What is the question?

DISCUSSION

E.T. JAYNES (*Washington University*):

It is always interesting to recall the arguments that Jeffreys used to find priors. The case recounted by Zellner is a typical example where it appears at first glance that we have nothing to go on; yet by thinking more deeply, Jeffreys finds something. He shows an uncanny ability to see intuitively the right thing to do, although the rationalization he offers is sometimes, as Laplace said of Bayes' argument, "fine et très ingénieuse, quoiqu'un peu embarrassée". It was from studying these flashes of intuition in Jeffreys that I become convinced that there must exist a general formal theory of determination of priors by logical analysis of the prior information-and that to develop it is today the top priority research problem of Bayesian theory.

Pragmatically, the actual results of the Jeffreys-Zellner-Siow and Bernardo tests seem quite reasonable; without considerable analysis one could hardly say how or whether we should want them any different. Likewise, there is little to say about the mathematics, since once the premises are accepted, all else seems to follow in a rather straightforward and inevitable way. So let us concentrate on the premises; more specifically, on the technical problems encountered in both works, caused by putting that lump of prior probability on a single point $\lambda = 0$.

1. *The problem*

In most Bayesian calculations the same prior appears in numerator and denominator, and any normalization constant cancels out. Usually, passage to the limit of an "uninformative" improper prior is then uneventful; i.e., our conclusions are very robust with respect to the exact prior range. But in Jeffreys' significance test this robustness is lost, since $K = p(D|H_o)/p(D|H_1)$ contains in the denominator an uncancelled factor which is essentially the prior density $\pi(\lambda)$ at $\lambda = \bar{x}$. Then in the limit of an improper prior we have $K \to \infty$ independently of the data D, a result given by Jeffreys (1939, p. 194, Eq. 10), and since rediscovered many times. Note that the difficulty is not due solely to the different dimensionality of the parameter spaces; it would appear in any problem where we think of H_o as specifying a definitive, fixed prior range, but fail to do the same for H_1.

Jeffreys (1961) dealt with this and other problems by using a Cauchy prior $\pi(\lambda|\sigma)$ scaled on σ in the significance test, although he would have used a uniform prior $\pi(\lambda) = 1$ in the same model H_1 had he been estimating λ. But then a question of principle rears up. To paraphrase Lindley's rhetorical question: Why should our prior knowledge, or ignorance, of λ depend on the question we are asking about it? Even

more puzzling: why should it depend on another parameter σ, which is itself unknown? One feels the need for a clearer rationalization.

Furthermore, the difficulty was not really removed, but only concealed from view, by Jeffreys' procedure. All his stated conditions on the prior would have been met equally well had he chosen a Cauchy distribution with interquartile span 4σ instead of σ; but then all his K-values would have been quadrupled, leading to indifference at a very different value of the t-statistic [see Eq. (5-13) below]. We do not argue that Jeffreys made a bad choice; quite the contrary. Our point is rather that in his choice there were elements of arbitrariness, arising from a still unresolved question of principle. Pending that resolution, one is not in a position to say much about the "uniqueness" or "objectivity" of the test beyond the admitted virtue of yielding results that seem reasonable.

Bernardo comes up against just the same problem, but deals with it more forthrightly. Finding again that the posterior probability P_o of the null hypothesis H_o increases with the prior variance σ_1 in a disconcerting way, he takes what I should describe as a meat-axe approach to the difficulty, and simply chops away at its prior probability p until $P_o = pk/(pk + 1\text{-}p)$ is reduced to what he considers reasonable (from the Jeffreys-Zellner-Siow standpoint he chops a bit too much, since his P_o tends only to 1/2 on prolonged sampling when H_o is true). This approach has one great virtue: whereas the Jeffreys results tended to be analytically messy, calling for tedious approximations, Bernardo emerges triumphantly (in the limit of large σ_1) with a beautifully neat expression (Eq. (11)) which has also, intuitively, a clear ring of truth to it.

But for this nice result, Bernardo pays a terrible price in unBayesianity. He gets it only by making p vary with the sample size n, calling for another obvious paraphrase of Lindley. This elastic quality of his prior is rationalized by an information-theoretic argument; it is, in a sense, the prior for which one would expect (before seeing the data) to learn the most from the experiment. But is this the property one wants?

