PATRICK SUPPES

TESTING THEORIES AND THE FOUNDATIONS OF STATISTICS

1. HISTORICAL PERSPECTIVE

In this paper I examine the extent to which problems in the foundations of probability are relevant to the testing of theories, and what view towards probability, if any, can be inferred from the problems and practices found in the scientific literature. I start with a historical perspective and then consider particular examples in contemporary science. In this latter discussion I confront some of the issues made salient by Bayesians.

Far and away the most serious quantitative scientific treatise in ancient times that uses both mathematics and data in a systematic way is Ptolemy's *Almagest*. What is surprising is that in Ptolemy's *Almagest* and in other astronomical treatises of ancient times there is no evidence of a quantitative theory of error; in fact, there is little evidence of any theory of error at all. This is in marked contrast to ancient astronomy's mathematical and observational sophistication. It might be thought that the lack of such a systematic theory of error is simply a reflection of the absence of any developments of a quantitative sort in probability theory in Hellenistic science, and consequently, an explanation for the absence of such an analysis in Ptolemy is easily found.

The story, however, is much more complicated, because what is true of Ptolemy's *Almagest* is also true of Newton's *Principia*. There is, I believe, not one single computation of a quantitative error term in Newton's *Principia*. It contains a few remarks about errors, but all of them are of a qualitative and incidental character. Again, it might be thought that this is simply a consequence of the fact that the theory of probability was just being developed in quantitative form in the seventeenth century, and as a result, detailed applications could hardly be expected. That more complicated explanations of the absence of such a theory of error are needed is testified to by the absence of such systematic computations of errors in Laplace's *Celestial Mechanics*. Of all the

Harper and Hooker (eds.), Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II, 437–455.

classical treatises in which one would expect to find such systematic computations, Laplace's is the one. In spite of the fact that Laplace more than anyone else contributed to the development of the theory of probability in the eighteenth century and the early part of the nineteenth century, and in spite of the fact that he discusses in his treatise on the theory of probability the analysis of data from a probabilistic standpoint in order to determine evidence for 'constant causes', there is in the systematic treatise on the solar system no detailed analysis of error terms or any application directly of a quantitative theory of error.

This Ptolemaic tradition did not end with Laplace, but also is found in Maxwell's treatise on electricity and magnetism. There is little numerical confrontation between data and theory in Maxwell, and certainly no analysis of problems of errors of measurement. In fact, from the standpoint of the confrontation between data and theory, there would seem to be some downhill sliding from the time of Ptolemy to that of Maxwell.

Within astronomy proper, reporting error terms in the analysis of astronomical data did become common in the nineteenth century. I hope on another occasion to trace that history. At the present time, however, my understanding of it is too poor to enter into the details. It is certainly true that, from the standpoint of physical theory, the more important development of electromagnetic theory does not reflect a corresponding development of a sophisticated theory and practice of data analysis at the level characteristic of astronomy in the last half of the nineteenth century.

The conceptual point of importance for this paper is that the verification of the historically important theories of physical phenomena has practically never used a detailed statistical theory of confirmation to test empirical adequacy. In thinking about the ways in which statistics and probability are and should be used in science, it seems to me that this historical fact is important to keep in mind so as not to create a simplified theory of how theories may be tested. By this statement I do not mean to suggest that I am against the use of statistical methods in the verification of theories. I only enter the cautionary note that the verification of theories is a complex business, and any simple view of how to apply statistical methods is bound to be inaccurate.

Numerous other examples from the seventeenth, eighteenth, and

nineteenth centuries can easily be found. It might be thought, however, that these historical examples have been superseded by the statistical sophistication of the twentieth century. After all it may properly be claimed that, in spite of the kind of developments begun by Laplace, much of the development of explicit procedures of statistical inference and estimation dates from the second or third decade of the twentieth century, and that to get a true assessment of the situation, we must examine some twentieth-century theories.

