You might imagine that you have done up to say a thousand trials and if that is the sort of thing you had in mind before you started experimenting you will probably be satisfied to use as a distribution when p is not a half, something roughly uniform, though possibly concentrated in a narrow interval covering $p = \frac{1}{2}$. I do not think you can use a uniform distribution going the whole way from $p = 0$ to $p = 1$, if it is a question of the bias of a coin; for example, you might use something uniform in a rather narrow range or something like $p^{\alpha}(1-p)^{\alpha}$ to make it smooth. But at the back of your mind you have the idea that you are going to do an experiment of reasonable size. However, if you were told that the experiment might become enormously large, and if you can imagine some possible results of an experiment of that size, you may decide that you would accept E.S.P. even if p were very close to $\frac{1}{2}$. Now if the sigma-age were greater than, say, 10, or something like that, you would have to think awfully carefully. If you were really doing this experiment you would have to think of a great many possible results of the experiment to make sure that you were being consistent; and if you did that, then it may well be that you would decide to use a very curious sharply peaked prior distribution. But I think you might well come round to advance the view that if on tail area probabilities the chance was as small as 10^{-10} this would still not be evidence in favour of E.S.P. But after it really happened, you might begin to doubt your original judgements. So you must try to think out in advance and decide on a prior distribution which would enable you to be consistent whatever happens. That is in theory. It might be very difficult. You do not need more than one test depending on the intentions of the experimenter. In principle you must think of all possibilities and then decide on a single test which will depend on a single prior distribution.

Mr C. B. WINSTEN: What I was going to say is so closely related to what Dr Good was saying that I hasten to follow him as closely as I can. I, too, want to emphasize that one often may learn about 'initial probabilities' from final probabilities, and I feel this affects the argument quite considerably. Sometimes, as in simple urn experiments, one deduces final probabilities from initial probabilities. On the other hand, one can imagine a situation like that Dr Good has just described

in which one has a set of hypotheses which we can call H_1, H_2, H_3, say, and one can suppose a set of observations producing likelihoods, l_1, l_2, l_3. Then one can imagine an observer being given a set of likelihoods, and then being asked which ratio of experimentally obtained likelihoods for hypotheses 2 and 3 he would accept as establishing these hypotheses as having about equal credence, or acceptability, or posterior probability. As a result of this procedure one is establishing the 'prior probabilities', if one can call them that. The content of Bayes's theorem in this situation is, however, completely different from that in the urn case; indeed, it seems to me mistaken even to pose the whole thing as being an application of Bayes's theorem. Instead of saying that the posterior probability is proportional to prior probability times likelihood, one is deducing from the observer's rating of the likelihood scales what weights are needed to establish equal posterior belief.

The term 'weight' is preferable to the term 'probability' because if one is going to use the term probability for something which you obtain from this merging of the likelihood scales, then one must be visualizing carrying out a further experiment later. The numbers one is going to obtain from the weights and the likelihood ratios of the present experiment are then going to be used as weights for the likelihoods of the next experiment. And only in that situation is it in fact worthwhile to try and set up what one might call an analogue of Bayes's theorem. Otherwise it seems to me that one simply tries to discover somebody's degrees of belief from his scaling of the likelihood function. In that situation it seems to me that one should not really even mention Bayes's theorem. One should mention the corresponding formula as a possible summary of the ways in which people treat a summing up of a likelihood choice criterion.

I do not know whether in some situations one could get intermediate cases. I wonder in the light of this whether Professor Barnard's distinction between acceptabilities and probabilities is concerned with whether one can carry out a particular sort of numerical analysis on the choices between likelihoods.

BARNARD: To come back to this point about likelihood and normalization, and in a way back to the general issue, Professor Savage, as I

understood him, said earlier that a difference between likelihoods and probabilities was that probabilities would normalize because they integrate to one, whereas likelihoods will not. Now probabilities integrate to one only if all possibilities are taken into account. This requires in its application to the probability of hypotheses that we should be in a position to enumerate all possible hypotheses which might explain a given set of data. Now I think it is just not true that we ever can enumerate all possible hypotheses. We must always leave it open that someone with more imagination, or more knowledge, or more information can come along later and suggest an explanation of the fact with which we are confronted that we just had not thought of at all. If this is so we ought to allow that in addition to the hypotheses that we really consider we should allow something that we had not thought of yet, and of course as soon as we do this we lose the normalizing factor of the probability, and from that point of view probability has no advantage over likelihood. This is my general point, that I think while I agree with a lot of the technical points, I would prefer that this is talked about in terms of likelihood rather than probability. I should like to ask what Professor Savage thinks about that, whether he thinks that the necessity to enumerate hypotheses exhaustively, is important.

