Bayes again!

E. T. Jaynes

Washington University, St. Louis

December 1963

Is it permitted for a physicist to join the fun in this Neo-Bayesian vs. orthodox fight? Recently, my attention was called to the remarkable article, "Linguistic Analysis at a Statistical Controversy" by Irwin D. J. Bross, in the February 1963 issue [1]. Of course, what we need is not any linguistic analysis, but a *mathematical* analysis of the situation.

For about eight years I have been making constant routine use of Bayesian methods in statistical problem of physics and engineering, and comparing their results with those obtained by orthodox methods. I believe that the issues can removed entirely from this realm of futile verbal exchange and reduced to definite questions of mathematical fact; and that a glimpse of what results when we do this might be more useful to readers of this journal than another round of polemics.

Let us start, as did Bross, by quoting the words of J. W. Pratt (issue of April, 1962 [2]): "Now that I have ceased pretending to be impartial, I may point art that no connected argument leading to orthodox methods has ever been advanced. Neyman and Pearson contributed vitally to our understanding by their *formulation* of statistical problems, but they have never claimed their *methods* were more than ad hoc procedures with some pleasant properties. Their methods, while extremely ingenious and useful, are not completely satisfactory, let alone uniquely objective and scientific."

Since Bross has written an article whose sole purpose, as far as I can see, is to attack this statement, fairness requires that equal time he granted for its defence. I am unable to find any "Neo-Bayesian jargon" or "in-congruencies" in it. Pratt has stated a very important truth, and he has stated it in clear and simple terms. But, since there are apparently statisticians who simply refuse to see this, let us simplify Pratt's statement with some details.

Bross objects to the remark that "no connected argument" has been produced for the orthodox methods, on the grounds that it is a very sweeping statement; and so it is. I believe it is also a correct one, since in spite of much literature searching, I have never been able to find any derivation of these methods from first principles. Bross specifically mentions significance tests and confidence intervals, so let us do likewise.

Now where is the connected argument leading to the chi-squared test? Here in how Cramèr [3] introduces it: "It will then be in conformity with the general principle of least

squares to adopt as measure of deviation en expression of the form $\sum c_I(n_I/N - p_I)^2$ where the coefficients c_I may be chosen mere or less arbitrarily. It was shown by K. Pearson that if we take $c_I = N/p_I$, we shall obtain a deviation measure with particularly simple properties." In other words, chi-squared is adapted, not because of any theoretical justification, but because it has, in Pratt's words, "some pleasant properties."

If there is any more "connected argument" leading to chi-squared, that fact can be established once and for all by *producing* the argument. In the meantime, no amount of pointing to the admittedly first-rate mathematicians who have worked this field quite answers the question raised by Pratt.

Actually, there is a very good theoretical justification for chi-squared; but it is Bayesian. We want to test hypothesis H_1 against H_2; both belong to the same "Bernoulli class" B_r (i.e., r possible outcomes at each trial; independence of different trials). Having performed the random experiment N times and observed the sample populations $\{n_1 \cdots n_r\}$, calculate χ_1^2 and χ_2^2. Let the prior probabilities of H_1 and H_2 be w_1, w_2; and the posterior probabilities (after observing the sample) W_1, W_2. Bayes' theorem then yields, to a good approximation

$$\log(W_2/W_1) = \log(w_2/w_1) + \frac{1}{2}\chi_1^2 - \frac{1}{2}\chi_2^2 \tag{1}$$

Thus, χ_1^2 summaries all the information in the sample that is relevant for testing hypothesis H_1 against any alternative in the same Bernoulli class.

But Eq. 1 is only an approximation valid when χ^2 is sufficiently small for both hypotheses. The exact equation is of the form (1) with $\frac{1}{2}\chi^2$ replaced by

$$\psi = \sum_{I=1}^{r} n_I \log(n_I/Np_I) \geq 0 \tag{2}$$

This is the quantity which, according to Bayes' theorem, precisely summaries the sample data for purposes of testing the "null hypothesis" H for which the probabilities at each trial are $\{p_1 \cdots p_r\}$ against the class of alternatives B_r. Its meaning is best described by the following easily proved theorem: *Given an hypothesis H_1 and the sample data calculate ψ_1. Then it is possible to find an alternative hypothesis H_2 in B_r for which*

$$\log(W_2/W_1) - \log(w_2/w_1) = D \tag{3}$$

where $D \leq \psi_1$. There is no H_2 in B_r for which $D > \psi_1$.

