DISCUSSION

I.J. GOOD *(Virginia Polythecnic and State University):*

I shall restrict my discussion to the historical aspects of the paper presented by Professors Giron and Rios.

J.M. Keynes (1921) argued that not all logical probabilities could be compared. B.O. Koopman (1940a, b), acknowledged Keynes's influence and laid down fairly convincing but complicated axioms for partially ordered "intuitive" probabilities, where "intuitive" I think meant either logical or subjective. I propounded the simplest possible acceptable theory of partially ordered subjective probabilities in Good (1950, p. 119) and pointed out that such a theory is identical with the use of upper and lower probabilities provided that it is agreed that we can imagine perfectly shuffled packs of cards. I extended the theory to include utilities in an obvious manner in Good (1952) or see Good (1954). At a 1960 conference in Stanford (Good, 1962) I showed that this simplest possible theory of partially ordered probabilities implies formal axioms connecting upper and lower probabilities. Cedric Smith (1961) justified my theory by using arguments analogous to those used by Savage (1954) for the theory of sharp

probabilities. For this justification he made use of convex sets of prior distributions. Smith said he left some loose ends and presumably these have received the attention of Rios and Giron. Whether this is so I have unfortunately not had time to check.

Most of my historical research was concerned with finding the publications where I had mentioned the partially ordered theory of subjective probability. I have found fifty such publications, or perhaps only 49, ranging from 1949 to 1979. I have given the list to the authors, to prove to them that I have emphasized the partially ordered theory perhaps *ad nauseam*, but in the references to this discussion I have listed only Good (1950, 1952, 1962, 1976 and 1977). For example, in Good (1976, p. 137) I pointed out that my theory is a Bayes/non Bayes compromise, as Rios and Giron have now recognized.

It may be helpful to mention that the theory of partially-ordered probabilities (and utilities) is sometimes called a theory of comparative or qualitative probabilities (and utilities). The subjective version could reasonably be called Good's theory or the Doogian theory or the comparative or qualitative or partially-ordered Bayesian theory, and "quasi-Bayesian" is yet another name for the same thing.

Although I have always accepted this theory, in practice I often prefer to use sharp probabilities and utilities for the sake of simplicity, as an approximation to the partially ordered theory.

On a point of terminology, I think the expression "confidence interval" should be restricted to the Neyman-Pearson sense. In the Bayesian theory one can use the expression "Bayesian estimation interval".

Turning now to Professor Hills' paper, the word "paradox" has at least two distinct meanings which can be distinguished by talking about apparent and true paradoxes. If I thought there were any true paradoxes in the theory of subjective probabilities that I support, then I would be forced to abandon rationality. I am not yet prepared to do that.

Perhaps the common denominator of all Bayesian statistics is the product law, $P(A \,\&\, B) = P(A).P(B|A)$, meaning that if two of the probabilities mean anything then so does the third, and this is so even if the probabilities are merely constrained by inequalities. Have we any reason to doubt this product law, in the light of the various apparent paradoxes mentioned by Dr. Hill? I think these paradoxes arise, at least in part, through performing limiting operations in the wrong order. For example, the limit of *P.D.* $(x|y_0 < y < y_0 + \delta y_0)$ as δy_0 tends to zero is not necessarily equal to *P.D.* $(x|y = y_0)$ when $P\,(y = y_0) = 0$. (Here *P.D.* stands for "probability density"). To assume otherwise is equivalent to assuming that all Jacobians are equal to 1. Also, in the problem of the distribution on a sphere, there is a difference between a random great circle on a sphere rather than a great circle known to pass through a known fixed point (the North Pole). These two comments I believe remove the paradox from the example of the density on a sphere and the density on a longitude.

We all know that improper priors can sometimes be used if the limiting operations are performed in the right order. But one interesting example where an improper prior is definitely ruled out occurs in some work on Bayesian significance testing for multinomials and contingency tables (Good, 1965, 1967, 1976; Good and Crook, 1974; Crook and Good, 1979). In this work there is a Bayes factor $F(k)$ depending on a non-

negative hyperparameter k such that the null hypothesis corresponds to $k = \infty$ and such that $F(k)$ tends to 1 when k tends to infinity. If a hyperprior $\Phi(k)$ is assumed for k such that $\int_0^\infty d\Phi(k)$ is divergent, then the resultant Bayes factor $F = \lim_{K \to \infty} \int_0^K F(k)\, d\Phi(k) / \int_0^K d\Phi(k) = 1$. In other words the evidence against the null hypothesis is completely annihilated by a prior that is "improper at infinity". Satisfactory results were obtained in the applications by using a proper prior that approximates the Jeffreys-Haldane improper prior of density $1/k$. The proper prior chosen for this purpose was a log-Cauchy with appropriate hyperhyperparameters.

Now consider Lester Dubins' problem (de Finetti, 1972, p. 205). An integer n has been selected by one of two procedures A or B. In procedure A the probability of a specific value of n is 2^{-n} ($n = 1, 2, \ldots$), whereas in procedure B the probability is uniformly distributed. Thus

$$P(n \text{ is definable in less than } 10^{1000000} \text{ years} \mid B) = 0.$$

So if B is true we can never get evidence for it. (The universe is only about 10^{15} years old). But it was assumed that n *has* been defined. Therefore $P(B) = 0$. If we had originally judged that $P(B) = 0.5$, then we must change our minds in view of this additional thinking. I don't regard this as an inconsistency, in fact I have argued the value of what I call "dynamic probability". According to this theory we must admit that probabilities can change without new empirical information. See Good (1977).

Regarding the drunken-sailor problem, I don't see the advantage of explaining it in two dimensions rather than in one dimension. I think the problem then reduces to one discussed at the Waterloo conference on statistical inference, following a paper by Fraser.