2. Twentieth-century theories

A. *Quantum Mechanics*

The most impressive and most extensively tested theory in this century is surely quantum mechanics. When one turns to the evidence in support of quantum mechanics, and the kind of confrontations between data and theory that are used to support the theory, some surprises are in store. First of all, in the standard treatises, no attempt is made to present empirical data or to point out in what respects discrepancies exist between theoretical predictions and empirical data. To support this statement, I casually looked in my own library at three treatises on quantum mechanics: P. A. M. Dirac, *Quantum Mechanics* (3rd ed., 1947); L. D. Landau and E. M. Lifshitz, *Quantum Mechanics: Non-relativistic Theory* (1958); Albert Messiah, *Quantum Mechanics*, Volumes 1 and 2 (1961). By reporting these negative results, I am not suggesting that there is no place one can find experimental evidence. Rather, unlike the tradition of Newton and Laplace, in contemporary treatments of quantum mechanics, there is often no attempt to present supporting data and to examine discrepancies between theory and data. Although data can be found in the experimental literature, and one can track down the examination of data in the classical experiments that are ordinarily cited in support of quantum mechanics, I believe it is fair to say that there exists no book in which these data are brought together in a systematic way and in which a careful examination from a statistical standpoint of the relation between the data and the theory is considered. One classical text, Leonard Schiff's *Quantum Mechanics* (1949), does list in the discussion on the physical basis of quantum mechanics the relations between the classical experiments, running from Young's experiments at the beginning of the

nineteenth century on diffraction to the Stern-Gerlach experiment in 1922. But there is in Schiff's book no detailed discussion of the relation between theory and data, but only a development of the theory. On the other hand, turning to one of the more data-oriented books that do not develop the theory of quantum mechanics, for example, F. K. Richtmyer and E. H. Kennard, *Introduction to Modern Physics* (4th ed., 1947), one does find the experimental data and the comparisons with theoretical predictions. However, I believe that this classical text contains not a single statistical inference, nor even a statistically descriptive statistic. Although I have not systematically surveyed the original experimental literature, from my experience what I have said about Richtmyer and Kennard also holds for this literature in almost all cases.

To some extent, these references are to the older experimental literature in physics. Perusal of current issues of *Physical Review* indicates that the actual use of standard statistical tests can be found in a variety of experimental articles. Yet the main thrust of my remark is, I think, still correct. In the testing of highly structured theories of the kind characteristic of physics, there is little use of the vast apparatus of modern statistics.

B. *Econometrics*

Perhaps the sharpest scientific contrast to quantum mechanics and to other parts of physics can be found in economics. Among social scientists, econometricians are probably the most statistically sophisticated and the most careful in their use of statistical procedures. The analysis of data is superb from a statistical standpoint in most of the major work. Because it is true that economists deal with nonexperimental data, there is even more reason to be statistically explicit about the analysis and inferences made. In this respect, economics compares more directly with astronomy or meteorology than with quantum mechanics, where the data are almost all experimental in character. As an example of the kind of theory used in econometrics, I have selected a recent article by Chiswick and Mincer (1972). What is striking about this and other serious applications of mathematical concepts in economics is that when data are involved the model is usually of a relatively simple character without substantial theoretical deductions from the model itself. I quote from the second and third pages of the article (35–36) in which the mathematical model used for the analysis is stated.

The relation between gross earnings and investment in human capital for the ith person in year j can be written as

$$(1) \qquad E_{ji} = E_{oi} + \sum_{t=1}^{j-1} r_{ti} C_{ti},$$

where the gross earnings (E_{ji}) are a function of the 'original' endowment (E_{oi}) and the sum of the returns on previous investments (C_{ti}), r_{ti} being the average rate of return to the investment in the tth year. In this expression, earnings are a linear function of dollars of investment.

An alternative specification of the relation between gross earnings and investment can be obtained by expressing C_{ti} as a fraction of E_{ti} (that is, $C_{ti} = k_{ti} E_{ti}$). If the original endowment is assumed constant across years and individuals (E_0), we can write

$$(2) \qquad E_{ji} = E_0 + \sum_{t=1}^{j-1} r_{ti} k_{ti} E_{ti} = E_0 \prod_{t=1}^{j-1} (1 + r_{ti} k_{ti}).$$

By taking the natural log of both sides of Equation (2), since $r_t k_t$ is small, we obtain (approximately)

$$(3) \qquad \ln(E_{ji}) = \ln E_0 + \sum_{t=1}^{j-1} r_{ti} k_{ti}.$$

What is proposed is a simple linear model that the effects of returns on previous investments, as well as the function of 'the original' endowment, can satisfactorily be expressed in a linear way. This linear model is not derived from any more fundamental assumptions, nor is it the consequence of elementary qualitative assumptions or of some deeper running formulation of economic theory. This kind of regression model is characteristic of applications in econometrics, and the efforts that have been made to understand thoroughly the statistical pitfalls of making inferences by use of such models are thoroughly explored in the literature, for example, in Malinvaud's classic work (1966). Although Malinvaud's book takes us beyond the kind of linear model defined by Chiswick and Mincer, it does not take us far.

