable contrary fact – he would have done so. Yet with a single exception, discussed below (Item 13), he does not even mention any of the substantive issues raised. Instead he favors us with pleasantries about my choice of words and phrases.

Kempthorne complains, with some justice, that I did not criticize orthodox methods as they exist today, but rather as they existed before publication of his recent book (Kempthorne and Folks, 1971; hereafter referred to as KF). But then what shall one say about reaching back to Boole (1854) and Venn (1866) for criticisms of Bayesian methods as they existed over 100 years ago; thereby ignoring not only my recent work on these problems (1968, 1971), but also that of Jeffreys (1939)? I can well imagine the howls of outrage and cries of 'unfair' that would issue forth if I went back only 37 years to quote Karl Pearson (1936) as my authority in criticizing maximum likelihood. Now let us take up some of Kempthorne's more specific comments.

(2) My topic was the relative merits of orthodox and Bayesian methods, and not how they correlate with intelligence. Not having studied the latter topic, I have nothing more to add to the conclusions already reported by professional statisticians, viz:

> I believe, for example, that it would be very difficult to persuade an intelligent physicist that current statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes' theorem.
>
> G. E. P. Box (1962)

> A student of statistical methods tends to be one of two types; either he accepts the technique in its entirety and applies it to every conceivable situation, or he is more intelligent and questions the applicability at all.
>
> O. Kempthorne (1952)

With regard to the other remark, I think an historical study would show that the reasons for the interest of both Laplace and Jeffreys in probability theory arose from the problem of extracting 'signals' (i.e., new systematic effects) from the 'noise' of imperfect observations, in astronomy and geophysics respectively. The procedures would today be called 'significance tests', and I wish every one who has not already done so, would read Jeffreys' (1939) beautiful and comprehensive chapters on significance tests, then compare them from the standpoint of solid content and usefulness in real problems, with any work ever written on the subject from the orthodox point of view.

Likewise, my own interest in statistics arose from problems of extracting signals from noise in several applications ranging from optimum design of radar receivers and magnetic resonance probes, to land mine detectors. I am on record (Jaynes, 1963) as claiming that there is no area of physics, from elementary particle theory to cyclotron design, in which the phenomenon of noise does not present itself.

In view of all this, one can imagine my consternation at the suggestion that "there has been little attention to unavoidable noise" in physics. Physicists were actively studying noise and, thanks to Laplace, knew the proper way to deal with it, long before there was any such thing as a Statistician.

(3) (a) Of course, by 'the orthodox solution' I mean the particular one *which I am describing*; and likewise for 'the Bayesian solution'. Of course, there are many different orthodox solutions to a given problem – but I think that is the last thing a defender of orthodoxy would wish to bring to our attention.

Dirac did not in any way suggest that "working statisticians would estimate a probability to be a negative number", as a reading of his lecture will show. On the other hand, it *is* a matter of documentable fact that some orthodox statisticians suggest estimating a parameter known to be positive by an estimator which can become negative for some samples [KF, p. 203, Equation (7.42)].

(b) It is really discouraging to find – 25 years after the birth of information theory (Shannon, 1948), 17 years after its bearing on the prior probability problem was shown (Jaynes, 1957), ten years after the generalization to continuous distributions (Jaynes, 1963), six years after the resulting functional analysis generalization of Gibb's work to irreversible statistical mechanics was given (Jaynes, 1967), five years after it was shown that the theory becomes parameter-independent if one uses the entropy relative to the invariant measure on the parameter space (Jaynes, 1968), and two years after the frequency interpretation of that invariant measure was demonstrated (Jaynes, 1971) – that an eminent worker in statistics is still writing that attempts to produce prior distributions by logical analysis have 'failed'.

It is true that the principles of maximum entropy and transformation groups have not yet led to the solutions of every conceivable statistical problem; and I know that there are some who reject the entire program just for that reason. Presumably these same critics do not condemn the use of insulin on the grounds that it will not cure all diseases. The point is that we have solved *some* problems, in a way which I believe will be recognized by history as the final answer; and in fact we have succeeded in a wide enough class of problems to cover perhaps 90% of current applications. Criticisms of Bayesian methods on the grounds that we still have unsolved problems, come with particularly ill grace from those who have in the past, by their discouraging negative attitude, done everything in their power to prevent these problems from being solved.

I would think that anyone might recognize that a meaningful comparison of Bayesian and orthodox solutions must use the Bayesian solution which makes use of the *same* information as does the orthodox solution. A Bayesian solution which makes use of extra prior information that the orthodox method cannot use at all, will of course be superior for that reason alone; it is more instructive – and in a sense fairer – to make comparisons using a Bayesian solution based on a noninformative prior. Now, a noninformative prior is one which is uniform, not necessarily with respect to Lebesgue measure for any particular choice of the parameter, but with respect to the invariant measure defined by the transformation group on the parameter space. As explained in my work referred to, this is just the mathematical statement of the basic desideratum of consistency: in two problems where we have the same prior information, we should assign the same prior probabilities.

My previous work (1968) shows how to construct priors for location and scale parameters, the rate constant of a Poisson process, and the parameter of a binomial distribution, by logical analysis. Evidently, the point needs to be made repeatedly and with more examples; so let me show briefly how to find the prior in the parameter space (α, β) of the standard regression problem $y=\alpha+\beta x$, by logical analysis, for the case that x, y are variables of the same kind (for example, the departure from average barometric pressure at New York and Boston), so that it is as natural to consider regression of (x on y) as (y on x). Given any proposed element of prior probability $f(\alpha, \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta$, interchange x and y. The estimated line becomes $x=\alpha'+\beta' y$, with a prior probability element $g(\alpha', \beta')\,\mathrm{d}\alpha'\,\mathrm{d}\beta'$. From the Jacobian of the transformation $\alpha' = -\beta^{-1}\alpha$, $\beta'=\beta^{-1}$, we find $g(\alpha'\beta')=\beta^3 f(\alpha, \beta)$. This transformation equation holds whatever the function f.

Now if we are ‘completely ignorant’ of (α, β), the interchange of (x, y) shouldn't matter; we are also ‘completely ignorant’ of (α', β'). But consistency demands that in two problems where we have the same state of knowledge, we must assign the same probabilities. Therefore f and g must be the same function; i.e., the prior density representing ‘complete ignorance’ must satisfy the functional equation $\beta^3 f(\alpha, \beta)=f(-\beta^{-1}\alpha, \beta^{-1})$, which has the solution $f(\alpha, \beta)=(1+\beta^2)^{-3/2}$. Thus, setting $\beta=\tan\theta$, the invariant measure of the parameter space is

$$\mathrm{d}\mu = \mathrm{d}\alpha\,\mathrm{d}\sin\theta\,.$$

Why is this not uniformly distributed in θ rather than in $\sin\theta$? Answer: it is uniform in $\sin\theta$ only for fixed α; but under rotations of the (x, y) plane α also varies [indeed, under any Euclidean transformation $(x, y) \to (x', y')$, where $x = x' \cos\phi - y' \sin\phi + x_0$, $y = y' \cos\phi + x' \sin\phi + y_0$, the estimated line $y = \alpha + \beta x$ goes into $y' = \alpha' + \beta' x'$, where $\alpha' = (\alpha - y_0 + \beta x_0)/(\cos\phi + \beta \sin\phi)$, $\beta' = (\beta \cos\phi - \sin\phi)/(\cos\phi + \beta \sin\phi) = \tan\theta'$; and we readily verify the invariance: $\mathrm{d}\alpha' \,\mathrm{d} \sin\theta' = \mathrm{d}\alpha \,\mathrm{d} \sin\theta$, while $\mathrm{d}\alpha \,\mathrm{d}\theta$ is not invariant].