SAVAGE: Surely, as you say, we cannot always enumerate hypotheses so completely as we like to think. The list can, however, always be completed by tacking on a catch-all 'something else'. In principle, a person will have probabilities given 'something else' just as he has probabilities given other hypotheses. In practice, the probability of a specified datum given 'something else' is likely to be particularly vague – an unpleasant reality. The probability of 'something else' is also meaningful of course, and usually, though perhaps poorly defined, it is definitely very small. Looking at things this way, I do not find probabilities unnormalizable, certainly not altogether unnormalizable.

Whether probability has an advantage over likelihood seems to me like the question whether volts have an advantage over amperes. The meaninglessness of a norm for likelihood is for me a symptom of the great difference between likelihood and probability. Since you question that symptom, I shall mention one or two others.

First, if we have a probability density of a parameter α, say $\rho(\alpha)$, and reparameterize using, for example, $\beta = \alpha^3$ as the new parameter, then the density of β at the value corresponding to α is $\frac{1}{3}\rho(\alpha)/\alpha^2$. But if $\Pr(x|\alpha)$ is a likelihood in α, the likelihood in β at $\beta = \alpha^3$ is simply $\Pr(x|\alpha)$. Again suppose that x is known to have a Poisson distribution with mean α^{-1} and that $x = 0$ is observed. The likelihood is then $\exp(-\alpha^{-1})$, and it is hard to see how that function, which approaches 1 as $\alpha \to \infty$, could be interpreted as a probability density. The essence of the example is preserved, and the idea of continuous distribution is avoided, if α is assumed to be confined to positive integral values.

On the more general aspect of the enumeration of all possible hypotheses, I certainly agree that the danger of losing serendipity by binding oneself to an over-rigid model is one against which we cannot be too alert. We must not pretend to have enumerated all the hypotheses in some simple and artificial enumeration that actually excludes some of them. The list can however be completed, as I have said, by adding a general 'something else' hypothesis, and this will be quite workable, provided you can tell yourself in good faith that 'something else' is rather improbable. The 'something else' hypothesis does not seem to make it any more meaningful to use likelihood for probability than to use volts for amperes.

Let us consider an example. Offhand, one might think it quite an acceptable scientific question to ask, 'What is the melting point of californium?' Such a question is, in effect, a list of alternatives that pretends to be exhaustive. But, even specifying which isotope of californium is referred to and the pressure at which the melting point is wanted, there are alternatives that the question tends to hide. It is possible that californium sublimates without melting or that it behaves like glass. Who dare say what other alternatives might obtain? An attempt to measure the melting point of californium might, if we are serendipitous, lead to more or less evidence that the concept of melting point is not directly applicable to it. Whether this happens or not, Bayes's theorem will yield a posterior probability distribution for the melting point given that there really is one, based on the corresponding prior conditional probability and on the likelihood of the observed reading of the thermometer as a function of each possible melting point. Neither the prior probability that there is no melting

point, nor the likelihood for the observed reading as a function of hypotheses alternative to that of the existence of a melting point enter the calculation. The distinction between likelihood and probability seems clear in this problem, as in any other.

BARNARD: Professor Savage says in effect, 'add at the bottom of the list $H_1, H_2, \ldots$ "something else" '. But what is the probability that a penny comes up heads given the hypothesis 'something else'. We do not know. What one requires for this purpose is not just that there should be some hypotheses, but that they should enable you to compute probabilities for the data, and that requires very well defined hypotheses. For the purpose of applications, I do not think it is enough to consider only the conditional posterior distributions mentioned by Professor Savage.

LINDLEY: I am surprised at what seems to me an obvious red herring that Professor Barnard has drawn across the discussion of hypotheses. I would have thought that when one says this posterior distribution is such and such, all it means is that among the hypotheses that have been suggested the relevant probabilities are such and such; conditionally on the fact that there is nothing new, here is the posterior distribution. If somebody comes along tomorrow with a brilliant new hypotheses, well of course we bring it in.

BARTLETT: But you would be inconsistent because your prior probability would be zero one day and non-zero another.