L. PICCINATO *(University of Rome):*

In principle I have some difficulty to understand fully what "complete ignorance" is, and I would prefer a slightly different approach. The model of professors Giron and Rios generalizes the usual model for decision problems in that it considers that we have not just one probability distribution on the states but that such law is known to belong to a given set. This generalization could be seen in a different way: the standard bayesian model is an ideal paradigm and it is surely useful to have some flexibility when we turn to practical applications (the case mentioned of several decision makers is an example). Therefore the perspective I would like to suggest is that of sensitivity analysis, or robustness, with respect to the choice of the prior.

The paper gives useful indications about how to proceed in this type of analysis. Anyway I am inclined to think that all Bayesians act sometimes as quasi-Bayesians in the sense of this paper: when we use conjugate distributions we are actually dealing with a problem which is in an intermediate position between total ignorance and a fully Bayesian approach where only one probability distribution on the states is requested. But in that case the use of classes of priors is only a matter of formal generality, which is attainable without any practical complication.

I think that these concepts about quasi-Bayesian procedures and the related

mathematical aspects could be remarkably interesting in the framework of practical statistical analysis, I mean when K^* is suitably chosen in order to provide a better understanding of some decision problem. Some good examples of this kind were given e.g. by M. Skibinski and L. Cote (1963). The tools proposed by Giron and Rios could then usefully experimented along similar lines.

I suppose that professors Giron and Rios are substantially in agreement with me about the fact that the Bayesian approach provides an "ideal paradigm", in fact they essentially apply the Bayesian scheme in the rigourous classical way in correspondence with each element of K^*. This makes a remarkable difference with the approach by Skibinski and Cote who, unfortunately in my opinion, do not avoid integrations over the sample space. Let me recall about this that quasi-Bayesian procedures are sometimes imbedded in non-Bayesian frameworks, so that they can be misleading. For example, when two decision functions are compared in the standard non-Bayesian way (i.e. through the risk functions) one can find that decision d_1 is better than decision d_2 when θ belongs to a given subset Ω' of the state space Ω. It is then implicit that if you have a partial information that θ belongs to that subset (K^* could be a class of distributions with a support contained in Ω') you can say that d_1 is simply better than d_2, and hold that further information about θ is irrelevant. However this is not true, in general, also from an "objective" viewpoint (i.e. without using a specific prior) and in standard cases you could find experimental outcomes such that the terminal decision provided by d_1 is worse (in terms of losses) than the terminal decision provided by d_2, for every $\theta \epsilon \Omega'$. This depends of course on the fact that the risk functions are not admissible tools for a Bayesian analysis in post-experimental situations.

In conclusion let me say that I agree firmly with professors Giron and Rios in their attitude to relax usual assumptions without losing the basic aspects of the Bayesian approach, that is logical coherence.

A problem I would raise in connection with professor Hill's paper is the following: how to deal with statistical models if we must get rid with conglomerability? Of course the ground for accepting or refusing conglomerability (or complete additivity) is a logical one, and must be independent from the mentioned question. However, even if I agree that a logically sound approach needs essentially finite additivity only (so that complete additivity becomes a mathematical simplification to be used with care), it seems to me not irrelevant to seek what kind of implications this attitude has with respect to the standard statistical practice.

For "statistical model" I mean as usual a set of probability distributions on the space of possible outcomes, possibly indexed by a parameter whose actual value is unknown. Suppose that all conditional distributions are equal: then the value of the parameter seems irrelevant, at least from an intuitive viewpoint. In fact, if conglomerability holds, we can easily predict the future observations without knowing anything about that value. But, if conglomerability does not hold, it seems that something does not work well with our model, and our usual way of thinking, for example about the role and use of identifiability.

It is clear that a possible answer is that also statistical models must be handled with care, just as the assumption of complete additivity. For example, I think that this view is maintained by de Finetti, who dislikes such things as "statistical hypotheses" and so

on; in fact he often suggests to deal only (or at least preferably) with well defined events, which could be actually falsified or verified; a specific quotation could be de Finetti (1971).

Nevertheless, as a statistician, I find that statistical models are quite useful, at least as a communication tool among the various kinds of researchers involved in a given joint work, and that often some relevant theoretical information can be imbedded in the model.

So, let us come back to the initial question: can we live with non-conglomerability (as professor Hill is proposing) without giving up with statistical models?

C. VILLEGAS *(Simon Fraser University):*

I will comment only on B.M. Hill's paper.

Finitely additive probability measures are conceptually important and are certainly useful in the foundations of probability and statistics. As a matter of fact I have used finitely additive probability measures in two papers (Villegas 1964, 1967). But for technical reasons it is usually better to assume countable additivity.

In recent years increased interest has been shown in the use of betting schemes for the analysis of statistical inferences. In those betting schemes the statistician plays the role of a bookie that posts odds on a family of events, and has to withstand the bets of a gambler. The odds posted by the statistician are said to be coherent (relatively to the betting scheme) if the statistician cannot be made a sure loser. In this betting context support for countable additivity comes from Theorem 6 of Heath and Sudderth (1972). Roughly speaking, the theorem says that if, in the absence of data, the gambler is allowed to make countably many bets, then the posted odds will be coherent if and only if they are based on a countably additive probability measure.

Similar results hold when data are available. Thus, Corollary 1 of Heath and Sudderth (1978) says that, if the gambler is allowed to make only a finite number of bets, then the posted conditional odds will be coherent if and only if they are based on a posterior distribution corresponding to a finitely additive, proper prior.

These results can be extended to obtain a justification for countably additive proper priors. Thus, in a future paper, I will prove that, if the gambler is allowed to make countably many bets, then the posted conditional odds are coherent if and only if they are based on a posterior distribution which corresponds to a proper, countably additive prior.

A new conditional frequency interpretation of statistical inferences has been offered in Villegas (1977a). Future repetitions used in frequency interpretations are not real but hypothetical or simulated, and they should be considered only as a means for learning from the data. In Villegas (1977a) it is argued that better inferences may be obtained if only future hypothetical samples similar to the actual data are considered, because in this way the noise may hopefully be reduced, and we may get a better picture of what the actual sample has to say about the population.