This invariance of the measure $\mathrm{d}\mu$ means that, however we draw the x and y axes, the prior $\mathrm{d}\mu = \mathrm{d}\alpha \,\mathrm{d} \sin\theta$ expresses exactly the same state of prior knowledge about the position of the regression line. It thus leaves the entire decision to the subsequent evidence of the sample – which, of course, is exactly what Fisher insisted that a method of inference ought to do. But as we see, if this is the property we want to have, the goal is not achieved by closing our eyes to the very existence of a prior. It can be achieved only by logical analysis showing us *which* prior has the desired property. If we do have relevant prior information, it can now be incorporated into the problem by finding the probability measure $\mathrm{d}p$ that maximizes the entropy relative to $\mathrm{d}\mu$: $H = -\int \mathrm{d}p \log(\mathrm{d}p/\mathrm{d}\mu)$, subject to whatever constraint the prior information imposes on $\mathrm{d}p$; if the constraints take the form of mean values, this reduces to the canonical ensemble formalism of statistical mechanics of J. Willard Gibbs.

Now the simple facts, made understandable by Cox's theorems, illustrated in my presentation and in many other examples throughout the Bayesian literature, explain what we have observed throughout the history of orthodox statistics; every advance in orthodox practice has brought the actual procedures back closer and closer to the original methods of Laplace. The rise of decision theory was, in fact, the main spark that touched off the present 'Bayesian Revolution'. Other examples are Fisher's introduction of conditioning, discussed below, and his introduction of notion of sufficiency.

The discovery of sufficiency was, of course, a great advance *in orthodox statistics*; because in an important class of problems it removed the ambiguity in deciding which statistic should be used; if a sufficient statistic for θ exists, it is rather hard to justify using any other for inference about θ, for reasons illustrated in my Example 5 and explained under 'What Went Wrong?' But in Bayesian statistics there never was any ambiguity of this type to resolve. Fisher's definition of sufficiency can be stated more succinctly (and in my view, more meaningfully) as: If the posterior distribution of θ depends on the sample $(x_1 \dots x_n)$ only through the value of a certain function $\theta^*(x_1 \dots x_n)$, then θ^* is a sufficient statistic for θ. Evidently, if a sufficient statistic exists, application of Bayes' theorem will lead us to it automatically without our having to take any special note of the idea. But Bayes' theorem will lead us to the optimum inference whether or not any sufficient statistic exists; i.e., sufficiency is a convenience affecting the amount of calculation but not the quality of the inference.

I am afraid that to castigate Bayesian methods, but not orthodox ones, on grounds of lack of uniqueness, is to get it exactly backwards. It is orthodox statistics that offers us many different solutions to a single problem, (i.e., given prior information, sampling distribution, and sample), depending on whose school of thought, whose textbook within that school, and even which chapter of that textbook, you read. An estimator ought to be unbiased, efficient, consistent, etc.; but in general orthodoxy gives us no criterion as to the relative importance of these, nor any method by which a 'best' estimator can be constructed. The use of an unbiased estimator or a shortest confidence interval will lead us to different conclusions with different choices of parameters. KF (p. 316) cannot make up their minds about whether to accept the principle of conditioning, and advocate significance tests in which the conclusions depend on the arbitrary ordering you or I might assign to data sets *which were not observed!* Indeed, there is scarcely any problem of inference for which KF offer any definite preferred solution; in most cases there is an inconclusive discussion that terminates abruptly with the remark that 'it is all very difficult', leaving the reader in utter confusion as to which method should be used. But with all this ambiguity, orthodox methods provide no means for taking prior information into account.

In sharp contrast to this, for a given sampling distribution and sample, different Bayesian results correspond, as rational inferences should, to and *only* to, differences in the prior information. When priors are determined by the principles of maximum entropy and transformation groups, Bayesian methods achieve complete invariance under parameter changes (Jaynes, 1968).

(4) We are now told that even to utter the words 'Fisher-Neyman-Pearson theory' is a calumny on Fisher's views (but apparently not on

Neyman's or Pearson's); and again for the 'benefit' (precious little) of readers not present at the Conference, may I state that I first heard this phrase from the lips of Professor Oscar Kempthorne, shortly before my talk was given. I repeated it only to say that I would follow common practice by using the word 'orthodox' as an approximate synonym.

However, since the issue has been raised, I would like to state that the term 'Fisher-Neyman-Pearson approach' appears to me as an entirely accurate and appropriate term for a certain area of statistical thought. To use it is in no way to ignore, much less deny, the fact that there were differences between Fisher on the one hand, and Neyman-Pearson on the other. However, this should not blind us to the fact that there is a very much larger area of agreement; i.e., a corpus of ideas which are not in Bayesian statistics, but are common to the Fisher and Neyman-Pearson points of view and which therefore characterize their union. I refer to the ideas that (1) the word 'probability' must be used only in the sense of 'frequency in a random experiment', (2) inference requires that we find sampling distributions of some 'statistics' in addition to the direct sample distribution $p(\mathrm{d}x \mid \theta)$, (3) the conclusions we draw from an experiment can depend on the probabilities of data sets which were not observed, or the psychological state of mind of the experimenter (optional stopping), (4) we can improve the precision of our results by throwing away relevant information instead of taking it into account (the procedure euphemistically called 'randomization'), (5) the attempt to dispense with prior probabilities.

Recalling the difference between the Fisher and Neyman-Pearson camps over confidence intervals vs fiducial probabilities, let's just see how great this calumny is. Given a basic sample distribution $p(\mathrm{d}x \mid \theta)$, choose two 'statistics' $\theta_1(x_1 \dots x_n)$, $\theta_2(x_1 \dots x_n)$ such that $\mathrm{prob}(\theta_1 < \theta < \theta_2) = P$; this defines a $100\,P$ percent confidence interval. Letting $\theta_1 \to \theta_{\min}$, the lower bound of the parameter space, we have $\mathrm{prob}(\theta < \theta_2) = P$, which is Fisher's definition (Collected works, 27.253) of the fiducial distribution of θ, based on the statistic θ_2. As we see, the deep, profound difference in basic approach is fully as great as that between Tweedledee and Tweedledum.