LINDLEY: No, it is not zero. My prior probability for other hypotheses may be ϵ. All I am saying is that conditionally on the other $1-\epsilon$, the distribution is as it is.

BARNARD: Yes, but your normalization factor is now determined by ϵ. Of course ϵ may be anything up to 1. Choice of letter has an emotional significance.

LINDLEY: I do not care what it is as long as it is not one.

BARNARD: In that event two things happen. One is that the normalisation has gone west, and hence also this alleged advantage over likelihood. Secondly, you are not in a position to say that the posterior probability which you attach to an hypothesis from an experiment with these unspecified alternatives is in any way comparable with

another probability attached to another hypothesis from another experiment with another set of possibly unspecified alternatives. This is the difficulty over likelihood. Likelihood in one class of experiments may not be comparable to likelihood from another class of experiments, because of differences of metric and all sorts of other differences. But I think that you are in exactly the same difficulty with conditional probabilities just because they are conditional on your having thought of a certain set of alternatives. It is not rational in other words. Suppose I come out with a probability of a third that the penny is unbiased, having considered a certain set of alternatives. Now I do another experiment on another penny and I come out of that case with the probability one third that it is unbiased, having considered yet another set of alternatives. There is no reason why I should agree or disagree in my final action or inference in the two cases. I can do one thing in one case and another in another, because they represent conditional probabilities leaving aside possibly different events.

LINDLEY: All probabilities are conditional.

BARNARD: I agree.

LINDLEY: If there are only conditional ones, what is the point at issue?

Professor E. S. PEARSON: I suggest that you start by knowing perfectly well that they are conditional and when you come to the answer you forget about it.

BARNARD: The difficulty is that you are suggesting the use of probability for inference, and this makes us able to compare different sets of evidence. Now you can only compare probabilities on different sets of evidence if those probabilities are conditional on the same set of assumptions. If they are not conditional on the same set of assumptions they are not necessarily in any way comparable.

LINDLEY: Yes, if this probability is a third conditional on that, and if a second probability is a third, conditional on something else, a third still means the same thing. I would be prepared to take my bets at 2 to 1.

BARNARD: Only if you knew that the condition was true, but you do not.

GOOD: Make a conditional bet.

BARNARD: You can make a conditional bet, but that is not what we are aiming at.

WINSTEN: You are making a cross comparison where you do not really want to, if you have got different sets of initial experiments. One does not want to be driven into a situation where one has to say that everything with a probability of a third has an equal degree of credence. I think this is what Professor Barnard has really said.

BARNARD: It seems to me that likelihood would tell you that you lay 2 to 1 in favour of H_1 against H_2, and the conditional probabilities would be exactly the same. Likelihood will not tell you what odds you should lay in favour of H_1 as against the rest of the universe. Probability claims to do that, and it is the only thing that probability can do that likelihood cannot.

SAVAGE: I agree very much with Mr Lindley in this discussion. As I said in my remarks [on p. 80], in so far as I am interested in probabilities conditional on 'not something else', neither the probability of 'something else' nor the probabilities conditional on this hypothesis are relevant. Also, it is not precluded that I should have probabilities given the hypothesis 'something else'; the operational meaning of such probabilities is the same as that of any others, though they are likely to be particularly intuitive as opposed to reasoned.

COX: I wish to make a technical comment on the idea of a simple test of a null hypothesis. Suppose that our simple null hypothesis says that the density of the observations is $f_0(x)$, and that the test consists in calculating the function $t(x)$ and regarding large values of $t(x)$ as evidence against a null hypothesis. Suppose we consider the following family of hypotheses:

$$f_\theta(x) = f_0(x)\, e^{\theta t(x)} / \int f_0(x)\, e^{\theta t(x)}\, dx.$$

That is a family of hypotheses depending on the parameter θ; when $\theta = 0$ it reduces to the null hypothesis. Clearly the uniformly most powerful test of $\theta = 0$ is based on large values of t. Thus the choice of the statistic t is mathematically equivalent to postulating a family of alternative hypotheses. Correspondingly, this general class of alterna-

tives for all t leads to a class of simple tests of significance. So I suggest that the distinction between setting up families of alternatives and using a simple test of significance is primarily a verbal distinction. It may still be important, but there is no working difference between the two in the end; of course the argument cuts both ways.

