VOL. 70, NO. 3 MAY 1963

PSYCHOLOGICAL REVIEW

BAYESIAN STATISTICAL INFERENCE FOR PSYCHOLOGICAL RESEARCH [1]

WARD EDWARDS, HAROLD LINDMAN, AND LEONARD J. SAVAGE

University of Michigan

Bayesian statistics, a currently controversial viewpoint concerning statistical inference, is based on a definition of probability as a particular measure of the opinions of ideally consistent people. Statistical inference is modification of these opinions in the light of evidence, and Bayes' theorem specifies how such modifications should be made. The tools of Bayesian statistics include the theory of specific distributions and the principle of stable estimation, which specifies when actual prior opinions may be satisfactorily approximated by a uniform distribution. A common feature of many classical significance tests is that a sharp null hypothesis is compared with a diffuse alternative hypothesis. Often evidence which, for a Bayesian statistician, strikingly supports the null hypothesis leads to rejection of that hypothesis by standard classical procedures. The likelihood principle emphasized in Bayesian statistics implies, among other things, that the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.

The main purpose of this paper is to introduce psychologists to the Bayesian outlook in statistics, a new fabric with some very old threads. Although this purpose demands much repetition of ideas published elsewhere, even Bayesian specialists will find some remarks and derivations hitherto unpublished and perhaps quite new. The empirical scientist more interested in the ideas and implications of Bayesian statistics than in the mathematical details can safely skip almost all the equations; detours and parallel verbal explanations are provided. The textbook that would make all the Bayesian procedures mentioned in this paper readily available to experimenting psychologists does not yet exist, and perhaps it cannot exist soon; Bayesian statistics as a coherent body of thought is still too new and incomplete.

Bayes' theorem is a simple and fundamental fact about probability

[1] Work on this paper was supported in part by the United States Air Force under Contract AF 49(638)-769 and Grant AF-AFOSR-62-182, monitored by the Air Force Office of Scientific Research of the Air Force Office of Aerospace Research (the paper carries Document No. AFOSR-2009); in part under Contract AF 19(604)-7393, monitored by the Operational Applications Laboratory, Deputy for Technology, Electronic Systems Division, Air Force Systems Command; and in part by the Office of Naval Research under Contract Nonr 1224(41). We thank H. C. A. Dale, H. V. Roberts, R. Schlaifer, and E. H. Shuford for their comments on earlier versions.

that seems to have been clear to Thomas Bayes when he wrote his famous article published in 1763 (recently reprinted), though he did not state it there explicitly. Bayesian statistics is so named for the rather inadequate reason that it has many more occasions to apply Bayes' theorem than classical statistics has. Thus, from a very broad point of view, Bayesian statistics dates back at least to 1763.

From a stricter point of view, Bayesian statistics might properly be said to have begun in 1959 with the publication of *Probability and Statistics for Business Decisions*, by Robert Schlaifer. This introductory text presented for the first time practical implementation of the key ideas of Bayesian statistics: that probability is orderly opinion, and that inference from data is nothing other than the revision of such opinion in the light of relevant new information. Schlaifer (1961) has since published another introductory text, less strongly slanted toward business applications than his first. And Raiffa and Schlaifer (1961) have published a relatively mathematical book. Some other works in current Bayesian statistics are by Anscombe (1961), de Finetti (1959), de Finetti and Savage (1962), Grayson (1960), Lindley (1961), Pratt (1961), and Savage et al. (1962).

The philosophical and mathematical basis of Bayesian statistics has, in addition to its ancient roots, a considerable modern history. Two lines of development important for it are the ideas of statistical decision theory, based on the game-theoretic work of Borel (1921), von Neumann (1928), and von Neumann and Morgenstern (1947), and the statistical work of Neyman (1937, 1938b, for example), Wald (1942, 1955, for example), and others; and the personalistic definition of probability, which Ramsey (1931) and de Finetti (1930, 1937) crystallized. Other pioneers of personal probability are Borel (1924), Good (1950, 1960), and Koopman (1940a, 1940b, 1941). Decision theory and personal probability fused in the work of Ramsey (1931), before either was very mature. By 1954, there was great progress in both lines for Savage's *The Foundations of Statistics* to draw on. Though this book failed in its announced object of satisfying popular non-Bayesian statistics in terms of personal probability and utility, it seems to have been of some service toward the development of Bayesian statistics. Jeffreys (1931, 1939) has pioneered extensively in applications of Bayes' theorem to statistical problems. He is one of the founders of Bayesian statistics, though he might reject identification with the viewpoint of this paper because of its espousal of personal probabilities. These two, inevitably inadequate, paragraphs are our main attempt in this paper to give credit where it is due. Important authors have not been listed, and for those that have been, we have given mainly one early and one late reference only. Much more information and extensive bibliographies will be found in Savage et al. (1962) and Savage (1954, 1962a).

We shall, where appropriate, compare the Bayesian approach with a loosely defined set of ideas here labeled the classical approach, or classical statistics. You cannot but be familiar with many of these ideas, for what you learned about statistical inference in your elementary statistics course was some blend of them. They have been directed largely toward the topics of testing hypotheses and interval estimation, and they fall

roughly into two somewhat conflicting doctrines associated with the names of R. A. Fisher (1925, 1956) for one, and Jerzy Neyman (e.g. 1937, 1938b) and Egon Pearson for the other. We do not try to portray any particular version of the classical approach; our real comparison is between such procedures as a Bayesian would employ in an article submitted to the *Journal of Experimental Psychology*, say, and those now typically found in that journal. The fathers of the classical approach might not fully approve of either. Similarly, though we adopt for conciseness an idiom that purports to define *the* Bayesian position, there must be at least as many Bayesian positions as there are Bayesians. Still, as philosophies go, the unanimity among Bayesians reared apart is remarkable and an encouraging symptom of the cogency of their ideas.

In some respects Bayesian statistics is a reversion to the statistical spirit of the eighteenth and nineteenth centuries; in others, no less essential, it is an outgrowth of that modern movement here called classical. The latter, in coping with the consequences of its view about the foundations of probability which made useless, if not meaningless, the probability that a hypothesis is true, sought and found techniques for statistical inference which did not attach probabilities to hypotheses. These intended channels of escape have now, Bayesians believe, led to reinstatement of the probabilities of hypotheses and a return of statistical inference to its original line of development. In this return, mathematics, formulations, problems, and such vital tools as distribution theory and tables of functions are borrowed from extrastatistical probability theory and from classical statistics itself. All the elements of Bayesian statistics, except perhaps the personalistic view of probability, were invented and developed within, or before, the classical approach to statistics; only their combination into specific techniques for statistical inference is at all new.

The Bayesian approach is a common sense approach. It is simply a set of techniques for orderly expression and revision of your opinions with due regard for internal consistency among their various aspects and for the data. Naturally, then, much that Bayesians say about inference from data has been said before by experienced, intuitive, sophisticated empirical scientists and statisticians. In fact, when a Bayesian procedure violates your intuition, reflection is likely to show the procedure to have been incorrectly applied. If classically trained intuitions do have some conflicts, these often prove transient.

ELEMENTS OF BAYESIAN STATISTICS

Two basic ideas which come together in Bayesian statistics, as we have said, are the decision-theoretic formulation of statistical inference and the notion of personal probability.

Statistics and decisions. Prior to a paper by Neyman (1938a), classical statistical inference was usually expressed in terms of justifying propositions on the basis of data. Typical propositions were: Point estimates; the best guess for the unknown number μ is m. Interval estimates; μ is between m_1 and m_2. Rejection of hypotheses; μ is not 0. Neyman's (1938a, 1957) slogan "inductive behavior" emphasized the importance of action, as opposed to assertion, in the face of uncertainty. The decision-theoretic, or economic, view of statistics was advanced with particular vigor by Wald (1942). To illustrate, in the decision-theoretic outlook a

point estimate is a decision to act, in some specific context, as though μ were m, not to assert something about μ. Some classical statisticians, notably Fisher (1956, Ch. 4), have hotly rejected the decision-theoretic outlook.

While Bayesian statistics owes much to the decision-theoretic outlook, and while we personally are inclined to side with it, the issue is not crucial to a Bayesian. No one will deny that economic problems of behavior in the face of uncertainty concern statistics, even in its most "pure" contexts. For example, "Would it be wise, in the light of what has just been observed, to attempt such and such a year's investigation?" The controversial issue is only whether such economic problems are a good paradigm of all statistical problems. For Bayesians, all uncertainties are measured by probabilities, and these probabilities (along with the here less emphasized concept of utilities) are the key to all problems of economic uncertainty. Such a view deprives debate about whether all problems of uncertainty are economic of urgency. On the other hand, economic definitions of personal probability seem, at least to us, invaluable for communication and perhaps indispensable for operational definition of the concept.

A Bayesian can reflect on his current opinion (and how he should revise it on the basis of data) without any reference to the actual economic significance, if any, that his opinion may have. This paper ignores economic considerations, important though they are even for pure science, except for brief digressions. So doing may combat the misapprehension that Bayesian statistics is primarily for business, not science.

Personal probability. With rare exceptions, statisticians who conceive of probabilities exclusively as limits of relative frequencies are agreed that uncertainty about matters of fact is ordinarily not measurable by probability. Some of them would brand as nonsense the probability that weightlessness decreases visual acuity; for others the probability of this hypothesis would be 1 or 0 according as it is in fact true or false. Classical statistics is characterized by efforts to reformulate inference about such hypotheses without reference to their probabilities, especially initial probabilities.

These efforts have been many and ingenious. It is disagreement about which of them to espouse, incidentally, that distinguishes the two main classical schools of statistics. The related ideas of significance levels, "errors of the first kind," and confidence levels, and the conflicting idea of fiducial probabilities are all intended to satisfy the urge to know how sure you are after looking at the data, while outlawing the question of how sure you were before. In our opinion, the quest for inference without initial probabilities has failed, inevitably.

You may be asking, "If a probability is not a relative frequency or a hypothetical limiting relative frequency, what is it? If, when I evaluate the probability of getting heads when flipping a certain coin as .5, I do not mean that if the coin were flipped very often the relative frequency of heads to total flips would be arbitrarily close to .5, then what do I mean?"

We think you mean something about yourself as well as about the coin. Would you not say, "Heads on the next flip has probability .5" if and only if you would as soon guess heads as not, even if there were some important reward for being right? If so,

your sense of "probability" is ours; even if you would not, you begin to see from this example what we mean by "probability," or "personal probability." To see how far this notion is from relative frequencies, imagine being reliably informed that the coin has either two heads or two tails. You may still find that if you had to guess the outcome of the next flip for a large prize you would not lift a finger to shift your guess from heads to tails or vice versa.

Probabilities other than .5 are defined in a similar spirit by one of several mutually harmonious devices (Savage, 1954, Ch. 1–4). One that is particularly vivid and practical, if not quite rigorous as stated here, is this. For you, now, the probability $P(A)$ of an event A is the price you would just be willing to pay in exchange for a dollar to be paid to you in case A is true. Thus, rain tomorrow has probability 1/3 for you if you would pay just \$.33 now in exchange for \$1.00 payable to you in the event of rain tomorrow.

A system of personal probabilities, or prices for contingent benefits, is inconsistent if a person who acts in accordance with it can be trapped into accepting a combination of bets that assures him of a loss no matter what happens. Necessary and sufficient conditions for consistency are the following, which are familiar as a basis for the whole mathematical theory of probability:

$$0 \leqq P(A) \leqq P(S) = 1,$$
$$P(A \bigcup B) = P(A) + P(B),$$

where S is the tautological, or universal, event; A and B are any two incompatible, or nonintersecting, events; and $A \bigcup B$ is the event that either A or B is true, or the union of A and B. Real people often make choices that reflect violations of these rules, especially the second, which is why personalists emphasize that personal probability is orderly, or consistent, opinion, rather than just any opinion. One of us has presented elsewhere a model for probabilities inferred from real choices that does not include the second consistency requirement listed above (Edwards, 1962b). It is important to keep clear the distinction between the somewhat idealized consistent personal probabilities that are the subject of this paper and the usually inconsistent subjective probabilities that can be inferred from real human choices among bets, and the words "personal" and "subjective" here help do so.

Your opinions about a coin can of course differ from your neighbor's. For one thing, you and he may have different bodies of relevant information. We doubt that this is the only *legitimate source of difference of* opinion. Hence the personal in personal probability. Any probability should in principle be indexed with the name of the person, or people, whose opinion it describes. We usually leave the indexing unexpressed but underline it from time to time with phrases like "the probability for you that H is true."

Although your initial opinion about future behavior of a coin may differ radically from your neighbor's, your opinion and his will ordinarily be so transformed by application of Bayes' theorem to the results of a long sequence of experimental flips as to become nearly indistinguishable. This approximate merging of initially divergent opinions is, we think, one reason why empirical research is called "objective." Personal probability is sometimes dismissed with the assertion that scientific knowledge cannot be mere opinion. Yet, obviously, no sharp lines separate the

conjecture that many human cancers may be caused by viruses, the opinion that many are caused by smoking, and the "knowledge" that many have been caused by radiation.

Conditional probabilities and Bayes' theorem. In the spirit of the rough definition of the probability $P(A)$ of an event A given above, the conditional probability $P(D|H)$ of an event D given another H is the amount you would be willing to pay in exchange for a dollar to be paid to you in case D is true, with the further provision that all transactions are canceled unless H is true. As is not hard to see, $P(D \cap H)$ is $P(D|H)P(H)$ where $D \cap H$ is the event that D and H are both true, or the intersection of D and H. Therefore,

$$P(D|H) = \frac{P(D \cap H)}{P(H)}, \quad [1]$$

unless $P(H) = 0$.

Conditional probabilities are the probabilistic expression of learning from experience. It can be argued that the probability of D for you—the consistent you—after learning that H is in fact true is $P(D|H)$. Thus, after you learn that H is true, the new system of numbers $P(D|H)$ for a specific H comes to play the role that was played by the old system $P(D)$ before.

Although the events D and H are arbitrary, the initial letters of Data and Hypothesis are suggestive names for them. Of the three probabilities in Equation 1, $P(H)$ might be illustrated by the sentence: "The probability for you, now, that Russia will use a booster rocket bigger than our planned Saturn booster within the next year is .8." The probability $P(D \cap H)$ is the probability of the joint occurrence of two events regarded as one event, for instance: "The probability for you, now, that the next manned space capsule to enter space will contain three men and also that Russia will use a booster rocket bigger than our planned Saturn booster within the next year is .2." According to Equation 1, the probability for you, now, that the next manned space capsule to enter space will contain three men, given that Russia will use a booster rocket bigger than our planned Saturn booster within the next year is .2/.8 = .25.

A little algebra now leads to a basic form of Bayes' theorem:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}, \quad [2]$$

provided $P(D)$ and $P(H)$ are not 0. In fact, if the roles of D and H in Equation 1 are interchanged, the old form of Equation 1 and the new form can be expressed symmetrically, thus:

$$\frac{P(D|H)}{P(D)} = \frac{P(D \cap H)}{P(D)P(H)} = \frac{P(H|D)}{P(H)}, \quad [3]$$

which obviously implies Equation 2. A suggestive interpretation of Equation 3 is that the relevance of H to D equals the relevance of D to H.

Reformulations of Bayes' theorem apply to continuous parameters or data. In particular, if a parameter (or set of parameters) λ has a prior probability density function $u(\lambda)$, and if x is a random variable (or a set of random variables such as a set of measurements) for which $v(x|\lambda)$ is the density of x given λ and $v(x)$ is the density of x, then the posterior probability density of λ given x is

$$u(\lambda|x) = \frac{v(x|\lambda)u(\lambda)}{v(x)}. \quad [4]$$

There are of course still other possibilities such as forms of Bayes'

theorem in which λ but not x, or x but not λ, is continuous. A complete and compact generalization is available and technically necessary but need not be presented here.

In Equation 2, D may be a particular observation or a set of data regarded as a datum and H some hypothesis, or putative fact. Then Equation 2 prescribes the consistent revision of your opinions about the probability of H in the light of the datum D—similarly for Equation 4.

In typical applications of Bayes' theorem, each of the four probabilities in Equation 2 performs a different function, as will soon be explained. Yet they are very symmetrically related to each other, as Equation 3 brings out, and are all the same kind of animal. In particular, all probabilities are really conditional. Thus, $P(H)$ is the probability of the hypothesis H for you conditional on all you know, or knew, about H prior to learning D; and $P(H|D)$ is the probability of H conditional on that same background knowledge together with D.

Again, the four probabilities in Equation 2 are personal probabilities. This does not of course exclude any of them from also being frequencies, ratios of favorable to total possibilities, or numbers arrived at by any other calculation that helps you form your personal opinions. But some are, so to speak, more personal than others. In many applications, practically all concerned find themselves in substantial agreement with respect to $P(D|H)$; or $P(D|H)$ is public, as we say. This happens when $P(D|H)$ flows from some simple model that the scientists, or others, concerned accept as an approximate description of their opinion about the situation in which the datum was obtained. A traditional example of such a statistical model is that of drawing a ball from an urn known to contain some balls, each either black or white. If a series of balls is drawn from the urn, and after each draw the ball is replaced and the urn thoroughly shaken, most men will agree at least tentatively that the probability of drawing a particular sequence D (such as black, white, black, black) given the hypothesis that there are B black and W white balls in the urn is

$$\left(\frac{B}{B+W}\right)^b\left(\frac{W}{B+W}\right)^w,$$

where b is the number of black, and w the number of white, balls in the sequence D.

Even the best models have an element of approximation. For example, the probability of drawing any sequence D of black and white balls from an urn of composition H depends, in this model, only on the number of black balls and white ones in D, not on the order in which they appeared. This may express your opinion in a specific situation very well, but not well enough to be retained if D should happen to consist of 50 black balls followed by 50 white ones. Idiomatically, such a datum convinces you that this particular model is a wrong description of the world. Philosophically, however, the model was not a description of the world but of your opinions, and to know that it was not quite correct, you had at most to reflect on this datum, not necessarily to observe it. In many scientific contexts, the public model behind $P(D|H)$ may include the notions of random sampling from a well-defined population, as in this example. But precise definition of the population may be difficult or impossible, and a sample whose randomness would thoroughly satisfy you, let alone your

neighbor in science, can be hard to draw.

In some cases $P(D|H)$ does not command general agreement at all. What is the probability of the actual seasonal color changes on Mars if there is life there? What is this probability if there is no life there? Much discussion of life on Mars has not removed these questions from debate.

Public models, then, are never perfect and often are not available. Nevertheless, those applications of inductive inference, or probabilistic reasoning, that are called statistical seem to be characterized by tentative public agreement on some model and provisional work within it. Rough characterization of statistics by the relative publicness of its models is not necessarily in conflict with attempts to characterize it as the study of numerous repetitions (Bartlett, in Savage et al., 1962, pp. 36–38). This characterization is intended to distinguish statistical applications of Bayes' theorem from many other applications to scientific, economic, military, and other contexts. In some of these nonstatistical contexts, it is appropriate to substitute the judgment of experts for a public model as the source of $P(D|H)$ (see for example Edwards, 1962a, 1963).

The other probabilities in Equation 2 are often not at all public. Reasonable men may differ about them, even if they share a statistical model that specifies $P(D|H)$. People do, however, often differ much more about $P(H)$ and $P(D)$ than about $P(H|D)$, for evidence can bring initially divergent opinions into near agreement.