If a prior is to incorporate the *prior information* we had about λ before the sample was observed, it cannot depend on the sample. The difficulty is particularly acute if the test is conducted sequentially; must we go back to the beginning and revise our prior as each new data point comes in? Yet after all criticisims I like the general tone of Bernardo's result, and deplore only his method of deriving it.

The common plot of these two scenarios is: we (1) start to apply Bayes' theorem in what seems a straightforward way; (2) discover that the result has an unexpected dependence on the prior; (3) patch things up by tampering with the prior until the expected kind of result emerges. The Jeffreys and Bernardo tamperings are similar in effect, although they offer very different rationalizations for what they do. But in both cases the tampering has a mathematical awkwardness and the rationalization a certain contrived quality, that leads one to ask whether some important point has been missed.

Now, why should that first result have been unexpected? If, according to H_1, we know initially only that λ is in some very wide range $2\sigma_1$, and we then receive data showing that it is actually within $\pm\sigma/\sqrt{n}$ of the value predicted by H_o, -as a physicist would put it, "the data agree with H_o to within experimental error"- that is indeed very strong evidence in favor of H_o. Such data ought to yield a likelihood ratio $K = \sqrt{n}\sigma_1/\sigma$

increasing with σ_1, just as Bernardo finds. This first result is clearly the correct answer to the question Q_1 that was being asked.

If we find that answer disconcerting, it can be only because we had in the back of our minds a different, unenunciated question Q_2. On this view, the tampering is seen as a mutilation of equations originally designed to answer Q_1, so as to force them to answer instead Q_2.

The higher-level question: "Which question should we ask?" does not seem to have been studied explicitly in statistics, but from the way it arises here, one may suspect that the answer is part of the necessary "software" required for proper use of Bayesian theory. That is, just as a computer stands ready to perform any calculation we ask of it, our present theory of Bayesian inference stands ready to answer any question we put to it. In both cases, the machine needs to be programmed to tell it which task to perform. So let us digress with some general remarks on question-choosing.

2. *Logic of Questions*

For many years I have called attention to the work on foundations of probability theory by R.T. Cox (1946,1961) which in my view provides the most fundamental and elegant basis for Bayesian theory. We are familiar with the Aristotelian deductive logic of propositions; two propositions are equivalent if they say the same thing, from a given set of them one can construct new propositions by conjunction, disjunction, etc. The probability theory of Bernouilli and Laplace included Aristotelian logic as a limiting form, but was a mathematical extension to the intermediate region $(0<p<1)$ between proof and disproof where, of necessity, virtually all our actual reasoning takes place. While orthodox doctrine was rejecting this as arbitrary, Cox proved that it is the only consistent extension of logic in which degrees of plausibility are represented by real numbers.

Now we have a new work by Cox (1978) which may prove to be of even more fundamental importance for statistical theory. Felix Klein (1939) suggested that questions, like propositions, might be used as logical elements. Cox shows that in fact there is an exactly parallel logic of questions: two questions are equivalent if they ask the same thing, from a given set of them one can construct new questions by conjunction (ask both), disjunction (ask either), etc. All the "Boolean algebra" of propositions may be taken over into a new symbolic algebra of questions. Every theorem of logic about the "truth value" of propositions has a dual theorem about the "asking value" of questions.

Presumably, then, besides our present Bayesian statistics -a formal theory of optimal inference telling us which propositions are most plausible- there should exist a parallel formal theory of optimal inquiry, telling us which questions are most informative. Cox makes a start in this direction, showing that a given question may be defined in many ways by the set of its possible answers, but the question possesses an entropy independent of its defining set, and the entropies of different questions obey algebraic rules of combination much like those obeyed by the probabilities of propositions.

The importance of such a theory, further developed, for the design of experiments and the choosing of procedures for inference, is clear. For over a century we have

argued over which *ad hoc* statistical procedures ought to be used, not on grounds of any demonstrable properties, but from nothing more than ideological committments to various preconceived positions. There is still a great deal of this in my exchanges with Margaret Maxfield and Oscar Kempthorne in Jaynes (1976), and even a little in the exchange with Dawid, Stone, and Zidek over marginalization in Jaynes (1980). A formal theory of optimal inquiry might resolve differences of opinion in a way that Wald-type decision theory and Shannon-type information theory have not accomplished.