It might be thought that I am pushing a kind of conservation thesis: the more theory the less statistics, and the less theory the more statistics. From an empirical standpoint there is something to be said for this. It is even true of the mathematics, for example. Although the mathematical requirements are rather different, the mathematical level of a treatise like Malinvaud's is, in my judgment, about comparable to the treatises on quantum mechanics I mentioned above, although Malinvaud's treatise would satisfy mathematical standards of rigor more explicitly than would the physical treatises.

Thus far I have picked two extremes of theories – one, the highly structured and developed theories of quantum mechanics, and the other, the very simple regression models characteristic of much of econometrics. It is natural to ask if these two examples, each extreme in its own way, represent the whole story. I do not think this is the case, and I want to turn to still a third class of theories, theories that do not have the depth of structure and development of quantum mechanics, but that have a fundamental theory and consequences that lead to more elaborate structures.

C. *Psychological Theories*

This third kind of case is drawn from psychology, which also is more like physics than econometrics in that the tests of theories are basically experimental rather than nonexperimental in character. Although psychological theories of learning, as an example, are much shallower than the great physical theories of the twentieth century, they can be given a precise formulation in general terms and can lead to rigorous deductions of particular predictions for particular experiments. Surprisingly, even in relatively simple applications of the theory quite intractable stochastic processes arise for which the application of standard statistical procedures of estimation of parameters is essentially hopeless. Without going into detail, let me mention one or two examples. A typical simple learning experiment that constitutes a test of an underlying theoretical model, itself derived from a more general qualitative theory, involves estimating parameters in a chain of infinite order. In other words, the mathematical model itself is a stochastic process that is a chain of infinite order. In most cases the chain of infinite order is an ergodic process, but explicit maximum-likelihood or Bayesian estimate of the parameters is strictly out of the question, and in practice some relatively rough-and-ready approximation to a maximum-likelihood estimate, or sometimes a minimum chi-squared estimate, is used. In almost all cases, the tests are not actually maximum likelihood nor minimum chi-square, but tests that approximate these, and whose characteristics as tests have not in any sense been thoroughly investigated.

In most experiments that test such psychological models, the number of observations is huge, for example, upwards from two or three thousand to twenty or thirty thousand. Given the large number of observations, the relative statistical crudeness of a pseudo-maximum-likelihood estimate

is not disturbing to anyone, for it is clear that refinements of statistical procedure will have little effect on the summary estimate of the goodness of fit of the theory to data. Because of computational difficulties in many applications of the suitable maximum-likelihood function, the complete function is computed rather than seeking a solution to the derivative to find the maximum. The graphing of the complete function has been instructive in a number of cases, because when we look at the complete function we see that a fairly wide variation in the value of the parameter estimated makes little difference in the fit of the theory to the data. An example of this for a simple linear learning model (that is, a chain of infinite order) in the observables is shown in Figure 1 (drawn from Suppes

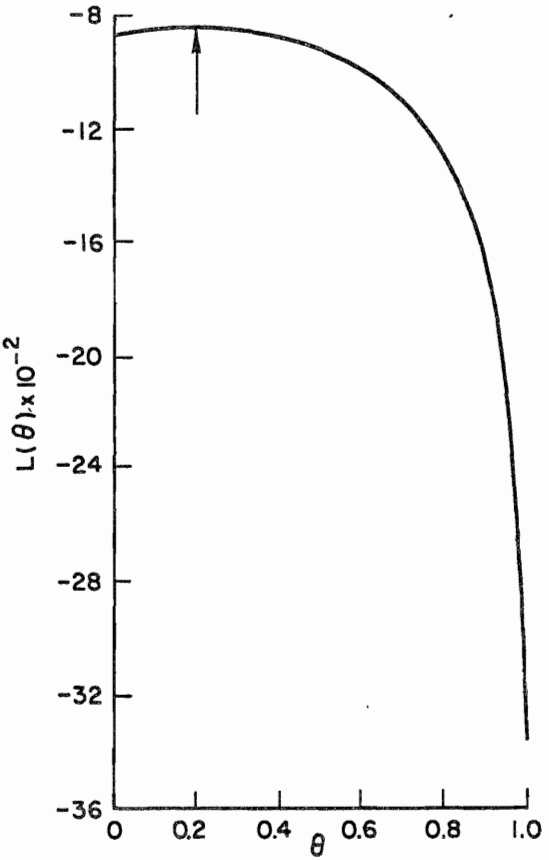

Fig. 1. The pseudo-likelihood function, $L^*(\theta)$, for the linear model.

and Atkinson, 1960, p. 218). Note that the suitable likelihood function is quite flat, between 0.02 and 0.30, and consequently, any value of the learning parameter θ lying in this interval will give about as good a fit as another. This kind of flatness of suitable likelihood function is another reason that working scientists will not take too seriously refinements of statistical concepts, at least insofar as they are billed as offering something of importance and significance to the scientist in evaluating the relation between theory and experiment.