The probability $P(D)$ is usually of little direct interest, and intuition is often silent about it. It is typically calculated, or eliminated, as follows. When there is a statistical model, H is usually regarded as one of a list, or partition, of mutually exclusive and exhaustive hypotheses H_i such that the $P(D|H_i)$ are all equally public, or part of the statistical model. Since $\Sigma_i P(H_i|D)$ must be 1, Equation 2 implies that

$$P(D) = \Sigma_i P(D|H_i)P(H_i).$$

The choice of the partition H_i is of practical importance but largely arbitrary. For example, tomorrow will be "fair" or "foul," but these two hypotheses can themselves be subdivided and resubdivided. Equation 2 is of course true for all partitions but is more useful for some than for others. As a science advances, partitions originally not even dreamt of become the important ones (Sinclair, 1960). In principle, room should always be left for "some other" explanation. Since $P(D|H)$ can hardly be public when H is "some other explanation," the catchall hypothesis is usually handled in part by studying the situation conditionally on denial of the catchall and in part by informal appraisal of whether any of the explicit hypotheses fit the facts well enough to maintain this denial. Good illustrations are Urey (1962) and Bridgman (1960).

In statistical practice, the partition is ordinarily continuous, which means roughly that H_i is replaced by a parameter λ (which may have more than one dimension) with an initial probability density $u(\lambda)$. In this case,

$$P(D) = \int P(D|\lambda)u(\lambda)d\lambda.$$

Similarly, $P(D)$, $P(D|H_i)$, and $P(D|\lambda)$ are replaced by probability densities in D if D is (absolutely) continuously distributed.

$P(H|D)$ or $u(\lambda|D)$, the usual output of a Bayesian calculation, seems

to be exactly the kind of information that we all want as a guide to thought and action in the light of an observational process. It is the probability for you that the hypothesis in question is true, on the basis of all your information, including, but not restricted to, the observation D.

PRINCIPLE OF STABLE ESTIMATION

Problem of prior probabilities. Since $P(D|H)$ is often reasonably public and $P(H|D)$ is usually just what the scientist wants, the reason classical statisticians do not base their procedures on Equations 2 and 4 must, and does, lie in $P(H)$, the prior probability of the hypothesis. We have already discussed the most frequent objection to attaching a probability to a hypothesis and have shown briefly how the definition of personal probability answers that objection. We must now examine the practical problem of determining $P(H)$. Without $P(H)$, Equations 2 and 4 cannot yield $P(H|D)$. But since $P(H)$ is a personal probability, is it not likely to be both vague and variable, and subjective to boot, and therefore useless for public scientific purposes?

Yes, prior probabilities often are quite vague and variable, but they are not necessarily useless on that account (Borel, 1924). The impact of actual vagueness and variability of prior probabilities differs greatly from one problem to another. They frequently have but negligible effect on the conclusions obtained from Bayes' theorem, although utterly unlimited vagueness and variability would have utterly unlimited effect. If observations are precise, in a certain sense, relative to the prior distribution on which they bear, then the form and properties of the prior distribution have negligible influence on the posterior distribution. From a practical point of view, then, the untrammeled subjectivity of opinion about a parameter ceases to apply as soon as much data become available. More generally, two people with widely divergent prior opinions but reasonably open minds will be forced into arbitrarily close agreement about future observations by a sufficient amount of data. An advanced mathematical expression of this phenomenon is in Blackwell and Dubins (1962).

When prior distributions can be regarded as essentially uniform. Frequently, the data so completely control your posterior opinion that there is no practical need to attend to the details of your prior opinion. For example, consider taking your temperature.

Headachy and hot, you are convinced that you have a fever but are not sure how much. You do not hold the interval 100.5°–101° even 20 times more probable than the interval 101°–101.5° on the basis of your malaise alone. But now you take your temperature with a thermometer that you strongly believe to be accurate and find yourself willing to give much more than 20 to 1 odds in favor of the half-degree centered at the thermometer reading.

Your prior opinion is rather irrelevant to this useful conclusion but of course not utterly irrelevant. For readings of 85° or 110°, you would revise your statistical model according to which the thermometer is accurate and correctly used, rather than proclaim a medical miracle. A reading of 104° would be puzzling—too inconsistent with your prior opinion to seem reasonable and yet not obviously absurd. You might try again, perhaps with another thermometer.

It has long been known that, under suitable circumstances, your actual

posterior distribution will be approximately what it would have been had your prior distribution been uniform, that is, described by a constant density. As the fever example suggests, prior distributions need not be, and never really are, completely uniform. To ignore the departures from uniformity, it suffices that your actual prior density change gently in the region favored by the data and not itself too strongly favor some other region.

But what is meant by "gently," by "region favored by the data," by "region favored by the prior distribution," and by two distributions being approximately the same? Such questions do not have ultimate answers, but this section explores one useful set of possibilities. The mathematics and ideas have been current since Laplace, but we do not know any reference that would quite substitute for the following mathematical paragraphs; Jeffreys (1939, see Section 3.4 of the 1961 edition) and Lindley (1961) are pertinent. Those who would skip or skim the mathematics will find the trail again immediately following Implication 7, where the applications of stable estimation are informally summarized.

Under some circumstances, the posterior probability density

$$u(\lambda|x) = \frac{v(x|\lambda)u(\lambda)}{\int v(x|\lambda')u(\lambda')d\lambda'} \quad [5]$$

can be well approximated in some senses by the probability density

$$w(\lambda|x) = \frac{v(x|\lambda)}{\int v(x|\lambda')d\lambda'}, \quad [6]$$

where λ is a parameter or set of parameters, λ' is a corresponding variable of integration, x is an observation or set of observations, $v(x|\lambda)$ is the probability (or perhaps probability density) of x given λ, $u(\lambda)$ is the prior probability density of λ, and the integrals are over the entire range of meaningful values of λ. By their nature, u, v, and w are nonnegative, and unless the integral in Equation 6 is finite, there is no hope that the approximation will be valid, so these conditions are adopted for the following discussion.

Consider a region of values of λ, say B, which is so small that $u(\lambda)$ varies but little within B and yet so large that B promises to contain much of the posterior probability of λ given the value of x fixed throughout the present discussion. Let α, β, γ, and φ be positive numbers, of which the first three should in practice be small, and are formally taken to be less than 1. In these terms, three assumptions will be made that define one set of circumstances under which $w(\lambda|x)$ does approximate $u(\lambda|x)$ in certain senses, for the given x.

Assumption 1:

$$\int_{\bar{B}} w(\lambda|x)d\lambda \leqq \alpha \int_B w(\lambda|x)d\lambda,$$

where $\bar{B}$ means, as usual, the complement of B. (That is, B is highly favored by the data; α might be 10^{-4} or less in everyday applications.)

Assumption 2: For all $\lambda \epsilon B$,

$$\varphi \leqq u(\lambda) \leqq (1 + \beta)\varphi.$$

(That is, the prior density changes very little within B; .01 or even .05 would be good everyday values for β. The value of φ is unimportant and is not likely to be accurately known.)

Assumption 3:

$$\int_{\bar{B}} u(\lambda|x)d\lambda \leqq \gamma \int_B u(\lambda|x)d\lambda.$$

(That is, B is also highly favored by the posterior distribution; in applications, γ should be small, yet a γ as large as 100α, or even $1{,}000\alpha$, may have to be tolerated.)

Assumption 3 looks, at first, hard to verify without much knowledge of $u(\lambda)$. Consider an alternative:

Assumption 3′: $u(\lambda) \leqq \theta\varphi$ for all λ,

where θ is a positive constant. (That is, u is nowhere astronomically big compared to its nearly constant values in B; a θ as large as 100 or 1,000 will often be tolerable.)

Assumption 3′ in the presence of Assumptions 1 and 2 can imply 3, as is seen thus.

$$\int_{\bar{B}} u(\lambda|x)d\lambda \Big/ \int_B u(\lambda|x)d\lambda$$

$$= \int_{\bar{B}} v(x|\lambda)u(\lambda)d\lambda \Big/ \int_B v(x|\lambda)u(\lambda)d\lambda$$

$$\leqq \theta\varphi \int_{\bar{B}} v(x|\lambda)d\lambda \Big/ \varphi \int_B v(x|\lambda)d\lambda$$

$$\leqq \theta\alpha.$$

So if $\gamma \geqq \theta\alpha$, Assumption 3′ implies Assumption 3.

Seven implications of Assumptions 1, 2, and 3 are now derived. The first three may be viewed mainly as steps toward the later ones. The expressions in the large brackets serve only to help prove the numbered assertions.

Implication 1: $\int v(x|\lambda)u(\lambda)d\lambda$

$$\left[\geqq \int_B v(x|\lambda)u(\lambda)d\lambda \geqq \varphi \int_B v(x|\lambda)d\lambda\right]$$

$$\geqq \frac{\varphi}{1+\alpha} \int v(x|\lambda)d\lambda.$$

Implication 2: $\int v(x|\lambda)u(\lambda)d\lambda$

$$\left[= \int_B v(x|\lambda)u(\lambda)d\lambda + \int_{\bar{B}} v(x|\lambda)u(\lambda)d\lambda\right.$$

$$\left.\leqq (1+\gamma) \int_B v(x|\lambda)u(\lambda)d\lambda\right]$$

$$\leqq (1+\gamma)(1+\beta)\varphi \int v(x|\lambda)d\lambda.$$

With two new positive constants δ and ϵ defined by the context, the next implication follows easily.

Implication 3: $(1-\delta) = \dfrac{1}{(1+\beta)(1+\gamma)}$

$$\leqq \frac{u(\lambda|x)}{w(\lambda|x)} \leqq (1+\beta)(1+\alpha) = (1+\epsilon)$$

for all λ in B, except where numerator and denominator of $u(\lambda|x)/w(\lambda|x)$ both vanish. (Note that if α, β, and γ are small, so are δ and ϵ.)

Let $u(C|x)$ and $w(C|x)$ denote $\int_C u(\lambda|x)d\lambda$ and $\int_C w(\lambda|x)d\lambda$, that is, the probabilities of C under the densities $u(\lambda|x)$ and $w(\lambda|x)$.

Implication 4: $u(B|x) \geqq 1-\gamma$, and for every subset C of B,

$$1-\delta \leqq \frac{u(C|x)}{w(C|x)} \leqq 1+\epsilon.$$

Implication 5: If t is a function of λ such that $|t(\lambda)| \leqq T$ for all λ, then

$$\left|\int t(\lambda)u(\lambda|x)d\lambda - \int t(\lambda)w(\lambda|x)d\lambda\right|$$

$$\left[\leqq \int_B |t(\lambda)|\,|u(\lambda|x) - w(\lambda|x)|d\lambda\right.$$

$$+ \int_{\bar{B}} |t(\lambda)|u(\lambda|x)d\lambda + \int_{\bar{B}} |t(\lambda)|w(\lambda|x)d\lambda$$

$$\left.\leqq T\int_B \left|\frac{u(\lambda|x)}{w(\lambda|x)} - 1\right| w(\lambda|x)d\lambda + T(\gamma+\alpha)\right]$$

$$\leqq T[\max(\delta,\epsilon) + \gamma + \alpha].$$

Implication 6: $|u(C|x) - w(C|x)|$
$\leqq \max(\delta,\epsilon) + \gamma + \alpha$ for all C.

It is sometimes important to evaluate $u(C|x)$ with fairly good percentage accuracy when $u(C|x)$ is small but not nearly so small as α or γ, thus.

Implication 7: $(1-\delta)\left(1 - \dfrac{\alpha}{w(C|x)}\right)$

$$\left[\leqq (1-\delta)\frac{w(C\cap B|x)}{w(C|x)} \leqq \frac{u(C\cap B|x)}{w(C|x)}\right]$$

$$\leqq \frac{u(C|x)}{w(C|x)}\left[\leqq \frac{u(C\cap B|x)+\gamma}{w(C|x)}\right.$$

$$\left.\leqq (1+\epsilon)\frac{w(C\cap B|x)}{w(C|x)} + \frac{\gamma}{w(C|x)}\right]$$

$$\leqq (1+\epsilon) + \frac{\gamma}{w(C|x)}.$$

What does all this epsilontics mean for practical statistical work? The overall goal is valid justification for proceeding as though your prior distribution were uniform. A set of three assumptions implying this justification was pointed out: First, some region B is highly favored by the data. Second, within B the prior density changes very little. Third, most of the posterior density is concentrated inside B. According to a more stringent but more easily verified substitute for the third assumption, the prior density nowhere enormously exceeds its general value in B.

Given the three assumptions, what follows? One way of looking at the implications is to observe that nowhere within B, which has high posterior probability, is the ratio of the approximate posterior density to the actual posterior density much different from 1 and that what happens outside B is not important for some purposes. Again, if the posterior expectation, or average, of some bounded function is of interest, then the difference between the expectation under the actual posterior distribution and under the approximating distribution will be small relative to the absolute bound of the function. Finally, the actual posterior probability and the approximate probability of any set of parameter values are nearly equal. In short, the approximation is a good one in several important respects—given the three assumptions. Still other respects must sometimes be invoked and these may require further assumptions. See, for example, Lindley (1961).

Even when Assumption 2 is not applicable, a transformation of the parameters of the prior distribution sometimes makes it so. If, for example, your prior distribution roughly obeys Weber's law, so that you tend to assign about as much probability to the region from λ to 2λ as to the region from 10λ to 20λ, a logarithmic transformation of λ may well make Assumption 2 applicable for a considerably smaller β than otherwise.

We must forestall a dangerous confusion. In the temperature example as in many others, the measurement x is being used to estimate the value of some parameter λ. In such cases, λ and x are measured in the same units (degrees Fahrenheit in the example) and interesting values of λ are often numerically close to observed values of x. It is therefore imperative to maintain the conceptual distinction between λ and x. When the principle of stable estimation applies, the normalized function $v(x|\lambda)$ as a function of λ, not of x, approximates your posterior distribution. The point is perhaps most obvious in an example such as estimating the area of a circle by measuring its radius. In this case, λ is in square inches, x is in inches, and there is no temptation to think that the form of the distribution of x's is the same as the form of the posterior distribution of λ's. But the same point applies in all cases. The function $v(x|\lambda)$ is a function of both x and λ; only by coincidence will the form or the parameters of $v(x|\lambda)$ considered as a function of λ be the same as its form or parameters considered as a function of x. One such coincidence occurs so often that it tends to mislead intuition. When your statistical model leads you to expect that a set of observations will be normally distributed, then the posterior distribution of the mean of the quantity being observed will, if stable estimation applies, be normal with the mean equal to the mean of the observations. (Of course it will have a smaller standard deviation than the standard deviation of the observations.)

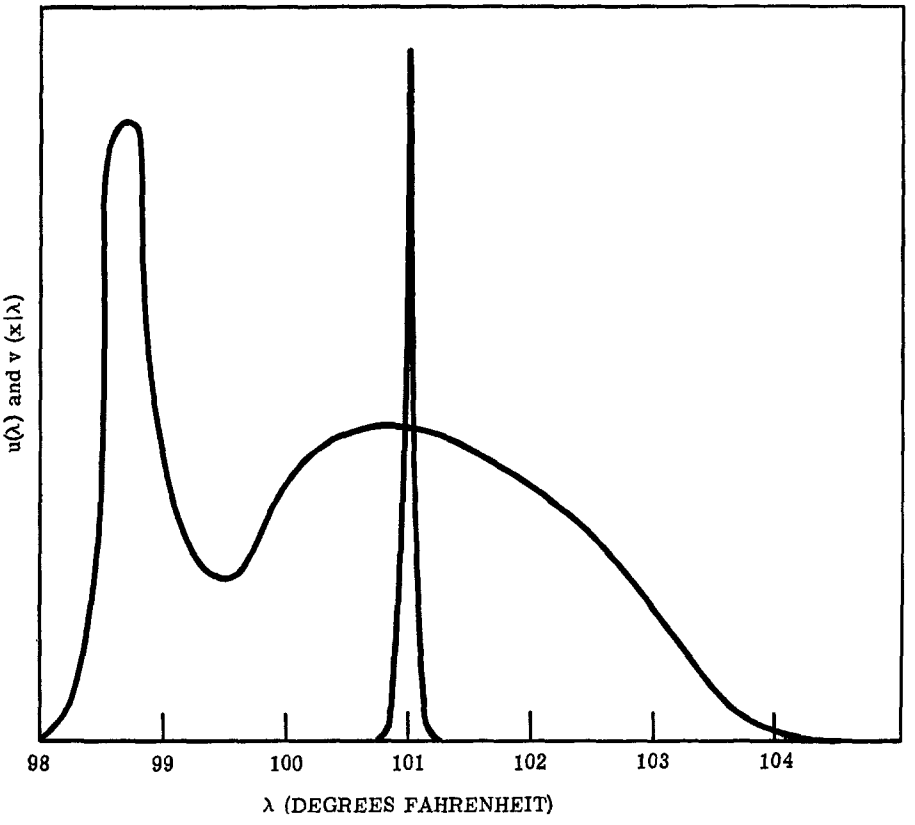

FIG. 1. $u(\lambda)$ and $v(x|\lambda)$ for the fever thermometer example. (Note that the units on the y axis are different for the two functions.)

Numerically, what can the principle of stable estimation do for the fever-thermometer example? Figure 1 is a reasonably plausible numerical picture of the situation. Your prior distribution in your role as invalid has a little bump around 98.6°, because on other occasions you have taken your temperature when feeling out of sorts and found it depressingly normal. Still, you really think you have a fever, so most of your density is spread over the region 99.5°–104.5°. It gets rather low at the high end of that interval, since you doubt that you could have so much as a 104° fever without feeling even worse than you do.

The thermometer has a standard deviation of .05° and negligible systematic error—this is reasonable for a really good clinical thermometer, the systematic error of which should be small compared to the errors of procedure and reading. For convenience and because it is plausible as an approximation, we assume also that the thermometer distributes its errors normally. The indicated reading will, then, lie within a symmetric region .1° wide around the true temperature with probability a little less than .7. If the thermometer reading is 101.0°, we might take the region B to extend from 100.8° to 101.2°—four standard deviations on each side of the observation. According to tables of the normal distribution, α is then somewhat less than 10^{-4}.

The number φ should be thought of as the smallest value of $u(\lambda)$ within B, but its actual value cancels out of all important calculations and so is immaterial. For the same reason, it is also immaterial that the two functions $v(101.0|\lambda)$ and $u(\lambda)$ graphed in Figure 1 are not measured in the same units and therefore cannot meaningfully share the same vertical scale; in so drawing them, we sin against logic but not against the calculation of $u(\lambda|x)$ or $w(\lambda|x)$. Figure 1 suggests that β is at most .05, and we shall work with that value, but it is essential to give some serious

justification for this crucial assumption, as we shall later.

We justify Assumption 3 by way of Assumption 3′. The figure, drawn for qualitative suggestion rather than accuracy, makes a θ of 2 look reasonable, but since you may have a very strong suspicion that your temperature is nearly normal, we take $\theta = 100$ for safety. The real test is whether there is any hundredth, say, of a degree outside of B that you initially held to be more than 100 times as probable as the initially least probable hundredth in B. You will not find this question about yourself so hard, especially since little accuracy is required.

Actually, the technique based on θ could fail utterly without really spoiling the program. Suppose, for example, you really think it pretty unlikely that you have a fever and have unusually good knowledge of the temperature that is normal for you (at this hour). You may then have as much probability as .95 packed into some interval of .1° near normal, but in no such short interval in B are you likely to have more than one fiftieth of the residual probability. This leads to a θ of at least $.95/(.05 \times .02) = 950$. Fortunately, different, but somewhat analogous, calculations show that even very high concentrations of initial probability in a region very strongly discredited by the data do not interfere with the desired approximation. This alternative sort of calculation will be made clear by later examples about hypothesis testing.