Our present problem involves a special case of this. If, seeing the answer to question Q_1 we are unhappy with it, what alternative question Q_2 did we have, unconsciously, in the back of our minds? Is there a question Q_3 that is the optimal one to ask for the purpose at hand? Since the conjectured formal theory of inquiry is still largely undeveloped, we try to guess some of its eventual features by studying this example.

Note that the issue is not which question is "correct". We are free to ask of the Bayesian formalism any question we please, and it will always give us the best answer it can, based on the information we have put into it. But still, we are in somewhat the position of a lawyer at a courtroom trial. Even when he has on the stand a witness who knows all the facts of the case and is sworn to tell the truth, the information he can actually elicit from this witness still depends on his adroitness in asking the right questions.

If his witness is unfriendly, he will not extract any information at all unless he knows the right questions to force it out, phrasing them as sharp leading questions and demanding unequivocal "yes" or "no" answers. But if a witness is friendly and intelligent, one can get all the information desired more quickly by asking simply, "Please tell us in your own words what you know about the case?" Indeed, this may bring out unexpected new facts for which one could not have formulated any specific question.

Significance tests which specify a sharply defined hypothesis and preassigned significance level, and demand to know whether the hypothesis does or does not pass at that level, therefore in effect treat probability theory as an unfriendly witness and automatically preclude any possibility of getting more information than that one bit demanded.

Suppose we try instead the opposite tactic, and regard Bayesian formalism as a friendly witness, ready and willing to give us all the pertinent information in our problem even information that we had not realized was pertinent if we only allow if the freedom to do so. Instead of demanding the posterior probability of some sharply formulated null hypothesis H_0, suppose we ask of it only, "Please tell us in your own words what you know about λ?" Perhaps by asking a less sharp and restrictive question, we shall elicit more information.

3. *Information from questions*

Evidently, to deal with such problems one ought to be an information theorist, and not only in the narrow sense of One-Who-Uses-Entropy. In the present problem we are concerned not only with the range of possible answers, as measured by the

entropy of a question, but also with the specific kind information that the question can elicit. In the following we use the word "information" in this semantic sense rather than the entropy sense.

All statistical procedures are in the last analysis prescriptions for information processing: what information have we put into our mathematical machine, and what information are we trying to get out of it? In these terms, what is the difference -if any- between significance testing and estimation? Having put certain information (model, prior, and data) into our hopper, we may carry out either, by asking different questions. But the answers to different questions do not necessarily convey different information.

The tests considered by Zellner and Bernardo sought information that can help us decide whether to adopt a new hypothesis H_1 with a value of λ different from its currently supposed value $\lambda = 0$. Presumably, any procedure which yields the same information would be equally acceptable for this purpose, even though current pedagogy might not call it a "significance test".

Now this information criterion establishes an ordering of different procedures, or "tests", rather like the notion of admissibility. If test B (which answers question Q_B) always gives us the same information as test A, and sometimes more, then B may be said to dominate A in the sense of information yield, or question Q_B dominates Q_A in "asking power"; and if B requires no more computation, on what grounds could one ever prefer A?

In my work of 1976 (p. 185 and p. 219), I showed that the original Bayesian sigificance test of Laplace, which asks for the posterior probability P_1 of a one-sided alternative hypothesis, dominates the traditional orthodox t-test and F-test in just this sense. That is, given P_1 we know what the verdict would be, at any significance level, for all three of the corresponding orthodox tests (one equal-tails and two one-sided; but the veredict of any one orthodox test is far from determining P_1. Thanks to Cox, we have now a much broader view of this phenomenon.

Let us call a question *simple* if its answer is a single real number; or in Cox's terminology, if its irreducible defining set is a set of real numbers. For example: "What is the probability that λ, or some function of λ, lies in a certain region R?"

In any problem involving a single parameter λ for which there is a single sufficient statistic u, then given any simple question Q_A about λ, the answer will be, necessarily, some function $a(u)$. Given any two such questions Q_A, Q_B and any fixed prior information, the answers $a(u)$, $b(u)$, being functions of a single variable u, must obey some functional relation $a = f(b)$. If $f(b)$ is single-valued, then the answer to Q_B tells us everything that the answer to Q_A does. As Cox puts it, "An assertion answering a question answers every implicate of that question". If the inverse function $b = f^{-1}(a)$ is not single-valued, then Q_B dominates Q_A.