Use of ψ instead of χ^2 as a measure of goodness of fit has thus an evident theoretical advantage, and the practical advantage, that, being exact, it requires no grouping of categories for which the sample numbers n_I are small. Of course, when we replace $\frac{1}{2}\chi^2$ by ψ in Eq. 1, we have exactly Wald's probability-ratio test and so we have a "connected" argument which discloses the intimate relation between Wald's sequential testing method, where a definite alternative is stated, and χ^2, which makes no reference to any specific alternative. We also see that χ^2 is not a measure of goodness of fit relative to all conceivable alternatives; but only relative to those in the same Bernoulli class. This fact has, no

doubt, always been understood implicitly; but to the best of my knowledge it has not been explicitly pointed out in the orthodox statistical literature. By a straightforward generalisation of the above argument, we can construct the appropriate ψ which measures goodness of fit relative to any well-defined class of alternatives.

This is a good example of what, I suggest, is the general situation; the Bayesian approach to statistics supplies the missing theoretical basis for, and often improvements on, orthodox methods which had long been, just as Pratt says, "ad hoc procedures with some pleasant properties."

In the Neyman-Pearson decision criterion, we fix the probability of one type of error at some small value δ, and subject to this constraint, adapt the decision rule which minimises the probability of the other type of error. This ad hoc procedure has the pleasant property that it is very easy to use in practice, involving in effect one less degree of freedom than we would otherwise have to deal with. But where is the connected argument giving it a theoretical justification? Again, to the best of my knowledge, only the Bayesian approach gives us this. One finds that the Neyman-Pearson decision rule is included as a special case of the Bayesian, for particular prior probabilities and loss functions. The Bayesian approach, far from conflicting with the Neyman-Pearson method, completes that method by supplying what was previously missing; a clear statement of the class of problems for which it *is* the optimum procedure.

The situation is rather different with the principle of confidence intervals. This method is not only ad hoc — it is ambiguous. Furthermore, it does not have particularly pleasant properties; the attempt to calculate confidence intervals leads to some of the more dreary and messy mathematical problems in statistics. Consider the simplest case of Bernoulli trials B_2; we observe n successes in N trials, and are 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 simply a problem of parameter estimation, not different in principle from any other. Bayes' theorem solves it in three lines giving, in the case of uniform prior probability density for f, a posterior distribution proportional to $f^n(1-f)^{N-n}$, with mean value $\bar{f} = (n+1)/(N+2)$ and variance $\sigma^2 = \bar{f}(1-\bar{f})/(N+3)$, a result given by Laplace in 1774 [4]. The $\bar{f} \pm \sigma$ thus found provide a good statement of the "best" estimate of f, and an interval within which the true value is reasonably likely to be. The full posterior distribution of f yields more detailed statements for large N (the only case in which accurate estimates are possible at all) it goes into a normal distribution from which these my be read off by inspection. When we treat this same problem by confidence intervals, we find that it is no longer a homework problem, but a research project [5]. The final results are so complicated that, for practical use, they must be expressed in graphs!

At this point, a little dose of pragmatism will do wonders in restoring this argument to a constructive level. In all of probability theory there is no principle which has been subject to more sneering abuse than Laplace's rule of succession, just mentioned. Instant denunciation or this rule has become an automatic reflex action. But suppose we control this reflex just long enough to take a glimpse at the final results, comparing, say, Neyman's 90 percent confidence belts with the Bayesian 90 percent posterior probability belts. in.

Bayesian belts lie just barely inside the Neyman belts; the difference is visible graphically only for very wide belts for which, in any event, no accurate estimate was possible.