3. Foundations of statistics

I do not for a moment want to defend the careless statistical practices of physicists; in fact, I think much information in physical experiments is lost because of the lack of care in analysis. A good example is the lack of the analysis of randomness in experiments on radioactive decay of a given substance. Such radioactive decay is often cited by both physicists and philosophers as a prime example of randomness in nature. Personally I tend to accept this view, but I would be much more secure in it if I had at hand the kind of massive analysis of data from radioactive decay that is characteristic of the kind of tests statisticians have given such data in other domains. It is well known that it is extraordinarily difficult to produce data that illustrate the appropriate randomness features in all respects. It would be curious indeed to find deviations from what was expected in the matter of radioactive decay, and only by the application of refined statistical techniques are we at all likely to find such deviations if they do exist, or to confirm the view that randomness all the way is the story.

In spite of this example, and in spite of my willingness to criticize the statistical procedures of physicists, I think we need a way of looking at the foundations of statistics that takes account of the kind of rough-and-ready theoretical tests I have described earlier in this paper. What we need among our concepts of logical inference and statistical inference are appropriate ways of closing the gap between commonsense conclusions that the evidence is indeed decisive and that no refined tests are required.

It may be possible to say on some occasions that it is merely a matter of routine work to produce the statistical distributions of data required

to make the inference explicit and firm, but of course this is not the way the matter mainly works. If one examines the discussion of these matters in tests like those of Richtmyer and Kennard, it is evident that the data are not even thought of in a way that would permit a statistical inference to be made. A Bayesian inference could be made only in the crudest and most subjective way, and yet we all admit and agree to the solidity of the evidence in many cases and on the extent to which it supports the theoretical predictions.

There are several ways of expressing skepticism about the realism of objective or Bayesian approaches to this problem. Apart from the sophisticated problem of assuming probability distributions for data structures, there can be and should be proper skepticism about assigning a subjective or prior probability to the agreement between an experimental result and a theoretical prediction. To indicate some of the problems in a quantitative way, let E be the experimental result (which we may think of as a random variable) and let T be the theoretical prediction (which we may also think of as a random variable). We might begin by expressing our confidence in the agreement between the two by the following probabilistic inequality:

$$(1) \qquad P(E=T)>1-\varepsilon.$$

The difficulty with inequality (1) is that we almost never expect the experimental result and the experimental prediction to agree exactly if the random variables are thought of as having an underlying continuous distribution or, put another way, if the empirical quantities in question are assumed to be essentially continuous in nature. A more realistic expression of our confidence in the essential agreement between E and T may be expressed in the following way:

$$(2) \qquad P(|E-T|<\varepsilon_1)>1-\varepsilon_2.$$

In both inequalities (1) and (2) we expect ε, ε_1, and ε_2 to be small numbers, but already inequality (2) has assumed a rather formidable character and seems too elaborate for the purpose at hand.

Inequality (2) has a surface resemblance to a confidence-interval statement, but the lack of an underlying theory of distributions makes development of a theory of confidence intervals for the kind of situation I am talking about even more unrealistic than the use of inequalities like (2).

What seems natural is to analyze these situations of qualitative judgment in terms of a qualitative theory of probability and belief. Thus, in qualitative terms we replace inequality (2) by the qualitative statement (3).

(3) The event that $|E-T|$ is small $\approx X$.

In (3), the relation $\approx$ is the relation for indistinguishability and X is the certain event. (We might think of X as the sample space, but I agree with the Bayesians in being suspicious of having one definite reference space, and I prefer simply to let X be a certain event – I am even willing to slide between saying *a* certain event X and *the* certain event X.)

In the classical theory of qualitative probability, the relation $\approx$ of indistinguishability would be reflexive, symmetric, and transitive. Here I ask only that it be reflexive and symmetric. For judgments of strictly greater qualitative probability, I use a semiorder that is a relation $\succ$ that satisfies the following three axioms:

1. Not $x \succ x$.
2. If $x \succ y$ and $y \succ z$ then either $x \succ w$ or $w \succ y$.
3. If $x \succ y$ and $z \succ w$ then either $x \succ w$ or $z \succ y$.