In the case of a single sufficient statistic, then, any simple question whose answer is a strict monotonic function of u, yields all the information that we can elicit about λ, whatever question we ask; and it dominates any simple question whose answer is not a strict monotonic function of u. But this is just the case discussed by Bernardo; he considers σ known, and consequently $\bar{x}$ is sufficient statistic for λ. Since his odds ratio $K(\bar{x})$ is not a strict monotonic function of $\bar{x}$, we know at once that Bernardo's test is

dominated by another.

The Jeffreys-Zellner-Siow tests are more subtle in this respect, since σ is unknown, and consequently there are two jointly sufficient statistics $(\bar{x},s)$. Given two simple questions Q_A, Q_B with answers $a(\bar{x},s)$, $b(x,s)$, the condition that they ask essentially the same thing, leading to a functional relation $a=f(b)$, is that the Jacobian $J = \partial(a,b)/\partial(\bar{x},s)$ should vanish. If $J \neq 0$, then neither questions can dominate the other and no simple question can dominate both. But any two simple questions for which $(\bar{x},s)$ are uniquely recoverable as single-valued functions $\bar{x}(a,b)$, $s(a,b)$ will jointly elicit all the information that any question can yield, and thus their conjunction dominates any simple question.

We may, therefore, conclude the following. Since Jeffreys' test asks a simple question, whose answer is the odds ratio $K(\bar{x},s)$, it can be dominated by a compound question, the conjuction of two simple questions. Indeed, since K depends only on the magnitude of the statistic t, it is clear that Jeffreys' question is dominated by any one simple question whose answer is a strict monotonic function of t.

These properties generalize effortlessly to higher dimensions and arbitrary sets. Whenever sufficient statistics exist, the most searching questions for any statistical procedure, -whatever current pedagogy may call it- are those (simple or compound) from whose answers the sufficient statistics may be recovered; and all such questions elicit just the same information from the data.

As soon as I realized this, it struck me that this is exactly the kind of result that Fisher would have considered intuitively obvious from the start; however, a search of his collected works failed to locate any passage where such an idea is stated. Perhaps others may recall instances where he made similar remarks in private conversation; it is difficult to believe that he was unaware of it.

With these things in mind, let us re-examine the rationale of the Jeffreys-Zellner-Siow and Bernardo tests.

4. *What is our rationale?*

In pondering this -trying to see where we have confused two different questions and what the question Q_2 is- I was struck by the constrast between the reasoning used in the proposed tests and the reasoning that physicists use, in everyday practice, to decide such matters. We cite one case history; recent memory would yield a dozen equally good, which make the same point.

In 1958, Cocconi and Salpeter proposed a new theory H_1 of gravitation, which predicted that the inertial mass of a body is a tensor. That is, instead of Newton's $F=Ma$, one had $F_i = \Sigma M_{ij} a_j$. For terrestrial mechanics the principal axes of this tensor would be determined by the distribution of mass in our galaxy, such that with the x-axis directed toward the galactic center, $M_{xx}/M_{yy} = M_{xx}/M_{zz} = (1+\lambda)$. From the approximately known galactic mass and size, one could estimate (Weisskopf, 1961) a value $\lambda \cong 10^{-8}$.

Such a small effect would not have been noticed before, but when the new hypothesis H_1 was brought forth it became a kind of challenge to experimental physicists: devise an experiment to detect this effect, if it exists, with the greatest possible sensitivity. Fortunately, the newly discovered Mössbauer effect provided a test

with sensitivity far beyond one's wildest dreams. The experimental verdict (Sherwin, *et.al,* 1960) was that λ, if it exists, cannot be greater than $|\lambda| < 10^{-15}$. So we forgot about H_1 and retained our null hypothesis: H_o = Einstein's theory of gravitation, in which $\lambda = 0$.