Looking only at future samples which are similar to the actual one means conditioning on the future hypothetical sampling belonging to a compact set. Within the context of a betting scheme this means that all bets are off if the observation does not belong to a compact set.

Using this form of conditioning, the results of Heath and Sudderth can be extended to obtain justifications for improper priors. Thus, in a future paper, I will prove that, if the gambler is allowed to make countably many bets, but all bets are off when the observation is outside a compact set (which may be chosen by the gambler), then the posted conditional odds are coherent if and only if they are based on a posterior distribution which corresponds to a possibly improper, countably additive prior.

Objections against the use of improper priors have been raised also from the point of view of admissibility. Thus, in the estimation of a location parameter, the Bayes estimators based on a uniform prior may be inadmissible (Stein, 1956), However, results will in general be different if we condition on the observed value belonging to a compact set. Then the risks become conditional risks, and a new concept of conditional admissibility emerges. In a future paper it will be shown that it is not difficult to modify C.R. Blyth's (1951) proof of admissibility of Bayes estimators based on improper priors are conditionally admissible in the above mentioned sense. And, according to the new frequency interpretation of Villegas (1977a), this is all that is needed in statistical inference.

It should be recognized that there are two lines of development for Bayesian statistics: one is the personalistic line, based on personal, subjective priors, and the other is the logical probability line, based on logical priors that represent ignorance. The second line is not so well developed as the first one, but some progress has already been made (Villegas, 1977b).

Logical priors are usually invariant under a given group. Therefore they are not only relative to a given model, but even more, they are relative to a group that is given as an integral part of the model. Fraser's structural models are useful from a logical probability viewpoint. Stone's example becomes a structural model if the group with two generators is considered as an integral part of the model. In that case the logical prior is the uniform prior. But the story of the lady and the sailor brings other considerations which favor the selection of the other prior. Since the likelihood principle ignores the possibility that a group may be given as an integral part of a model, it is not valid in a logical probability approach to statistical inference.

J.M. DICKEY *(University College of Wales Aberystwyth):*

I find the paper by Professors Giron and Rios intriguing, especially the idea of working with "extremal" posterior distributions to surround, so to speak, the coherent inferences of persons whose prior distributions lie within a range of distributions. This harmonizes closely with my idea of "scientific reporting" as a reporting of the prior-to-posterior transformation over a class of prior distributions conceived as containing the reasonable uncertainties of a population of scientists (Dickey, 1973). Various graphical methods are available for reporting such a distribution-valued functional. Bounding methods are also proposed in both papers.

The idea of Giron and Rios seems simple and straightforward, and in view of the long story of statistical theorists saying they could not know their prior distributions, one would have expected this idea to have developed much earlier. The authors have

done a great service in carefully setting out the theory. I look forward to seeing more applications.

An obvious direction of generalization which may interest the authors is to replace the set K^* of permitted, equally acceptable, prior distributions by a new distribution of distributions, an expression of uncertainty concerning uncertainty. This could be used to generate sets K^*, for example, by setting thresholds on some density for the new distribution in function space. My own paper in this meeting investigates the form such a distribution might take and its use in the problem of assessing (choosing) a subjective probability distribution. See also Dickey and Freeman (1975). There are, of course, logical difficulties with the meaning of such a second-order belief distribution, and in both our settings one would need to resist the temptation to marginalize by taking the second-order average of first-order beliefs.

Finally, I should like to complain that the term "agreement set" for K^* or its convex closure could be misleading. Presumably, the decision makers agree in having their opinions fall in the set. But then they **disagree** on which distribution is appropriate *within the set*.

There are many diverse issues raised in Proffesor Hill's paper. The main point for me is that he argues with De Finetti in favour of merely finite additivity, and consequent nonconglomerability. In the sphere example this would mean that *all* the great circles through the poles could have uniform distributions within a circle, while the two-dimensional probability on the sphere could also be uniform. This conflicts with the conditional distribution that would be obtained by a limiting argument conditioning on an observed small interval of longitudes.

I am grateful to Professor Hill for personal conversations in which he informed me that his issue in the sphere example is not the same issue as brought forward by Kolmogorov (1933, Ch. V, Sec. 2). Kolmogorov cites Borel for what I have called the Borel-Kolmogorov nonuniqueness, whereby a conditional distribution obtained in the usual way from a joint density will depend on the conditioning variable used to define the conditioning event, rather than just on the conditioning event itself. In the sphere example, a different experiment which slices the earth by parallel planes will produce uniform distributions within the circles produced.

Apparently, Hill is not thinking of any experiment at all when he asks for the distribution within a great circle, but wants to base a conditional distribution on the purely logical statement that a particular great circle obtains. He wants finite additivity "in part for purely logical reasons". He also claims to need it for practical reasons, since "one will often find it advisable to make approximations using infinite models".

I simply do not understand the practical need for merely finite additivity. When I make approximations to finitistic situations using infinite models I shall not restrict myself to using only a few logical statements to obtain a mathematical model. I shall look at the real-world problem and the real uncertainties involved. For example, just because some exercises in textbooks fail to give information distinguishing between equal-length intervals would not be enough to tempt me in a real-world problem to use a uniform pseudodensity over the whole real line. It seems to me that countable additivity, conglomerability, and proper integrable distributions enable us to treat real problems realistically, without worrying that the mathematics itself will deal us an

unpleasant surprise. I should like to hear further about the practical issues. Mervyn Stone's lazy-Bayesian examples over the years have only served to warn us against nonintegrable distributions, which were already ruled out by the axioms of coherent behaviour.

M.H. DEGROOT *(Carnegie-Mellon University):*

In the paper by Rios and Giron, partial information about a prior distribution is represented by simply dividing all distributions in Ω^* into a set K^* of possible prior distributions and the complementary set of impossible prior distributions. Wouldn't it be more reasonable to assign probabilities to the distributions in Ω^* ; i.e., to assign a probability distribution P^{**} to the set Ω^* . In turn, one might then assign a distribution P^{***} to the set Ω^{**} of all distributions P^{**} , etc. In brief, why not develop a hierarchical model?