The indistinguishability relation is just the nontransitive indifference relation that can be defined in terms of strict probability preference by the following:

(4) $A \approx B$ iff not $A \succ B$ and not $B \succ A$.

Although one can add to the axioms introduced and write axioms that, in the case of a finite number of events, will lead to the existence of a probability measure, that is not my purpose here. The matter has been studied elsewhere, and good results for this particular case have been given by Domotor and Stelzer (1971). Rather, in the present discussion I want to pursue the kind of apparatus I have introduced.

As I see it, it is a mistake to make the Bayesian move and to ask the investigator for a subjective estimate that the agreement, for example, between experimental and theoretical results is more accurate or less accurate than the accuracy with which the velocity of light is measured, or the specific heat of sodium. In other words, the natural Bayesian thing would be to ask the investigator to 'calibrate' his judgment of the quality of the result by comparing it with other results for corresponding physical experiments or appropriate experiments in the domain in question. In

practice, this is exactly what we do not do. We expect the investigator to present the numerical results, and in general we expect some discrepancy between the experimental result and the theoretical prediction. A qualitative comment may be made by the investigator, such as, 'the agreement is pretty good', or 'the agreement is not so good as we may ultimately expect to obtain but the results are encouraging'. The reader and colleague is left to draw his own conclusion, and there is a clear restraint from offering a more detailed or more complete analysis.

It is this practical sense of leaving things vague and qualitative that needs to be dealt with and made explicit. In my judgment to insist that we assign sharp probability values to all of our beliefs is a mistake and a kind of Bayesian intellectual imperialism. I do not think this corresponds to our actual ways of thinking, and we have been seduced by the simplicity and beauty of some of the Bayesian models. On the other hand, a strong tendency exists on the part of practicing statisticians to narrow excessively the domain of statistical inference, and to end up with the view that making a sound statistical inference is so difficult that only rarely can we do so, and usually only in the most carefully designed and controlled experiments.

I want to trod a qualitative middle path between these two extremes of optimism and pessimism about the use of probabilistic and statistical concepts. On a previous occasion I have expressed my skepticism about drawing a sharp and fundamental difference between logical inference and statistical inference, if only because this distinction is not present in the ordinary use of language and ordinary thinking (Suppes, 1966). The kind of gap I have attempted to stress in the present discussion is one that lies between the present explicit theory of logical inference and the explicit theory of statistical inference. The appropriate place to look for the theory to close this gap is in the semantics of ordinary language, for although I have been concerned with statements made in the summary about the tests of theories, I think that the character of these statements is by and large consistent with statements of ordinary language about ordinary experience, and that we should not invoke a special language and a special apparatus. What we have left is a residue of common sense and ordinary language, and it is my philosophical belief that not only will this residue remain, but it will also remain robust in the discussion of scientific theories and their verification. To expect these robust uses of commonsense judg-

ments to be eliminated from our judgments of scientific theory is a mistaken search for precision. In fact, as I have attempted to argue, in many ways the stronger the theory and the better the evidence, the less tendency to use any defined statistical apparatus to evaluate the predictions of the theory. I see no reason to think that this broad generalization will not continue to hold.

This means that the intellectual task of closing the gap is almost identical with the task of giving a proper semantics for such ordinary language statements as:

> Almost all observations are in agreement with the experiment.
> Most of the observations are in agreement with the theoretical predictions.
> The agreement between prediction and theory is pretty good.
> See for yourself. The results are not bad.

The tools for providing such a semantical analysis are now being developed by a number of people, and the prospect for having a well-developed theory of these matters in the future looks bright. In the meantime, there are many simpler ways of improving on the situation that have relevance both to the foundations of statistics and to the testing of theories. In the next section I discuss one such approach, which can also be used in the deeper semantical analysis still to be developed in detail.

4. Upper and lower probabilities

The first step in escaping some of the misplaced precision of standard statistics is to replace the concept of probability by that of upper and lower probability. The first part of what I have to say is developed in more detail in Suppes (1974) and I shall only sketch the results here. (References to the earlier literature are to be found in this article.)

To begin with, let X be the sample space, $\mathfrak{F}$ an algebra of events on X, and A and B events, i.e., elements of $\mathfrak{F}$. The three essential properties we expect upper and lower measures on the algebra $\mathfrak{F}$ to satisfy are the following:

(I) $P_*(A) \geqslant 0.$

(II) $P_*(X) = P^*(X) = 1.$

(III) If $A \cap B = \emptyset$ then

$$P_*(A)+P_*(B) \leqslant P_*(A \cup B) \leqslant P_*(A)+P^*(B) \leqslant P^*(A \cup B) \leqslant P^*(A)+P^*(B).$$

From these properties we can easily show that

$$P_*(A)+P^*(\neg A)=1.$$

Surprisingly enough, quite simple axioms on qualitative probability can be given that lead to upper and lower measures that satisfy these properties. The intuitive idea is to introduce standard events that play the role of standard scales in the measurement of weight. Examples of standard events would be the outcomes of flipping a fair coin n number of times for some fixed n.