The difference is not in the approach, but in the perception with which it was used. Fisher, with his vastly greater intuitive understanding, saw at once something which still does not seem generally recognized by others; that all this is valid only when we are using sufficient statistics. Even in the Fisher obituary notice, Kendall (1963, p. 4) questions the need for sufficiency. My Example 5 was intended to make Fisher's point by demonstrating just what can happen when we use a confidence interval not based on a sufficient statistic. Obviously, anyone who rejects fiducial probability, but endorses the use of confidence intervals, is not doing so on grounds of their actual performance.

Surprisingly, after protesting *my* calumny of Fisher's views, we find KF (p. 380) taking a dim view of fiducial probability, saying: "If a fiducial distribution is merely a restatement of a test of significance, we see no need for it". They might better have said: "Since a fiducial distribution of θ is a simultaneous statement of *all* tests of significance concerning θ, we see no need for the separate significance tests". While we may not have an 'equal distribution of ignorance', we have a more than equal distribution of calumny.

(5) It is quite true the foundations of modern physics are in a shambles, and in this area we also have controversy arising from unsolved problems. Being deeply involved in those also, I can report that current controversial issues in physics are orders of magnitude more complicated mathematically, and more subtle conceptually, than the trivia that we are quibbling about here. Indeed, the simple facts about probability theory that I am trying to point out were seen at once by the great mathematical physicists – Laplace, Maxwell, Gibbs, Poincaré, Jeffreys. For many years I have found it a refreshing rest to take off a few hours from the problems of physics, and work out another Bayesian-orthodox comparison.

(6) Professor Kempthorne objects very strongly to my use of the term 'common sense'. May we assume, then, that he denounces with equal force Fisher's use of the term (Collected works, 26.47) in appealing, three times in one page, to common sense rather than mathematical properties, to justify his 'information' measure?

I do indeed have a very great and insuperable difficulty in reconciling quantum physics with common sense, and am on recent record as having said exactly that (Jaynes, 1973). In fact, I would note that orthodox statistics and the 'Copenhagen' interpretation of quantum theory are just two different manifestations of a single intellectual disease, closely related to logical positivism, which has debilitated every area of theoretical science in this century. The symptoms of this disease are the loss of conceptual discrimination; i.e., the inability to distinguish between probability and frequency, between reality and our knowledge of reality, between meaning and method of testing, etc.

It is true that a 'publicly agreed verdict' may well be terribly wrong. This is just what happens when the public has been misled by false indoctrination of exactly the kind that I am trying to correct here. But to throw out the notions of 'common sense' and a 'publicly agreed verdict' is to forfeit the only visible means by which this controversy could be resolved. Although the temptation is strong. I will refrain from quoting Section 17.5 of KF, entitled 'Publicly Agreed Probabilities'.

(7) We apparently agree that a statistical method should be judged by the results it gives in practice. Well and good. However, I categorically deny that "the Bayesian idea was rejected by Boole, Venn, Fisher and Neyman" on these grounds. It is just the weakness of their work that they rejected Bayesian methods on purely philosophical or ideological grounds, *without* examining their actual performance.

Since the case of Boole and Venn has been brought up, let us examine the work of these gentlemen and see for ourselves the validity of their actual criticisms, and the accuracy with which their work is reported today in the orthodox literature. I believe that Boole, like most other critics of Laplace, failed to comprehend fully his definition of probability. Since Laplace has been quoted out of context so many times in this and other matters, let us take the trouble to quote his definition in full. The first volume of his *Théorie Analytique* is concerned with mathematical preliminaries, and the actual development of probability theory begins in Volume 2. The first sentence of Volume 2 is: "The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible".

This definition has stated only the finite discrete case, but we know how to generalize it. The point is that Laplace defined probability in a way which clearly represents *a state of knowledge*; and not a frequency. Of course, as Laplace demonstrates over and over again, connections between probability and frequency appear later, as mathematical consequences of the theory. I claim that these derivable connections (the limit theorems of Jacob Bernoulli and de Moivre-Laplace, Laplace's rule of succession, the de Finetti exchangeability theorem, etc.) include all the ones actually used in applications.

If one has no prior knowledge other than enumeration of the possibilities (i.e., specification of the sample space), then to assign equal probabilities is clearly the only honest way one can describe that state of, knowledge. This can be formalized more completely than Laplace did, by the aforementioned desideratum of consistency: if we were to assign any distribution other than the uniform one it would be possible, by a mere permutation of labels, to exhibit a second problem in which our state of knowledge is exactly the same, but in which we are assigning different probabilities. But in this case Laplace surely considered the argument and result so obvious that he would insult the reader's intelligence by mentioning them. The only serious error Laplace made was overestimating the intelligence of his readers.

Boole (1854), not perceiving this, rejected Laplace's work on the ground that the prior was 'arbitrary', i.e., not determined by the data. He did *not* reject it in the ground of the actual performance of Laplace's results in the case of uniform prior because he, like Laplace's other critics, never bothered to examine the actual performance under these conditions, much less to compare it with alternative methods. Had he done so, he might have discovered the real facts about performance, presented 85 years later by Jeffreys. Curiously, Boole, after criticizing Laplace's prior distribution based on the principle of indifference, then invokes that principle to defend his own methods against the criticisms of Wilbraham (see several articles in *Phil. Mag,* Vols. vii and viii. 1854).

This brings up another matter that needs to be mentioned. Boole's unjust criticism of Laplace has been quoted approvingly, over and over again, in the orthodox literature, Fisher (1956) being a very generous contributor. But in that same literature, a conspiracy of silence hides the fact that Boole's own work on probability theory (Boole, 1854, Chapters 16–21) contains ludicrous errors, far worse than any committed by Laplace. Some were noted by Wilbraham (1854), McColl (1897) and Keynes (1921). See his Example 6, page 286, where by a confusion of propositions [taking the probability of the proposition: 'If X is true, Y is true' as the conditional probability $p(Y \mid X)$] he arrives at the conclusion that two propositions with the same truth value can have different probabilities. He not only fails to see the absurdity of this, but even calls it to the reader's attention as something which 'deserves to be specially noticed'. Or his solution to another problem, page 324, Equation (10), which reduces to an absurdity in the special cases $c_1=c_2=1$ and $c_1= =p_1=1$. While Laplace considered real problems and got scientifically useful answers, Boole invented artificial school-room type problems, and often gave absurd answers. Finally, it is mathematically trivial to show that all of 'Boolean algebra' was contained already in the rules of probability theory given by Laplace – in the limit as all probabilities go to zero or unity, any equation of Laplace's 'Calculus of Inductive Reasoning' reduces to one of Boolean algebra.

Now let's turn to the case of Venn (1866), who expresses his disdain for mathematical demonstration very clearly throughout his book and its preface. Venn's Chapter 6 is an attack on Laplace's rule of succession, so viciously unfair that even Fisher (1956) was impelled to come to Laplace's defense on this issue. Fisher questions whether Venn was even aware of the fact that Laplace's rule had a mathematical basis, and like other mathematical theorems has 'stipulations specific for its validity'. He proceeds to give examples in which, unlike those of the 'great thinker' Venn, the stipulations are satisfied, and Laplace's rule is the correct one to use.