After all philosophical arguments and polemics, there remains one simple mathematical fact, which anyone can verify for himself; stated more generally, *the person who estimates a parameter and places limits of reasonable accuracy on that estimate by the method of Bayes and Laplace starting from a uniform prior probability density, arrives at final conclusions which are for all practical purposes identical to those obtained by confidence intervals; and he does this by a calculation that is an order of magnitude shorter.*

Even the $(+1)$ and $(+2)$ of Laplace's formula come back to haunt us if we take as our "best" estimates the centre of a confidence interval at the 84 percent confidence level (Ref. [3], Eq. 34.2.5). The person who, once aware of these things, still persists in saying that it is wrong to use the Bayes-Laplace methods and right to use confidence intervals, places himself in a very curious logical position. From a purely pragmatic standpoint, it is just impossible to see what all the shouting is about.

Of course, from another standpoint, we all understand perfectly well what the shouting is about. The person who wishes to attack the Bayesian methods very wisely refrains from comparison of final results. The objection to Bayesian methods, which have filled the statistical literature for two generations, have nothing to do with their success or failure in applications, but instead attack their philosophical basis. If one wishes to argue this on the philosophical level, then I think there is by now an abundance of good arguments supporting the Bayesian view. But I don't think any argument on this level is ever going to convince anybody who doesn't want to be convinced. If we are ever to resolve this issue, we will have to stick to the more mundane level of comparing the mathematics.

When we do this, we find many more interesting things. For example, perhaps the strangest but most persistent objection to the Bayesian methods concerns the assignment of prior probabilities $p(\theta|X)$ to to a continuously variable parameter θ, where X stands for the prior information. Bayes suggested that, if we have no prior information about the value of θ, we may express this fact by assigning a uniform prior probability density: $p(\theta|X) = \text{const}$. But, goes the standard argument, if I know nothing about θ, then I also know nothing about θ^2 but assigning uniform prior probability to θ in not the same thing as assigning uniform prior probability to θ^2. How are we to decide which is right? Who is to say? How can we apply a theory which is ambiguous?

This in an important issue, which deserves a more complete discussion than we have time for here; but we can easily put the matter into proper perspective. Our ultimate purpose in assigning prior probabilities to θ is, of course, to make some estimate of θ or some related quantity, and probability theory cannot give us the required posterior probability distribution $p(\theta|EX)$ conditional on the new evidence E, until we put in the prior probabilities. This is simply the mathematical expression of an obvious common-sense fact: before you can answer the question, "what do you know about θ after observing evidence E?" you have to be able to answer the question, "What did you know about θ before observing E?" Any principle of parameter estimation or hypothesis testing which refuses to acknowledge the relevant of prior information, or fails to provide any means of

taking it into account, is by that fact alone proved to be, in Pratt's words, "not completely satisfactory, let alone uniquely objective and scientific." For, to seize upon one piece of evidence E and ignore another X, is to commit the most obvious inconsistency.

In this connection, there is one "normative axiom" which I think will be generally accepted: *If you reject method B on the grounds that it has property P, then it would be utterly preposterous to advocate that it be replaced by another method which also has property P.* So, before we have the full story about this problem of parameter changes, we must also examine the "orthodox" methods with this in mind.

In order to avoid the use of prior probabilities, the orthodox statistician introduces new principles not contained in basic probability theory — bias, efficiency, confidence intervals, likelihood, etc. But if you square an unbiased estimate of θ, you will not have an unbiased estimate of θ^2; there will be a positive bias equal to the variance of the estimator. If you "correct" for this bias in each case, your final conclusions will depend on how you have defined your parameters. Similarly, the square of an efficient estimate of θ is not an efficient estimate of θ^2; indeed, the very definition of efficiency [3] is parameter dependent in such a way that if an efficient estimator of θ exists, then (Ref. [3] p. 481) an efficient estimator of θ^2 does *not* exist. If you find a shortest confidence interval for θ, and a similar one for θ^2, the two procedures will end up placing θ in different intervals.

By any at the criteria of estimation, the orthodox statistician's final conclusions are going to depend on how he has defined his parameters. If instead of θ he decides to work with θ^n, then for sufficiently large n his conclusions will be wildly different from those he gets by use of θ. How are we to decide which is right? Who is to say? How can we use a theory which is ambiguous?