BARNARD: That would suggest that Daniel Bernoulli was concerned with hypotheses which said that the probability of getting particular configurations of the poles of the planets was some sort of function $e^{\theta\omega}$, where ω is the area of the smallest circle on the sphere which will enclose them all. Now this is clearly not what he had in mind, is it?

SEVERAL SPEAKERS: But it leads to an identical answer.

BARNARD: All he had in mind it seems to me was that if the planets really lie close together, that is something which could probably be explained dynamically, and he very legitimately said, before we start doing this, before we construct alternatives, let us see if we need to. Let us try the simple single hypothesis first. If the data do not fit that, then it is worth while going ahead. If it is consistent with the data let us not waste our time.

PEARSON: But he had a certain kind of alternative in mind. I do not think you need be able to define the hypotheses precisely. You can choose the test without that. If he had in mind the alternative that there was some sort of repulsion, so that the poles would have got as far apart as possible, he would probably have used another kind of test. So the alternatives were affecting the test he used.

BARNARD: Yes, I quite agree with that, but the alternatives which were affecting the test were not statements of probabilistic hypotheses. Therefore I think we in fact agree that significance tests are sensible things to do.

BARTLETT: I think this is a point that Professor Anscombe has made also. If you have rather vague alternatives you can justify classical tests of significance.

WINSTEN: I would like to return to the question of Dr Good's and my remarks. Is measuring prior probability from how different people react to different likelihoods different from proceeding in Professor Savage's way, before the experiment starts?

SAVAGE: It is not different in the sense of referring to different kinds of probability. But it is very valuable to be reminded that if one takes consistency very seriously it is equally legitimate to argue in either direction.

Mr R. SYSKI: I would like to add that the use of the Bayes approach was defended by the Polish mathematician H. Steinhaus as early as 1950. Since then, he and his followers have published several papers dealing with fundamentals and industrial applications (Steinhaus, 1950, 1954; Rajski, 1954, 1958).

On the lighter side of the subject it may be of interest to mention that behind the Iron Curtain Bayes's hypothesis has been mixed up with political implications. Probability Theory as such presents ideological difficulties for communism. See, for example, a curious statement by Gnedenko and Kolmogorov (1954, p. 1), which reads: 'In fact, all epistemologic value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict, non-random regularity.' Using Bayes's hypothesis, Steinhaus and others overcame this 'official' interpretation, and thus provided possibilities for the unhampered development of Probability Theory.

Finally, I wish to ask how far the theory of Subjective Probability is modified, if at all, when events are specified by abstract valued random variables. There are here several intrinsic difficulties and much depends on the topology of the range space.

SAVAGE: Your final question is a mathematical one rather apart from the main themes of discussion here. To say something about it, de Finetti has always maintained that countable additivity and the attendant restriction of measures to σ-algebras of events are not an essential part of the probability concept. He makes a good case for the idea that probabilities should in principle be thought of as defined for all events. In consequence, many of the mathematical inconveniences of strange range spaces that have been discovered in recent years seem to drop away as side issues.

BARNARD: Can I follow that with a question about de Finetti's attitude to the non-simply additive random distribution on the sphere. I mean Hausdorff's example (Borel, 1926) in which almost the whole

sphere is divided into three mutually exclusive sets, A, B and C, such that A is congruent to B (in the sense that a rigid rotation of the sphere will make A coincide with B), and B to C. The extraordinary feature is that the set A is also congruent to the union of B and C. This shows that you cannot have a random distribution on the sphere which is even finitely additive. What does de Finetti say about that?

SAVAGE: I think he would say something like this. Suppose we are trying to make a mathematical model of someone's opinions about where on the earth a certain meteorite is. The person may be so rash as to blurt out that he always regards congruent sets on the surface of a sphere as equally probable. But Hausdorff's example shows that the person's opinions cannot really have this property. In short, a person who had opinions about all sets on the sphere would have to assign unequal probabilities to some pairs of congruent sets.

For my own part, it makes me dizzy to talk about all the subsets of a sphere; that is an awful lot of sets. From a practical point of view, it is enough to know the probabilities of polyhedral sets. Certainly it is more than enough to know the probabilities of all Borel sets. While agreeing with de Finetti that there is no absolute place to draw the line and that no class of sets should be regarded as not having probabilities, I would underline that in practical computations the probabilities of only a relatively few and simple sets are actually used.

GOOD: I think you need to equate probability with exterior measure, if you are going to allow non-measurable sets.