D.A.S. FRASER *(University of Toronto):*

I wish to discuss three points connected with Professor Hill's paper: how the Stone example provides a strong counter example to the Strong Likelihood Principle; how the modelling of the internal variable of the Stone example leads to the overriding probability statements; and how information concerning a realization from such an internal variable must satisfy certain requirements as to how it was produced in order to be acceptable for probability calculations.

The Stone example A has seemed to me to be a very striking counter example to the Strong Likelihood Principle. Professor Hill has doubts and discusses the distinctions between the full parameter and two interesting component parameters. The full parameter for the model is $\theta = p$, the path from the origin to the treasure; a derived parameter of interest is $\theta_1 = \theta_1(\theta) = \overrightarrow{x}$, the last directed segment of p; a further derived parameter of interest is $\theta_2 = \theta_2(\theta) = x$, the end point of the path p. These parameters are not the same and yet, given a data-point $\hat{p}$ (the path to the sailor), the possible values for them fall into a one-one equivalence. The observed likelihood function is a function of the full parameter θ; as presented it is not a likelihood for either component parameter but does of course provide information concerning each. The full parameter space is $\Omega = \{p\}$, the free group on two generators.

A salient feature of the Stone example is the striking contrast between the following two results: the likelihood function from data assigns equal likelihood ($\frac{1}{4}$) to each of four possible paths to the treasure; direct probability arguments based on an internal variable put an operational 3/4 probability on a preferred one of the four possible paths. Thus, likelihood says the four possibilities are on a par one-with-another, whereas an internal variable nominates one of the four possibilities as a 75% favourite. The example seems to make clear that likelihood does not contain all the needed information.

Perhaps some further details can add emphasis to this result. For the Strong Likelihood Principle my own preference is a prescription in the following form: from a statistical investigation use only the observed likelihood function. An alternative form closer to that proposed by Birnbaum is the following: if the likelihood function from a

first model + data-point is the same as the likelihood function from a second model + data-point then the inferences should be the same in the two cases. For this we note that the likelihood function is a nonnegative function on the parameter space Ω left indeterminate to a positive multiplicative constant; that is, it is a positive ray from the origin in the vector space R^n. The equality, then, of two likelihood functions requires the same parameter space Ω and the same ray in R^n.

Is the probability imbalance and the constant likelihood on four parameter points, a necessary consequence of the unusual parameter space? Or could we find another model + data-point that yields an identical likelihood function but with a different probability imbalance or more simply with say symmetry on the four possible parameter values? We examine this latter possibility.

For this suppose we start with some particular likelihood function obtained from the Stone example with a data-point; let $\hat{p}_0$ be the data point and $\theta^1, \theta^2, \theta^3, \theta^4$ be the four possible parameter values consistent with $\hat{p}_0$. For a second model we take the same parameter space Ω, the same sample space $S = \Omega$, and the following very special probability structure:

$$
\begin{aligned}
&P(\hat{p}_0|\theta^i) = \tfrac{1}{4}, p(e|\theta^i) = \tfrac{3}{4} && i = 1,\ldots,4\\
&P(e|\hat{p}) = 1\\
&P(\theta|\theta) = \tfrac{1}{4}, P(e|\theta) = \tfrac{3}{4} && \theta \neq \theta^i, \theta \neq \hat{p}
\end{aligned}
$$

where e is the identity element. The likelihood function from the sample point $\hat{p}_0$ is the same as that from the Stone example and yet the model treats the four parameter values symmetrically. This provides the formal contradiction to the Strong Likelihood Principle.

Clearly the likelihood function alone is not enough. Of course many statisticians do not accept the Strong Likelihood Principle, usually on the good grounds that many fruitful statistical results are available outside the Principle. The Stone example however is direct: the likelihood function alone omits an essential probability property.

The Stone example contains a primary random system - the spinning of the woman at the end of the taut thread. Based on this process, there is an overriding 3/4 probability that the path is extended, and correspondingly an overriding 3/4 probability that the last path segment comes from the treasure. This seems to provide the motivation for Stone's "classical statistician" although details are not given. A formal version of the preceding appears in my Comments on the Stone paper but was sidestepped in Stone's elusive rejoinder. The recognition of the fundamental importance of primary or internal random systems seems long overdue in contrast with the intensive activity in some areas of contemporary statistics.

Prof. Hill also considers the system in which a point is selected uniformly on the surface of a sphere with a designated north and south pole; an investigator is given the exact longitude of the point. Prof. Hill seems to show preference for a uniform distribution for the point on the given great circle of longitude. This is in conflict with a basic probability position, both classical and Bayesian, that marginal and conditional probabilities go together to give joint probabilities. For we note that the standard

conditional distribution given that longitude equals the recorded value has density proportional to the cosine of the latitude.

What is the key element in the preceding conflict? We have a situation where there is information concerning a realization from a random system, and yet the information does not fully identify the realization. Discussions of conditional probability show that we need to know not only the information as to possible values for a realization **but also** how that information was produced; see for example Fraser (1976, Ch. 4), Fraser and Brenner (1979).

Most discussions of conditional probability overlook the need to know how the information is produced concerning the possible values for the concealed realization. Without it, contradictions are obtained and various "paradoxes" are to be found in the literature. Information without knowledge concerning its production does not support probabilities. This is a very fundamental argument against the Bayesian position.

S. FRENCH *(University of Manchester):*

I wish to comment upon Girón and Ríos's paper. First, a few points of a technical nature. The authors have to use topological properties of Ω and ideas of continuity in case (a) of their theory. I wonder if these assumptions can be weakened by using the approach of Krantz et al. (1971). These latter authors have avoided the use of topological assumptions in their measurement systems instead relying on weaker solvability conditions applied to the underlying qualitative orders. Perhaps Girón and Ríos could generalise their results similarly.