Returning from the digression, continue with $\theta = 100$. The comment after Assumption 3′ leads to $\gamma = \theta\alpha = 10^{-4} \times 10^{2} = .01$.

Explore now some of the consequences of the theory of stable estimation for the example: $w(\lambda|101.0)$ is normal about 101° with a standard deviation of .05°. If the region B is taken to be the interval from 100.8° to 101.2°, then $\alpha = 10^{-4}$, $\beta = .05$, and $\gamma = .01$. Therefore, $\delta = 1 - [(1+\beta)(1+\gamma)]^{-1} < .06$, and $\epsilon = (1+\beta)(1+\alpha) - 1 < .051$. According to Implication 4, for any C in B, $u(C|101.0)$ differs by at most about 6% from the explicitly computable $w(C|101.0)$. For any C, whether in B or not, Implication 6 guarantees $|u(C|101.0) - w(C|101.0)| \leqq .068$. An especially interesting example for C is the outside of some interval that has, say, 95% probability under $w(\lambda|101.0)$ so that $w(C|101.0) = .05$. Will $u(C|101.0)$ be moderately close to 5%? Implications 4 and 6 do not say so, but Implication 7 says that $(.94)(.0499) = .0470 \leqq u(C|101.0) \leqq (1.050)(.05) + .01 = .0625$. This is not so crude for the sort of situation where such a $u(C|101.0)$ might be wanted. Even if $w(C|101.0)$ is only .01, we get considerable information about $u(C|101.0)$; $.0093 \leqq u(C|101.0) \leqq .021$. For $w(C|101.0) = .001$, $.000849 \leqq u(C|101.0) \leqq .011$. At this stage, the upper bound has become almost useless, and when $w(C|101.0)$ is as small as 10^{-4}, the lower bound is utterly useless.

Implication 5, and extensions of it are also applicable. If, for example, you record what the thermometer says, the mean error and the root-mean-squared error of the recorded value, averaged according to your own opinion, should be about 0° and about .05°, respectively, according to a slight extension of Implication 5.

To re-emphasize the central point, those details about your initial opinion that were not clear to you yourself, about which you might not agree with your neighbor, and that would have been complicated to keep track of anyway can be neglected after a fairly good measurement.

A vital matter that has been postponed is to adduce a reasonable value for β. Like θ, β is an expression of personal opinion. In any application, β must be large enough to be an expression of actual opinion or, in "public" applications, of "public" opinion. If your opinion were perfectly clear or if the public were of one mind, you could determine β by dividing the maximum of your $u(\lambda)$ in B by its minimum and subtracting 1; but the most important need for β arises just when clarity or agreement is lacking. For unity of discussion, permit us to focus on the problem imposed by lack of clarity.

One way to express the lack of clarity, or the vagueness, of an actual set of opinions about λ is to say that many somewhat different densities portray your opinion tolerably well. In assuming that .05 was a sufficiently large β for the fever example, we were assuming that you would reject as unrealistic any initial density $u(\lambda)$ whose maximum in the interval B from 100.8° to 101.2° exceeds its minimum in B by as much as 5%. But how can you know such a thing about yourself? Still more, how could you hope to guess it about another?

To begin with, you might consider pairs of very short intervals in B and ask how much more probable one is than the other, but this will fail in realistic problems. To see why it fails, ask yourself what odds Ω you would offer (initially) for the last hundredth of a degree in B against the first hundredth; that is, imagine contracting to pay $\$\Omega$ if λ is in the first hundredth of a degree of B, to receive

$1.00 if it is in the last hundredth, and to be quits otherwise. If, for instance, you are feeling less sick than 101°, then you will be clear that $u(\lambda)$ is decreasing throughout B, that Ω is less than 1, and that $1 - \Omega$ would be the smallest valid value for β. However, you are likely to be highly confused about Ω. Doubtless Ω is very little less than 1. Is .9999 much too large or .91 much too small? We find it hard to answer when the question is put thus, and so may you.

As an entering wedge, consider an interval much longer than B, say from 100° to 102°. Perhaps you find $u(\lambda)$ to decrease even throughout this interval and even to decrease moderately perceptibly between its two end points. The ratio $u(101)/u(102)$ while distinctly greater than 1 may be convincingly less than 1.2. If the proportion by which $u(\lambda)$ diminished in every hundredth of a degree from 100° to 102° were the same—more formally, if the logarithmic derivative of $u(\lambda)$ were constant between 100° and 102°—then $u(101.2)/u(100.8)$ would be at most $(1.2)^{.4/2} = (1.2)^{.2} = 1.037$. Of course the rate of decrease is not exactly constant, but it may seem sufficiently generous to round 1.037 up to 1.05, which results in the β of .05 used in this example. Had you taken your temperature 25 times (with random error but negligible systematic error), which would not be realistic in this example but would be in some other experimental settings, then the standard error of the measurements would have been .01, and B would have needed to be only .08° instead of .4° wide to take in eight standard deviations. Under those circumstances, β could hardly need to be greater than .01, that is, $(1.05)^{.08/.4} - 1$.

How good should the approximation be before you can feel comfortable about using it? That depends entirely on your purpose. There are purposes for which an approximation of a small probability which is sure to be within fivefold of the actual probability is adequate. For others, an error of 1% would be painful. Fortunately, if the approximation is unsatisfactory it will often be possible to improve it as much as seems necessary at the price of collecting additional data, an expedient which often justifies its cost in other ways too. In practice, the accuracy of the stable-estimation approximation will seldom be so carefully checked as in the fever example. As individual and collective experience builds up, many applications will properly be judged safe at a glance.

Far from always can your prior distribution be practically neglected. At least five situations in which detailed properties of the prior distribution are crucial occur to us:

1. If you assign exceedingly small prior probabilities to regions of λ for which $v(x|\lambda)$ is relatively large, you in effect express reluctance to believe in values of λ strongly pointed to by the data and thus violate Assumption 3, perhaps irreparably. Rare events do occur, though rarely, and should not be permitted to confound us utterly. Also, apparatus and plans can break down and produce data that "prove" preposterous things. Morals conflict in the fable of the Providence man who on a cloudy summer day went to the post office to return his absurdly low-reading new barometer to Abercrombie and Fitch. His house was flattened by a hurricane in his absence.

2. If you have strong prior reason to believe that λ lies in a region for which $v(x|\lambda)$ is very small, you may be unwilling to be persuaded by the evidence to the contrary, and so again may violate Assumption 3. In this situation, the prior distribution might consist primarily of a very sharp spike, whereas $v(x|\lambda)$, though very low in the region of the prior spike, may be comparatively gentle everywhere. In the previous paragraph, it was $v(x|\lambda)$ which had the sharp spike, and the prior distribution which was near zero in the region of that spike. Quite often it would be inappropriate to discard a good theory on the basis of a single opposing experiment. Hypothesis testing situations discussed later in this paper illustrate this phenomenon.

3. If your prior opinion is relatively diffuse, but so are your data, then Assumption 1 is seriously violated. For when your data really do not mean much compared to what you already know, then the exact content of the initial opinion cannot be neglected.

4. If observations are expensive and you have a decision to make, it may not pay to collect enough information for the principle of stable estimation to apply. In such situations you should collect just so much information that the expected value of the best course of action available in the light of the information at hand is greater than the expected value of any program that involves collecting more observations. If you have strong prior opinions about the parameter, the amount of new information available when you stop collecting more may well be far too meager to satisfy the principle. Often, it will not pay you to collect any new information at all.

5. It is sometimes necessary to make decisions about sizable research commitments such as sample size or experimental design while your knowledge is still vague. In this case, an extreme instance of the former one, the role of prior opinion is particularly conspicuous. As Raiffa and Schlaifer (1961) show, this is one of the most fruitful applications of Bayesian ideas.

Whenever you cannot neglect the details of your prior distribution, you have, in effect, no choice but to determine the relevant aspects of it as best you can and use them. Almost always, you will find your prior opinions quite vague, and you may be distressed that your scientific inference or decision has such a labile basis. Perhaps this distress, more than anything else, discouraged statisticians from using Bayesian ideas all along (Pearson, 1962). To paraphrase de Finetti (1959, p. 19), people noticing difficulties in applying Bayes' theorem remarked "We see that it is not secure to build on sand. Take away the sand, we shall build on the void." If it were meaningful utterly to ignore prior opinion, it might presumably sometimes be wise to do so; but reflection shows that any policy that pretends to ignore prior opinion will be acceptable only insofar as it is actually justified by prior opinion. Some policies recommended under the motif of neutrality, or using only the facts, may flagrantly violate even very confused prior opinions, and so be unacceptable. The method of stable estimation might casually be described as a procedure for ignoring prior opinion, since its approximate results are acceptable for a wide range of prior opinions. Actually, far from ignoring prior opinion, stable estimation exploits certain well-defined features of prior opinion and is acceptable only insofar as those features are really present.

A SMATTERING OF BAYESIAN DISTRIBUTION THEORY

The mathematical equipment required to turn statistical principles into practical procedures, for Bayesian as well as for traditional statistics, is distribution theory, that is, the theory of specific families of probability distributions. Bayesian distribution theory, concerned with the interrelation among the three main distributions of Bayes' theorem, is in some respects more complicated than classical distribution theory. But the familiar properties that distributions have in traditional statistics, and in the theory of probability in general, remain unchanged. To a professional statistician, the added complication requires little more than possibly a

shift to a more complicated notation. Chapters 7 through 13 of Raiffa and Schlaifer's (1961) book are an extensive discussion of distribution theory for Bayesian statistics.

As usual, a consumer need not understand in detail the distribution theory on which the methods are based; the manipulative mathematics are being done for him. Yet, like any other theory, distribution theory must be used with informed discretion. The consumer who delegates his thinking about the meaning of his data to any "powerful new tool" of course invites disaster. Cookbooks, though indispensable, cannot substitute for a thorough understanding of cooking; the inevitable appearance of cookbooks of Bayesian statistics must be contemplated with ambivalence.

Conjugate distributions. Suppose you take your temperature at a moment when your prior probability density $u(\lambda)$ is not diffuse with respect to $v(x|\lambda)$, so your posterior opinion $u(\lambda|x)$ is not adequately approximated by $w(\lambda|x)$. Determination and application of $u(\lambda|x)$ may then require laborious numerical integrations of arbitrary functions. One way to avoid such labor that is often useful and available is to use conjugate distributions. When a family of prior distributions is so related to all the conditional distributions which can arise in an experiment that the posterior distribution is necessarily in the same family as the prior distributions, the family of prior distributions is said to be conjugate to the experiment. By no means all experiments have nontrivial conjugate families, but a few ubiquitous kinds do. Examples: Beta priors are conjugate to observations of a Bernoulli process, normal priors are conjugate to observations of a normal process with known variance. Several other conjugate pairs are discussed by Raiffa and Schlaifer (1961).

Even when there is a conjugate family of prior distributions, your own prior distribution could fail to be in or even near that family. The distributions of such a family are, however, often versatile enough to accommodate the actual prior opinion, especially when it is a bit hazy. Furthermore, if stable estimation is nearly but not quite justifiable, a conjugate prior which approximates your true prior even roughly may be expected to combine with $v(x|\lambda)$ to produce a rather accurate posterior distribution.

Should the fit of members of the conjugate family to your true opinion be importantly unsatisfactory, realism may leave no alternative to something as tedious as approximating the continuous distribution by a discrete one with many steps, and applying Bayesian logic by brute force. Respect for your real opinion as opposed to some handy stereotype is essential. That is why our discussion of stable estimation, even in this expository paper, emphasized criteria for deciding when the details of a prior opinion really are negligible.

An example: Normal measurement with variance known. To give a minimal illustration of Bayesian distribution theory, and especially of conjugate families, we discuss briefly, and without the straightforward algebraic details, the normally distributed measurement of known variance. The Bayesian treatment of this problem has much in common with its classical counterpart. As is well known, it is a good approximation to many other problems in statistics. In particular, it is a good approximation to the case of 25 or more normally distributed observations of unknown variance,

with the observed standard error of the mean playing the role of the known standard deviation and the observed mean playing the role of the single observation. In the following discussion and throughout the remainder of the paper, we shall discuss the single observation x with known standard deviation σ, and shall leave it to you to make the appropriate translation into the set of $n \geqq 25$ observations with mean $\bar{x}(=x)$ and standard error of the mean $s/\sqrt{n}(=\sigma)$, whenever that translation aids your intuition or applies more directly to the problem you are thinking about. Much as in classical statistics, it is also possible to take uncertainty about σ explicitly into account by means of Student's t. See, for example, Chapter 11 of Raiffa and Schlaifer (1961).

Three functions enter into the problem of known variance: $u(\lambda)$, $v(x|\lambda)$, and $u(\lambda|x)$. The reciprocal of the variance appears so often in Bayesian calculations that it is convenient to denote $1/\sigma^2$ by h and call h the precision of the measurement. We are therefore dealing with a normal measurement with an unknown mean μ but known precision h. Suppose your prior distribution is also normal. It has a mean μ_0 and a precision h_0, both known by introspection. There is no necessary relationship between h_0 and h, the precision of the measurement, but in typical worthwhile applications h is substantially greater than h_0. After an observation has been made, you will have a normally distributed posterior opinion, now with mean μ_1 and precision h_1.

$$\mu_1 = \frac{\mu_0 h_0 + xh}{h_0 + h}$$

and

$$h_1 = h_0 + h.$$

The posterior mean is an average of the prior mean and the observation weighted by the precisions. The precision of the posterior mean is the sum of the prior and data precisions. The posterior distribution in this case is the same as would result from the principle of stable estimation if in addition to the datum x, with its precision h, there had been an additional measurement of value μ_0 and precision h_0.

If the prior precision h_0 is very small relative to h, the posterior mean will probably, and the precision will certainly, be nearly equal to the data mean and precision; that is an explicit illustration of the principle of stable estimation. Whether or not that principle applies, the posterior precision will always be at least the larger of the other two precisions; therefore, observation cannot but sharpen opinion here. This conclusion is somewhat special to the example; in general, an observation will occasionally increase, rather than dispel doubt.

In applying these formulas, as an approximation, to inference based on a large number n of observations with average $\bar{x}$ and sample variance s^2, x is $\bar{x}$ and h is n/s^2. To illustrate both the extent to which the prior distribution can be irrelevant and the rapid narrowing of the posterior distribution as the result of a few normal observations, consider Figure 2. The top section of the figure shows two prior distributions, one with mean -9 and standard deviation 6 and the other with mean 3 and standard deviation 2. The other four sections show posterior distributions obtained by applying Bayes' theorem to these two priors after samples of size n are taken from a distribution with mean 0 and standard deviation 2. The samples are artificially selected to have exactly the mean 0. After 9, and still more after

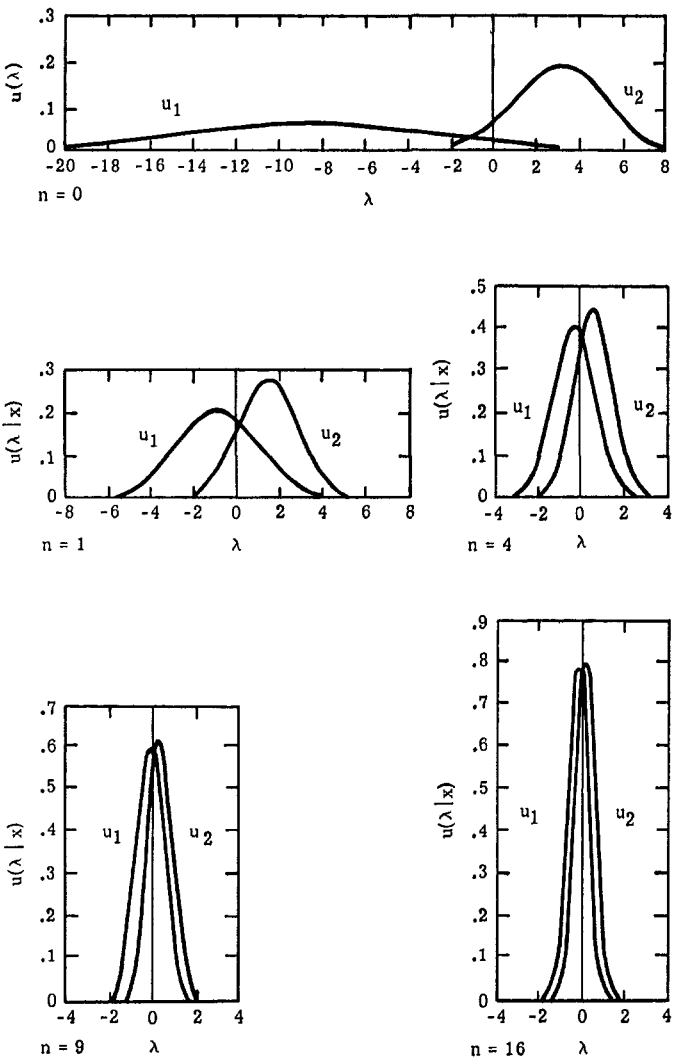

FIG. 2. Posterior distributions obtained from two normal priors after n normally distributed observations.

16, observations, these markedly different prior distributions have led to almost indistinguishable posterior distributions.

Of course the prior distribution is never irrelevant if the true parameter happens to fall in a region to which the prior distribution assigns virtually zero probability. A prior distribution which has a region of zero probability is therefore undesirable unless you really consider it impossible that the true parameter might fall in that region. Moral: Keep the mind open, or at least ajar.

Figure 2 also shows the typical narrowing of the posterior distribution with successive observations. After 4 observations, the standard deviation of your posterior distribution is less

than one half the standard deviation of a single observation; after 16, less than one fourth; and so on. In planning experiments, it sometimes seems distressing that the standard deviation decreases only as the square root of the number of observations, so a threefold improvement by sheer force of numbers, if possible at all, costs at least a ninefold effort. But subjects in unpublished experiments by W. L. Hays, L. D. Phillips, and W. Edwards are unwilling to change their diffuse initial opinions into sharp posterior ones, even after exposure to overwhelming evidence. This reluctance to extract from data as much certainty as they permit may be widespread. If so, explicit application of Bayes' theorem to information processing tasks now performed by unaided human judgment may produce more efficient use of the available information (for a proposal along these lines, see Edwards, 1962a, 1963).

When practical interest is focused on a few of several unknown parameters, the general Bayesian method is to find first the posterior joint distribution of all the parameters and from it to compute the corresponding marginal distribution of the parameters of special interest. When, for instance, n observations are drawn from a normal distribution of unknown mean μ and standard deviation σ, stable estimation applied to the two parameters μ and $\ln \sigma$ followed by elimination of $\ln \sigma$ leads to approximation of the posterior distribution of μ in terms of Student's t distribution with $n - 1$ degrees of freedom, in somewhat accidental harmony with classical statistics. (For those who have not encountered it before, the symbol ln stands for natural logarithm, or logarithm to the base e.)

Frequently, however, too little is known about the distribution from which a sequence of observations is drawn to express it confidently in terms of any moderate number of parameters. These are the situations that have evoked what is called the theory of nonparametric statistics. Ironically, a main concern of nonparametric statistics is to estimate the parameters of unknown distributions. The classical literature on nonparametric statistics is vast; see I. R. Savage (1957, 1962) and Walsh (1962). Bayesian counterparts of some of it are to be expected but are not yet achieved. To hint at some nonparametric Bayesian ideas, it seems reasonable to estimate the median of a largely unknown distribution by the median of the sample, and the mean of the distribution by the mean of the sample; given the sample, it will ordinarily be almost an even-money bet that the population median exceeds the sample median; and so on. Technically, the "and so on" points toward Bayesian justification for the classical theory of joint nonparametric tolerance intervals.