How is it possible for one human mind to reject Laplace's rule of succession; and then advocate a frequency definition of probability? Anybody who assigns a probability to an event equal to its observed frequency in many trials, is doing just what Laplace's rule tells him to do. In my Example 4, we examined Laplace's calculation underlying this rule, and learned that anybody who rejects Laplace's methods in favor of confidence intervals for the binomial, is certainly not doing so on grounds of actual performance.

I would like to plead here for a greater concern for historical accuracy, in writing on these matters. For over a century, there has been a conspiracy in the statistical literature to rewrite history and denigrate Laplace, first in the Boole-Venn manner, then by denying him credit when his principles were rediscovered (examples below). An *ad hominem* attack on Laplace (as 'a consummate politician') has even befouled the air of this Conference. I have long since learned never to accept the word of a biased source (Boole, Venn, Von Mises, Fisher, E. T. Bell, Cramér, Feller, etc.) on *any* question of what Laplace did or did not do. When working in my study, Laplace's *Théorie Analytique* is always at my elbow; and when any question about him comes up, I go straight to the original source. It is for this reason that my judgment of Laplace differs so radically from that presented in the literature from Boole on.

Not only those who are ignorant of history, but also those who will not profit by its lessons, are doomed to repeat it. Starting with Condorcet and his omelette, those who scorned Laplace's outlook and methods – whether in science or politics – and tried to do things differently, have shared a common experience.

(A) In George Gamow's book, *The Biography of the Earth* (1941), Laplace's theory of the origin of the solar system is torn to shreds. But in 1944, Weiszäcker pointed out a few things that Laplace's critics had overlooked; and the 1948 edition of Gamow's book had a new 15 page section entitled, '*Laplace was right after all!*'

(B) Abraham Wald, in his mimeographed course notes of 1941, rejected Laplace's methods of parameter estimation and hypothesis testing and asserted that such problems cannot be solved by the principles of probability theory. During the 1940's Wald sought a new foundation for statistics based on the idea of rational decisions, which had the aim of avoiding the mistakes of Laplace; but in Wald's final 1950 book, *Statistical Decision Functions,* the fundamental place of 'Bayes strategies' is finally recognized. As it turned out, Wald's life work was to prove, very much against his will, that the original methods developed by Laplace in the 18'th century, which he and many other statisticians had scorned for years, were in fact the unique solution to the problem of rational decisions. *Laplace was right after all.*

(C) I had the same experience. In 1951, I somehow came to the conclusion that Bayes' theorem did not adequately represent the full variety of inductive reasoning, and sought to develop a two-valued theory of probability, very much like the one presented here by Shafer, except that my numbers corresponded to the sum and difference of his. I even expounded this in a Round Table Discussion at one of the Berkeley Statistical Symposiums. However, I then made the tactical error of trying to apply this theory to some real problems. At about the third attempt, the scales fell from my eyes and I saw that a two-valued theory contains nothing that is not already given by Laplace's original one-valued theory, by going to a deeper sample space. In other words, the defects that

I thought I saw in Laplace's theory were my own defects, in not having the ingenuity to invent an adequate model. *Laplace was right after all.*

Now, I don't know how many other people are doomed to follow this path – already far more man-years of potentially useful talent have been wasted on futile attempts to evade Laplace's principles, than were ever invested in circle-squaring and perpetual motion machines. But just as Lindemann's proof put an end to circle-squaring for all who could see its implications, so Cox's theorems (1946) ought to have put an end, twenty-five years ago, to these unceasing efforts to evade what cannot be evaded. The situation is described in more detail in my review of Cox (1961). This is why I can say the following to latter-day Don Quixotes:

Many of us have already explored the road you are following, and we know what you will find at the end of it. It doesn't matter how many new words you drag into this discussion to avoid having to utter the word 'probability' in a sense different from frequency: likelihood, confidence, significance, propensity, support, credibility, acceptability, indiffidence, consonance, tenability, – and so on, until the resources of the good Dr Roget are exhausted. All of these are attempts to represent degrees of plausibility by real numbers, and they are covered automatically by Cox's theorems. It doesn't matter which approach you happen to like philosophically – by the time you have made your methods fully consistent, you will be forced, kicking and screaming, back to the ones given by Laplace. Until you have achieved mathematical equivalence with Laplace's methods, it will be possible, by looking at specific problems with Galileo's magnification, to exhibit the defects in your methods.

Here are two typical examples of the kind of factual distortion that we find in the literature. KF (p. 314) quote approvingly a statement of Fisher (1956, p. 4) that: "So early as Darwin's experiments on growth rate the need was felt for some sort of a test of whether an apparent effect might reasonably be due to chance". More specifically, Fisher (p. 81) then states that the 'Student' *t*-test was "the first exact test of significance." Neither book makes any mention of the historical fact that Laplace developed many significance tests to determine whether discrepancies between prediction and observation 'might reasonably be due to chance' and used them to decide which astronomical problems were worth working on: a bit of wisdom that might well be noted by scientists today. Laplace also illustrates the use of these tests, including two-way classifications, in many other problems of geodesy, meteorology, population statistics, etc. As I hope to show in detail elsewhere, Laplace's significance tests were in no way inferior – and were in some cases demonstrably superior – to tests advocated in the orthodox literature today.

Likewise, both KF and Fisher denounce the use of Bayes' theorem and uphold the 'student' t-test as a great advance in statistical practice; but of course neither mentions the fact that precisely the same result follows in two lines *from* Bayes' theorem; given the data $D=\{x_1 \ldots x_n\}$, the likelihood function is $L(\mu, \sigma)=\sigma^{-n} \exp(-nQ/2\sigma^2)$, where $Q=s^2+(\bar{x}-\mu)^2$. Integrating out σ with respect to Jeffreys' prior, the posterior density of μ is $\sim Q^{-n/2}$, which but for notation is just the t-distribution. Students reading these works obtain a completely false picture of both the historical and mathematical facts about significance tests.

As a second example, KF (p. 305) consider tests of a simple hypothesis M_1 against a simple alternative M_2, on data D. The likelihood ratio in favor of M_1 is $L(D)=p(D \mid M_1)/p(D \mid M_2)$. KF note that, if M_2 is true, the expected value $E_2(L)$ is unity, and conclude that the 'Bayesian process' has bad operating characteristics. But of course, this is not the proper criterion, because it is $C=\log L$, and not L, that has equal positive and negative change for equal strength of evidence for and against M_1. The inequalities $E_2(C)\leqslant 0$, $E_1(C)\geqslant 0$ [with equality if and only if $p(D \mid M_1)= =p(D \mid M_2)$ for all D] then establish what they would regard as 'good' operating characteristics. Twenty pages later, KF are back to the same problem; only now they remember to take the logarithm, represent it as an orthodox test, and have no cause to complain of the operating characteristics of the statistic C. And so the indoctrination goes on; I could cite at least twenty more examples of these tactics from recent textbooks.