From this and other case histories in which other conclusions were drawn, we can summarize the procedure of the physicist's significance test as follows: (A) Assume the alternative H_1, which contains a new parameter λ, true as a working hypothesis. (B) On this basis, devise an experiment which can measure λ with the greatest possible precision. (C) Do the experiment. (D) Analyze the data as a pure estimation problem-Bayesian, orthodox, or still more informal, but in any event leading to a final "best" estimate and a statement of the accuracy claimed: $(\lambda)_{est} = \lambda' \pm \delta\lambda$. It is considered good form to claim an accuracy $\delta\lambda$ corresponding to at least two, preferably three, standard deviations. (E) Let λ_o be the correct value according to the null hypothesis H_o (we supposed $\lambda_o = 0$ above, but it is now best to bring it explicitly into view), and define the "statistic" $t \equiv (\lambda' - \lambda_o)/\delta\lambda$. Then there are three possible outcomes:

If $\|t\| < 1$, retain H_o,	STATUS QUO
If $\|t\| >> 1$, accept H_1,	AWARD NOBEL PRIZES
If $1 < \|t\| < 3$, withhold judgment	SEEK BETTER EXPERIMENTS

That is, to within the usual poetic license, the reasoning format in which the progress of physics takes place.

You see why I like the actual results reported here by Zellner and Bernardo, although I find their rationalizations puzzling. They did indeed find, as the criterion for accepting H_1, that the estimated deviation $|\lambda' - \lambda_o|$ should be large compared to the accuracy of the measurement, considered known ($\sigma/\sqrt{n}$) in Bernardo's problem, and estimated from the data in the usual way ($s/\sqrt{n}$) in Zellner's.

It is in the criterion for retaining H_o that we seem to differ; contrast the physicist's rationale with that usually advanced by statisticians, Bayesian or otherwise. When we retain the null hypothesis, our reason is not that it has emerged from the test with a high posterior probability, or even that it has accounted well for the data. H_o is retained for the totally different reason that if the most sensitive available test fails to detect its existence, the new effect $(\lambda - \lambda_o)$ can have no observable consequences. That is, we are still free to adopt the alternative H_1 if we wish to; but then we shall be obliged to use a value of λ so close to the previous λ_o that all our resulting predictive distributions will be indistinguishable from those based on H_o.

In short, our rationale is not probabilistic at all, but simply pragmatic; having nothing to gain in predictive power by switching to the more complicated hypothesis H_1, we emulate Ockham. Note that the force of this argument would be in no way diminished even if H_o had emerged from some significance test with an extremely low posterior probability; we would still have nothing to gain by switching. Our acceptance of H_1 when $|t| >> 1$ does, however, have a probabilistic basis, as we shall see presently.

Today, most physicists have never heard the term "significance test". Nevertheless, the procedure just described derives historically from the original tests devised by Laplace in the 18'th Century, to decide whether observational data indicate

the existence of new systematic effects. Indeed, the need for such tests in astronomy was the reason why the young Pierre Simon developed an interest in probability theory, forty-five years before he became the *Marquis de Laplace*. This problem is therefore the original one, out of which "Bayesian statistics" grew.

As noted also by E.C. Molina (1963) in introducing the photographic reproduction of Bayes' paper, even the result that we call today "Bayes' theorem" was actually given not by Bayes but by Laplace (the only valid reason I have found for calling it "Bayes' theorem" was provided at this meeting; "There's no theorem like Laplace's theorem" does not set well to Irving Berlin's music). Molina also offers some penetrating remarks about Boole's work, showing that those who have quoted Boole in support of their criticisms of Bayes and Laplace may have mistaken Boole's intention.

Now, although Laplace's tests were thoroughly "Bayesian" in the sense just elucidated, they encountered no such difficulty as those found by Jeffreys and Bernardo; he always got clear-cut decisions from uniform priors without tampering. To see how this was managed, let us examine the simplest of all Laplacian significance tests.

As soon as fairly extensive birth records were kept, it was noticed that there were almost always slightly more boys than girls, the ratio for large samples lying usually in the range $1.04 < (n_b/n_g) < 1.06$. Today we should, presumably, reduce this to some hypothesis about a difference in properties of X and Y chromosomes (for example, the smaller Y chromosome, leading to a boy, would be expected to migrate more rapidly). But for Laplace, knowing nothing of such things, the problem was much simpler. Making no reference to any causal mechanism, he took the model of Bernoulli trials with parameter λ = probability of a boy.