POINT AND INTERVAL ESTIMATION

Measurements are often used to make a point estimate, or best guess, about some quantity. In the fever-thermometer example, you would want, and would spontaneously make, such an estimate of the true temperature. What the best estimate is depends on what you need an estimate for and what penalty you associate with various possible errors, but a good case can often be made for the posterior mean, which minimizes the posterior mean squared error. For general scientific reporting there seems to be no other serious contender (see Savage, 1954, pp. 233–234). When the principle of stable estimation applies, the maximum-likelihood esti-

mate is often a good approximation to the posterior mean.

Classical statistics has also stressed interval, as opposed to point, estimates. Just what these are used for is hard to formulate (Savage, 1954, Section 17.2); they are, nonetheless, handy in thinking informally about specific applications of statistics. The Bayesian theory of interval estimation is simple. To name an interval that you feel 95% certain includes the true value of some parameter, simply inspect your posterior distribution of that parameter; any pair of points between which 95% of your posterior density lies defines such an interval. We call such intervals credible intervals, to distinguish them from the confidence intervals and fiducial intervals of classical statistics.

Of course, somewhat as for classical interval estimates, there are an unlimited number of different credible intervals of any specified probability. One is centered geometrically on the posterior mean; one, generally a different one, has equal amounts of probability on each side of the posterior median. Some include nearly all, or all, of one tail of the posterior distribution; some do not. The choice, which is seldom delicate, depends on the application. One choice of possible interest is the shortest credible interval of a specified probability; for unimodal, bilaterally symmetric posterior distributions, it is centered on the posterior mean, and median. In the fever example, in which an observation with standard deviation .05° made the principle of stable estimation applicable, the region $101° \pm 1.96\sigma = 101° \pm .098$ is the shortest interval containing approximately 95% of the posterior probability; 100.83° to 101.08° and 100.92° to ∞ are also 95% credible intervals, though asymmetric ones.

In certain examples like this one, the smallest credible interval of a specified credibility corresponds closely to the most popular of the classical confidence intervals having confidence level equal to that credibility. But in general credible intervals will differ from confidence intervals.

INTRODUCTION TO HYPOTHESIS TESTING

No aspect of classical statistics has been so popular with psychologists and other scientists as hypothesis testing, though some classical statisticians agree with us that the topic has been overemphasized. A statistician of great experience told us, "I don't know much about tests, because I have never had occasion to use one." Our devotion of most of the rest of this paper to tests would be disproportionate, if we were not writing for an audience accustomed to think of statistics largely as testing.

So many ideas have accreted to the word "test" that one definition cannot even hint at them. We shall first mention some of the main ideas relatively briefly, then flesh them out a bit with informal discussion of hypothetical substantive examples, and finally discuss technically some typical formal examples from a Bayesian point of view. Some experience with classical ideas of testing is assumed throughout. The pinnacle of the abstract theory of testing from the Neyman-Pearson standpoint is Lehmann (1959). Laboratory thinking on testing may derive more from R. A. Fisher than from the Neyman-Pearson school, though very few are explicitly familiar with Fisher's ideas culminating in 1950 and 1956.

The most popular notion of a test is, roughly, a tentative decision between two hypotheses on the basis of data,

and this is the notion that will dominate the present treatment of tests. Some qualification is needed if only because, in typical applications, one of the hypotheses—the null hypothesis—is known by all concerned to be false from the outset (Berkson, 1938; Hodges & Lehmann, 1954; Lehmann, 1959; I. R. Savage, 1957; L. J. Savage, 1954, p. 254); some ways of resolving the seeming absurdity will later be pointed out, and at least one of them will be important for us here.

The Neyman-Pearson school of theoreticians, with their emphasis on the decision-theoretic or behavioral approach, tend to define a test as a choice between two actions, such as whether or not to air condition the ivory tower so the rats housed therein will behave more consistently. This definition is intended to clarify operationally the meaning of decision between two hypotheses. For one thing, as Bayesians agree, such a decision resembles a potential dichotomous choice in some economic situation such as a bet. Again, wherever there is a dichotomous economic choice, the possible values of the unknown parameters divide themselves into those for which one action or the other is appropriate. (The neutral zone in which both actions are equally appropriate is seldom important and can be dealt with in various ways.) Thus a dichotomous choice corresponds to a partition into two hypotheses. Nonetheless, not every choice is like a simple bet, for economic differences within each hypothesis can be important.

Sometimes the decision-theoretic definition of testing is expressed as a decision to act as though one or the other of the two hypotheses were believed, and that has apparently led to some confusion (Neyman, 1957, p. 16). What action is wise of course depends in part on what is at stake. You would not take the plane if you believed it would crash, and would not buy flight insurance if you believed it would not. Seldom must you choose between exactly two acts, one appropriate to the null hypothesis and the other to its alternative. Many intermediate, or hedging, acts are ordinarily possible; flying after buying flight insurance, and choosing a reasonable amount of flight insurance, are examples.

From a Bayesian point of view, the special role of testing tends to evaporate, yet something does remain. Deciding between two hypotheses in the light of the datum suggests to a Bayesian only computing their posterior probabilities; that a pair of probabilities are singled out for special attention is without theoretical interest. Similarly, a choice between two actions reduces to choosing the larger of two expected utilities under a posterior distribution. The feature of importance for the Bayesian was practically lost in the recapitulation of general classical definitions. This happened, in part, because the feature would seem incidental in a general classical theory though recognized by all as important in specific cases and, in part, because expression of the feature is uncongenial to classical language, though implicitly recognized by classical statisticians.

In many problems, the prior density $u(\lambda)$ of the parameter(s) is often gentle enough relative to $v(x|\lambda)$ to permit stable estimation (or some convenient variation of it). One important way in which $u(\lambda)$ can fail to be sufficiently gentle is by concentrating considerable probability close to some point (or line, or surface, or the like). Certain practical devices can render the treatment of such a concentration of probability relatively

public. These devices are, or should be, only rather rarely needed, but they do seem to be of some importance and to constitute appropriate Bayesian treatment of some of the scientific situations in which the classical theory of hypothesis testing has been invoked. At least occasionally, a pair of hypotheses is associated with the concentration of probability. For example, if the squirrel has not touched it, that acorn is almost sure to be practically where it was placed yesterday. For vividness and to maintain some parallelism with classical expressions, we shall usually suppose concentration associated with a null hypothesis, as in this example; it is straightforward to extend the discussion to situations where there is not really such a pair of hypotheses. The theory of testing in the sense of dealing with concentrated probability as presented here draws heavily on Jeffreys (1939, see Ch. 5 and 6 of the 1961 edition) and Lindley (1961).

Examples. Discussion of a few examples may bring out some points associated with the various concepts of testing.

Example 1. Two teaching-machine programs for sixth-grade arithmetic have been compared experimentally.

For some purposes each program might be characterized by a single number, perhaps the mean difference between pretest and posttest performance on some standardized test of proficiency in arithmetic. This number, an index of the effectiveness of the program, must of course be combined with economic and other information from outside the experiment itself if the experiment is to guide some practical decision.

If one of the two programs must be adopted, the problem is one of testing in the sense of the general decision-theoretic definition, yet it is likely to be such that practicing statisticians would not ordinarily call the appropriate procedure a test at all. Unless your prior opinion perceptibly favored one of the two programs, you should plainly adopt that one which seemed, however slightly, to do better in the experiment. The classical counterpart of this simple conclusion had to be discovered against the tendency to invoke "significance tests" in all testing situations (Bahadur & Robbins, 1950).

But suppose one program is much more expensive to implement than the other. If such information about costs is available, it can be combined with information provided by the experiment to indicate how much proficiency can be bought for how many dollars. It is then a matter of judgment whether to make the purchase. In principle the judgment is simply one of the dollar value of proficiency (or equivalently of the proficiency value of dollars); in practice, such judgments are often difficult and controversial.

If the experiment is indecisive, should any decision be risked? Of course it should be if it really must be. In many actual situations there are alternatives such as further experimentation. The choice is then really at least trichotomous but perhaps with dichotomous emphasis on continuing, as opposed to desisting from, experimentation. Such suggestions as to continue only if the difference is not significant at, say, the 5% level are sometimes heard. Many classical theorists are dissatisfied with this approach, and we believe Bayesian statistics can do better (see Raiffa & Schlaifer, 1961, for some progress in this direction).

Convention asks, "Do these two programs differ at all in effectiveness?" Of course they do. Could any real

difference in the programs fail to induce at least some slight difference in their effectiveness? Yet the difference in effectiveness may be negligible compared to the sensitivity of the experiment. In this way, the conventional question can be given meaning, and we shall often ask it without further explanation or apology. A closely related question would be, "Is the superiority of Method A over Method B pointed to by the experiment real, taking due account of the possibility that the actual difference may be very small?" With several programs, the number of questions about relative superiority rapidly multiplies.

Example 2. Can this subject guess the color of a card drawn from a hidden shuffled bridge deck more or less than 50% of the time?

This is an instance of the conventional question, "Is there any difference at all?" so philosophically the answer is presumably "yes," though in the last analysis the very meaningfulness of the question might be challenged. We would not expect any such ostensible effect to stand up from one experiment to another in magnitude or direction. We are strongly prejudiced that the inevitable small deviations from the null hypothesis will always turn out to be somehow artifactual—explicable, for instance, in terms of defects in the shuffling or concealing of the cards or the recording of the data and not due to Extra-Sensory Perception (ESP).

One who is so prejudiced has no need for a testing procedure, but there are examples in which the null hypothesis, very sharply interpreted, commands some but not utter credence. The present example is such a one for many, more open minded about ESP than we, and even we can imagine, though we do not expect, phenomena that would shake our disbelief.

Example 3. Does this packed suitcase weigh less than 40 pounds? The reason you want to know is that the airlines by arbitrary convention charge overweight for more. The conventional weight, 40 pounds, plays little special role in the structure of your opinion which may well be diffuse relative to the bathroom scale. If the scale happens to register very close to 40 pounds (and you know its precision), the theory of stable estimation will yield a definite probability that the suitcase is overweight. If the reading is not close, you will have overwhelming conviction, one way or the other, but the odds will be very vaguely defined. For the conditions are ill suited to stable estimation if only because the statistical model of the scale is not sufficiently credible.

If the problem is whether to leave something behind or to put in another book, the odds are not a sufficient guide. Taking the problem seriously, you would have to reckon the cash cost of each amount of overweight and the cash equivalent to you of leaving various things behind in order to compute the posterior expected worth of various possible courses of action.

We shall discuss further the application of stable estimation to this example, for this is the one encounter we shall have with a Bayesian procedure at all harmonious with a classical tail-area significance test. Assume, then, that a normally distributed observation x has been made, with known standard deviation σ, and that your prior opinion about the weight of your suitcase is diffuse relative to the measurement. The principle of stable estimation applies,

so, as an acceptable approximation,

$$P(\lambda \leqq 40|x) = \Phi\left(\frac{x - 40}{\sigma}\right) = \Phi(t),$$

in case $|t|$ is not too great. In words, the probability that your suitcase weighs at most 40 pounds, in the light of the datum x, is the probability to the left of t under the standard normal distribution. Almost by accident, this is also the one-tailed significance level of the classical t test for the hypothesis that $\lambda \leqq 40$. The fundamental interpretation of $\Phi(t)$ here is the probability for you that your suitcase weighs less than 40 pounds; just the sort of thing that classical statistics rightly warns us not to expect a significance level to be. Problems in which stable estimation leads exactly to a one-tailed classical significance level are of very special structure. No Bayesian procedure yet known looks like a two-tailed test (Schlaifer, 1961, p. 212).

Classical one-tailed tests are often recommended for a situation in which Bayesian treatment would call for nothing like them. Imagine, for instance, an experiment to determine whether schizophrenia impairs problem solving ability, supposing it all but inconceivable that schizophrenia enhances the ability. This is classically a place to use a one-tailed test; the Bayesian recommendations for this problem, which will not be explored here, would not be tail-area tests and would be rather similar to the Bayesian null hypothesis tests discussed later. One point recognized by almost all is that if schizophrenia can do no good it must then do some harm, though perhaps too little to perceive.

Before putting the suitcase on the bathroom scales you have little expectation of applying the formal arithmetic of the preceding paragraphs. At that time, your opinion about the weight of the suitcase is diffuse. Therefore, no interval as small as 6 or 8 σ can include much of your initial probability. On the other hand, if $|t|$ is greater than 3 or 4, which you very much expect, you will not rely on normal tail-area computations, because that would put the assumption of normality to unreasonable strain. Also Assumption 2 of the discussion of stable estimation will probably be drastically violated. You will usually be content in such a case to conclude that the weight of the suitcase is, beyond practical doubt, more (or less) than 40 pounds.

The preceding paragraph illustrates a procedure that statisticians of all schools find important but elusive. It has been called the interocular traumatic test;[2] you know what the data mean when the conclusion hits you between the eyes. The interocular traumatic test is simple, commands general agreement, and is often applicable; well-conducted experiments often come out that way. But the enthusiast's interocular trauma may be the skeptic's random error. A little arithmetic to verify the extent of the trauma can yield great peace of mind for little cost.

BAYESIAN HYPOTHESIS TESTING

Odds and likelihood ratios. Gamblers frequently measure probabilities in terms of odds. Your odds in favor of the event A are (aside from utility effects) the amount that you would just be willing to pay if A does not occur in compensation for a commitment from someone else to pay you one unit of money if A does occur. The odds $\Omega(A)$ in favor of A are thus related to the probability $P(A)$ of A

[2] J. Berkson, personal communication, July 14, 1958.

and the probability $1 - P(A)$ of not A, or $\bar{A}$, by the condition,

$$\Omega(A)[1 - P(A)] = P(A).$$

Odds and probability are therefore translated into each other thus,

$$\Omega(A) = \frac{P(A)}{1 - P(A)} = \frac{P(A)}{P(\bar{A})};$$

$$P(A) = \frac{\Omega(A)}{1 + \Omega(A)}.$$

For example, odds of 1, an even-money bet, correspond to a probability of 1/2; a probability of 9/10 corresponds to odds of 9 (or 9 to 1), and a probability of 1/10 corresponds to odds of 1/9 (or 1 to 9). If $P(A)$ is 0, $\Omega(A)$ is plainly 0; and if $P(A)$ is 1, $\Omega(A)$ may be called ∞, if it need be defined at all.

From a Bayesian standpoint, part of what is suggested by "testing" is finding the posterior probability $P(A \mid D)$ of the hypothesis A in the light of the datum D, or equivalently, finding the posterior odds $\Omega(A \mid D)$.

According to Bayes' theorem

$$P(A \mid D) = \frac{P(D \mid A)P(A)}{P(D)}, \quad [7]$$

$$P(\bar{A} \mid D) = \frac{P(D \mid \bar{A})P(\bar{A})}{P(D)}. \quad [8]$$

Dividing each side of Equation 7 by the corresponding side of Equation 8, canceling the common denominators $P(D)$, and making evident abbreviations leads to a condensation of Equations 7 and 8 in terms of odds;

$$\Omega(A \mid D) = \frac{P(D \mid A)}{P(D \mid \bar{A})}\Omega(A)$$

$$= L(A; D)\Omega(A). \quad [9]$$

In words, the posterior odds in favor of A given the datum D are the prior odds multiplied by the ratio of the conditional probabilities of the datum given the hypothesis A and given its negation. The ratio of conditional probabilities $L(A; D)$ is called the likelihood ratio in favor of the hypothesis A on the basis of the datum D.

Plainly, and according to Equation 9, D increases the odds for A, if and only if D is more probable under A than under its negation $\bar{A}$ so that $L(A; D)$ is greater than 1.

If D is impossible under $\bar{A}$, Equation 9 requires an illegitimate division, but it can fairly be interpreted to say that A has acquired probability 1 unless $\Omega(A) = 0$, in which case the problem is ill specified. With that rather academic exception, whenever $\Omega(A)$ is 0 so is $\Omega(A \mid D)$; roughly, once something is regarded as impossible, no evidence can reinstate its credibility.

In actual practice, $L(A; D)$ and $\Omega(A)$ tend to differ from person to person. Nonetheless, statistics is particularly interested in examining how and when Equation 9 can lead to relatively public conclusions, a theme that will occupy several sections.

Simple dichotomy. It is useful, at least for exposition, to consider problems in which $L(A; D)$ is entirely public. For example, someone whose word you and we trust might tell us that the die he hands us produces 6's either (A) with frequency 1/6 or ($\bar{A}$) with frequency 1/5. Your initial opinion $\Omega(A)$ might differ radically from ours. But, for you and for us, the likelihood ratio in favor of A on the basis of a 6 is (1/6)/(1/5) or 5/6, and the likelihood ratio in favor of A on the basis of a non-6 is (5/6)/(4/5) or 25/24. Thus, if a 6 appears when the die is rolled, everyone's confidence in A will diminish slightly; specifically, odds in favor of A will be diminished by 5/6. Similarly, a non-6 will augment $\Omega(A)$ by the factor 25/24.

If such a die could be rolled only once, the resulting evidence $L(A; D)$ would be negligible for almost any purpose; if it can be rolled many times, the evidence is ultimately sure to become definitive. As is implicit in the concept of the not necessarily fair die, if $D_1, D_2, D_3, \cdots$ are the outcomes of successive rolls, then the same function $L(A; D)$ applies

to each. Therefore Equation 9 can be applied repeatedly, thus:

$$\begin{aligned}
\Omega(A \mid D_1) &= L(A; D_1)\Omega(A) \\
\Omega(A \mid D_2, D_1) &= L(A; D_2)\Omega(A \mid D_1) \\
&= L(A; D_2)L(A; D_1)\Omega(A) \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
\Omega(A \mid D_n, \cdots, D_1) &= L(A; D_n)\Omega(A \mid D_{n-1}, \\
&\qquad D_{n-2}, \cdots, D_1) \\
&= L(A; D_n)L(A; D_{n-1}) \cdots \\
&\qquad L(A; D_1)\Omega(A) \\
&= \Pi_{j=1}^{n} L(A; D_j)\Omega(A).
\end{aligned}$$

This multiplicative composition of likelihood ratios exemplifies an important general principle about observations which are independent given the hypothesis.

For the specific example of the die, if x 6's and y non-6's occur (where of course $x + y = n$), then

$$\Omega(A \mid D_n, \cdots, D_1) = \left(\frac{5}{6}\right)^x \left(\frac{25}{24}\right)^y \Omega(A).$$

For large n, if A obtains, it is highly probable at the outset that x/n will fall close to 1/6. Similarly, if A does not obtain x/n will probably fall close to 1/5. Thus, if A obtains, the overall likelihood $(5/6)^x(25/24)^y$ will probably be very roughly

$$\begin{aligned}
\left(\frac{5}{6}\right)^{n/6} \left(\frac{25}{24}\right)^{5n/6} &= \left[\left(\frac{5}{6}\right)^{1/6} \left(\frac{25}{24}\right)^{5/6}\right]^n \\
&= (1.00364)^n \\
&= 10^{0.00158n}.
\end{aligned}$$

By the time n is 1,200 everyone's odds in favor of A will probably be augmented about a hundredfold, if A is in fact true. One who started very skeptical of A, say with $\Omega(A)$ about a thousandth, will still be rather skeptical. But he would have to start from a very skeptical position indeed not to become strongly convinced when n is 6,300 and the overall likelihood ratio in favor of A is about 10 billion.