Now let's come to Kempthorne's statement that "a Bayesian interval has no predictive verifiability". I suggest that the main message has totally escaped him. If the optimum confidence interval is mathematically identical with the Bayesian interval based on a noninformative prior distribution, it is a bit difficult to understand how the Bayesian result could fail to have whatever 'predictive verifiability' – or any other property – is possessed by the confidence interval.

Unfortunately, there is a serious wandering of the mind in connection with the test according to which the Bayesian will 'lose his shirt'. First we are told that the confidence interval advocate will assert at 19 to 1 that the interval contains the unknown parameter value. Now, is the Bayesian required to accept this in every case? For which confidence level is this asserted? Does the width of that interval enter into the judging of the game? It is not a matter of mathematics that an undefined claim can be sustained.

Then the game appears to change suddenly; we now learn that it is the Bayesian who is making the probability statements. It is averred that the Bayesian will lose his shirt and be consistently wrong – excuse me, coherently wrong – in a case where some individual *C* challenges the assertions. But now is this individual *C* to challenge all assertions wherever they may be? At what odds? I suggest that if Professor Kempthorne will try to back up his position by producing a specific, well defined situation instead of making assertions about undefined generalities, the mathematical situation will force him to see that his claim simply is not true.

Indeed, the contest proposed by Kempthorne has already been carried out in the Monte Carlo experiments of A. Zellner (1965) and H. Thornber (1965). The results were, in the words of H. V. Roberts (1965); "Using sampling-theory criteria, the Bayesian estimators appeared better in all examples, the margin being substantial for Zellner's experiment and modest for Thornber's". Roberts proceeds to explain why this must be so; by the time all necessary provisions for a 'fair' contest have been incorporated into the experiment, all the ingredients of the Bayesian theory (prior distribution, loss function, etc.) will necessarily be present. As Roberts concludes; "The simulation can only demonstrate the mathematical theorem".

My sixth example, on the Cauchy distribution, demonstrated (and I thought rather cogently) that the 'long-run performance' of a statistical procedure is *not* the proper criterion of its usefulness. But Professor Kempthorne simply ignores this, and continues to argue long-run performance as the criterion ('The Bayesian will lose his shirt', etc). So comtemplate this example, given by David Forney (1972):

THE WEATHERMAN'S JOB

In a certain city, the joint frequencies of the actual weather and the weatherman's predictions are given by:

		Actual	
		Rain	Shine
Predicted	Rain	$\frac{1}{4}$	$\frac{1}{2}$
	Shine	0	$\frac{1}{4}$

An enterprising fellow trained in orthodox statistics (but not in meteorology) notices that, while the weatherman is right only 50% of the time, a prediction of 'shine' everyday would be right 75% of the time, and applies for the weatherman's job. Should he get it? Which would you rather have in your city?

The weatherman is delivering useful information at a rate $I=$(entropy of distribution of predictions) + (entropy of actual weather distribution) $-$(entropy of joint distribution)$=(0.562+0.562-1.040)/\ln 2=0.123$ bits/day. As explained previously (Jaynes, 1968) this means that in the course of a year the weatherman's information has reduced the number of reasonably probable weather sequences by a factor of $W=\exp(0.123\times\times 365\times\ln 2)=2.92\times 10^{13}$. With the weatherman on the job, you will never be caught out in an unpredicted rain; with the orthodox statistician this would happen to you one day out of four.

As this example one more forces one to recognize, the value of an inference lies in its usefulness *in the individual case*, and not in its long-run frequency of success; they are not necessarily even positively correlated. The question of how often a given situation would arise is utterly irrelevant to the question how we should reason when it *does* arise. I don't know how many times this simple fact will have to be pointed out before statisticians of 'frequentist' persuasions will take note of it; but I think it is important that we keep trying.

(8) The book by Kempthorne and Folks (1971) does indeed mention the works of Good, Savage and Jeffreys, unlike so many orthodox textbooks. That is, these works are included in a list of references. This leaves to be desired only that their contents had also been noted.

(9) Here and elsewhere, Professor Kempthorne seems to regard L. J. Savage (1954) as the official spokesman for Bayesian theory; and implies that if I state anything differently from Savage, then I must not have read his book. By that reasoning, I believe we have an even stronger case for inferring that someone else has not read it. It is true that Savage, probably

more than any other person, was the one who stimulated new thought on these issues (although to me personally, the arguments of Jeffreys (1939) and R. T. Cox (1946, 1961) have always seemed far more cogent). But a great deal has happened in Bayesian statistics since 1954, and I think that at present the only thing which Bayesian statistics and Savage's personalistic theory have in common is that they both use Bayes' theorem, without apology or embarrassment. Today, very few if any Bayesians would give full support to Savage's notion of 'personalistic probability', and I am on record (Jaynes, 1968) as taking my stand with orthodox statisticians on this matter; i.e., the notion of personalistic probability belongs to the field of psychology, and not to statistics.

(10) I cannot see the point of this comment. In my paper I stated very explicitly that, while there are some differences of opinion, most would hold that the proper method for the problem is the confidence interval. I believe that is a clear and accurate statement of fact. Of course, one cannot necessarily have confidence in confidence intervals; that is just the point I thought I was making in demonstrating that there are cases in which one can have zero confidence in a confidence interval.

(11) The impelling urge to find fault rather than to understand rules the situation here. In comment No. 6, Professor Kempthorne objects to the idea of 'publicly agreed verdict', but now he apparently wishes to speak with approval of a "requirement of interpersonal validation of subjectively formed decisions". But an interpersonal validation (which would amount to a publicly agreed verdict) can only take place through the common sense judgments of different people who are all exposed to the same system of facts. I am under the impression that the comments of Bross were refuted by specific factual counter-examples, in addition to a general proof, demonstrating the opposite of what Bross claimed.

(12) Whether any particular problem should be called technically a significance test, a test of goodness of fit, an acceptance test, an hypothesis test, or a decision problem, is a matter of pedantry on which orthodox statisticians are themselves in disagreement; so why can't we just call it 'a test' and get on with the substantive issues? I believe my presentation made it clear in each case: (1) what was the problem? (2) How was it handled? And that should be enough.

(13) This comment brings to mind an older controversy, with more than one similarity to our present one. Protestant countries long refused

to accept the Gregorian calendar, in spite of its clearly superior performance. England held out for 170 years after it had been adopted by Catholic countries, leading Voltaire to quip that the British "would rather disagree with the sun than agree with the Pope". It appears that some would rather disagree with common sense than agree with Bayes.