Early in their Paper, Girón and Ríos discuss partial orders derived from convex cones in $\mathbb{R}^n$. I wonder if they have seen the recent work of Hartley (1978). His approach seems to give the weakest set of conditions available for playing with such orders. Also for a practical illustration of the use of such cone-orders in the sensitivity analysis of a decision problem, the authors have referred to Fishburn (1964). His paper (1965) in Operations Research is also of relevance and, perhaps, easier to find.

Turning now to what I believe to be a more important question. The authors consider a decision maker who knows his utility function perfectly and his subjective probabilities imperfectly. Is this a reasonable model? It says essentially that he can locate for each possible consequence an exactly equivalent gamble based upon some auxillary experiment. Is it feasible to suggest that he can do this, yet be unable to locate a gamble based on the auxillary experiment equivalent to a gamble based upon an unknown state of nature? The problem of measuring subjective probability is just as easy, or difficult, as that of measuring utility. In terms of axiom systems my point is this. In assuming the existence of a utility function $u(\cdot)$ the authors are hiding under their decision space another decision space in which the ordering of decision rules is complete.

Finally, since I see the primary use of this theory to be in the area of sensitivity analysis, perhaps the following suggestion is appropriate. I have seen papers in which, as here, the utilities are known and the probabilities only partially known and also papers in which the probabilities are known and the utilities partially known. I wonder if duality theorems of mathematical programming can give us a means of allowing both quantities to be partially known? Perhaps the authors know of a reference in this area.

D.V. LINDLEY *(University College London):*

I have a brief comment on the paper by Girón and Ríos. How does a partially ignorant person act? Bayesian decision theory is a recipe for the selection of a single act: Bayesian inference provides all the information about the unknowns in the problem needed to select the act. The authors' theory ends with a class of acts: if this class contains more than one member, how is a unique act to be selected in cases where no more data is available? A possible application of this theory is to multiple decision problems where several opinions are present, but again there is the difficulty of the choice of a single act.

Turning now to Hill's paper, Kolmogoroff (1933 Ch. 5), makes the point that conditional probability is either defined with respect to an event of non-zero probability, or for a random variable $x(w)$ defined over a space of values of w, and not for the single event $x(w) = x_0$ when this has probability zero. My understanding is that Kolmogoroff would want to know what random variable gave longtitude 30; was it longtitude, or was it some other variable? This seems right to me and I'd welcome Hill's comments on this. It constrasts with the likelihood principle since it requires knowing not just that the longtitude was 30 but what other values (like 25) one might have had. What are the "gaping holes" - mentioned in the first paragraph - in a sigma-additive theory using proper distributions?

REPLY TO THE DISCUSSION

F.J. GIRON *(Universidad de Malaga)* and S. RIOS *(Universidad de Madrid):*

We would like to start by paraphrasing Dempster, quoted by Bernardo (1979): "In the area of statistical inference, there must be little that any one has thought about that Dr. Good has not written about, to the point that a computerized information retrieval system would be very helpful to scholars in the area".

Our paper does not intend to be a historical paper nor a paper on the history of partially ordered probabilities, and explicit reference to previous ideas on the subject are mentioned in section 1.

With respect to the priority claimed by Professor Good, it is worthwhile mentioning here that the idea of approximating sharp probabilities by means of an interval is to be found in an early paper by Fréchet. Unfortunately we have not been able to trace back the appropriate reference thought it might be found in Econometrica. To what extent early ideas influence a theory is always a controversial subject. As an example some french authors and others refer to the Kolmogorov axioms as the Fréchet-Kolmogorov axiomatic set up.

We agree with Professor Piccinato that the Bayesian approach is the "ideal paradigm". Yet to contemplate the quasi-Bayesian theory merely as a sensitivity analysis approach is, we believe, to focus just on a particular aspect of the model. Its interest resides in that the hypothesis of the model are more general than that of the Bayesian model; more mathematically tractable than other former approaches (the one mentioned by Professor Piccinato of Skibinski and Cote (1963) could be an example); and above all in the main theorem that establishes an equivalence between the ideas of partial ordering of decision rules and partial information in terms of probability

measures. On the other hand, the interpretation of the theory from the point of view of sensitivity analysis also stressed by Dr. French in his contribution to the discussion, allows for a unified and systematic treatment of the problem of sensitivity analysis in Bayesian decision making.

With respect to the problem of non-admissibility of quasi-bayesian procedures that Professor Piccinato mentions nearly at the end of his contribution, the situation here is exactly the same as in the Bayesian case. Problems of admissibility in post-experimental situations depend on three facts: 1^{ro}, prior partial information may be incompatible with some experimental outcomes; 2^{nd}, the support of distributions of K^* may be a proper subset Ω' of Ω, thus discarding some states of Nature; 3^{rd}, the judicious use of Fubini's theorem.

We are grateful to Professor Dickey for his comments and, like him, we would also like to see more applications of the theory. We have taken up his complaint and have change the term "agreement set" into the more innocuous term, and we believe it more apt, "feasible set".

The generalization suggested by Professor Dickey, which is also pointed out by De Groot in his contribution, of developing a hierarchical model seems interesting, specially the idea of setting thresholds on some distribution of distributions (the second stage in the hierarchical model) to generate sets K^* of first-orders beliefs. This idea is also closely related to the paper by De Robertis and Hartigan (submitted for publication to the Annals of Statistics) about ranges of measures as an expression of partial ignorance.

Professor De Groot's suggestion of developing a hierarchical model is discussed at length in the paper by Good at this conference. However as he presents the hierarchical model we would have in the first stage a complete ordering given by the probability measure P^{**}. In the second stage, we would now have as new states of Nature the set of all probability measures on Ω^*, that is Ω^{**}, on which a new distribution P^{***}, could be assigned, and so on; so that this would drive to a complete ordering of decision rules by marginalizing on succesive stages unless in any of the stages the probabilities assigned were partially ordered (cf. Good, p. 7, line 12 of his revised manuscript) and thus the final ordering of decision rules would only be partial.