The arithmetic for $\bar{A}$ is:

$$\begin{aligned}
&\left[\left(\frac{5}{6}\right)^{1/6} \left(\frac{25}{24}\right)^{4/5}\right]^n \\
&\qquad = (0.9962)^n = 10^{-0.00165n}.
\end{aligned}$$

So the rate at which evidence accumulates against A, and for $\bar{A}$, when $\bar{A}$ is true is in this case a trifle more than the rate at which it accumulates for A when A is true.

Simple dichotomy is instructive for statistical theory generally but must be taken with a grain of salt. For simple dichotomies —that is, applications of Equation 9 in which everyone concerned will agree and be clear about the values of $L(A; D)$—rarely, if ever, occur in scientific practice. Public models almost always involve parameters rather than finite partitions.

Some generalizations are apparent in what has already been said about simple dichotomy. Two more will be sketchily illustrated: Decision-theoretic statistics, and the relation of the dominant classical decision-theoretic position to the Bayesian position. (More details will be found in Savage, 1954, and Savage et al., 1962, indexed under simple dichotomy.)

At a given moment, let us suppose, you have to guess whether it is A or $\bar{A}$ that obtains and you will receive $\$I$ if you guess correctly that A obtains, $\$J$ if you guess correctly that $\bar{A}$ obtains, and nothing otherwise. (No real generality is lost in not assigning four arbitrarily chosen payoffs to the four possible combinations of guess and fact.) The expected cash value to you of guessing A is $\$IP(A)$ and that of guessing $\bar{A}$ is $\$JP(\bar{A})$. You will therefore prefer to guess A if and only if $\$IP(A)$ exceeds $\$JP(\bar{A})$; that is, just if $\Omega(A)$ exceeds J/I. (More rigorous treatment would replace dollars with utiles.)

Similarly, if you need not make your guess until after you have examined a datum D, you will prefer to guess A if, and only if, $\Omega(A \mid D)$ exceeds J/I. Putting this together with Equation 9, you will prefer to guess A if, and only if,

$$L(A; D) > \frac{J}{I\Omega(A)} = \Lambda,$$

where your critical likelihood ratio Λ is defined by the context.

This conclusion does not at all require that the dichotomy between A and $\bar{A}$ be simple, or public, but for comparison with the classical approach to the same problem continue to assume that it is. Classical statisticians were the first to conclude that there must be some Λ such that you will guess A if $L(A; D) > \Lambda$ and guess $\bar{A}$ if $L(A; D) < \Lambda$. (For this sketch, it is excusable to neglect the possibility that $\Lambda = L(A; D)$.) By and large, classical statisticians say that the choice of Λ is an entirely subjective one which no one but you can make (e.g., Lehmann, 1959, p. 62). Bayesians agree; for according to Equation 9, Λ is inversely proportional to your current odds for A, an aspect of your personal opinion.

The classical statisticians, however, have overlooked a great simplification, namely that

your critical Λ will not depend on the size or structure of the experiment and will be proportional to J/I. Once the Bayesian position is accepted, Equation 9 is of course an argument for this simplification, but it can also be arrived at along a classical path, which in effect derives much, if not all, of Bayesian statistics as a natural completion of the classical decision-theoretic position. This relation between the two views, which in no way depends on the artificiality of simple dichotomy here used to illustrate it, cannot be overemphasized. (For a general demonstration, see Raiffa & Schlaifer, 1961, pp. 24–27.)

The simplification is brought out by the set of indifference curves among the various probabilities of the two kinds of errors (Lehmann, 1958). Of course, any reduction of the probability of one kind of error is desirable if it does not increase the probability of the other kind of error, and the implications of classical statistics leave the description of the indifference curves at that. But the considerations discussed easily imply that the indifference curves should be parallel straight lines with slope $-[J/I\Omega(A)]$. As Savage (1962b) puts it:

> the subjectivist's position is more objective than the objectivist's, for the subjectivist finds the range of coherent or reasonable preference patterns much narrower than the objectivist thought it to be. How confusing and dangerous big words are [p. 67]!

Classical statistics tends to divert attention from Λ to the two conditional probabilities of making errors, by guessing A when $\bar{A}$ obtains and vice versa. The counterpart of the probabilities of these two kinds of errors in more general problems is called the operating characteristic, and classical statisticians suggest, in effect, that you should choose among the available operating characteristics as a method of choosing Λ, or more generally, your prior distribution. This is not mathematically wrong, but it distracts attention from your value judgments and opinions about the unknown facts upon which your preferred Λ should directly depend without regard to how the probabilities of errors vary with Λ in a specific experiment.

There are important advantages to recognizing that your Λ does not depend on the structure of the experiment. It will help you, for example, to choose between possible experimental plans. It leads immediately to the very important likelihood principle, which in this application says that the numerical value of the likelihood ratio of the datum conveys the entire import of the datum. (A later section is about the likelihood principle.) Wolfowitz (1962) dissents.

Approaches to null hypothesis testing. Next we examine situations in which a very sharp, or null, hypothesis is compared with a rather flat or diffuse alternative hypothesis. This short section indicates general strategies of such comparisons. None of the computations or conclusions depend on assumptions about the special initial credibility of the null hypothesis, but a Bayesian will find such computations uninteresting unless a non-negligible amount of his prior probability is concentrated very near the null hypothesis value.

For the continuous cases to be considered in following sections, the hypothesis A is that some parameter λ is in a set that might as well also be called A. For one-dimensional cases in which the hypothesis A is that λ is almost surely negligibly far from some specified value λ_0, the odds in favor of A given the datum D, as in Equation 9, are

$$\begin{aligned}\Omega(A \mid D) &= \frac{P(A \mid D)}{P(\bar{A} \mid D)} \\ &= \frac{v(D \mid \lambda_0)}{\int v(D \mid \lambda) u(\lambda \mid \bar{A}) d\lambda} \quad \Omega(A) \\ &= L(A ; D)\Omega(A).\end{aligned}$$

Natural generalizations apply to multidimensional cases. The numerator $v(D \mid \lambda_0)$ will in usual applications be public. But the denominator, the probability of D under the alternative hypothesis, depends on the usually far from public prior density under the alternative hypothesis. Nonetheless, there are some relatively public methods of appraising the

denominator, and much of the following discussion of tests is, in effect, about such methods. Their spirit is opportunistic, bringing to bear whatever approximations and bounds offer themselves in particular cases. The main ideas of these methods are sketched in the following three paragraphs, which will later be much amplified by examples.

First, the principle of stable estimation may apply to the datum and to the density $u(\lambda|\bar{A})$ of λ given the alternative hypothesis $\bar{A}$. In this case, the likelihood ratio reflects no characteristics of $u(\lambda|\bar{A})$ other than its value in the neighborhood favored by the datum, a number that can be made relatively accessible to introspection.

Second, it is relatively easy, in any given case, to determine how small the likelihood ratio can possibly be made by utterly unrestricted and artificial choice of the function $u(\lambda|\bar{A})$. If this rigorous public lower bound on the likelihood ratio is not very small, then there exists no system of prior probabilities under which the datum greatly detracts from the credibility of the null hypothesis. Remarkably, this smallest possible bound is by no means always very small in those cases when the datum would lead to a high classical significance level such as .05 or .01. Less extreme (and therefore larger) lower bounds that do assume some restriction on $u(\lambda|\bar{A})$ are sometimes appropriate; analogous restrictions also lead to upper bounds. When these are small, the datum does rather publicly greatly lower the credibility of the null hypothesis. Analysis to support an interocular traumatic impression might often be of this sort. Inequalities stated more generally by Hildreth (1963) are behind most of these lower and upper bounds.

Finally, when $v(D|\lambda)$ admits of a conjugate family of distributions, it may be useful, as an approximation, to suppose $u(\lambda|\bar{A})$ restricted to the conjugate family. Such a restriction may help fix reasonably public bounds to the likelihood ratio.

We shall see that classical procedures are often ready severely to reject the null hypothesis on the basis of data that do not greatly detract from its credibility, which dramatically demonstrates the practical difference between Bayesian and classical statistics. This finding is not altogether new. In particular, Lindley (1957) has proved that for any classical significance level for rejecting the null hypothesis (no matter how small) and for any likelihood ratio in favor of the null hypothesis (no matter how large), there exists a datum significant at that level and with that likelihood ratio.

To prepare intuition for later technical discussion we now show informally, as much as possible from a classical point of view, how evidence that leads to classical rejection of a null hypothesis at the .05 level can favor that null hypothesis. The loose and intuitive argument can easily be made precise (and is, later in the paper). Consider a two-tailed t test with many degrees of freedom. If a true null hypothesis is being tested, t will exceed 1.96 with probability 2.5% and will exceed 2.58 with probability .5%. (Of course, 1.96 and 2.58 are the 5% and 1% two-tailed significance levels; the other 2.5% and .5% refer to the possibility that t may be smaller than -1.96 or -2.58.) So on 2% of all occasions when true null hypotheses are being tested, t will lie between 1.96 and 2.58. How often will t lie in that interval when the null hypothesis is false? That

depends on what alternatives to the null hypothesis are to be considered. Frequently, given that the null hypothesis is false, all values of t between, say, -20 and $+20$ are about equally likely for you. Thus, when the null hypothesis is false, t may well fall in the range from 1.96 to 2.58 with at most the probability $(2.58 - 1.96)/[+20 - (-20)] = 1.55\%$. In such a case, since 1.55 is less than 2 the occurrence of t in that interval speaks mildly for, not vigorously against, the truth of the null hypothesis.

This argument, like almost all the following discussion of null hypothesis testing, hinges on assumptions about the prior distribution under the alternative hypothesis. The classical statistician usually neglects that distribution—in fact, denies its existence. He considers how unlikely a t as far from 0 as 1.96 is if the null hypothesis is true, but he does not consider that a t as close to 0 as 1.96 may be even less likely if the null hypothesis is false.

A Bernoullian example. To begin a more detailed examination of Bayesian methods for evaluating null hypotheses, consider this example:

We are studying a motor skills task. Starting from a neutral rest position, a subject attempts to touch a stylus as near as possible to a long, straight line. We are interested in whether his responses favor the right or the left of the line. Perhaps from casual experience with such tasks, we give special credence to the possibility that his long-run frequency p of "rights" is practically $p_0 = 1/2$. The problem is here posed in the more familiar frequentistic terminology; its Bayesian translation, due to de Finetti, is sketched in Section 3.7 of Savage (1954). The following discussion applies to any fraction p_0 as well as to the specific value 1/2. Under the null hypothesis, your density of the parameter p is sharply concentrated near p_0, while your density of p under the alternative hypothesis is not concentrated and may be rather diffuse over much of the interval from 0 to 1.

If n trials are undertaken, the probability of obtaining r rights given that the true frequency is p is of course $C_r^n p^r(1 - p)^{n-r}$. The probability of obtaining r under the null hypothesis that p is literally p_0 is $C_r^n p_0{}^r(1 - p_0)^{n-r}$. Under the alternative hypothesis, it is

$$\int_0^1 C_r^n p^r(1 - p)^{n-r} u(p \mid H_1) dp,$$

that is, the probability of r given p averaged over p, with each value in the average weighted by its prior density under the alternative hypothesis. The likelihood ratio is therefore

$$L(p_0; r, n) = \frac{p_0{}^r(1 - p_0)^{n-r}}{\int_0^1 p^r(1 - p)^{n-r} u(p \mid H_1) dp}. \quad [10]$$

The disappearance of C_r^n from the likelihood ratio by cancellation is related to the likelihood principle, which will be discussed later. Had the experiment not been analyzed with a certain misplaced sophistication, C_r^n would never have appeared in the first place. We would simply have noted that the probability of any specific sequence of rights and lefts with r rights and $n - r$ lefts is, given p, exactly $p^r(1 - p)^{n-r}$. That the number of different sequences of this composition is C_r^n is simply irrelevant to Bayesian inference about p.

One possible way to reduce the

denominator of Equation 10 to more tractable form is to apply the principle of stable estimation, or more accurately certain variants of it, to the denominator. To begin with, if $u(p|H_1)$ were a constant u', then the denominator would be

$$\int_0^1 p^r(1-p)^{n-r}u(p|H_1)dp = u'\int_0^1 p^r(1-p)^{n-r}dp = \frac{u'}{(n+1)C_r^n}. \quad [11]$$

The first equality is evident; the second is a known formula, enchantingly demonstrated by Bayes (1763). Of course u cannot really be a constant unless it is 1, but if r and $n-r$ are both fairly large $p^r(1-p)^{n-r}$ is a sharply peaked function with its maximum at r/n. If $u(p|H_1)$ is gentle near r/n and not too wild elsewhere, Equation 11 may be a satisfactory approximation, with $u' = u(r/n|H_1)$. This condition is often met, and it can be considerably weakened without changing the conclusion, as will be explained next.

If the graph of $u(p|H_1)$ were a straight, though not necessarily horizontal, line then the required integral would be

$$\int_0^1 p^r(1-p)^{n-r}u(p|H_1)dp = \frac{u\left(\frac{r+1}{n+2}|H_1\right)}{(n+1)C_r^n}. \quad [12]$$

This is basically a standard formula like the latter part of Equation 11, and is in fact rather easily inferred from that earlier formula itself. Consequently, for large r and $n-r$, Equation 12 can be justified as an approximation with $u' = u[(r+1)/(n+2)|H_1]$ whenever $u(p|H_1)$ is nearly linear in the neighborhood of $(r+1)/(n+2)$, which under the assumed conditions is virtually indistinguishable from r/n.

In summary, it is often suitable to approximate the likelihood ratio thus:

$$L(p_0; r, n) = \frac{n+1}{u'}C_r^n p_0^r(1-p_0)^{n-r} = \frac{(n+1)P(r|p_0, n)}{u'} \quad [13]$$

where $u' = u(r/n|H_1)$ or $u[(r+1)/(n+2)|H_1]$.

Does this approximation apply to you in a specific case? If so, what value of u' is appropriate? Such subjective questions can be answered only by self-interrogation along lines suggested by our discussion of stable estimation. In particular, u' is closely akin to the φ of our Condition 2 for stable estimation. In stable estimation, the value of φ cancels out of all calculations, but here, u' is essential. One way to arrive at u' is to ask yourself what probability you attach to a small, but not microscopic, interval of values of p near r/n under the alternative hypothesis. Your reply will typically be vague, perhaps just a rough order of magnitude, but that may be enough to settle whether the experiment has strikingly confirmed or strikingly discredited the null hypothesis.

In principle, any positive value of u' can arise, but values between .1 and 10 promise to predominate in practice. The reasons for this are complex and not altogether clear to us, but something instructive can be said about them here. To begin with, since the integral of $u(p|H_1)$ is 1, $u(p|H_1)$ cannot exceed 10 throughout an interval as long as 1/10. Therefore, if $u(r/n|H_1)$ is much greater than 10, $u(p|H_1)$ must undergo great diminution quite close to r/n, and the approximation will not be applicable unless $v(r|p, n)$ is very violent indeed, which can happen only if r and $n-r$ are very large, perhaps several thousands.

Typically, $u(p|H_1)$ attains its maximum at p_0, or at any rate is rather substantial near there—its maximum is necessarily at least 1,

because its integral from 0 to 1 is 1. Therefore, should the null hypothesis obtain, $u(r/n|H_1)$ is most unlikely to be as small as 1/10. Under the alternative hypothesis, you must, according to a simple mathematical argument, attach probability less than 1/10 to the set of those values of p for which $u(p|H_1)$ is less than 1/10. Under a reasonably diffuse alternative hypothesis, the probability of an r for which $u(r/n|H_1)$ is at most 1/10 is much the same as the probability of a p for which $u(p|H_1)$ is at most 1/10. Thus, under either hypothesis, you are unlikely to encounter an r for which $u(r/n|H_1) < 1/10$. You are actually much more unlikely yet to encounter such an r for which the approximation is applicable.

In this particular example of a person aiming at a line with a stylus, structuring your opinion in terms of a sharp null hypothesis and a diffuse alternative is rather forced. More realistically, your prior opinion is simply expressed by a density with a rather sharp peak, or mode, at $p_0 = 1/2$, and your posterior distribution will tend to have two modes, one at p_0 and the other about at r/n. Nonetheless, an arbitrary structuring of the prior density as a weighted average, or probability mixture, of two densities, one practically concentrated at p_0 and the other somewhat diffuse, may be a useful approach.

Conversely, even if the division is not artificial, the unified approach is always permissible. This may help emphasize that determining the posterior odds is seldom the entire aim of the analysis. The posterior distribution of p under the alternative hypothesis is also important. This density $u(p|r, n, H_1)$ is determined by Bayes' theorem from the datum (r, n) and the alternative prior density $u(p|H_1)$; for this, what the hypothesis H_0 is, or how probable you consider it either before or after the experiment are all irrelevant. As in any other estimation problem, the principle of stable estimation may provide an adequate approximation for $u(p|r, n, H_1)$. If in addition, the null hypothesis is strongly discredited by the datum, then the entire posterior density $u(p|r, n)$ will be virtually unimodal and identifiable with $u(p|r, n, H_1)$ for many purposes. In fact, the outcome of the test in this case is to show that stable estimation (in particular our Assumption 3) is applicable without recourse to Assumption 3′.

The stable-estimation density for this Bernoullian problem is of course $p^r(1 - p)^{n-r}$ multiplied by the appropriate normalizing constant, which is implicit in the second equality of Equation 11. This is an instance of the beta density of indices a and b,

$$\frac{(a + b - 1)!}{(a - 1)!\,(b - 1)!}\,p^{a-1}(1 - p)^{b-1}.$$

In this case, $a = r + 1$ and $b = (n - r) + 1$.

In view of the rough rule of thumb that u' is of the order of magnitude of 1, the factor $(n + 1)P(r|p_0, n)$ is at least a crude approximation to $L(p_0; r, n)$ and is of interest in any case as the relatively public factor in $L(p_0; r, n)$ and hence in $\Omega(H_0|r, n)$. The first three rows of Table 1 show hypothetical data for four different experiments of this sort (two of them on a large scale) along with the corresponding likelihood ratios for the uniform alternative prior. The numbers in Table 1 are, for illustration, those that would, for the specified number of observations, barely lead to rejection of the null hypothesis, $p = .5$, by a classical two-tailed test at the .05 level.

How would a Bayesian feel about the numbers in Table 1? Remember that a likelihood ratio greater than 1 leaves one more confident of the null hypothesis than he was to start with, while a likelihood ratio less than 1 leaves him less confident of it than

TABLE 1

LIKELIHOOD RATIOS UNDER THE UNIFORM ALTERNATIVE PRIOR AND MINIMUM LIKELIHOOD RATIOS FOR VARIOUS VALUES OF n AND FOR VALUES OF r JUST SIGNIFICANT AT THE .05 LEVEL

	Experiment number				
	1	2	3	4	∞
n	50	100	400	10,000	(very large)
r	32	60	220	5,098	$(n + 1.96\sqrt{n})/2$
$L(p_0; r, n)$	.8178	1.092	2.167	11.689	$.11689\sqrt{n}$
$L_{\min}$	.1372	.1335	.1349	.1465	.1465

he was to start with. Thus Experiment 1, which argues against the null hypothesis more persuasively than the others, discredits it by little more than a factor of 1.27 to 1 (assuming $u' = 1$) instead of the 20 to 1 which a naive interpretation of the .05 level might (contrary to classical as well as Bayesian theory) lead one to expect. More important, Experiments 3 and 4, which would lead a classical statistician to reject the null hypothesis, leave the Bayesian who happens to have a roughly uniform prior, more confident of the null hypothesis than he was to start with. And Experiment 4 should reassure even a rather skeptical person about the truth of the null hypothesis. Here, then, is a blunt practical contradiction between conclusions produced by classical and Bayesian rules for statistical inference. Though the Bernoullian example is special, particularly in that it offers relatively general grounds for u' to be about 1, classical procedures quite typically are, from a Bayesian point of view, far too ready to reject null hypotheses.