The formal setup is as follows. The basic structures to which the axioms apply are quadruples $\langle X, \mathfrak{F}, \mathscr{S}, \geqslant \rangle$, where X is a nonempty set, $\mathfrak{F}$ is an algebra of subsets of X, that is, $\mathfrak{F}$ is a nonempty family of subsets of X and is closed under union and complementation, $\mathscr{S}$ is a similar algebra of sets, intuitively the events that are used for standard measurements, and I shall refer to the events in $\mathscr{S}$ as *standard* events S, T, etc. The relation $\geqslant$ is the familiar ordering relation on $\mathfrak{F}$. I use standard abbreviations for equivalence and strict ordering in terms of the weak ordering relation. (A weak ordering is transitive and strongly connected, i.e., for any events A and B, either $A \geqslant B$ or $B \geqslant A$.)

DEFINITION. *A structure $\mathfrak{X} = \langle X, \mathfrak{F}, \mathscr{S}, \geqslant \rangle$ is a finite approximate measurement structure for beliefs if and only if X is a nonempty set, $\mathfrak{F}$ and $\mathscr{S}$ are algebras of sets on X, and the following axioms are satisfied for every A, B, and C in $\mathfrak{F}$ and every S and T in $\mathscr{S}$:*

AXIOM 1. *The relation $\geqslant$ is a weak ordering of $\mathfrak{F}$;*

AXIOM 2. *If $A \cap C = \emptyset$ and $B \cap C = \emptyset$ then $A \geqslant B$ if and only if $A \cup C \geqslant B \cup C$;*

AXIOM 3. $A \geqslant \emptyset$;

AXIOM 4. $X > \emptyset$;

AXIOM 5. *$\mathscr{S}$ is a finite subset of $\mathfrak{F}$;*

AXIOM 6. *If $S \neq \emptyset$ then $S > \emptyset$;*

AXIOM 7. *If $S \geqslant T$ then there is a V in $\mathscr{S}$ such that $S \approx T \cup V$.*

From these axioms the following theorem can be proved.

THEOREM 1. *Let $\mathscr{X} = \langle X, \mathfrak{F}, \mathscr{S}, \geqslant \rangle$ be a finite approximate measurement structure for beliefs. Then*

(i) *there exists a probability measure P on $\mathscr{S}$ such that for any two standard events S and T*

$$S \geqslant T \text{ if and only if } P(S) \geqslant P(T),$$

(ii) *the measure P is unique and assigns the same positive probability to each minimal event of $\mathscr{S}$,*

(iii) *if we define P_* and P^* as follows:*

(a) *for any event A in $\mathfrak{F}$ equivalent to some standard event S,*

$$P_*(A) = P^*(A) = P(S),$$

(b) *for any A in $\mathfrak{F}$ not equivalent to some standard event S, but lying in the minimal open interval (S, S') for standard events S and S'*

$$P_*(A) = P(S) \quad \text{and} \quad P^*(A) = P(S'),$$

then P_* and P^* satisfy conditions (I)–(III) for upper and lower probabilities on $\mathfrak{F}$, and

(c) *if n is the number of minimal elements in $\mathscr{S}$ then for every A in $\mathfrak{F}$*

$$P^*(A) - P_*(A) \leqslant \frac{1}{n},$$

(iv) *if we define for A and B in $\mathfrak{F}$*

$$A *> B \text{ if and only if } \exists S \text{ in } \mathscr{S} \text{ such that } A > S > B,$$

then $*>$ is a semiorder on $\mathfrak{F}$, if $A *> B$ then $P_*(A) \geqslant P^*(B)$, and if $P_*(A) \geqslant P^*(B)$ then $A \geqslant B$.

Following an earlier suggestion of Good (1962), events whose upper probability is 1 can be said to be *almost certain*, and events whose lower probability is 0 can be said to be *almost impossible*.