His problem was then: given specific data $D = \{n_b, n_g\}$, do these data indicate the existence of some systematic cause favoring boys? Always direct and straightforward in his thinking, for him the proper question to ask of the theory was simply: Q_L = "Conditional on the data, what is the probability that $\lambda > (1-\lambda)$?" With uniform prior, answer was

$$P_L = \frac{(n+1)!}{n_b!\, n_g!} \int_{\lambda_0}^{1} \lambda^{n_b}(1-\lambda)^{n_g} d\lambda$$

with $n = n_b + n_g$, $\lambda_0 = 1/2$. In this *Essai Philosophique* Laplace reports many results from this, and in the *Theorie Analytique* (Vol. 2, Chap. 6) he gives the details of his rather tedious methods for numerical evaluation.

Needless to say, Laplace was familiar with the normal approximation to $p(d\lambda \mid D)$, the inverse of the de Moivre-Laplace limit theorem. But Laplace also realized that the normal approximation is valid only within a few standard deviations of the peak, and when the numbers n_b, n_g become very large, it can lead easily to errors of a factor of 10^{100} in $P_L/(1-P_L)$; hence his tedious methods.

Bernardo's example of Mrs. Stewart's telepathic powers, where the null hypothesis value $\lambda_0 = 0.2$ is about 24 standard deviations out, is another instance where the normal approximation leads to enormous numerical errors in K (many millions, by my estimate).

But pragmatically, once it is estimated that an odds ratio is about 10^{130}, it hardly matters if the exact value is really only 10^{120}. Once it is clear that the evidence is overwhelmingly in favor of H_1, nobody cares precisely how overwhelming it is. After Laplace's time, physicists lost interest in his accurate but tedious evaluations of P_L; for the criterion that we have overwhelming evidence in favor of a positive effect ($\lambda > \lambda_o$), is just that the overwhelmingly greater part of the mass of the posterior distribution $p(d\lambda|D)$ shall lie to the right of λ_o. In the above example, the peak and standard deviation of $p(d\lambda|D)$ are $\lambda' = n_b/n$, $\delta\lambda = [\lambda'(1-\lambda')/n]^{1/2}$ and this criterion reduces to the aforementioned $t = (\lambda' - \lambda_o)/\delta\lambda >> 1$, of the modern physicist's significance test-just the same criterion that Jeffreys and Bernardo arrive at in their different ways.

We have noted above that the orthodox t-test and F-test are dominated by Laplace's, and argued that the Jeffreys and Bernardo tests must also be dominated by some other. Let us now compare their specific tests with the ones Laplace would have used in their problems.

5. *Comparisons with Laplace*

In Bernardo's problem we have a normal sampling distribution $p(dx|\lambda,\sigma) \sim N(\lambda,\sigma)$ with σ known. Hypothesis H_o specifies $\lambda = \lambda_o$, H_1 a normal prior $\pi(d\lambda|H_1) \sim N(\mu_1,\sigma_1)$, leading to a normal posterior distribution $p(d\lambda|D,H_1) \sim \mathrm{N}(\lambda',\delta\lambda)$ where

$$(\delta\lambda)^{-2} = n\sigma^{-2} + \sigma_1^{-2} \tag{5.1}$$

$$\lambda' = n(\delta\lambda/\sigma)^2 \bar{x} + (\delta\lambda/\sigma_1)^2\mu_1 \tag{5.2}$$

Laplace, asking for the probability of a positive effect, would calculate

$$P_L = p(\lambda > \lambda_o|D,H_1) = \Phi(t) \tag{5.3}$$

where $\Phi(t)$ is the cumulative normal distribution, and as always, $t \equiv (\lambda' - \lambda_o)/\delta\lambda$.

Bernardo (Eq. 9) finds for the posterior odds ratio

$$K_B = p(H_o|D)/p(H_1|D) = \exp(-R/2) \tag{5.4}$$

where

$$R = \frac{(\bar{x}-\lambda_o)^2}{\sigma^2/n} - \frac{(\bar{x}-\mu_1)^2}{\sigma_1^2 + \sigma^2/n} . \tag{5.5}$$