Before getting too indignant about that high (92%) significance level indicated by the Bayesian test and denying that the evidence is that clear, let's first do that quick, short-cut calculation right. The standard error of the difference of means should be estimated not by $\sqrt{7.48^2+6.48^2}=9.90$; but by (Fisher, 1958, p. 116; Hoel, 1971, p. 134):

$$\sqrt{\frac{7.48^2}{9}+\frac{6.48^2}{4}}=4.09\,.$$

If this standard error were known, rather than estimated, it would correspond to a significance level, not of 92%, but of 97.5%.

(15) *'Improper' Priors*. Let me try to explain the situation. 'Complete initial ignorance' of a scale parameter σ corresponds formally to use of the Jeffreys prior $\mathrm{d}\sigma/\sigma=\mathrm{d}\log\sigma$. But as noted before (Jaynes, 1968), to apply this within infinite limits $(-\infty<\log\sigma<\infty)$ would not represent any realistic state of prior information. For example, if x is a measured length of some material object on the earth, we surely know that the standard error σ_x of the measurement cannot be less than the size of one atom, $\sim 10^{-8}$ cm; or greater than the size of the earth, $\sim 10^9$ cm. So we know in advance that $(-8<\log_{10}\sigma_x<+9)$. Outside this range, the prior density must be zero.

Similarly, if x is the measured breaking stress of some structural material, we know in advance that σ_x surely cannot be less than the pressure of sound waves, ~ 1 dyne cm^{-2}, due to people talking in the room; nor greater than 10^{14} dynes cm^{-2}, which is 1000 times the tensile strength of any known material. So the prior density must be all contained in $(0<<\ln\sigma_x<33)$. If x is a time interval measured in seconds, we can be pretty sure in advance that $(-12<\log_{10}\sigma_x<18)$.

Generally, thinking about any problem in this way will lead one to specify prior limits $\sigma_{\min}$, $\sigma_{\max}$ within which the unknown value surely lies; within this interval the invariance arguments leading to the form $\mathrm{d}\sigma/\sigma$ still apply if there is no other prior information (Jaynes, 1968). Therefore, the prior is normalizable, and we have a well-behaved mathematical problem.

Now if our final conclusions depend appreciably on the exact prior limits chosen, then obviously we should analyze our prior information more carefully than I did above, to get more reliable numerical values for $\sigma_{\min}$, $\sigma_{\max}$. But it just wouldn't be very intelligent to go to all that work, only to discover that $\sigma_{\min}$, $\sigma_{\max}$ cancel out of the expressions representing our final conclusions (which might be the first few moments, or the quartiles, of a posterior distribution). So it will be good strategy to work through the solution first for general limits, whereupon the mathematics will tell us under just what conditions the prior limits matter; and when they don't.

Having thus formulated the problem, the conclusion is fairly obvious: if the likelihood function is sufficiently concentrated (i.e., if the experiment is a sufficiently informative one), then the prior limits cannot matter appreciably as long as they are outside the region of appreciable likelihood. To put it in a way somewhat crude, but not really wrong: if the amount of likelihood [integral of $L(\sigma)$] lying outside the limits $(\sigma_1<\sigma<$ $<\sigma_2)$ is less than 10^{-6} of the total likelihood, then as long as our prior limits are still wider $(\sigma_{\min}<\sigma_1<\sigma_2<\sigma_{\max})$, the exact values of $\sigma_{\min}$, $\sigma_{\max}$ can't make more than about one part in 10^6 difference in our conclusions. If, then, we don't worry about them, and just take the limiting form of the solution as $\sigma_{\min}\to 0$, $\sigma_{\max}\to\infty$ for mathematical convenience, we are committing no worse a sin than does the person who laboriously determines the proper values of $\sigma_{\min}$, $\sigma_{\max}$, works out the exact solution based on them – and then rounds off his final result to six significant figures. We are only getting that result with an order of magnitude less labor.

If, on the other hand, we should encounter a non-normalizable posterior distribution in this limit, the theory is telling us that the experiment is so uninformative that our exact state of prior information is still important, and must be taken into account explicitly. This phenomenon, far from being a defect of Bayesian methods, is a valuable safety device that warns us when an experiment is too uninformative to justify, by itself, any definite inferences. If someone ignores the warning, and gets into trouble with 'improper priors', what we are witnessing is not a failure, but only a misapplication, of Bayesian methods.

Finally, let us keep in mind that we are really concerned here with relative value judgments; and so if anyone attacks Bayesian methods because of the possible situation just described, fairness demands that he also takes note of what happens to orthodox methods in the same problems. Now one of the substantive factual issues illustrated in my presentation, is this: orthodox methods, when improved to the maximum possible degree, reduce ultimately to procedures that are mathematically identical with applying Bayes' theorem *with just the noninformative improper prior* about which Professor Kempthorne expresses such alarm! We saw this phenomenon in Examples 2, 3, 5 and 6. As we have just seen, this causes difficulty only when the experiment is so uninformative that our final conclusions must, necessarily, still depend strongly on our prior knowledge. The Bayesian can correct this at once by using a realistic prior, leading to the inferences that *are* justified by the total information at hand; but the orthodoxian cannot, because his ideology forbids him to recognize the existence of any prior which is not also a known frequency.

In fact, we had just this situation in the first part of my Example 3, where we took no note of the actual failure times. If all units tested fail, the test provides no evidence against the hypothesis of arbitrarily large λ. The Bayesian test (6) based on a uniform prior then yields a non-normalizable posterior distribution $p(\mathrm{d}\lambda \mid n, r, t) \sim (1-e^{-\lambda t})^n \,\mathrm{d}\lambda$, which tells us that λ is almost certainly greater than $(t^{-1} \log n)$, but gives no upper limit. In this way, the safety device warns us that our prior information concerning the possibility of very large λ, remains relevant; by taking it explicitly into account, rational inferences about λ are still possible, as I showed by the maximum entropy prior.

But we saw that the orthodox ST test was, in the absence of such pathology, mathematically identical with this Bayesian test; so what happens to it? Well, this is just the case already noted where the ST test breaks down entirely, telling us to reject at all significance levels. In problems where the Bayesian cannot use the approximation of an improper prior, orthodox methods give no warning, but simply yield absurd results; and only the alertness and common sense of the user can save him from the consequences. As we see, it is the orthodoxian, and not the Bayesian, who is going to be in trouble in cases where 'improper priors' cannot be used.

Note the treatment of an almost identical problem in KF (p. 203, Equation 7.42). Here they suggest use of an estimator which estimates the mean life to be infinite if we observe one failure, to be negative if we observe no failures, and which has infinite variance unless we observe 3 or more failures! Again, I think common sense renders a rather clear verdict in this comparison. If Professor Kempthorne thinks that the Bayesian solution to this problem is open to criticism, I wonder how he would defend the solution proposed in his book.

(17) KF do indeed advocate reporting critical significance levels, and for this enlightenment over most previous treatments we can be grateful. We could be even more grateful if the enlightenment had persisted to the end of the book, where KF reproduce the same old tables, so arranged that critical levels cannot be located.