In some fifteen years of studying the statistical literature, I have found perhaps fifty passages in which some orthodox statistician sneers at the Bayes uniform probability assignment because it is not invariant under a change of parameters, but fails to add that his own criteria of bias, efficiency, and shortest confidence interval suffer from exactly the same lack of invariance.

Now, which of the orthodox methods *did* achieve invariance? To the best of my knowledge, there are just two principles which are independent of the choice of parameters; sufficiency and maximum likelihood. But, and this is the most amusing thing of all, these are just the two principles that *do* have a simple justification in Bayesian terms. The definition of sufficiency can be stated as: if the posterior distribution of θ depends on the observed sample values only through a single function $f(x_1 \cdots x_n)$ of the sample values x_i, then f is a sufficient statistic for estimation of θ. This definition of sufficiency is easily shown to be mathematically equivalent to Fisher's, and I think it is more succinct and intuitively meaningful.

Likewise, as Fisher has stressed, the square of a maximum-likelihood estimate of θ is a maximum-likelihood estimate of θ^2. But, there is that embarrassing little mathematical fact that the method of maximum likelihood is mathematically identical with applying Bayes' theorem *with the Bayes uniform prior probability assignment,* then choosing the mode of the posterior distribution as our estimate.

The same thing is found when we look at hypothesis testing. After many years of rejecting Bayesian methods here, orthodox statisticians hailed the great advance provided by Wald's sequential procedure based on the probability-ratio test. After considerable mathematical labours, it was proved that this procedure is the long-sought optimum one, in the sense of requiring, on the average, fewer tests than any other for a given probability of error. But after a few years, another embarrassing little mathematical fact became obvious to everyone: *Wald's method is mathematically identical with applying Bayes' theorem with uniform prior probabilities, then deciding that an hypothesis is true if the posterior probability reaches a certain pre-assigned level.* Now this is just the way Laplace was handling decision problems of physics in the 18'th century; and, of course, just what orthodox statisticians (including Wald himself in his Columbia University course notes of about 1941) assured us were completely wrong.

It is this last fact, more than anything else, that has many of us to take another look at the whole situation and ask ourselves, "All right, now just what is it that was so awful about the Bayesian approach? Can you really maintain that a viewpoint which leads in three lines to the same results that cost Wald several years of mathematical labour, is wrong? What are we to think of a doctrinaire school or thought which has denied to two generations of scientific workers the use of statistical methods which were finally proved to be the optimum ones after all? Gentlemen, *shall we start talking sense?*

The issue is not whether the Bayesian methods are 100 percent satisfactory — of course, no methods are. Both camps still face many ambiguities and unresolved questions of principle. In particular, the problem of invariance under parameter chances not resolved above — I have only pointed out that this same problem permeates all of statistics, so that an honest man cannot use it as a club to beat do either theory in favour of the other. The only question of real substance at this time in not whether the Bayesian methods are perfect, but only whether they are better or worse than the orthodox ones. I have tried to give here the bare outline of a comparative analysis [6] which considers not only the philosophical issues, but also the more important matter of their similarities and differences in actual practice.

We have by now same very compelling and "connected" arguments [6, 7, 8] supporting the view that there is a *general* set of rules for consistent inductive inference, which includes all the procedures of orthodox statistics as special cases or good approximations thereto, and which can be applied in many problems (which are of great importance in current physics and engineering), where orthodox statistics has no procedures at all. This set of rules is just the original Bayes-Laplace theory, which orthodox statistics rejects. Although I personally find the arguments for the Bayesian approach entirely convincing, it is an empirical fact that orthodox statisticians are immune to them; so it is a waste of effort to repeat them. Also, these arguments are, admittedly, still heuristic from the mathematical side. In view of this, how can we keep future discussion an a constructive level where there is hope of making progress?

There is, I think, only one way. We must continue to examine the specific mathematical steps, and final conclusions, which Bayesian and orthodox methods lead to when applied

to specific problems; just the procedure scientists use to test any physical theory. In spite of all evidence now in, it is of course still a logical possibility that Bayesian methods *are*, as Fisher always claimed, "founded upon an error," which the orthodox methods avoid. If this is true, it could, in my opinion, be demonstrated once and for all by producing a *single* example of a statistical problem where the orthodox methods give a satisfactory result, but the Bayesian ones do not. It cannot be proved by linguistics.