Approximation in the spirit of stable estimation is by no means the last word on evaluating a likelihood ratio. Sometimes, as when r or $n - r$ are too small, it is not applicable at all, and even when it might otherwise be applicable, subjective haze and interpersonal disagreement affecting u' may frustrate its application. The principal alternative devices known to us will be at least mentioned in connection with the present example, and most of them will be explored somewhat more in connection with later examples.

It is but an exercise in differential calculus to see that $p^r(1 - p)^{n-r}$ attains its maximum at $p = r/n$. Therefore, regardless of what $u(p)$ actually is, the likelihood ratio in favor of the null hypothesis is at least

$$L_{\min} = \frac{p_0{}^r(1 - p_0)^{n-r}}{\left(\frac{r}{n}\right)^r \left(1 - \frac{r}{n}\right)^{n-r}}.$$

If this number is not very small, then everyone (who does not altogether reject Bayesian ideas) must agree that the null hypothesis has not been greatly discredited. For example, since $L_{\min}$ in Table 1 exceeds .05, it is impossible for the experiments considered there that rejection at the 5% significance level should ever correspond to a nineteenfold diminution of the odds in favor of the null hypothesis. It is mathematically possible but realistically preposterous for $L_{\min}$ to be the actual likelihood ratio. That could occur only if your $u(p|H_1)$ were concentrated at r/n,

and prior views are seldom so prescient.

It is often possible to name, whether for yourself alone or for "the public," a number u^* that is a generous upper bound for $u(p|H_1)$, that is, a u^* of which you are quite confident that $u(p|H_1) < u^*$ for all p (in the interval from 0 to 1). A calculation much like Equations 11 and 13 shows that if u^* is substituted for u' in Equation 13, the resultant fraction is less than the actual likelihood ratio. If this method of finding a lower bound for L is not as secure as that of the preceding paragraph, it generally provides a better, that is, a bigger, one. The two methods can be blended into one which is always somewhat better than either, as will be illustrated in a later example.

Upper, as well as lower, bounds for L are important. One way to obtain one is to paraphrase the method of the preceding paragraph with a lower bound rather than an upper bound for $u(p)$. This method will seldom be applicable as stated, since $u(p)$ is likely to be very small for some values of p, especially values near 0 or 1. But refinements of the method, illustrated in later examples, may be applicable.

Another avenue, in case $u(p|H_1)$ is known with even moderate precision but is not gentle enough for the techniques of stable estimation, is to approximate $u(p|H_1)$ by the beta density for some suitable indices a and b. This may be possible since the two adjustable indices of the beta distribution provide considerable latitude and since what is required of the approximation is rather limited. It may be desirable, because beta densities are conjugate to Bernoullian experiments. In fact, if $u(p|H_1)$ is a beta distribution with indices a and b, then $u(p|r, n, H_1)$ is also a beta density, with indices $a + r$ and $b + (n - r)$. The likelihood ratio in this case is

$$\frac{(a-1)!\,(b-1)!\,(n+a+b-1)!}{(a+r-1)!\,(b+n-r-1)!\,(a+b-1)!} \times p_0^r(1-p_0)^{n-r}.$$

These facts are easy consequences of the definite integral on which Equation 11 is based. More details will be found in Chapter 9 of Raiffa and Schlaifer (1961).

A one-dimensional normal example. We examine next one situation in which classical statistics prescribes a two-tailed t test. As in our discussion of normal measurements in the section on distribution theory, we will consider one normally distributed observation with known variance; as before, this embraces by approximation the case of 25 or more observations of unknown variance and many other applications such as the Bernoullian experiments.

According to Weber's Law, the ratio of the just noticeable difference between two sensory magnitudes to the magnitude at which the just noticeable difference is measured is a constant, called the Weber fraction. The law is approximately true for frequency discrimination of fairly loud pure tones, say between 2,000 and 5,000 cps; the Weber fraction is about .0020 over this fairly wide range of frequencies. Psychophysicists disagree about the nature and extent of interaction between different sense modalities. You might, therefore, wonder whether there is any difference between the Weber fraction at 3,000 cps for subjects in a lighted room and in complete darkness. Since search for such interactions among modalities has failed more often than it has succeeded, you might give considerable initial credence to the null hypothesis that there will be no (appreciable) difference between the Weber fractions obtained in light and in darkness. However, such effects might possibly be substantial. If they are, light could facilitate or could hinder frequency discrimination. Some work on arousal might lead you to expect facilitation; the idea of visual stimuli competing with auditory stimuli for attention might lead you to expect hindrance. If the null hypothesis is false, you might consider any value between .0010 and .0030 of the Weber fraction obtained in darkness to be roughly as plausible as any other value in that range. Your instruments and procedure permit determination of the Weber fraction with a standard deviation of 3.33×10^{-5} (a standard deviation of .1 cps at 3,000 cps, which is not too implausible if your procedures permit repeated measurements and are in other ways extremely accurate). Thus the range of plausible values is 60 standard deviations wide—quite large compared with similar numbers in other parts of experimental psychology, though small compared with many analogous numbers in physics or chemistry. Such a small standard deviation relative to the range of plausible values is not indispensable to the example, but it is convenient and helps make the example congenial to both physical and social scien-

tists. If the standard deviation were more than 10^{-4}, however, the eventual application of the principle of stable estimation to the example would be rather difficult to justify.

A full Bayesian analysis of this problem would take into account that each observation consists of two Weber fractions, rather than one difference between them. However, as classical statistics is even too ready to agree, little if any error will result from treating the difference between each Weber fraction determined in light and the corresponding Weber fraction determined in darkness as a single observation. In that formulation, the null hypothesis is that the true difference is 0, and the alternative hypothesis envisages the true difference as probably between $-.0010$ and $+.0010$. The standard deviation of the measurement of the difference, if the measurements in light and darkness are independent, is $1.414 \times 3.33 \times 10^{-5} = 4.71 \times 10^{-5}$. Since our real concern is exclusively with differences between Weber fractions and the standard deviation of these differences, it is convenient to measure every difference between Weber fractions in standard deviations, that is to multiply it by 21,200 ($= 1/\sigma$). In these new units, the plausible range of observations is about from -21 to $+21$, and the standard deviation of the differences is 1. The rest of the discussion of this example is based on these numbers alone.

The example specified by the last two paragraphs has a sharp null hypothesis and a rather diffuse symmetric alternative hypothesis with good reasons for associating substantial prior probability with each. Although realistically the null hypothesis cannot be infinitely sharp, calculating as though it were is an excellent approximation. Realism, and even mathematical consistency, demands far more sternly that the alternative hypothesis not be utterly diffuse (that is, uniform from $-\infty$ to $+\infty$); otherwise, no measurement of the kind contemplated could result in any opinion other than certainty that the null hypothesis is correct.

Having already assumed that the distribution of the true parameter or parameters under the null hypothesis is narrow enough to be treated as though it were concentrated at the single point 0, we also assume that the distribution of the datum given the parameter is normal with moderate variance. By moderate we mean large relative to the sharp null hypothesis but (in most cases) small relative to the distribution under the alternative hypothesis of the true parameter.

Paralleling our treatment of the Bernoullian example, we shall begin, after a neutral formulation, with an approximation akin to stable estimation, then explore bounds on the likelihood ratio L that depend on far less stringent assumptions, and finally explore normal prior distributions.

Without specifying the form of the prior distribution under the alternative hypothesis, the likelihood ratio in the Weber-fraction example under discussion is

$$L(\lambda_0; x) = \frac{\frac{1}{\sigma}\varphi\left(\frac{x-\lambda_0}{\sigma}\right)}{\int \frac{1}{\sigma}\varphi\left(\frac{x-\lambda}{\sigma}\right) u(\lambda \mid H_1)d\lambda}. \quad [14]$$

The numerator is the density of the datum x under the null hypothesis; σ is the standard deviation of the measuring instrument. The denominator is the density of x under the alternative hypothesis. The values of λ are the possible values of the actual difference under the alternative hypothesis, and λ_0 is the null value, 0. $\varphi[(x-\lambda)/\sigma]$ is the ordinate of the standard normal density at the point $(x-\lambda)/\sigma$. Hereafter, we will use the familiar statistical abbreviation $t = (x-\lambda_0)/\sigma$ for the t of the classical t test. Finally, $u(\lambda \mid H_1)$ is the prior probability density of λ under the alternative hypothesis.

If $u(\lambda \mid H_1)$ is gentle in the neighborhood of x and not too violent else-

where, a reasonable approximation to Equation 14, akin to the principle of stable estimation, is

$$L(\lambda_0; x) = \frac{\varphi(t)}{\sigma u(x)}. \qquad [15]$$

According to a slight variation of the principle, already used in the Bernoullian example, near linearity may justify this approximation even better than near constancy does. Since σ is measured in the same units as x or λ, say, degrees centigrade or cycles per second, and $u(x)$ is probability per degree centigrade or per cycle per second, the product $\sigma u(x)$ (in the denominator of Equation 15) is dimensionless. Visualizing $\sigma u(x)$ as a rectangle of base σ, centered at x, and height $u(x)$, we see $\sigma u(x)$ to be approximately your prior probability for an interval of length σ in the region most favored by the data.

Consider an example. If $\lambda_0 = 0$ and $\sigma = 1$, then an observation of 2.58 would be significantly different from the null hypothesis at the .01 level of a classical two-tailed t test. If your alternative density were uniform over the range -21 to $+21$, then its average height would be about .024. But it is not uniform, and it is presumably somewhat higher near 0 than it is farther away. Perhaps under the alternative hypothesis, you would distinctly not attach more than 1/20 prior probability to any region one unit wide, and do attach about that much prior probability to such intervals in the immediate vicinity of the null value. According to the table of normal ordinates, $\varphi(2.58) = .0143$, so the likelihood ratio is about .286. Thus for the Bayesian, as for the classical statistician, the evidence here tells against the null hypothesis, but the Bayesian is not nearly so strongly persuaded as the classical statistician appears to be. The datum 1.96 is just significant at the .05 level of a two-tailed test. But the likelihood ratio is 1.17. This datum, which leads to a .05 classical rejection, leaves the Bayesian, with the prior opinion postulated, a shade more confident of the null hypothesis than he was to start with. The overreadiness of classical procedures to reject null hypotheses, first illustrated in the Bernoullian example, is seen again here; indeed, the two examples are really much the same in almost all respects. This sort of calculation, incidentally, is a more rigorous equivalent of the intuitive argument given just before the discussion of the Bernoullian example.

Lower bounds on L. An alternative when $u(\lambda \mid H_1)$ is not diffuse enough to justify stable estimation is to seek bounds on L. Imagine all the density under the alternative hypothesis concentrated at x, the place most favored by the data. The likelihood ratio is then

$$L_{\min} = \frac{\varphi(t)}{\varphi(0)} = e^{-\frac{1}{2}t^2}.$$

This is of course the very smallest likelihood ratio that can be associated with t. Since the alternative hypothesis now has all its density on one side of the null hypothesis, it is perhaps appropriate to compare the outcome of this procedure with the outcome of a one-tailed rather than a two-tailed classical test. At the one-tailed classical .05, .01, and .001 points, $L_{\min}$ is .26, .066, and .0085, respectively. Even the utmost generosity to the alternative hypothesis cannot make the evidence in favor of it as strong as classical significance levels might suggest. Incidentally, the situation is little different for a two-tailed classical test and a prior distribution for the alternative hypothesis concentrated symmetrically at a pair of points straddling the null value. If the prior distribution under the alternative hypothesis is required to be not only symmetric around the null value but also unimodal, which seems very safe for many problems, then the results are too similar to those obtained later for the smallest possible likelihood ratio obtainable with a symmetrical normal prior density to merit separate presentation here.

If you know that your prior density $u(\lambda \mid H_1)$ never exceeds some upper bound u^*, you can

improve, that is, increase, the crude lower bound $L_{\min}$. The prior distribution most favorable to the alternative hypothesis, given that it nowhere exceeds u^*, is a rectangular distribution of height u^* with x as its midpoint. Therefore

$$L(\lambda_0; x) \geqq \frac{\varphi(t)}{\sigma u^* \left[\Phi\left(\frac{1}{2\sigma u^*}\right) - \Phi\left(-\frac{1}{2\sigma u^*}\right)\right]} \geqq L_{\min} \quad [16]$$

where Φ is the standard normal cumulative function. Not only is this lower bound better than $L_{\min}$, no matter how large u^*, it also improves with decreasing σ, as is realistic. The improvement over $L_{\min}$ is negligible if $\sigma u^* \geqq 0.7$.

Either directly or by recognizing that the square bracket in Inequality 16 is less than 1, it is easy to derive a cruder but simpler bound, which is sometimes better than $L_{\min}$,

$$L(\lambda_0; x) \geqq \frac{\varphi(t)}{\sigma u^*}. \quad [17]$$

A counterpart of this more elementary bound was exhibited in the Bernoullian example. When σu^* is less than about .2, the square bracket in Inequality 16 is negligibly different from 1, so Inequality 16 reduces to Inequality 17.

In the present example, perhaps assignment of a probability as high as .1 to any interval as short as one standard deviation, given that light does materially affect frequency discrimination, may be distinctly contrary to your actual opinion. If so, you are entitled to apply Inequality 16 (and of course also Inequality 17) with $u^* = .1$ and $\sigma = 1$. The minimal likelihood ratios obtained from Inequality 16 (with $\sigma u^* = .1$) corresponding to values of t just significant at the .05, .01, and .001 levels by classical two-tailed tests are .58, .14, and .018, respectively. These bounds, though still not high, are considerably higher than $L_{\min}$.

Upper bounds on L. In order to discredit a null hypothesis, it is useful to find a practical upper bound on the likelihood ratio L, which can result in the conclusion that L is very small. It is impossible that $u(\lambda|H_1)$ should exceed some positive number for all λ, but you may well know plainly that $u(\lambda|H_1) \geqq u_* > 0$ for all λ in some interval, say of length 4, centered at x. In this case,

$$L(\lambda_0; x) \leqq \frac{\varphi(t)}{\int_{-2}^{+2} \varphi\left(\frac{x-\lambda}{\sigma}\right) u(\lambda|H_1) d\lambda} \leqq \frac{\varphi(t)}{\sigma u_* [\Phi(2) - \Phi(-2)]} \leqq \frac{1.05\, e^{-\frac{1}{2}t^2}}{\sqrt{2\Pi}\, \sigma u_*} \leqq \frac{0.42\, e^{-\frac{1}{2}t^2}}{\sigma u_*} = \frac{0.42\, L_{min}}{\sigma u_*}.$$

If, for example, you attach as much probability as .01 to the intervals of length σ near x, your likelihood ratio is at most 42 $L_{\min}$.

For t's classically significant at the .05, .01, and .001 levels, your likelihood ratio is correspondingly at most 10.9, 2.8, and .36. This procedure can discredit null hypotheses quite strongly; t's of 4 and 5 lead to upper bounds on your likelihood ratio of .014 and .00016, insofar as the normal model can be taken seriously for such large t's.

Normal alternative priors. Since normal densities are conjugate to normal measurements, it is natural to study the assumption that $u(\lambda|H_1)$ is a normal density. This assumption may frequently be adequate as an approximation, and its relative mathematical simplicity paves the way to valuable insights that may later be substantiated with less arbitrary assumptions. In this paper we explore not all normal alternative priors but only those symmetrical about λ_0, which seem especially important.

Let $u(\lambda|H_1)$, then, be normal with mean λ_0 and with some standard deviation τ. Equa-

tion 14 now specializes to

$$L(\lambda_0; x) = \frac{\frac{1}{\sigma}\varphi(t)}{\frac{1}{\sqrt{\sigma^2+\tau^2}}\varphi\left(\frac{x-\lambda_0}{\sqrt{\sigma^2+\tau^2}}\right)} = \frac{\varphi(t)}{\alpha\varphi(\alpha t)}, \quad [18]$$

where

$$\alpha = \frac{\sigma}{\sqrt{\sigma^2+\tau^2}} = \frac{1}{\sqrt{1+(\tau/\sigma)^2}}$$

Plainly, α is a function of σ/τ and vice versa; for small values of either, the difference between α and σ/τ is negligible. We emphasize α rather than the intuitively more appealing σ/τ because α leads to simpler equations. Of course, α is less than one, typically much less. Writing the normal density in explicit form,

$$L(\alpha, t) = \frac{1}{\alpha}\exp -\tfrac{1}{2}(1-\alpha^2)t^2. \quad [19]$$

Table 2 shows numerical values of $L(\alpha, t)$ for some instructive values of α and for values of t corresponding to familiar two-tailed classical significance levels. The values of α between .01 and .1 portray reasonably precise experiments; the others included in Table 2 are instructive as extreme possibilities. Table 2 again illustrates how classically significant values of t can, in realistic cases, be based on data that actually favor the null hypothesis.

For another comparison of Equation 18 with classical tests consider that (positive) value t_0 of t for which L is 1. If L is 1, then the posterior odds for the two hypotheses will equal the prior odds; the experiment will leave opinion about H_0 and H_1 unchanged, though it is bound to influence opinion about λ given H_1. Taking natural logarithms of Equation 19 for $t = t_0$,

$$\ln\frac{1}{\alpha} - \frac{1}{2}(1-\alpha^2)t_0^2 = 0,$$

$$t_0 = \left\{\frac{-\ln\alpha^2}{1-\alpha^2}\right\}^{\frac{1}{2}} \quad [20]$$

If α is small, say less than .1, then $1-\alpha^2$ is negligibly different from 1, and so $t_0 \simeq \sqrt{-\ln\alpha^2}$. The effect of using this approximation can never be very bad; for the likelihood ratio actually associated with the approximate value of t_0 cannot be less than 1 or greater than 1.202. Table 3 presents a few actual values of t_0 and their corresponding two-tailed significance levels. At values of t slightly smaller than the break-even values in Table 3 classical statistics more or less vigorously rejects the null hypothesis, though the Bayesian described by α becomes more confident of it than he was to start with.