(21) In comment No. 9 Professor Kempthorne infers, from a difference in my position and that of Savage, that I have not read Savage's book. Now he infers, from a similarity between my work and Fisher's, that I have not read Fisher. These orthodox inferences – with the conclusion independent of the evidence – leave me in despair. My work was checked by another statistician for just such matters. In the first version I called y an 'ancillary statistic' in Example 6; but he objected to this on the ground that I was not using it in quite the same sense as Fisher did, so I deleted the term. Now I find myself being criticized by one orthodox statistician for having followed the advice of another.

Here is the point: Fisher (Collected Works, 27.257) held – without explaining why – that the distribution of an ancillary statistic must be independent of the value of the parameter, as expressed in his allegory of the Problem of the Nile. Presumably, this was one of the many things he saw intuitively; but whatever his private reason for this independence may have been, it is easy to see what it in fact accomplishes. To avoid a possible paradox (Barnard, 1962), we understand the conditioned probability symbol $p(\mathrm{d}\theta^* \mid y, \theta)$ to be shorthand for the limit as $\mathrm{d}y \to 0$, of the well-defined

$$p(\mathrm{d}\theta^* \mid \mathrm{d}y, \theta) = \frac{p(\mathrm{d}\theta^* \, \mathrm{d}y \mid \theta)}{p(\mathrm{d}y \mid \theta)}.$$

If $p(\mathrm{d}y \mid \theta)$ is independent of θ, then the θ-dependence of the conditioned probability $p(\mathrm{d}\theta^* \mid y, \theta)$ is the same as that of the joint probability $p(\mathrm{d}\theta^* \, \mathrm{d}y \mid \theta)$. In other words, it is fundamentally the joint probability, and not the conditioned probability, that really matters – but of course, that is just what the likelihood principle has told us all along.

With this little bit of insight, it becomes clear that mathematically, Fisher's conditioning on ancillary statistics is just a roundabout way of restoring agreement with Bayes' theorem, without having to admit that one is using it. But conditioning is not a general method; and simple mathematics shows that if we just apply Bayes' theorem directly, there is no longer any reason for y to be independent of θ. We then have a method that works in all problems – and is guaranteed to give the same result as Fisher's, with less calculation, in cases where his conditioning is possible. I hope this excursion will clear me of the charge of not having read Fisher.

(26) Since the question is asked, I will answer by showing just how the 'neo-Bayesian prescription' works, in precisely the problem where KF deny it.

KF, p. 439, consider a standard problem of linear regression with both variables subject to error. The model is $Y_i = \alpha + \beta X_i$ with measured values $x_i = X_i + e_i$, $y_i = Y_i + f_i$, the errors e_i, f_i being independent and $N(0, \sigma_x)$, $N(0, \sigma_y)$ respectively; σ_x, σ_y unknown. We take data $D = \{(x_1, y_1); (x_2, y_2); \ldots, (x_n, y_n)\}$ and from this we are to make inferences about α and β.

At this point, KF assert that 'in antithesis to the likelihood principle', the likelihood function is (1) totally uninformative, (2) ill-behaved, becoming indeterminate when $x_i = X_i$ and (3) that further assumptions are needed (about equality of several X_i, or about σ_x, σ_y, etc.) to make progress on the problem. They then suggest a method in which we partition the data points into two sets, and take the line joining their centroids as our estimate of the 'true' line.

We have here one more example – perhaps the finest yet produced – of just the Canonical Procedure that I complained about in my paper; still another time, an orthodox textbook rejects the Bayesian solution, without bothering to look at it, for patently false reasons; and gives instead an orthodox method which is far weaker in its ability to extract information from the sample.

An undergraduate in a laboratory science course does better than the proposed solution of KF, without any statistical theory at all, simply by plotting his experimental points and drawing the straight line that, as judged by eye, fits them best. He can determine the accuracy of his estimates of α, β by noting how much this line can be shifted or tilted before the fit appears appreciably worse. Furthermore, if the standard errors σ_x, and/or σ_y were unknown, he would do this in the same way whether the errors were in x only, in y only, or in both; if the ratio $\lambda = \sigma_y / \sigma_x$

were known, and/or if the errors (e_i, f_i) had a known correlation coefficient ϱ, it would make no difference in the correct data reduction procedure whatever the values of λ, ρ.

The reason why these things cannot matter is that, whether the errors are represented by a one-dimensional or two-dimensional region of uncertainty about each data point, and whatever the shape and orientation of the concentration ellipse, the component of error parallel to the estimated line contributes nothing to the error of estimation of either α or β. As common sense tells us – and the Bayesian analysis confirms – in any of these circumstances the problem of inference about α, β takes the same form, with only a single unknown error component to consider. In other words, there are *not* ten basically different estimation problems, as is implied by the elaborate KF classification scheme $(yRE \mid xCN)$, etc. If the standard errors are unknown, there is only one linear model problem.

To prove these assertions, let us just sketch the Bayesian analysis, which KF declare to be impossible. The likelihood function, which they do not even write down, is

$$L(\alpha, \beta, \sigma_x, \sigma_y, X_i) =$$
$$= (\sigma_x \sigma_y)^{-n} \exp\left\{-\tfrac{1}{2} \sum_{i=1}^{n} \left[\frac{(x_i - X_i)^2}{\sigma_x^2} + \frac{(y_i - \alpha - \beta X_i)^2}{\sigma_y^2}\right\}\right].$$

Obviously, it is in no way 'ill-behaved' or 'indeterminate'. Now let's see just how 'uninformative' is, and whether further assumptions are needed to make progress. If we want to make inferences about α, β, then $\{\sigma_x, \sigma_y, X_1 \ldots X_n\}$ are 'nuisance parameters' that prevent orthodox statistics from making any headway on this problem. It is then interesting to see how much they deter a Bayesian.

Integrating $\{X_1 \ldots X_n\}$ out of L with respect to uniform prior, we obtain a function which depends on (α, β) only through the quadratic form

$$Q(\alpha, \beta) \equiv \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 .$$

Making the change of variables: $\{\sigma_x, \sigma_y\} \to \{\sigma, \lambda\}$, where $\sigma^2 = \sigma_y^2 + \beta^2 \sigma_x^2$, $\lambda = \sigma_y / \sigma_x$, the posterior distribution of (α, β) is found to be independent of the prior distribution of λ, confirming a previous remark.

Integrating out σ with respect to Jeffreys' prior, we obtain a 'quasi-likelihood' function

$$f(\alpha, \beta) \sim [Q(\alpha, \beta)]^{-n/2},$$

which, when multiplied by the prior density and normalized, gives the joint posterior distribution of α, β. This function summarizes all the information about α, β that is contained in the data; and so the optimal procedure for any inference or decision problem involving α, β – whether in the form of interval estimation, tests of any hypotheses concerning α, β, etc. – can then be found from it.