Our paper is an attempt to characterize these partial orderings which, of course, can be embeded in a hierarchical model, one of the stages of which at least corresponds to partially ordered probabilities.

Dr. French suggests a generalization of our paper by using the approach of Krantz et al. (1971). We believe this program can be carried out along their lines. Another possible generalization of the results of our paper for partial comparative probabilities, that also takes into account the role of experimentation, could be based on the works of Fine (1971, 1973). Yet we want to point out two facts: 1^{ro}, in the Krantz et al. approach the subjective probability derived is finitely additive as in case (b) of our paper, in which the only requirement is the existence of a bounded utility function; 2^{nd}, the topological assumptions of case (a) guarantee the σ-additivity of probability measures of set K^* and neither compacness of Ω nor continuity of acts can be dropped if one is seeking for σ-additive subjective probability measures. Further, this allows for a parallel and systematic treatment of cases (a) and (b) and renders the proofs of main

theorems almost trivial by using the topological dual of spaces $C(\Omega)$ and $B(\Omega)$, respectively.

Unfortunaltely the paper by Hartley (1978) French mentions has not reached our hands at the time of writing the rejoinders.

We are in agreement with Dr. French when he says that our model is not reasonable because it takes for granted that utilities are perfectly known and, in practice, both quantities, probabilities and utilities, are only partially know. However we know of no duality theorem of mathematical programming that can accomodate the case when both quantities are partially known, although we think this to be a very important issue in practical decision making.

The question Professor Lindley raises is a key one; namely, how does a partially decision-maker act? The answer is in the premises of the theory, precisely in the formulation of Axiom 1. If a partially ignorant person has only a limited amount of information, then he selects a class of non-dominated acts such that it is worth while betting on these acts against other acts. Usually, this class contains more than one act, and then it is not clear how a single act is to be selected. A possibility would be to randomize among these acts, but this would be equivalent to consider a hierarchical model and this, in turn, is equivalent to having your decisions linearly ordered.

On the other hand, Bayesian decision theory may also lead to a class of acts (when several decisions attain the same Bayes risk) and then it is not also clear how to randomize.

In short, if one is partially ignorant one cannot expect to be able to linearly order the set of possible decisions.

Quasi-Bayesian theory takes into account the possibility of partial-instead of total-information thus generalizing Bayesian theory. Then, it is proven that such a hypothesis is intimately related to partial ordering of decisions as opposed to the complete ordering of decisions in Bayesian theory. Which is more plausible is a question of applicability and even of taste.

B.M. HILL *(University of Utah and University of Michigan):*

I would like to thank all of the discussants for their comments. Before responding to individual discussants it may be helpful to make some general remarks. The primary purpose of my article was to focus attention on the axioms for Bayesian inference and decision theory. The de Finetti axioms are weaker than others in that they allow finitely additive distributions and non-conglomerability. It is hard to imagine satisfactory axioms for quantitative probability that are still weaker than those of de Finetti, and failure to abide by axioms 1 and 2 can subject one to sure loss. Should, however, these axioms be strengthened? Should, for example, one require that decision procedures be extended admissible, or perhaps even admissible. If there are serious arguments so to strengthen the de Finetti axioms, then there should exist telling examples clearly demonstrating the shortcomings of the finitely additive approach. The examples that I chose were those that seemed most clearly to suggest possible shortcomings , and I attempted to determine just how serious a case could be made to strengthen the axioms. Thus in Mervyn Stone's example, I think most of us will prefer the Bayesian solution based upon a uniform prior distribution for N, whether this is taken literally or as an

approximation using proper prior distributions. The de Finetti axioms, however, do not exclude the finitely additive prior distribution $\tilde{q}(\cdot)$ that leads to the Stoned Bayesian Posterior. So it seems natural to ask exactly what ill consequences will occur if one were to use this prior distribution. Stone suggested that over a long sequence of repetitions of the experiment the Stoned Bayesian would get the treasure less frequently than someone who used the confidence solution. My discussion of the sphere example was meant to suggest why his argument is not very convincing even within the frequentist theory. For it is circular. Only if you have already rejected finite additivity and non-conglomerability does the argument suggest an unambiguous frequency for obtaining the treasure.

Now let me turn to the individual discussants. Professor Good suggests that the paradoxes (if such they be) arise from incorrect limiting arguments. I do not think so. Indeed, there are no limiting arguments in my article, and I tried to avert such a misinterpretation by conditioning upon an *exact* great circle. Admittedly this is an idealization for real world problems. But so conditioned the problem is still logically meaningful, analogous idealizations are commonly made in statistics, and there can easily arise situations where the appropriate conditioning event is not specified, i.e., we are not told whether the measurement process restricts us to the region between two parallel planes, or between intersecting planes through the poles, or still other regions. (Such sensitivity to the precise form of the conditioning event is still another reason to argue for the freedom of the finitely additive approach). Would Professor Good, along with Professor Fraser, simply refuse to discuss the question in the absence of such information? Professor Good then refers to the distinction between a random great circle on a sphere and a great circle known to pass through a fixed point (See also my footnote # 3). He should then be able to point to the ill consequences from taking the point as uniform on the great circle in the latter case. But I suspect that he will only be able to demonstrate such consequences if he has already assumed countable additivity and consequently also conglomerability. With regard to Professor Good's discussion of the Dubin's problem, I find his argument that $P(B)=0$ even less convincing than my own that $Pr\{T_2\}=0$ in Example 3. First of all the age of the universe is not so terribly well known as he implies. Would Professor Good be greatly surprised if by the year 2,079 some new theory suggested that the age should be revised upwards to 10^{25} years, or whatever? Secondly, I am concerned with his emphasis on "definability". Suppose we are discussing the number of elementary subatomic particles in the universe, and for the sake of argument assume that there is a well-defined number. Then although under hypothesis B it will probably take awfully long to "define" this number, the number has been assumed to exist, and the finitely additive uniform distribution (at least in the upper tail) may represent ones' opinions much more adequately than any countably additive distribution. What if, for example, one simply cannot name a number such that the probability to the right of that number is less than 10^{-100}?