Moreover, let us consider events that are not exactly equivalent to any of the standard events. This restriction is easy to impose on our measure-

ment procedures if we follow procedures often used in physics by requiring that each nonstandard event measured be assigned to a minimal open interval of standard events. In terms of such properly measured events A and B, as I shall call them, we may define upper and lower conditional probabilities as follows:

$$P_*(A \mid B) = P_*(A \cap B)/P_*(B),$$
$$P^*(A \mid B) = P^*(A \cap B)/P^*(B),$$

provided $P_*(B) > 0$. We can then show that the upper and lower conditional probabilities satisfy properties (I) to (III) except for the condition on $P_*(A) + P^*(A)$. In particular, $P_*(A \mid B) \leqslant P^*(A \mid B)$.

Within this framework we can then develop a reasonable approximation to Bayes' theorem or to the method of maximum likelihood. The point is that we can develop a machinery of statistical inference that is approximate in character and consequently is closer to the ordinary talk of scientists dealing with their summary evaluations of experimental tests of theories.

Moreover, an important feature of the kind of setup I am describing is that it is not meaningful to ask for arbitrary precision in the assignment of upper and lower probabilities to events, but only an assignment in rational numbers to the scale of the finite net of standard events. Further questions about precision do not have a clear meaning.

Using the kind of apparatus outlined, we can then replace, if we so desire, the inequality expressed in (2) in the previous section by the following equation using upper probabilities:

$$P^*(|E - T| < \varepsilon_1) = 1.$$

We could also paraphrase this statement in the manner indicated above by the qualitative statement that almost certainly all deviations between the experimental data and the theoretical predictions are less than ε or, even more qualitatively, are small.

I should also emphasize that the particular upper and lower measures derived from qualitative structures satisfying the axioms given above do not have many of the pathological characteristics of arbitrary upper and lower measures. This may be seen already in the fairly complete theory of conditional upper and lower measures that follows.[1]

I believe that the approximation theory I have sketched forms a natural

bridge between the quantitative theory of statistics and the qualitative statements of ordinary scientific language. Further developments of the theory are needed to make clear whether my hopes are realistic or too sanguine.

Stanford University

NOTE

[1] The measurement-theoretic conception of upper and lower probabilities I use is quite different conceptually from the 'uncertainty' approach to such probabilities of Dempster (1967, 1968). Consequently, my approach to the theory of statistical inference is also formally and conceptually different from Dempster's.

BIBLIOGRAPHY

Chiswick, B. R. and Mincer, J., 'Time-Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985', *Journal of Political Economy* **80** (1972) 34–66.

Dempster, A. P., 'Upper and Lower Probabilities Induced by a Multivalued Mapping', *Annals of Mathematical Statistics* **38** (1967) 325–340.

Dempster, A. P., 'A Generalization of Bayesian Inference', *Journal of the Royal Statistical Society* **30** (1968) (Series B), 205–247.

Dirac, P. A. M., *Quantum Mechanics* (3rd ed.), Oxford University Press, London, 1947.

Domotor, Z. and Stelzer, J., 'Representation of Finitely Additive Semiordered Qualitative Probability Structures', *Journal of Mathematical Psychology* **8** (1971) 145–158.

Good, I. J., 'Subjective Probability as the Measure of a Non-Measurable Set', in E. Nagel, P. Suppes, and A. Tarski (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*, Stanford University Press, Stanford, 1962.

Landau, L. D. and Lifshitz, E. M., *Quantum Mechanics: Non-Relativistic Theory*, Pergamon Press, London, 1958.

Malinvaud, E., *Statistical Methods of Econometrics*, Rand McNally, Chicago, 1966.

Messiah, A., *Quantum Mechanics*, Vol. 1, 2, North-Holland, Amsterdam, 1961.

Richtmyer, F. K. and Kennard, E. H., *Introduction to Modern Physics* (4th ed.), McGraw-Hill, New York, 1947.

Schiff, L., *Quantum Mechanics*, McGraw-Hill, New York, 1949.

Suppes, P., 'Probabilistic Inference and the Concept of Total Evidence', in J. Hintikka and P. Suppes (eds.), *Aspects of Inductive Logic*, North-Holland, Amsterdam, 1966.

Suppes, P., 'The Measurement of Belief', *Journal of the Royal Statistical Society* **36** (1974) (Series B), 160–175.

Suppes, P. and Atkinson, R. C., *Markov Learning Models for Multiperson Interactions*, Stanford University Press, Stanford, 1960.

DISCUSSION

Commentator: Lindley: I have a delightful book at home which describes how they made wheels in the middle of the last century. These wheels were masterpieces of design and construction and yet no measuring instrument of any sort was used in their construction. Nowadays we turn out wheels more cheaply and in vastly greater numbers partly because we use precise measuring instruments. My point is that the fact that measurements were not used, does not mean that we will not be better off by using them.