To confirm another previous remark, consider the simpler regression problem where errors are only in Y. Then $\sigma_x = 0$, $X_i = x_i$ is known, and $n+1$ nuisance parameters drop out of the problem. The likelihood function reduces to

$$L(\alpha, \beta, \sigma_y) = \sigma_y^{-n} \exp\left\{-\frac{n}{2\sigma_y^2} Q(\alpha, \beta)\right\}.$$

Integrating out σ_y with Jeffreys prior, we get the quasi-likelihood $f(\alpha\ \beta) \sim$ $\sim [Q(\alpha, \beta)]^{-n/2}$, precisely the same as before. The nuisance parameters had *no effect at all* on the quality of inference about α, β.

Representing the sample means, variances, covariance, and correlation coefficient by $\bar{x}$, $\bar{y}$, $s_x^2 = (\overline{x^2} - \bar{x}^2)$, $s_y^2 = (\overline{y^2} - \bar{y}^2)$, $s_{xy} = (\overline{xy} - \bar{x}\bar{y})$, $r = s_{xy}/s_x s_y$ respectively, $f(\alpha, \beta)$ has its maximum at (α^*, β^*), where $\alpha^* = \bar{y} - \beta^* \bar{x}$, and $\beta^* = (s_y/s_x)\, r$. With uniform priors, further integrations yield the marginal posterior distributions of α, β: $g(\alpha) \sim [(\alpha - a^*)^2 + A^2]^{-m}$, $h(\beta) \sim [(\beta -$ $- \beta^*)^2 + B^2]^{-m}$, where $m = (n-1)/2$, and $B = \beta^*\, r^{-1}(1 - r^2)^{1/2} = A/(\overline{x^2})^{1/2}$. Evidently, α^*, β^* are the 'best' estimates of α, β by the criterion of any loss function which is a monotonic increasing function of the errors $|\alpha - \alpha^*|$, $|\beta - \beta^*|$, For $n > 2$, the marginal distributions are normalizable, leading to definite interval estimate statements. For $n = 3$ and $n = 4$, the (median $\pm$ interquartile) estimates of α are $(\alpha^* \pm A)$ and $= (\alpha^* \pm A/\sqrt{3})$ respectively; similarly for β. When $n > 4$, the second moments also conconverge, leading to the (mean)$\pm$(standard deviation) estimates $\alpha^* \pm$ $\pm A\sqrt{n-4}$, etc.

Thus, for example, if we need to measure β to an accuracy of $\pm 1\%$, the sample size and correlation coefficient must satisfy $(n-4)\, r^2/(1 - r^2) > 10^4$. With a correlation coefficient $r = 0.9$, this requires $n = 2350$ measurements, while with better data, $r=0.99$, $n=208$ measurements suffice, and with $r=0.999$, only $n=25$ data points are needed. A simple analysis shows that to attain the same accuracy by the method described by KF would require at least $(16/3)=5.3$ times as many data points, if they are distributed with roughly uniform density along the line. As will be shown elsewhere, the above results can be improved a bit more by use of the invariant prior $d\alpha \, d\sin\theta$, and a similar invariant prior for X_i, Y_i.

CONCLUSION

I suppose it is possible, without actual logical contradiction, to maintain that Bayesian methods are utterly wrong, but that through a series of fortuitous accidents they always happen to give the right answer in every particular problem. However, I cannot believe that anybody will want to take that position. Now the person who, after studying the evidence given here and in the rest of the Bayesian literature, still wishes to claim that orthodox methods are superior, must realize that, if he is to avoid being forced into exactly that position, mere linguistics and ideological slogans will no longer suffice. The burden of proof is squarely on him to show us specific problems, with mathematical details, in which orthodox methods give a satisfactory result and Bayesian methods do not. My own studies have convinced me that such a problem does not exist.

Whether I am right or wrong in this belief, we now have a large mass of factual evidence showing that (a) orthodox methods contain dangerous fallacies, and must in any event be revised; and (b) Bayesian methods are easier to apply and give better results. As a teacher, I therefore feel that to continue the time honored practice – still in effect in many schools – of reaching pure orthodox statistics to students, with only a passing sneer at Bayes and Laplace, is to perpetuate a tragic error which has already wasted thousands of man-years of our finest mathematical talent in pursuit of false goals. If this talent had been directed toward understanding Laplace's contributions and learning how to use them properly, statistical practice would be far more advanced today than it is.

REFERENCES

Note: The following list includes only those works not already cited in my main presentation or Kempthorne's reply.

Barnard, G. A., 'Comments on Stein's "A Remark on the Likelihood Principle"', *J. Roy. Stat. Soc.* (A) **125**, 569 (1962).

Cox, R. T., *Am. J. Phys.* **17**, 1 (1946).

Cox, R. T., *The Algebra of Probable Inference*, Johns Hopkins University Press, 1961; Reviewed by E. T. Jaynes, *Am. J. Phys.* **31**, 66 (1963).

Deming, W. E., *Statistical Adjustment of Data*, J. Wiley, New York (1943).

Fisher, R. A., *Contributions to Mathematical Statistics*, W. A. Shewhart, (ed.), J. Wiley and Sons, Inc. New York (1950); Referred to above as 'Collected Works'.

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

Fisher, R. A., *Statistical Methods for Research Workers*, Hafner Publishing Co., New York: Thirteenth Edition (1958).

Forney, G. D., *Information Theory*, (EE376 Course Notes, Stanford University, 1972); p. 26.

Hoel, P. G., *Introduction to Mathematical Statistics*, Fourth Edition, J. Wiley and Sons, Inc., New York (1971).

Jaynes, E. T., 'Review of *Noise and Fluctuations*', by D. K. C. MacDonald, *Am. J. Phys.* **31**, 946 (1963).

Kendall, M. G., 'Ronald Aylmer Fisher, 1890–1962', *Biometrika* **50**, 1–15 (1963); reprinted in *Studies in the History of Statistics and Probability*, E. S. Pearson and M. G. Kendall, (eds)., Hafner Publishing Co., Darien, Conn. (1970).

Mandel, J., *The Statistical Analysis of Experimental Data*, Interscience Publishers, New York (1964); p. 290.

McColl, H., 'The Calculus of Equivalent Statements', *Proc. Lond. Math. Soc.* **28**, p. 556 (1897).

Pearson, Karl, 'Method of Moments and Method of Maximum Likelihood', *Biometrika* **28**, 34 (1936).

Pratt, John W., 'Review of *Testing Statistical Hypothesis*' (Lehmann, 1959); *J. Am. Stat. Assoc.* Vol. **56**, pp. 163–166 (1961).

Roberts, Harry V., 'Statistical Dogma: One Response to a Challenge', Multilithed, University of Chicago (1965).

Thornber, Hodson, 'An Autoregressive Model: Bayesian Versus Sampling Theory Analysis', Multilithed, Dept. of Economics, University of Chicago, Chicago, Illinois (1965).

Wilbraham, H., *Phil. Mag. Series*, 4, Vol. **vii**, (1854).

Zellner, Arnold, 'Bayesian Inference and Simultaneous Equation Models', Multilithed, University of Chicago, Chicago, Illinois (1965).