If $t = 0$, that is, if the observation happens to point exactly to the null hypothesis, $L = \frac{1}{\alpha}$; thus support for the null hypothesis can be very strong, since α might well be about .01. In the example, you perhaps hope to confirm the null hypothesis to everyone's satisfaction, if it is in fact true. You will

TABLE 2

VALUES OF $L(\alpha, t)$ FOR SELECTED VALUES OF α AND FOR VALUES OF t CORRESPONDING TO FAMILIAR TWO-TAILED SIGNIFICANCE LEVELS

α	σ/τ	t and Significance level				
		1.645 .10	1.960 .05	2.576 .01	3.291 .001	3.891 .0001
.0001	.0001	2,585	1,465	362	44.6	5.16
.001	.0010	259	147	36.2	4.46	.516
.01	.0100	25.9	14.7	3.63	.446	.0516
.025	.0250	10.4	5.87	1.45	.179	.0207
.05	.0501	5.19	2.94	.731	.0903	.0105
.075	.0752	3.47	1.97	.492	.0612	.00718
.1	.1005	2.62	1.49	.375	.0470	.00556
.15	.1517	1.78	1.02	.260	.0336	.00408
.2	.2041	1.36	.791	.207	.0277	.00349
.5	.5774	.725	.474	.166	.0345	.00685
.9	2.0647	.859	.771	.592	.397	.264
.99	7.0179	.983	.972	.946	.907	.869

TABLE 3

VALUES OF t_0 AND THEIR SIGNIFICANCE LEVELS FOR NORMAL ALTERNATIVE PRIOR DISTRIBUTIONS FOR SELECTED VALUES OF α

α	t_0	Significance level
.1	2.157	.031
.05	2.451	.014
.01	3.035	.0024
.001	3.718	.00020
.0001	4.292	.000018

therefore try hard to make σ small enough so that your own α and those of your critics will be small. In the Weber-fraction example, $\alpha \simeq .077$ (calculated by assuming that 90% of the prior probability under the alternative hypothesis falls between -21 and $+21$; assuming normality, it follows that $\tau \simeq 12.9$). If $t = 0$, then L is 12.9—persuasive but not irresistible evidence in favor of the null hypothesis. For $\alpha = .077$, t_0 is 2.3—just about the .02 level of a classical two-tailed test. Conclusion: An experiment strong enough to lend strong support to the null hypothesis when $t = 0$ will mildly support the null hypothesis even when classical tests would strongly reject it.

If you are seriously interested in supporting the null hypothesis if it is true—and you may well be, valid aphorisms about the perishability of hypotheses notwithstanding—you should so design your experiment that even a t as large as 2 or 3 strongly confirms the null hypothesis. If α is .0001, L is more than 100 for any t between -3 and $+3$. Such small α's do not occur every day, but they are possible. Maxwell's prediction of the equality of the "two speeds of light" might be an example. A more practical way to prove a null hypothesis may be to investigate several, not just one of its numerical consequences. It is not clear just what sort of evidence classical statistics would regard as strong confirmation of a null hypothesis. (See however Berkson, 1942.)

What is the smallest likelihood ratio L_{normin} (the minimum L for a symmetrical normal prior) that can be attained for a given t by artificial choice of α? It follows from Equation 19 that L is minimized at $\alpha = |t|^{-1}$, provided $|t| \geqq 1$, and at the unattainable value $\alpha = 1$, otherwise.

$$L_{\text{normin}} = e^{\frac{1}{2}}|t|e^{-\frac{1}{2}t^2} = 1.65|t|e^{-\frac{1}{2}t^2} \text{ for } |t| \geqq 1$$
$$= 1 \text{ for } |t| \leqq 1.$$

With any symmetric normal prior, any $|t| \leqq 1$ speaks for the null hypothesis. So L_{normin} exceeds L_{min} in all cases and exceeds it by the substantial factor 1.65 $|t|$ if $|t| \geqq 1$. Values of t corresponding to familiar two-tailed significance levels and the corresponding values of L_{normin} are shown in Table 4.

From this examination of one-dimensional normally distributed observations, we conclude that a t of 2 or 3 may not be evidence against the null hypothesis at all, and seldom if ever justifies much new confidence in the alternative hypothesis. This conclusion has a melancholy side. The justification for the assumption of normal measurements must in the last analysis be empirical. Few applications are likely to justify using numerical values of normal ordinates more than three standard deviations away from the mean. And yet without those numerical values, the methods of this section are not applicable. In short, in one-dimensional normal cases, evidence that does not justify rejection of the null hypothesis by the interocular traumatic test is unlikely to justify firm rejection at all.

Haunts of χ^2 and F. Classical tests of null hypotheses invoking the χ^2, and closely related F, distributions are so familiar that something must be said here about their Bayesian counterparts. Though often deceptively oversimplified, the branches of statistics that come together here are

TABLE 4

VALUES OF L_{normin} AND OF L_{min} FOR VALUES OF t CORRESPONDING TO FAMILIAR TWO-TAILED SIGNIFICANCE LEVELS

t	Significance level	L_{normin}	L_{min}
1.960	.05	.473	.146
2.576	.01	.154	.0362
3.291	.001	.0241	.00445
3.891	.0001	.00331	.000516

immense and still full of fundamental mysteries for Bayesians and classicists alike (Fisher, 1925, see Ch. 4 and 5 in the 1954 edition; Green & Tukey, 1960; Scheffé, 1959; Tukey, 1962). We must therefore confine ourselves to the barest suggestions.

Much of the subject can be reduced to testing whether several parameters λ_i measured independently with known variance σ^2 have a specified common value. This multidimensional extension of the one-dimensional normal problem treated in the last section is so important that we shall return to it shortly.

As is well known, the statistical theory of multidimensional normal measurement embraces in a grand generalization that of normal regression and Model I analysis of variance (and covariance); a host of other topics can more or less faithfully be reduced to it by approximation (Cramér, 1946, Ch. 29; Fisher, 1925, see Ch. 5 in the 1954 edition; Raiffa & Schlaifer, 1961).

Approximation of multinomial by multidimensional normal measurements has also been the main approach to that large domain which classically evokes χ^2 tests of association and goodness of fit (Cramér, 1946, Ch. 30; Fisher, 1925, see Ch. 4 in the 1954 edition; Jeffreys, 1939, see Section 4.1 in the 1961 edition). We shall not attempt to enter into this topic here, but the suitably prepared reader will find the approximation, and the references just cited, helpful.

One prominent classical application of the F distributions is testing whether two variances of normally distributed measurements are equal, as in Model II analysis of variance. The interested reader will easily see what the Bayesian counterpart of this test is from examples of tests in earlier sections of this paper and from the discussion of Bayesian applications of the F distributions in Chapter 12 of Raiffa and Schlaifer (1961).

About the very important topic of Model III analysis of variance, that is, analysis of variance ostensibly justified by the randomized allocation of treatments, we can say only that it is by no means so straightforward as is sometimes believed (Savage et al., 1962, pp. 33, 34, 87–92, and references cited there).

Multidimensional normal measurements and a null hypothesis. For those who may be interested in some relatively technical and tentative suggestions, we return in this section to the basic multidimensional normal testing problem that was defined in the last section.

For simplicity, and with the same justification as in the one-dimensional case, we shall assume that the variance σ^2 is known. The extension to unknown variance, in which the multivariate normal distribution is replaced by multivariate t distributions and the χ^2 distributions are replaced by F distributions will be clear to many readers, especially on reference to Chapter 12 of Raiffa and Schlaifer (1961).

Let λ be an unknown vector in n-dimensional Euclidean space, and suppose that, given λ, the measurement x is a vector spherically normally distributed around λ with known variance σ^2. The likelihood ratio for the null hypothesis that $\lambda = \lambda_0$ is then evidently

$$L(\lambda_0; x) = \frac{\sigma^{-n}\varphi\left(\frac{x - \lambda_0}{\sigma}\right)}{\sigma^{-n}\int \varphi\left(\frac{x - \lambda}{\sigma}\right) u(\lambda \mid H_1) d\lambda}, \qquad [21]$$

where φ is the standard n dimensional normal density. Equation 21 simply does in n dimensions what Equation 14 did in one. The n dimensional generalizations of the suggestions already made for appraising L in the one-dimensional problem are so natural that we shall be able to indicate them very briefly, and we shall hardly introduce any essentially new suggestions here. There is one important practical change with increasing n; certain methods that would be frequently applicable for small n become increasingly useless with large n.

If $u(\lambda|H_1)$ is sufficiently gentle and is approximately equal to u' near λ_0, then, in analogy with Equation 15, the ideas of stable estimation permit the approximation

$$L(\lambda_0; x) = \frac{e^{-\frac{1}{2}\chi^2}}{\sigma^n u'} \qquad [22]$$

where χ^2 is written instead of t^2 for the square of the length of the vector $x - \lambda_0$ divided by σ^2, as is usual when n is not necessarily 1.

As n increases, conditions for the applicability of Equation 22 will be encountered more and more rarely. For one reason, the sphere about x within which $u(\lambda|H_1)$ has to be nearly constant has radius somewhat larger than $\sigma\sqrt{n}$, and the larger that sphere, the less plausible the assumption of constancy within it. Still worse, the spheres within which this density can reasonably be expected to remain nearly constant will typically actually decrease in radius with increasing n. For example, in a study of three factors, each at four levels, the first-order interactions are expressed by 27 parameters. To say that your opinion of these is diffuse with respect to some standard deviation σ implies, among other things, that even if you found out any 26 of the parameters you would not feel competent to guess the last one to within several σ's. Even given the hypothesis that the interactions have no tendency to be small, it is hard to envisage situations in which the implication would be realistic. This example serves incidentally to remind us that there are often many "null" hypotheses claiming some measure of our credence. For example: all interactions vanish; all that involve the first factor vanish; all above those of the first order vanish, and the first-order interactions are well explained by this or that simple theory; and so on. In principle, these problems of multiple decision are natural outgrowths of the two-hypothesis situation, but much work remains to be done on them.

For a specified x, the prior distribution most pessimistic toward the null hypothesis is once more concentrated at λ_0 and yields

$$L_{\min} = e^{-\frac{1}{2}\chi^2}.$$

If n is large, say at least 10, and the null hypothesis is true, then it is almost certain (before making the measurement x) that χ^2 will be roughly equal to n. So, for large n, $L_{\min}$ is very small indeed, even when compared with significance levels of classical tests applied to the same data. Therefore, $L_{\min}$ will be of almost no practical use in such cases.

A somewhat more realistic approach in the general spirit of $L_{\min}$ would be to consider that spherically symmetrical distribution which would most discredit the null hypothesis. This approach might be worth some exploration, but is mathematically rather intractable.

The subjective upper and lower bounds for L that were illustrated in one dimension are easy to generalize to n dimensions. They may well prove less serviceable as n increases, but they merit trial and study.

We close this section with a sketchy report of what happens when $u(\lambda|H_1)$ is itself a spherical normal distribution about λ_0, with variance τ^2. We do so with particular diffidence, because there is here even less justification than before in hoping to approximate $u(\lambda|H_1)$ by a normal distribution centered at λ_0, and because the assumption of spherical symmetry for this distribution will often be particularly unrealistic. Still, we hope that the exercise, regarded with caution, will be suggestive of truth which can later be verified in some more secure way.

TABLE 5

VALUES OF $n_{.001}(\alpha)$ FOR WHICH THE BREAK-EVEN VALUE χ_0^2 IS JUST SIGNIFICANT AT THE .001 LEVEL FOR SELECTED VALUES OF α

α	σ/τ	χ_0^2	$n_{.001}(\alpha)$
.01	.010	18.4	2
.1	.101	18.6	4
.2	.204	26.8	8
.5	.577	73.9	40
.8	1.333	470	379
.9	2.065	1,896	1,710

Letting α be, as before, $\sigma/\sqrt{\sigma^2 + \tau^2}$, Equation 21 becomes

$$L = \frac{1}{\alpha^n} \exp -\frac{1}{2}\chi^2(1 - \alpha^2).$$

For a fixed fraction α and very large n, χ^2 is initially almost certain, given H_0, to be within a few percent of n and, given H_1, within a few percent of n/α^2. As follows easily, it is initially almost sure that the experiment will firmly lead to a correct decision between H_0 and H_1, no matter how close α is to 1, provided n is sufficiently large. For this reason, if for no other, we are bound to be interested in values of α for large n that correspond to values of σ/τ so large that they would render the experiment worthless if n were small.

The value of χ^2 that speaks neither for nor against the null hypothesis for a specified α is

$$\chi_0^2 = \frac{-n \ln \alpha^2}{1 - \alpha^2},$$

an easy and natural generalization of Equation 20. For large n, it is not reasonable to approximate χ_0^2 by substituting 1 for $1 - \alpha^2$. Since the coefficient of n in χ_0^2 is larger than 1 for every fraction α and since the value of χ^2 that is just significant at say, the .001 level only slightly exceeds n for sufficiently large n, there is some first integer $n_{.001}(\alpha)$ at which the break-even value χ_0^2 is just significant at the .001 level. Some representative values are shown in Table 5. From the point of view of this model of the testing situation, which is of course not unobjectionable, the classical procedure is startlingly prone to reject the null hypothesis contrary to what would often be very reasonable opinion.

Paralleling the situation for $n = 1$, it is $\alpha = \sqrt{n/\chi^2}$ that is most pessimistic toward the null hypothesis for a specified value of χ^2. The likelihood for this artificial value of α is

$$L_{\text{normin}} = \left(\frac{e\chi^2}{n}\right)^{n/2} e^{-\chi^2/2}.$$

Table 6 shows the values of L_{normin} that correspond to the values of χ^2 just significant at the .01 and .001 levels for several values of n. Here as in the one-dimensional case L_{normin} is small, but not as small as classical significance levels might suggest. In all these cases α is unrealistically large.

This cursory glance at multidimensional normally distributed observations has the same general conclusions as our more detailed study of the unidimensional normal case. Although the statistical theory of multi-

TABLE 6

VALUES OF L_{normin} THAT CORRESPOND TO THE VALUES OF χ^2 JUST SIGNIFICANT AT THE .01 AND .001 LEVELS FOR SELECTED VALUES OF n

n	$\chi_{.01}^2$			$\chi_{.001}^2$		
	α	σ/τ	L_{normin}	α	σ/τ	L_{normin}
1	.388	.421	.1539	.339	.360	.0242
3	.514	.600	.1134	.429	.475	.0166
10	.656	.870	.0912	.581	.715	.0127
30	.768	1.198	.0806	.709	1.005	.0108
100	.858	2.075	.0742	.818	1.422	.0097
300	.913	2.238	.0712	.887	1.919	.0092
1,000	.950	3.059	.0696	.935	2.636	.0088
3,000	.971	4.047	.0680	.962	3.499	.0087
10,000	.984	5.488	.0675	.979	3.798	.0086
∞	1.000	∞	.0668	1.000	∞	.0084

dimensional observations (classical or Bayesian) is distressingly sketchy and incomplete, drastic surprises about the relation between classical and Bayesian multidimensional techniques have not turned up and now seem unlikely.

Some morals about testing sharp null hypotheses. At first glance, our general conclusion that classical procedures are so ready to discredit null hypotheses that they may well reject one on the basis of evidence which is in its favor, even strikingly so, may suggest the presence of a mathematical mistake somewhere. Not so; the contradiction is practical, not mathematical. A classical rejection of a true null hypothesis at the .05 level will occur only once in 20 times. The overwhelming majority of these false classical rejections will be based on test statistics close to the borderline value; it will often be easy to demonstrate that these borderline test statistics, unlikely under either hypothesis, are nevertheless more unlikely under the alternative than under the null hypothesis, and so speak for the null hypothesis rather than against it.

Bayesian procedures can strengthen a null hypothesis, not only weaken it, whereas classical theory is curiously asymmetric. If the null hypothesis is classically rejected, the alternative hypothesis is willingly embraced, but if the null hypothesis is not rejected, it remains in a kind of limbo of suspended disbelief. This asymmetry has led to considerable argument about the appropriateness of testing a theory by using its predictions as a null hypothesis (Grant, 1962; Guilford, 1942, see p. 186 in the 1956 edition; Rozeboom, 1960; Sterling, 1960). For Bayesians, the problem vanishes, though they must remember that the null hypothesis is really a hazily defined small region rather than a point.

The procedures which have been presented simply compute the likelihood ratio of the hypothesis that some parameter is very nearly a specified single value with respect to the hypothesis that it is not. They do not depend on the assumption of special initial credibility of the null hypothesis. And the general conclusion that classical procedures are unduly ready to reject null hypotheses is thus true whether or not the null hypothesis is especially plausible a priori. At least for Bayesian statisticians, however, no procedure for testing a sharp null hypothesis is likely to be appropriate unless the null hypothesis deserves special initial credence. It is uninteresting to learn that the odds in favor of the null hypothesis have increased or decreased a hundredfold if initially they were negligibly different from zero.

How often are Bayesian and classical procedures likely to lead to different conclusions in practice? First, Bayesians are unlikely to consider a sharp null hypothesis nearly so often as do the consumers of classical statistics. Such procedures make sense to a Bayesian only when his prior distribution has a sharp spike at some specific value; such prior distributions do occur, but not so often as do classical null hypothesis tests.

When Bayesians and classicists agree that null hypothesis testing is appropriate, the results of their procedures will usually agree also. If the null hypothesis is false, the interocular traumatic test will often suffice to reject it; calculation will serve only to verify clear intuition. If the null hypothesis is true, the interocular traumatic test is unlikely to be of much use in one-dimensional cases, but may be helpful in multidimensional ones. In at least 95% of cases

when the null hypothesis is true, Bayesian procedures and the classical .05 level test agree. Only in borderline cases will the two lead to conflicting conclusions. The widespread custom of reporting the highest classical significance level from among the conventional ones actually attained would permit an estimate of the frequency of borderline cases in published work; any rejection at the .05 or .01 level is likely to be borderline. Such an estimate of the number of borderline cases may be low, since it is possible that many results not significant at even the .05 level remain unpublished.

The main practical consequences for null hypothesis testing of widespread adoption of Bayesian statistics will presumably be a substantial reduction in the resort to such tests and a decrease in the probability of rejecting true null hypotheses, without substantial increase in the probability of accepting false ones.

If classical significance tests have rather frequently rejected true null hypotheses without real evidence, why have they survived so long and so dominated certain empirical sciences? Four remarks seem to shed some light on this important and difficult question.

1. In principle, many of the rejections at the .05 level are based on values of the test statistic far beyond the borderline, and so correspond to almost unequivocal evidence. In practice, this argument loses much of its force. It has become customary to reject a null hypothesis at the highest significance level among the magic values, .05, .01, and .001, which the test statistic permits, rather than to choose a significance level in advance and reject all hypotheses whose test statistics fall beyond the criterion value specified by the chosen significance level. So a .05 level rejection today usually means that the test statistic was significant at the .05 level but not at the .01 level. Still, a test statistic which falls just short of the .01 level may correspond to much stronger evidence against a null hypothesis than one barely significant at the .05 level. The point applies more forcibly to the region between .01 and .001, and for the region beyond, the argument reverts to its original form.

2. Important rejections at the .05 or .01 levels based on test statistics which would not have been significant at higher levels are not common. Psychologists tend to run relatively large experiments, and to get very highly significant main effects. The place where .05 level rejections are most common is in testing interactions in analyses of variance—and few experimenters take those tests very seriously, unless several lines of evidence point to the same conclusions.

3. Attempts to replicate a result are rather rare, so few null hypothesis rejections are subjected to an empirical check. When such a check is performed and fails, explanation of the anomaly almost always centers on experimental design, minor variations in technique, and so forth, rather than on the meaning of the statistical procedures used in the original study.

4. Classical procedures sometimes test null hypotheses that no one would believe for a moment, no matter what the data; our list of situations that might stimulate hypothesis tests earlier in the section included several examples. Testing an unbelievable null hypothesis amounts, in practice, to assigning an unreasonably large prior probability to a very small region of possible values of the true parameter. In such cases, the more

the procedure is biased against the null hypothesis, the better. The frequent reluctance of empirical scientists to accept null hypotheses which their data do not classically reject suggests their appropriate skepticism about the original plausibility of these null hypotheses.

LIKELIHOOD PRINCIPLE

A natural question about Bayes' theorem leads to an important conclusion, the likelihood principle, which was first discovered by certain classical statisticians (Barnard, 1947; Fisher, 1956).