Suppes: A parallel which strikes me here is the situation in which Laplacian determinism now finds itself: Just as we have now discovered that the universe is such that we can no longer carry out the Laplacian deterministic program (determine completely the state of the universe at some given instant of time) so we cannot carry out your suggestions – *we simply cannot get around in our universe, and function at the level at which we wish to function,* without carrying out an extensive program of measurement (among other things). (Indeed, there seems to have been a 'natural line of theological succession' from the early belief that God ran the universe in a definite fashion, to the Laplacian belief that the universe ran itself in a definite fashion to Lindley's belief that we all have access to a unique prior probability.) Just as we cannot carry out Laplace's program, we cannot carry out yours. The fact of the matter is that we all could carry around a suitable gambling apparatus which, upon the assumption of its independence from the rest of the universe, would serve to insure that we always realize Savage's axioms in a conduct of our lives. But the plain fact is that we do not wish to do that kind of thing, we are not constructed so as to operate with such a precise judgment of probabilities and in formulating theories of the sort that I have been discussing probability theorists have committed an error of misplaced precision.

Teller: I'd like to make two comments. You mentioned briefly the gross unrealism of Savage's use of a class of functions which are supposed to represent acts. These are, roughly, the set of all functions from possible present states of the world into the set of possible consequences and not

all of them could possibly represent acts. I should just like to comment that Jeffrey's system shows us how to get along without them very nicely.

Suppes: But not how to get along without structural axioms.

Teller: No, and that brings me to my second comment. You want to criticize the structural axioms as also being unrealistic. Now when I think of these axioms I view them as capturing something like the properties of the system of beliefs of an ideally rational person. In this case they are construed as normative, rather than empirical, principles and it is hard to see how it could be relevant to criticize them on the grounds of practical difficulty in their applicability. Indeed, it is hard to see why it would be relevant to criticize them on the basis of their applicability via some given, practical mechanism – such as the gambling device you spoke about.

Lindley: The situation is exactly analogous with that in Euclidean geometry. The theory is precise, elegant and complex. Yet there do not exist the ideal 'points' and 'lines' of Euclid. Nevertheless, the theory is eminently practical. The same is true of Savage's theory.

Suppes: No, but the case of Euclidean geometry is very different from Savage's axioms and this very example allows me to reply to the point. The fact is that we can take an indefinite period, in physics, to work out the complexities involved in applying Euclidean geometry within certain approximations to the testing of our physical theories. But a theory of rationality is in a very different situation, for it is of the essence of that theory that it should show us how we go about making rational decisions in the light of the fact that the world is constantly changing during the decision period, that delays have associated costs, that our lifetime is finite and so on – in this case considerations of time and costs are of the essence, whereas they are not in the case of Euclidean geometry and physics. If we were to take Savage's axioms seriously, we should have to give the whole theory of rationality a new cast, it would have to be allied with just such a theory of science outlining their practical applications as I have indicated (using the illustration of the gambling device); but if this were actually done I think it is clear that we would be much less enchanted by Savage's axioms, they would be much less appealing as the foundations of the theory of rationality.

Finch: Two comments. The first is that it seems to me that you are not really attacking the Bayesian position, but rather operating still within it and simply pointing out that its actual applications may be a great deal

more complicated than the usual formulation of it would suggest. Second comment is that there are actually some results which are of relevance to the question you raised concerning the necessary and sufficient conditions governing the transitions from the discrete to the continuous case (from the probability relation to the continuous probability measure) and these results emerge from some quite deep results in measure theory, they are effectively contained in the works of Manneheim.

Suppes: I certainly agree that I'm working within the Bayesian tradition here, but attacking it for its contemporary drive toward a misplaced precision....

Finch: Well, would you agree that what we really need is a calculus which shows us how to go from relations amongst prior probabilities to relations amongst posterior probabilities without having to attach precise, detailed numbers to them all in between.

Suppes: Exactly.

Giere: I should simply like to ask you to state clearly how you now conceive of the direction of your program of investigations in probability theory – do you want to push away from personalistic probability in favour of a physical notion, or objective notion, of probability?

Suppes: At the present time I'm somewhat dualistic on this issue. I feel there is a firm place for a personalistic concept of probability referring to beliefs, though I have spoken out here against demanding too much precision for that notion; on the other hand I have also spoken strongly in favour of a physical interpretation of the notion of probability, with the probabilities determined by the physical hypothesis just as in the case of any other physical concept.