But by algebraic rearrangement, we find this is equal to

$$R = t^2 - w^2 \tag{5.6}$$

where $w \equiv (\mu_1 - \lambda_o)/\sigma_1$ is independent of the data and drops out if $\mu_1 = \lambda_o$ or if $\sigma_1 \to \infty$. Bernardo would then find for the posterior probability of the null hypothesis

$$P_B = p(H_o|D) = [\exp(t^2/2) + 1]^{-1} \tag{5.7}$$

and comparing with (5.3) we have, as anticipated, a functional relation $P_B = f(P_L)$. To see the form of it, I plotted P_B against P_L and was surprised to find a quite accurate semicircle, almost as good as one could make with a compass. To all the accuracy one could use in a real problem, the functional relation is simply

$$P_B \cong [P_L(1-P_L)]^{1/2}, \qquad 0 \le P_L \le 1 \tag{5.8}$$

The error in (5.8) vanishes at five points ($0 \le P_L \le 1$).

Since $P_B = f(P_L)$ is single-valued while the inverse function is not, we have the result that Laplace's original significance test does, indeed, dominate Bernardo's. As stressed in Jaynes (1976), one-sided tests always dominate two-sided ones; gives P_L we know everything that Bernardo's K or P_B can tell us; and if $|t| >> 1$ we know in addition whether $\lambda > \lambda_o$ or $\lambda < \lambda_o$, which P_B does not give.

Of course, in this case one can determine that extra bit of information from a glance at the data; so the mere fact of domination is hardly a strong selling point. What is important is that Laplace's method achieves this without any elements of arbitrariness or unBayesianity.

In Jeffreys' problem we have the same sampling distribution, with the standard likelihood function $L(\lambda,\sigma) = \sigma^{-n} \exp[-ns^2Q^2(\lambda)/2\sigma^2]$, where

$$Q(\lambda) \equiv [1 + (\lambda-\bar{x})^2/s^2]^{1/2} \tag{5.9}$$

H_0 and H_1 assign common priors $d\sigma/\sigma$, but H_o specifies $\lambda = \lambda_o$, while H_1 assigns the Cauchy prior $p(d\lambda|\sigma,H_1) = \pi(\lambda|\sigma)d\lambda$ with the density

$$\pi(\lambda|\sigma) = \frac{a\sigma}{\pi(a^2\sigma^2 + \lambda^2)} \tag{5.10}$$

scaled on σ (Jeffreys takes $a = 1$, $\lambda_o = 0$, but we define the problem thus to bring out some points noted in Sec. 1). To analyze the import of the data, Jeffreys then calculates the likelihood ratio

$$K(\bar{x},s) = \frac{p(D|H_o)}{p(D|H_1)} = M^{-1} \int_0^\infty L(\lambda_o,\sigma)d\sigma/\sigma \tag{5.11}$$

while Laplace (if he used the same prior) would calculate instead the probability of a positive effect, given H_1:

$$P_L(\bar{x},s) = p(\lambda > \lambda_o|D,H_1) = M^{-1} \int_{\lambda_0}^\infty d\lambda \int_0^\infty d\sigma\, \sigma^{-1}\, \pi(\lambda|\sigma)L(\lambda,\sigma) \tag{5.12}$$

These expressions have a common denominator M, equal to the integral in (5.12) with $\lambda_o = -\infty$.

It is straightforward but lengthy to verify that Jeffreys and Laplace do not ask

exactly the same question; i.e., $J \equiv \partial(K_J,P_L)/\partial(x,s) \neq 0$. However, they are not very different, as we see on making the same approximation (large n) that Jeffreys makes. Doing the σ-integration in (5.12) approximately, the other integrals may be done exactly, leading to the approximate form

$$K_J \cong [\pi(n\text{-}1)/2]^{1/2}\, a(1+q^2)/Q^n(\lambda_o) \tag{5.13}$$

where $q \equiv (x/as)$. This reduces to Jeffreys' result [Zellner's Eq. (2.7) in this volume] when $a = 1$, $\lambda_o = 0$. In the same approximation, Laplace's result is the tail area of a t-distribution with n-2 degrees of freedom:

$$P_L \cong A_n \int_{\lambda_o}^{\infty} d\lambda/Q^{n-1}(\lambda) \tag{5.14}$$

where A_n is a normalization constant. Of course, if Laplace used a uniform prior for λ, he would find instead the usual "Student" result with (n-1) degrees of freedom.

In the limit of an improper prior ($a \to \infty$), K_J diverges as noted in Sec. 1, the original motivation for both the Jeffreys and Bernardo tamperings; but the arbitrary parameter a cancels out entirely from Laplace's leading term, appearing only in higher terms of relative order n^{-1}.