BARNARD: Then you will not have an additive system.

GOOD: That is all right, for measurable sets it comes to the same thing. One is never interested in non-measurable sets in practice.

COX: I would like Professor Savage to elaborate on remarks he made in his paper about the difficulty of justifying randomization from a strict Bayesian point of view. Part of the solution here may lie in attaching a particularly high utility to experiments for which many people can assign a reasonable prior distribution. If one thinks solely of a particular experiment desiring to produce closest possible estimates of a particular difference, then it seems reasonable sometimes not to randomize. One may do what Professor Savage said, namely

to think up every little bit of information available and put it all together, and do what seems most likely to produce a precise estimate. But such an experiment may have very little value to anyone else, because not being aware of all the particular technical details, it would not be at all clear that there is not a tremendous systematic error in the experiment. One important property of randomization is that it makes the data reasonably convincing to other people as well as to oneself. Of course this is only half the story; randomization may increase accuracy by removing unsuspected biases. This aspect is particularly important in large experiments where bias is more important than random error.

SAVAGE: I think you lay your finger on the objectives of randomization, to make the experiment useful to others and to guard against one's own subconscious. What remains delicate and tentative for me is to understand when, and to what extent, randomization really can accomplish these objectives.

My doubts were first crystallized in the summer of 1952 by Sir Ronald Fisher. 'What would you do,' I had asked, 'if, drawing a Latin square at random for an experiment, you happened to draw a Knut Vik square?' Sir Ronald said he thought he would draw again and that, ideally, a theory explicitly excluding regular squares should be developed. As I have learned since, other statisticians have had, and worked on, this same idea; see, for example, Jones (1958), Yates (1951a, b). This illustrates once more that one need not be a Bayesian to arrive at criticisms to which the Bayesian is led systematically.

The possibility of accidentally drawing a Knut Vik square or accidentally putting just the junior rabbits into the control group and the senior ones into the experimental group illustrates a flaw in the usual reference-set argument that sees randomization as injecting 'objective', or gambling-device probabilities into the problem of inference. If the randomization and the experiment were so executed by an automaton that no one knew which Latin square had been drawn or which animals had been put in the control group, the argument would, I suppose, apply. But, in fact, this information is not, and ought not to be, kept from the experimenter. And he ought not,

in principle, to withhold it from those to whom he communicates his results.

In practice, we may hope, if the experiment is rather large and so designed as to control the variables that (subjectively) look most important, then randomization will almost always lead to a layout that does not look excessively suspicious to any given observer. But this hope needs serious investigation. Perhaps randomization is even one of the most efficient ways to arrive at such widely acceptable layouts. (Such rumours as that artists can make more random-looking designs than random number generators can are a little disquieting to this suggestion.)

In any event, randomization does remove an important possibility of personal interference, for anyone who believes that the randomization did take place according to Hoyle.

Many statisticians agree that an analysis of an experiment ought not be chosen at random. We think it wrong, for example, to break ties at random or to try to escape from the Behrens-Fisher problem by artificially pairing observations. But it has been puzzling to understand why, if random choices can be advantageous in setting up an experiment, they cannot also be advantageous in its analysis. The discussion Dr Cox and I have been giving of randomization seems to lead to an answer to this question. In making an analysis, there is no need to resort to chance to find a compromise analysis that will nearly suit everybody, for each interested person can in principle make for himself the analysis he thinks best. Attempting a compromise can only lose some of the relevant knowledge won by the experiment. Nor can randomization defend against the dangers of subconscious or conscious bias present at the analysis stage.

The arguments against randomized analysis would not apply if the data were too extensive or complex to analyse thoroughly by the individuals concerned. In such a case study of the data might itself become an empirical study based on sampling. Monte Carlo methods might be used. Or one of many possible expensive analyses might be determined in part by randomization in the hope of nearly pleasing everyone.

It seems to me that, whether one is a Bayesian or not, there is still a good deal to clarify about randomization.

BARTLETT: I think this discussion does indicate a certain tendency to compromise both from the Bayesian point of view and from the frequency point of view. On the one hand those who work in terms of a frequency theory avoid certain possible designs because of notions and prior probabilities of what they might contain. On the other hand the fact that the Bayesian would not adopt a perfectly chosen and systematic design, for whatever reasons, seems to represent a certain compromise, in the direction of introducing objective probabilities.