I am sorry if the following sounds like still another challenge; but perhaps that is after all the surest way to get to the heart of an issue. If, after pondering my arguments, you still believe that Bayesian methods are erroneous and orthodox ones superior, why not try to prove your point in a way that transcends all polemics, by producing this example?

If you succeed, then I and the other Bayesians will learn something much to our advantage, and we can all get back to more worth-while things. If you fail, then you will learn something with a force that no arguments of mine could quite convey. If you finally conclude, as I have, that this problem does not exist, what will be the proper attitude? I suppose it is possible, without actual logical contradiction, to maintain that Bayesian methods are utterly false but that through a fortuitous accident, they always happen to lead to the right answers in every particular problem. But I don't think anyone will want to take that position. If you study these things long enough, I think you will suddenly find that you have become a Neo-Bayesian!

Finally, a comment about the article by Professor H. O. Hartley, which also appeared in the February 1963 issue of this journal [9]. The problem pointed out by Hartley is a real one, well recognised by those using Bayesian statistical methods. It is, however, not a problem of statistics, but of communication between statistician and client; and readers of this journal may be interested to know what is being done about it.

In Hartley's dialogue, the trouble was that Mr. Busy, the engineer, did not understand statistical principles; and Dr. Bayes, the statistician, did not understand engineering. Dr. Bayes' realized that Mr. Busy's prior probabilities were inconsistent, but did not know how to resolve that inconsistency; on the other hand, Mr. Busy was drawing upon an enormous mass of prior information and experience which he could not possibly explain to Dr. Bayes. (It is only fair to remark parenthetically that this prior information problem is one for which orthodox statistics offers no solution at all.)

The answer, or course, in education. Programs of teaching Bayesian statistical methods to engineers are being initiated in many places, and for several years I have been aiding and abetting this process by giving intensive short courses, thus far at six different Universities and three industrial laboratories. It is possible to do this because the Bayesian theory is incomparably simpler and more general than the orthodox approach, and it corresponds exactly to the engineer's innate common sense; he is delighted to see the kind of reasoning he has been doing all along in a qualitative way, reduced to simple quantitative terms. Instead of a seemingly endless series of separate ad hoc principles, each applicable to a narrow and imprecisely defined class of problems, the Bayesian approach gives him a single set of principle which has an obvious intuitive appeal, and covers the entire field.

In particular sequential testing, which a few years ago was taught only as an advanced

graduate course in statistics, is now being taught to undergraduate engineers as the *first*, and simplest possible, example of the use of Bayes' theorem.

The engineer who has been taught Bayesian statistics and has seen their application in his problems, is incredulous when told that there exists a school of thought which condemns all these methods without bothering to examine their final results. It is particularly hard to understand how it is possible to reject the use of uniform prior probabilities to express ignorance, and then advocate the orthodox methods because, an the above examples show, *refusal to use prior probabilities at all is mathematically the same thing as assigning uniform prior probabilities.*

I believe that the Mr. Busy of 1970 will have no need of the services of Dr. Bayes; or of any other statistician. It is very much the other way around; the statistician who fails to learn Bayesian methods, and finds himself confronted with a problem where the prior information can no longer be swept under the rug, will find Mr. Busy willing and able to help him out.

References

[1] Irwin D. J. Bross. Linguistic Analysis of a Statistical Controversy. *The American Statistician*, 17(1):18–21, 1963.

[2] John W. Pratt. Letter to the Editor. *The American Statistician*, 16(2):24–24, 1962.

[3] H. Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[4] P. S. Laplace. Mèmoire sur la probabilitè des causes par les èvènements. 1774.

[5] C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934.

[6] E. T. Jaynes. *Probability in Science and Engineering*. McGraw-Hill Book Co., Inc., New York (in press).

[7] R. T. Cox. *The Algebra of Probable Inference*. John Hopkins Press; Baltimore, 1961.

[8] R. T. Cox and E. T. Jaynes. The Algebra of Probable Inference. *American Journal of Physics*, 31(1):66–67, 1963.

[9] H. O. Hartley. In Dr. Bayes' Consulting Room. *The American Statistician*, 17(1):22–24, 1963.