I find Professor Good's discussion of "dynamic probability" intriguing. But I doubt that it is relevant to the Dubin's problem or Example 3. The reason for my doubt is that the alteration in $Pr\{B\}$ or $Pr\{T_2\}$ that he suggests would be made merely to avoid non-conglomerability, without having advanced any serious argument as to the need for conglomerability. Finally, I was sorry that Professor Good did not choose to

discuss the drunken-sailor problem. Although the one dimensional version has much in common with it, there are certainly real differences between the two versions, for example, the non-amenable free group on two generators, and in particular the finitely additive analysis of the problem in two dimensions would appear to be new.

Professor Dickey questions the practical need for merely finitely additive distributions. I think all three of the examples I discussed suggest such a need. In the drunken-sailor example Professor Dickey presumably would object to the uniform finitely additive distribution on N, and at best would view it only as an approximation for a proper countably additive distribution. Even so, is it not sometimes useful to have available such a simple approximation, rather than to labor over the fine details of ones prior distributions in a situation where there is little to be gained from such labor? Similarly for the problem on the sphere. What if Professor Dickey does not have available all the real-world information he would like, so that the shape of the region delimited by the actual measurement process is not known. Keeping in mind the possibility of parallel hyperplanes, would he exclude the uniform distribution on a great circle, even as an approximation? Would he simply ignore the problem, as so many non-Bayesians do with regard to any problem that doesn't fit into a neat Kolmogorov-frequentistic mold? Finally, improper prior distributions can often be given a finitely additive interpretation, so that they are in fact consistent with the de Finetti axioms for coherent behaviour. (See my footnote n° 4.)

Professors Dickey and Lindley both point out that in the Kolmogorov approach it is not sufficient to know the conditioning event, and that one also needs information regarding the conditioning variable used to obtain that event, at least when the event has probability zero. This is true, and seems to me to cast doubt upon the approach itself. As I argued above, does this mean one should say nothing when such information about the variable is not available? My notion of uniformity on the surface of the sphere includes not only the evaluation of probabilities as proportional to surface area for sets that have surface area, but also the notion that conditional upon the point being in any specified finite sets of points, all such points are equally likely, and conditional upon a great circle, probability is proportional to arc length. This strong notion of uniformity is not possible in the Kolmogorov approach, but is compatible with the de Finetti axioms. Why should such an opinion be excluded? The contrast between the Kolmogorov approach and the likelihood principle is itself one of the gaping holes. Conventional statistical models often assume the data to have probability zero, and within the model Bayesians are forced to consider their probabilities conditional upon an isolated event of zero probability, although Kolmogorov (1933, p. 51) wishes to exclude precisely this situation. Of course one can take refuge in a finitistic approach, but then we lose the advantages in simplicity that we obtain with conventional models. I think the situation is somewhat akin to that with regard to stopping rules and the likelihood principle. A conventional non-Bayesian analysis is not really possible without knowledge of the stopping rule, and since we rarely if ever know the true stopping rule, a conventional analysis could at best yield only certain inequalities. In the same way, a conventional Bayesian analysis in the Kolmogorov system is only possible if one knows the conditioning variables, and I submit that in most applications they too are unknown. But we will nonetheless draw

inference and make decisions. I believe that the arguments against such an approach are circular. They have force only if one has already accepted countable additivity.

Professor Villegas has an interesting alternative way to deal with inadmissibility, but it does not seem appropriate to discuss this here.

Now let me turn to Professor Fraser's comments. Despite my very best efforts Professor Fraser still regards the Stone example as convincing evidence against the likelihood principle. I cannot agree. First of all, the only kind of likelihood principle that can have any credibility at all is one compatible with Bayesian inference. For even a non-Bayesian would have to reject a version of the likelihood principle that was not compatible with Bayesian inference whenever he thought that the prior distribution had a frequency interpretation. This in turn implies that a data-dependent transformation of the original parameter must be excluded as evidence against the likelihood principle, since the transformed parameter would have a different "prior" distribution than the original parameter, as I hope my discussion of $\vec{E}$ and $\dot{E}$ makes clear. Professor Fraser apparently now accepts this but offers still another experiment to provide a "formal contradiction to the Strong Likelihood Principle" (nearly the same as my likelihood principle). In order for his new experiment to make sense we must assume that the new experiment consists in first performing the original experiment to obtain his data $\hat{p}_0$ (my $\hat{p}$), and then performing some additional experiment to generate his new likelihood function. (Note that this must be done for all possible $\hat{p}_0$, not just a particular realization). Even if he is correct that the modified experiment yields the same likelihood function as the original experiment the argument loses its force because whatever asymmetry is involved in the original experiment must then be reflected in Fraser's modification. But his purpose was to treat the four parameter values symmetrically.

Professor Fraser also argues against the likelihood principle on the grounds that it counters many "fruitful statistical results". It is of course counter to conventional significance testing, but Bayesians are hardly alone in regarding such tests with a great deal of skepticism.

Finally, Professor Fraser discusses the need to know how information is produced, as was raised by Professor Lindley and discussed above in my reply. This is presumably a much more fundamental issue for Professor Fraser than for Professor Lindley, and is at the root of much criticism of the Bayesian approach, dating back at least to Venn. Thus Professor Fraser presumably would have us do nothing without such knowledge, and also without knowledge of stopping times, etc. This perhaps restricts the applications of statistics to the empty set. I would also ask Professor Fraser exactly how we are to discriminate between the various forms of knowledge, i.e., between knowledge that can be (in his sense) validly represented by a probability distribution, and opinions that cannot be so represented?