SAVAGE: From my point of view, the exploitation, in personal relations, of the fact that many people coincide in certain judgements is not a compromise. It does of course point up the common sense behind the belief that objective probability is a definable notion.

GOOD: I think the purpose of randomization from the subjectivist point of view is to simplify the analysis by throwing away some of the evidence, deliberately.

SAVAGE: That is a terrible crime, to throw away evidence.

GOOD: But it is evidence which is subjectively judged to be irrelevant. If you had an experiment in which you had to randomize say a thousand objects, say cups of tea, you can never be sure that you had excluded everything that would not be eventually discovered by someone to contain some peculiarities. And your judgement would be the judgement to suppress all these details.

COX: I think Professor Savage's argument leads to what seems to me an acceptable practical conclusion, that randomization is very useful in large and moderate-sized experiments, but is not really very much good in very small single experiments.

WINSTEN: It means also that you should publish the actual design of the Latin square, or whatever it is you chose, so that people can see whether perhaps they have not got a hypothesis of the other sort that they can fill in.

Mr E. D. VAN REST: I am rather surprised that previous speakers have tended to minimize the importance of randomization. Randomiz-

ation seems to be useful whenever knowledge is absent, and I think that is in line with all the previous discussion. Professor Savage discussed the example of animals, recognized to be in two classes, senior and junior.

Directly you can recognize the different classes, they are not a subject for randomization. In other words, we experiment over most of those classes of which we have knowledge and randomize where we have no knowledge. Fisher and Professor Savage rejected a regular arrangement which turned out as the result of randomization. That is exactly explainable in the same way; after the randomization has been done a classification has been recognized. The only reason for throwing it away is that it has not been recognized before starting. You use randomization to perform an averaging function, the averaging out of errors, and it is therefore just as legitimate in small experiments as in large experiments, but it is not so effective. It still needs to be done even though it is not so effective.

BARTLETT: This is the point of view of the non-Bayesian, the usual Fisherian approach. Are you suggesting that your comments justify randomization from Professor Savage's point of view?

VAN REST: To me it does not seem to matter which point of view you take.

SAVAGE: Suppose we had, say, thirty fur-bearing animals of which some were junior and some senior, some black and some brown, some fat and some thin, some of one variety and some of another, some born wild and some in captivity, some sluggish and some energetic, and some long-haired and some short-haired. It might be hard to base a convincing assay of a pelt-conditioning vitamin on an experiment with these animals, for every subset of fifteen might well contain nearly all of the animals from one side or another of one of the important dichotomies. The analysis of covariance (or analysis of the experiment as an unbalanced incomplete multifactor experiment) might give some, but not enough, help.

Thus contrary to what I think I was taught, and certainly used to believe, it does not seem possible to base a meaningful experiment on a small heterogeneous group. In particular, the availability of

technically valid confidence intervals may not really enable us to make a convincing measurement.

BARNARD: I have often said that I agree with the Bayesian approach in many situations, especially in industrial problems. I would like, however, to comment on the type of Bayesian argument that hinges on the 'smoothness' of the prior distribution. It seems to me very important to recognize just how smooth the distributions sometimes have to be for this approach method to give good results. In the sampling inspection situation mentioned already, one is tempted to assume that the proportion defective has a very smooth prior distribution say of the β type. This is all right very often for deciding what you are going to do on the basis of a given sample, but very much not all right when deciding what size of sample you are going to take or what kind of sampling you are going in for. It may, for example, lead you to underestimate the tremendous advantage of sequential methods as compared with fixed sample size.

GOOD: I should like to mention a topic rather different from the ones we have been discussing previously. What it has in common with them is in showing that philosophy does have something to offer to practical statistics.

The question was raised by Popper of how 'corroboration' should be defined; see, for example, Popper (1959, p. 387). He proposed various desiderata for it, and suggested a formula, with the remark that better formulae may be found. It is a question of assigning a meaning to $C(H:E|G)$, meaning and pronounced 'the corroboration of H provided by E, given G'. I think Popper missed out a desideratum which narrows down the field of possible interpretations considerably. It is this:

If evidence is considered in two parts, E and F, then the corroboration of H is analytically determined by that provided by E, combined with that provided by F when E is known.

From this axiom, combined with other mild ones, it follows that $C(H:E|G)$ must be a function $f\{P(H|EG)-P(H|G)\}$, where $f(.)$ is a differentiable function.*

Two of the interpretations of $C(H:E|G)$ are then $I(H:E|G)$, the

* The detailed analysis has since been published (Good, 1960).