Two possible experimental outcomes D and D'—not necessarily of the same experiment—can have the same (potential) bearing on your opinion about a partition of events H_i, that is, $P(H_i|D)$ can equal $P(H_i|D')$ for each i. Just when are D and D' thus evidentially equivalent, or of the same import? Analytically, when is

$$[P(H_i|D) =] \frac{P(D|H_i)P(H_i)}{P(D)} = \frac{P(D'|H_i)P(H_i)}{P(D')} [= P(H_i|D')] \quad [23]$$

for each i?

Aside from such academic possibilities as that some of the $P(H_i)$ are 0, Equation 23 plainly entails that, for some positive constant k and for all i,

$$P(D'|H_i) = kP(D|H_i). \quad [24]$$

But Equation 24 implies Equation 23, from which it was derived, no matter what the initial probabilities $P(H_i)$ are, as is easily seen thus:

$$\begin{aligned} P(D') &= \Sigma P(D'|H_i)P(H_i) \\ &= k\Sigma P(D|H_i)P(H_i) \\ &= kP(D). \end{aligned}$$

This conclusion is the likelihood principle: Two (potential) data D and D' are of the same import if Equation 24 obtains.

Since for the purpose of drawing inference, the sequence of numbers $P(D|H_i)$ is, according to the likelihood principle, equivalent to any other sequence obtained from it by multiplication by a positive constant, a name for this class of equivalent sequences is useful and there is precedent for calling it the likelihood (of the sequence of hypotheses H_i given the datum D). (This is not quite the usage of Raiffa & Schlaifer, 1961.) The likelihood principle can now be expressed thus: D and D' have the same import if $P(D|H_i)$ and $P(D'|H_i)$ belong to the same likelihood—more idiomatically, if D and D' have the same likelihood.

If, for instance, the partition is twofold, as it is when you are testing a null hypothesis against an alternative hypothesis, then the likelihood to which the pair $[P(D|H_0), P(D|H_1)]$ belongs is plainly the set of pairs of numbers $[a, b]$ such that the fraction a/b is the already familiar likelihood ratio $L(H_0; D) = P(D|H_0)/P(D|H_1)$. The simplification of the theory of testing by the use of likelihood ratios in place of the pairs of conditional probabilities, which we have seen, is thus an application of the likelihood principle.

Of course, the likelihood principle applies to a (possibly multidimensional) parameter λ as well as to a partition H_i. The likelihood of D, or the likelihood to which $P(D|\lambda)$ belongs, is the class of all those functions of λ that are positive constant multiples of (that is, proportional to) the function $P(D|\lambda)$. Also, conditional densities can replace conditional probabilities in the definition of likelihood ratios.

There is one implication of the like-

lihood principle that all statisticians seem to accept. It is not appropriate in this paper to pursue this implication, which might be called the principle of sufficient statistics, very far. One application of sufficient statistics so familiar as almost to escape notice will, however, help bring out the meaning of the likelihood principle. Suppose a sequence of 100 Bernoulli trials is undertaken and 20 successes and 80 failures are recorded. What is the datum, and what is its probability for a given value of the frequency p? We are all perhaps overtrained to reply, "The datum is 20 successes out of 100, and its probability, given p, is $C_{20}^{100}p^{20}(1-p)^{80}$." Yet it seems more correct to say, "The datum is this particular sequence of successes and failures, and its probability, given p, is $p^{20}(1-p)^{80}$." The conventional reply is often more convenient, because it would be costly to transmit the entire sequence of observations; it is permissible, because the two functions $C_{20}^{100}p^{20}(1-p)^{80}$ and $p^{20}(1-p)^{80}$ belong to the same likelihood; they differ only by the constant factor C_{20}^{100}. Many classical statisticians would demonstrate this permissibility by an argument that does not use the likelihood principle, at least not explicitly (Halmos & Savage, 1949, p. 235). That the two arguments are much the same, after all, is suggested by Birnbaum (1962). The legitimacy of condensing the datum is often expressed by saying that the number of successes in a given number of Bernoulli trials is a sufficient statistic for the sequence of trials. Insofar as the sequence of trials is not altogether accepted as Bernoullian—and it never is—the condensation is not legitimate. The practical experimenter always has some incentive to look over the sequence of his data with a view to discovering periodicities, trends, or other departures from Bernoullian expectation. Anyone to whom the sequence is not available, such as the reader of a condensed report or the experimentalist who depends on automatic counters, will reserve some doubt about the interpretation of the ostensibly sufficient statistic.

Moving forward to another application of the likelihood principle, imagine a different Bernoullian experiment in which you have undertaken to continue the trials until 20 successes were accumulated and the twentieth success happened to be the one hundredth trial. It would be conventional and justifiable to report only this fact, ignoring other details of the sequence of trials. The probability that the twentieth success will be the one hundredth trial is, given p, easily seen to be $C_{19}^{99}p^{20}(1-p)^{80}$. This is exactly 1/5 of the probability of 20 successes in 100 trials, so according to the likelihood principle, the two data have the same import. This conclusion is even a trifle more immediate if the data are not condensed; for a specific sequence of 100 trials of which the last is the twentieth success has the probability $p^{20}(1-p)^{80}$ in both experiments. Those who do not accept the likelihood principle believe that the probabilities of sequences that might have occurred, but did not, somehow affect the import of the sequence that did occur.

In general, suppose that you collect data of any kind whatsoever—not necessarily Bernoullian, nor identically distributed, nor independent of each other given the parameter λ—stopping only when the data thus far collected satisfy some criterion of a sort that is sure to be satisfied sooner or later, then the import of the sequence of n data actually observed will be exactly the same as it would be had you

planned to take exactly n observations in the first place. It is not even necessary that you stop according to a plan. You may stop when tired, when interrupted by the telephone, when you run out of money, when you have the casual impression that you have enough data to prove your point, and so on. The one proviso is that the moment at which your observation is interrupted must not in itself be any clue to λ that adds anything to the information in the data already at hand. A man who wanted to know how frequently lions watered at a certain pool was chased away by lions before he actually saw any of them watering there; in trying to conclude how many lions do water there he should remember why his observation was interrupted when it was. We would not give a facetious example had we been able to think of a serious one. A more technical discussion of the irrelevance of stopping rules to statistical analysis is on pages 36–42 of Raiffa and Schlaifer (1961).

This irrelevance of stopping rules to statistical inference restores a simplicity and freedom to experimental design that had been lost by classical emphasis on significance levels (in the sense of Neyman and Pearson) and on other concepts that are affected by stopping rules. Many experimenters would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience. Classical statisticians (except possibly for the few classical defenders of the likelihood principle) have frowned on collecting data one by one or in batches, testing the total ensemble after each new item or batch is collected, and stopping the experiment only when a null hypothesis is rejected at some preset significance level. And indeed if an experimenter uses this procedure, then with probability 1 he will eventually reject any sharp null hypothesis, even though it be true. This is perhaps simply another illustration of the overreadiness of classical procedures to reject null hypotheses. In contrast, if you set out to collect data until your posterior probability for a hypothesis which unknown to you is true has been reduced to .01, then 99 times out of 100 you will never make it, no matter how many data you, or your children after you, may collect. (Rules which have nonzero probability of running forever ought not, and here will not, be called stopping rules at all.)

The irrelevance of stopping rules is one respect in which Bayesian procedures are more objective than classical ones. Classical procedures (with the possible exceptions implied above) insist that the intentions of the experimenter are crucial to the interpretation of data, that 20 successes in 100 observations means something quite different if the experimenter intended the 20 successes than if he intended the 100 observations. According to the likelihood principle, data analysis stands on its own feet. The intentions of the experimenter are irrelevant to the interpretation of the data once collected, though of course they are crucial to the design of experiments.

The likelihood principle also creates unity and simplicity in inference about Markov chains and other stochastic processes (Barnard, Jenkins, & Winsten, 1962), which are sometimes applied in psychology. It sheds light on many other problems of statistics, such as the role of unbiasedness and Fisher's concept of ancillary statistic. A principle so simple with consequences so pervasive is bound to be controversial. For dissents see Stein (1962), Wolfowitz

(1962), and discussions published with Barnard, Jenkins, and Winsten (1962), Birnbaum (1962), and Savage et al. (1962) indexed under likelihood principle.

IN RETROSPECT

Though the Bayesian view is a natural outgrowth of classical views, it must be clear by now that the distinction between them is important. Bayesian procedures are not merely another tool for the working scientist to add to his inventory along with traditional estimates of means, variances, and correlation coefficients, and the t test, F test, and so on. That classical and Bayesian statistics are sometimes incompatible was illustrated in the theory of testing. For, as we saw, evidence that leads to classical rejection of the null hypothesis will often leave a Bayesian more confident of that same null hypothesis than he was to start with. Incompatibility is also illustrated by the attention many classical statisticians give to stopping rules that Bayesians find irrelevant.

The Bayesian outlook is flexible, encouraging imagination and criticism in its everyday applications. Bayesian experimenters will emphasize suitably chosen descriptive statistics in their publications, enabling each reader to form his own conclusions. Where an experimenter can easily foresee that his readers will want the results of certain calculations (as for example when the data seem sufficiently precise to justify for most readers application of the principle of stable estimation) he will publish them. Adoption of the Bayesian outlook should discourage parading statistical procedures, Bayesian or other, as symbols of respectability pretending to give the imprimatur of mathematical logic to the subjective process of empirical inference.

We close with a practical rule which stands rather apart from any conflicts between Bayesian and classical statistics. The rule was somewhat overstated by a physicist who said, "As long as it takes statistics to find out, I prefer to investigate something else." Of course, even in physics some important questions must be investigated before technology is sufficiently developed to do so definitively. Still, when the value of doing so is recognized, it is often possible so to design experiments that the data speak for themselves without the intervention of subtle theory or insecure personal judgments. Estimation is best when it is stable. Rejection of a null hypothesis is best when it is interocular.

REFERENCES

ANSCOMBE, F. J. Bayesian statistics. *Amer. Statist.*, 1961, **15**(1), 21–24.

BAHADUR, R. R., & ROBBINS, H. The problem of the greater mean. *Ann. math. Statist.*, 1950, **21**, 469–487.

BARNARD, G. A. A review of "Sequential Analysis" by Abraham Wald. *J. Amer. Statist. Ass.*, 1947, **42**, 658–664.

BARNARD, G. A., JENKINS, G. M., & WINSTEN, C. B. Likelihood, inferences, and time series. *J. Roy. Statist. Soc.*, 1962, **125**(Ser. A), 321–372.

BAYES, T. Essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, 1763, **53**, 370–418. (Reprinted: *Biometrika*, 1958, **45**, 293–315.)

BERKSON, J. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Ass.*, 1938, **33**, 526–542.

BERKSON, J. Tests of significance considered as evidence. *J. Amer. Statist. Ass.*, 1942, **37**, 325–335.

BIRNBAUM, A. On the foundations of statistical inference. *J. Amer. Statist. Ass.*, 1962, **57**, 269–306.

BLACKWELL, D., & DUBINS, L. Merging of opinions with increasing information. *Ann. math. Statist.*, 1962, **33**, 882–886.

BOREL, E. La théorie du jeu et les équations intégrales à noyau symétrique. *CR Acad. Sci., Paris*, 1921, **173**, 1304–1308. (Trans. by L. J. Savage, *Econometrica*, 1953, **21**, 97–124)

BOREL, E. A propos d'un traité de probabilités. *Rev. Phil.*, 1924, **98**, 321–336. (Reprinted: In, *Valeur pratique et philosophie des probabilités*. Paris: Gauthier-Villars, 1939. Pp. 134–146)

BRIDGMAN, P. W. A critique of critical tables. *Proc. Nat. Acad. Sci.*, 1960, **46**, 1394–1401.

CRAMÉR, H. *Mathematical methods of statistics*. Princeton: Princeton Univer. Press, 1946.

DE FINETTI, B. Fondamenti logici del ragionamento probabilistico. *Boll. Un. mat. Ital.*, 1930, **9**(Ser. A), 258–261.

DE FINETTI, B. La prévision: Ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré*, 1937, **7**, 1–68.

DE FINETTI, B. La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti da vista. In, *Induzione e statistica*. Rome, Italy: Istituto Matematico dell'Universita, 1959.

DE FINETTI, B., & SAVAGE, L. J. Sul modo di scegliere le probabilità iniziali. In, *Biblioteca del "metron."* Ser. C, Vol. 1. *Sui fondamenti della statistica*. Rome: University of Rome, 1962. Pp. 81–154.

EDWARDS, W. Dynamic decision theory and probabilistic information processing. *Hum. Factors*, 1962, **4**, 59–73. (a)

EDWARDS, W. Subjective probabilities inferred from decisions. *Psychol. Rev.*, 1962, **69**, 109–135. (b)

EDWARDS, W. Probabilistic information processing in command and control systems. Report No. 3780-12-T, 1963, Institute of Science and Technology, University of Michigan.

FISHER, R. A. *Statistical methods for research workers*. (12th ed., 1954) Edinburgh: Oliver & Boyd, 1925.

FISHER, R. A. *Contributions to mathematical statistics*. New York: Wiley, 1950.

FISHER, R. A. *Statistical methods and scientific inference*. (2nd ed., 1959) Edinburgh: Oliver & Boyd, 1956.

GOOD, I. J. *Probability and the weighing of evidence*. New York: Hafner, 1950.

GOOD, I. J. Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *J. Roy. Statist. Soc.*, 1960, **22**(Ser. B), 319–331.

GRANT, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychol. Rev.*, 1962, **69**, 54–61.

GRAYSON, C. J., JR. *Decisions under uncertainty: Drilling decisions by oil and gas operators*. Boston: Harvard Univer. Press, 1960.

GREEN, B. J., JR., & TUKEY, J. W. Complex analysis of variance: General problems. *Psychometrika*, 1960, **25**, 127–152.

GUILFORD, J. P. *Fundamental statistics in psychology and education*. (3rd ed., 1956) New York: McGraw-Hill, 1942.

HALMOS, P. R., & SAVAGE, L. J. Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. math. Statist.*, 1949, **20**, 225–241.

HILDRETH, C. Bayesian statisticians and remote clients. *Econometrica*, 1963, **31**, in press.

HODGES, J. L., & LEHMANN, E. L. Testing the approximate validity of statistical hypotheses. *J. Roy. Statist. Soc.*, 1954, **16**(Ser. B), 261–268.

JEFFREYS, H. *Scientific inference*. (3rd ed., 1957) England: Cambridge Univer. Press, 1931.

JEFFREYS, H. *Theory of probability*. (3rd ed., 1961) Oxford, England: Clarendon, 1939.

KOOPMAN, B. O. The axioms and algebra of intuitive probability. *Ann. Math.*, 1940, **41**(Ser. 2), 269–292. (a)

KOOPMAN, B. O. The bases of probability. *Bull. Amer. Math. Soc.*, 1940, **46**, 763–774. (b)

KOOPMAN, B. O. Intuitive probabilities and sequences. *Ann. Math.*, 1941, **42**(Ser. 2), 169–187.

LEHMANN, E. L. Significance level and power. *Ann. math. Statist.*, 1958, **29**, 1167–1176.

LEHMANN, E. L. *Testing statistical hypotheses*. New York: Wiley, 1959.

LINDLEY, D. V. A statistical paradox. *Biometrika*, 1957, **44**, 187–192.

LINDLEY, D. V. The use of prior probability distributions in statistical inferences and decisions. In, *Proceedings of the fourth Berkeley symposium on mathematics and probability*. Vol. 1. Berkeley: Univer. California Press, 1961. Pp. 453–468.

NEYMAN, J. Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Roy. Soc.*, 1937, **236**(Ser. A), 333–380.

NEYMAN, J. L'estimation statistique, traitée comme un problème classique de probabilité. In, *Actualités scientifiques et industrielles*. Paris, France: Hermann & Cie, 1938. Pp. 25–57. (a)

NEYMAN, J. *Lectures and conferences on mathematical statistics and probability.* (2nd ed., 1952) Washington, D. C.: United States Department of Agriculture, 1938. (b)

NEYMAN, J. "Inductive behavior" as a basic concept of philosophy of science. *Rev. Math. Statist. Inst.*, 1957, **25**, 7–22.

PEARSON, E. S. In L. J. Savage et al., *The foundations of statistical inference: A discussion.* New York: Wiley, 1962.

PRATT, J. W. Review of *Testing Statistical Hypotheses* by E. L. Lehmann. *J. Amer. Statist. Ass.*, 1961, **56**, 163–167.

RAIFFA, H., & SCHLAIFER, R. *Applied statistical decision theory.* Boston: Harvard University, Graduate School of Business Administration, Division of Research, 1961.

RAMSEY, F. P. "Truth and probability" (1926), and "Further considerations" (1928). In, *The foundations of mathematics and other essays.* New York: Harcourt, Brace, 1931.

ROZEBOOM, W. W. The fallacy of the null-hypothesis significance test. *Psychol. Bull.*, 1960, **57**, 416–428.

SAVAGE, I. R. Nonparametric statistics. *J. Amer. Statist. Ass.*, 1957, **52**, 331–344.

SAVAGE, I. R. *Bibliography of nonparametric statistics.* Cambridge: Harvard Univer. Press, 1962.

SAVAGE, L. J. *The foundations of statistics.* New York: Wiley, 1954.

SAVAGE, L. J. The foundations of statistics reconsidered. In, *Proceedings of the fourth Berkeley symposium on mathematics and probability.* Vol. 1. Berkeley: Univer. California Press, 1961. Pp. 575–586.

SAVAGE, L. J. Bayesian statistics. In, *Decision and information processes.* New York: Macmillan, 1962. Pp. 161–194. (a)

SAVAGE, L. J. Subjective probability and statistical practice. In L. J. Savage et al., *The foundations of statistical inference: A discussion.* New York: Wiley, 1962. (b)

SAVAGE, L. J., et al. *The foundations of statistical inference: A discussion.* New York: Wiley, 1962.

SCHEFFÉ, H. *The analysis of variance.* New York: Wiley, 1959.

SCHLAIFER, R. *Probability and statistics for business decisions.* New York: McGraw-Hill, 1959.

SCHLAIFER, R. *Introduction to statistics for business decisions.* New York: McGraw-Hill, 1961.

SINCLAIR, H. Hiawatha's lipid. *Perspect. Biol. Med.*, 1960, **4**, 72–76.

STEIN, C. A remark on the likelihood principle. *J. Roy. Statist. Soc.*, 1962, **125** (Ser. A), 565–568.

STERLING, T. D. What is so peculiar about accepting the null hypothesis? *Psychol. Rep.*, 1960, **7**, 363–364.

TUKEY, J. W. The future of data analysis. *Ann. math. Statist.*, 1962, **33**, 1–67.

UREY, H. C. Origin of tektites. *Science*, 1962, **137**, 746.

VON NEUMANN, J. Zur Theorie der Gesellschaftsspiele. *Math. Ann.*, 1928, **100**, 295–320.

VON NEUMANN, J., & MORGENSTERN, O. *Theory of games and economic behavior.* (3rd ed., 1953) Princeton: Princeton Univer. Press, 1947.

WALD, A. On the principles of statistical inference. (Notre Dame Mathematical Lectures, No. 1) Ann Arbor, Mich.: Edwards, 1942. (Litho)

WALD, A. *Selected papers in statistics and probability.* New York: McGraw-Hill, 1955.

WALSH, J. E. *Handbook of nonparametric statistics.* Princeton, N. J.: Van Nostrand, 1962.

WOLFOWITZ, J. Bayesian inference and axioms of consistent decision. *Econometrica*, 1962, **30**, 470–479.

(Received January 15, 1962)