Professor Piccinato raises some intriguing questions regarding the use of conventional statistical models. As I see it finite additivity and non-conglomerability offer us some additional freedom in the probabilistic expression of our knowledge. In some applications it will be important to take advantage of that freedom, and in some it will not. As in my reply to Professors Dickey and Lindley, I think that in the sphere example it is important not to force oneself into the Kolmogorov mold, at least not

without careful consideration as to the knowledge that one wishes to express. But I do not think there is anything incompatible between the careful use of conventional statistical models and the de Finetti theory. It is true that conventional parameters can often be dispensed with, as for example in an exchangeable sequence of zero-one variables, and where this is possible it seems preferable to do so rather than to invent artificial parameters. (In Hill, (1969), it is shown how conventional linear models can also be dealt with in this way). But on the other hand there are many situations which cannot as yet be handled satisfactorily in terms of the observable variables, and parametric models offer a convenient flexible way of dealing with such situations. In any case it is not a question of incompatibility, but merely of seeing things in another light. Finally the question as to the case where the conditional distributions are the same, and so as Professor Piccinato suggests, the parameter might seem to be irrelevant, is indeed a paradox of non-conglomerability. But despite the intuitive plausibility of merely dispensing with the parameter, perhaps we should recall that we must have had some reason to view the situation as non-conglomerable in the first place, and then to choose as best we can between the conflicting intuitions.

REFERENCES IN THE DISCUSSION

BERNARDO, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. B* **41**, 113-147.

BLYTH, C.R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* **22**, 22-42.

BRENNER, D. and FRASER, D.A.S. (1979). Conditional probability and the resolution of statistical models. *Statistische Hefte.* (to appear).

CROOK, J.F. and GOOD, I.J. (1979). Part II of Good (1976). *Ann. Statist.* **20**, 148-159.

DE FINNETI, B. (1971). Probabilitá de una teoria e probabilitá dei fatti. *Studi di Probabilitá Statistica e Ricerca Operativa*, 86-101. Universitá di Roma: Istituto di Calcolo delle Probabilitá.

— (1972). *Probability, Induction, and Statistics.* New York: Wiley.

DeROBERTIS, L. and HARTIGAN, J.A. (1979). Ranges of prior measures. *Tech. Rep.* Yale University

DICKEY, J.M. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *J. Roy. Statist. Soc. B* **35**, 285-305.

DICKEY, J.M. and FREEMAN, P.R. (1975). Population-distributed personal probabilities. *J. Amer. Statist. Assoc.* **70**, 362-364.

FINE, T.L. (1971). Rational decision making with comparative probability. *Proc. IEEE Conf. on Decision and Control.* 355-356.

FISHBURN, P.C. (1964). *Decision and Value Theory.* New York: Wiley.

— (1965). Analysis of decisions with incomplete knowledge of probabilities. *Operations Research* **13**, 217-237.

FRASER, D.A.S. (1976). *Probability and Statistics, Theory and Applications.* Toronto: University of Toronto. Textbook Store.

GOOD, I.J. (1950). *Probability and the Weighing of Evidence.* New York: Hafners.

— (1952). Rational decisions. *J. Roy. Statist. Soc. B.* **14,** 107-114.

— (1962). Subjective probability as the measure of a non-measurable set. *Logic, Methodology, and Philosophy of Science: Proc. of the 1960 International Congress* (Stanford), 319-329.

— (1965). *The Estimation of Probabilities.* Cambridge, Mass.: M.I.T. Press.

— (1967). A Bayesian significance test for multinomial distributions (with discussion). *J. Roy. Statist. Soc. B.* **29,** 399-431.

— (1976a). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4,** 1159-1189.

— (1976b). The Bayesian influence, or how to sweep subjectivism under the carpet. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* (Hooker, C.A. & Harper, W., eds.) Vol. 2, 125-174, Dordrecht: D. Reidel.

— (1977). Dynamic probability, computer chess, and the measurement of knowledge. In *Machine Intelligence* **8** (Elcock, E.W. and Michie, D., eds.) 139-150, New York: Wiley.

GOOD, I.J. and CROOK, J.F. (1974). The Bayes/non Bayes compromise and the multinomial distribution. *J. Amer. Statist. Assoc.* **69,** 711-720.

HARTLEY, R. (1978). On cone-efficiency, cone-convexity and cone-compactness. *SIAM J. Appl. Math.* **37,** 211-222.

HEATH, D.C. and SUDDERTH, W.D. (1972). On a theorem of De Finetti, oddsmaking, and game theory. *Ann. Math. Statist.* **43,** 2072-2077.

— (1978). On finitely additive priors. *Ann. Statist.* **6,** 333-345.

HILL, B.M. (1969). Foundations for the theory of least squares. *J. Roy. Statist. Soc. B* **31,** 89-97.

KEYNES, J.M. (1921). *A Treatise on Probability.* London: MacMillan.

KOLMOGOROFF, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer. English translation as *Foundations of the Theory of Probability.* New York: Chelsea, (1950).

KOOPMAN, B.O. (1940a). The basis of probability. *Bull. Amer. Math. Soc.* **46,** 763-764.

— (1940b). The axioms and algebra of intuitive probability. *Ann. Math.* **41,** 269-292.

KRANTZ, D.H. & *et. al.* (1971). *Foundations of Measurement, Vol. 1.* New York: Academic Press.

SKIBINSKI, M. and COTE, L. (1963). On the inadmissibility of some standard estimates in the presence of prior information. *Ann. Math. Statist.* **34,** 539-548.

SMITH, C.A.B. (1961). Consistency in statistical inference and decision. *J. Roy. Statist. Soc. B* **23,** 1-25.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp.* **1.** 197-206.

VILLEGAS, C. (1964). On qualitative probability σ-algebras. *Ann. Math. Statist.* **35,** 1787-1796.

— (1967). On qualitative probability. *Amer. Math. Month.* **74,** 661-669.

— (1977a). Inner statistical inference. *J. Amer. Statist. Assoc.* **72,** 453-458.

— (1977b). On the representation of ignorance. *J. Amer. Statist. Assoc.* **72,** 651-654.