('unexpectated') amount of information concerning H, provided by E, given G, and $W(H:E|G)$, the weight of evidence concerning H provided by E, given G. Symbolically,

$$\begin{aligned} I(H:E|G) &= \log\{P(E|H.G)/P(E|G)\}, \\ W(H:E|G) &= \log\{P(E|H.G)/P(E|\bar{H}.G)\} \\ &= I(H:E|G) - I(\bar{H}:E|G) \\ &= \log\{O(H|E.G)/O(H|G)\}, \end{aligned}$$

where O stands for 'odds', $p/(1-p)$, where p is a probability, and the bar stands for negation.

If corroboration has to be a function of $P(E|H.G)$ and $P(E|\bar{H}.G)$ alone, then it can be proved to be an increasing function of the weight of evidence.

A reasonable aim in the design of an experiment would be the maximization of the expected corroboration, for a given cost in experimentation, where the corroboration is one of the additive kinds, such as information or weight of evidence. Which of these two is more sensible will presumably depend on the narrowness of the intervals within which we can judge the probabilities $P(E|\bar{H}.G)$. Lindley (1956) considered the use of expected amounts of information in the design of experiments; in my book (mentioned above) I implicitly took it for granted that expected weight of evidence was relevant.

In order that these remarks should not be misleading, I should add that I still consider, with Savage, that the basic principle of rational behaviour is the maximization of expected utility. I have not changed my opinion about this since reading a chapter by F. P. Ramsey over twenty years ago. But in applications the emphasis is often on the judgements that can be made with the greatest precision: sometimes this will be the probabilities, and sometimes the utilities, and sometimes a mixture.

Dr G. M. JENKINS:* It is surprising that one of the features which has been accepted without much discussion or disagreement in this symposium is the role played by the likelihood function in statistical

* Dr Jenkins was on leave of absence at the time of the meeting and subsequently sent in the following contribution.

theory in so far as it describes the properties of the sample. It is worth noting that some more discussion is required about the choice of the likelihood function as a starting point in any theory of inference.

Main interest has centred around the role played by Bayes's Theorem. Professors Bartlett and Pearson have indicated that they would not use Bayes's Theorem because either they do not recognize its validity or usefulness, or else, even if they were prepared to grant it recognition, choose not to use it. Professors Barnard and Savage and Mr Lindley accept the use and usefulness of Bayes's Theorem, but differ in the extent to which they would apply it. Thus Professor Barnard is prepared to use Bayes's Theorem if the prior distribution is capable of objective description in the sense that past records are available from which some quantitative evaluation may be made, whereas Professor Savage and Mr Lindley are prepared to use it when the prior probabilities involved are far more vague in origin.

I think that ignoring Bayes's Theorem has put much of modern statistics out of gear with scientific thinking; that one indeed very rarely collects observations without some prior probabilities or prior information. In this context, it is necessary to distinguish carefully between prior information in the form of approximate statements such as 'the distribution is normal with given variance', or 'the regression is linear', from prior distributions which are statements about the relative frequencies of a given parameter or set of parameters derived from previous experience or intuition. Prior information in the way of an assumption about the model and about the distribution or joint distributions of the errors is of course essential to the writing down of the likelihood function in the first place. Alternatively we may write down likelihood functions which are sufficiently robust with respect to the sort of inference that we are interested in making.

The distinction between the sort of information which should be fed into Bayes's Theorem which marks the difference of approach between Professor Barnard on the one hand and Professor Savage and Mr Lindley on the other is best illustrated by means of an example. In the design of sampling inspection schemes, it is now being accepted that those based on the use of prior distributions (usually referred to as *process curves*) are likely to lead to better results than the purely

subjective judgements which had been considered previously. On the other hand there are situations, e.g. when an inspection scheme is being designed for a new product, in which there is no process curve available apart from the few results which are inevitably collected during the process of research and development. In these situations, it would be foolish to ignore the engineering experience of those who developed the product and reasonable guesses about the possible quality of the product may be used in the design of initial schemes which can then be modified in the light of further evidence about the process curve.