Had we been estimating λ instead, we should find the result $(\lambda)_{est} = \lambda' \pm \delta\lambda$, where $\lambda' = x$, $\delta\lambda = s/\sqrt{n}$. But Laplace's result (5.14) is a function only of the statistic $t = (\lambda' - \lambda_o)/\delta\lambda$, and Jeffreys' (5.13) is too for all practical purposes (exactly so if $\lambda_o = 0$, as Jeffreys assumes). Therefore, while considering σ unknown has considerably complicated the mathematics, it does not lead to any real difference in the conclusions. Again, Laplace's test yields the same information as that of Jeffreys, and in addition tells us the sign of $(\lambda\text{-}\lambda_o)$. In all cases -Jeffreys, Bernardo, Laplace, and the modern physicist's test- the condition that the data indicate the existence of a real effect is that $|t| >> 1$.

6. *Where does this leave Q_1?*

In summary it should not, in my view, be considered "wrong" to ask the original question Q_1 = "What is the relative status of H_o and H_1 in the light of the data?" But the correct answer to that question depends crucially on the prior range of λ according to H_1; and so the question appears in the retrospect awkward.

Now the original motivation for asking Q_1, stated very explicitly by Jeffreys, was to provide a probabilistic justification for the process of induction in science, whereby sharply defined laws are accepted as universally valid. But as both Jeffreys and Bernardo note, H_o can never attain a positive posterior probability unless it is given some to start with; hence that "pump-priming" lump of prior probability on a single point $\lambda = 0$. It seems usually assumed that this step is the cause of the difficulty.

However, the question Q_1 is awkward in another, and I think more basic, respect. The experiment cannot distinguish differences in λ smaller than its "resolving power" $\delta\lambda = s/\sqrt{n}$. Yet Q_1 asks for a decision between H_o and H_1 even when $|\lambda\text{-}\lambda_o| < \delta\lambda$. On the other hand, the experiment is easily capable of telling us whether λ is probably greater

or less than λ_o (Laplace's question), but Q_1 does not ask this. In short, Q_1 asks for something which the experiment is fundamentally incapable of giving; and fails to ask for something that the experiment *can* give.

[Incidentally, a "reference prior" based on the Fisher information $i(\lambda)$ is basically a description of this resolving power $\delta\lambda$ of the experiment. That is, the reference prior could be defined equally well as the one which assigns equal probabilities to the "equally distinguishable" subregions of the parameter space, of size $\delta\lambda$. This property is quite distinct from that of being "uninformative", although they happen to coincide in the case of single location and scale parameters].

But what we noted in Sec. 4 above suggests a different view of this. Why does induction need a probabilistic justification if it has already a more compelling pragmatic one? It is for the departures from the previous line of induction (i.e., switching to H_1) that we need -and Laplace gave- a probabilistic justification. Bernardo seems to have sensed this also, in being content with the fact that his $p(H_o|D)$ tends only to 1/2 when H_o is true. Once we see that maintenance of the *status quo* requires no probabilistic justification, the original reason for asking Q_1 disappears.

7. *Conclusion*

What both the Jeffreys and Bernardo tamperings achieved is that they managed to extricate themselves from an awkward start and, in the end, succeeded in extracting the same information from the data (but for the sign of λ-λ_o) that Laplace's question Q_L = "What is the probability that there is a real, positive effect?" elicited much more easily. What, then, was that elusive question Q_2? It was not identical with Q_L, and perhaps does not need to be stated explicitly at all; but in Cox's terminology we may take Q_2 as *any implicate of Laplace's question whose answer is a strict monotonic function of* $|t|$.

We have seen how the answers to seemingly very different questions may in fact convey the same information. Laplace's original test elicits all the information that can be read off from Jeffreys' $K_J(\bar{x},s)$ or Bernardo's $K_B(\bar{x})$. And for all purposes that are useful in real problems, Laplace's P_L may in turn be replaced by the λ' and $\delta\lambda$ of a pure estimation problem. Because of this, I suggest that the distinction between significance testing and estimation is artificial and of doubtful value in statistics-indeed, negative value if it leads to needless duplication of effort in the belief that one is solving two different problems.