What is probably required is a new word to distinguish between prior probabilities of an objective nature and those of a more subjective or personal nature. To the latter might be ascribed the word 'hunches', although this is certain to meet with objections from some quarters. What is clear and obvious, however, is the fact that if the information which is fed into Bayes's Theorem is vague and possibly very imprecise, then the corresponding posterior probabilities or expected losses will reflect this imprecision.

Discussion has also been confined entirely to what may be described as *static* theories of inference. All these theories are concerned with statements about sets of statistical parameters which are assumed to be constants of the problem. The notion that one is sampling from finite or infinite populations specified by these 'fixed' parameters is one which has proved useful in the development of statistical theory up to the present. Reflection will perhaps serve to indicate that this is a restrictive assumption and one which may eventually prove to be of limited usefulness in the handling of experimental data.

It is worth noting that Fisher appears at no time to have attached great importance to this concept. Thus in the use of maximum likelihood, fiducial inference and more explicitly in the use of conditional inference, Fisher has always thought in terms of the possible values of the parameters which 'flow' from the estimate obtained from the sample. Thus in conditional inference, statements are made using reference sets generated from within the sample which refer only to parameters which are of direct interest in the hypothesis being examined. Thus in applying the conditional argument to the problem of testing for randomness in binary sequences (and more generally

about various hypotheses concerning Markov chains), the inference about independence is made conditional on the number of 1's or 0's in the observed sequence. As pointed out by Cox (1958b), there are two advantages to this sort of approach, viz. relevance and expediency. Thus, it is restrictive and unnecessary to assume that we are sampling from a population for which there is a fixed probability P for the occurrence of a 1. All that is required is that the proportion of 1's is not changing violently over the length of the sample. Furthermore, it is expedient to use the conditional approach in problems of a discrete nature such as those raised in inference about Markov chains since it leads to the elimination of all the nuisance parameters not relevant to the hypothesis being examined.

In conditional inference, it is possible to see the germ of what may be described as a *dynamic* theory of inference. By this is meant that the statistical parameters which are effectively regarded as constants in the classical theory are themselves regarded as being governed by a stochastic process (usually of a non-stationary type; if it were stationary then of course the problem could be redefined in terms of new parameters relating to the behaviour of the stationary process) in the dynamic theory. This is clearly more in keeping with the behaviour of empirical investigations than the static theory. Thus the inference problem is regarded as a game of strategy between the statistician and nature in which the quantities that are being estimated are themselves changing in an irregular or unpredicatable manner. As further evidence is obtained, new prior probabilities may be fed back into Bayes's Theorem and new posterior probabilities calculated only to be revised at a future date.

These ideas are implicit in the work of Dr G. E. P. Box and his associates at the Statistical Techniques Research Group at Princeton in connection with the optimisation of the mode of operation of chemical plants by means of the technique known as Evolutionary Operation. It would seem that what is now required is a formulation of these concepts in terms of a dynamic theory of inference which should draw on the existing ideas of the static theories and embody them in a framework in which there is a feed back of information via Bayes's Theorem, i.e. a framework drawing on the theory of servomechanisms.

SAVAGE:* Let me make explicit, and comment on, a number of questions that have been brought up during the discussion.

But what if I don't know my own prior probabilities?

In spite of the over-formal arguments that we should be able to know our own prior probabilities by asking ourselves what bets we would and would not make, we often do not really know them at all well. We are vague about specific probabilities, as Professor Pearson has particularly emphasized, and we may not even think of some important relevant hypotheses let alone assign probabilities to them, as Professor Barnard has emphasized. These imperfections in the theory of personal probability are real and render its conclusions imperfect. We must, therefore, use the theory circumspectly, checking it frequently with common sense. We must also be prepared to find that when the sample is, so to speak, too small, an experiment leaves us in a quandary. Not knowing what to conclude is a reality not to be escaped by adopting any so-called 'exact' theory or rule.

Is Bayesian statistics appropriate to some problems but inappropriate to others?

I have yet to see any statistical procedure that makes a durable appeal and cannot be better understood in terms of personal probabilities than in terms of their denial and, therewith, denial of the applicability of Bayes's theorem. Please understand; I am not saying that we Bayesians have the last word in statistical theory which surely would prove false, but rather that a dualistic view of statistics does not seem called for at the present time, and the Bayesian view does seem to have a good deal to offer for the present.

I no longer believe that there exists some alternative to turn to when the subjective method fails to give a satisfactory answer so that there are two qualitatively different kinds of statistical situations. I used to be cowed by critics who said, with apparent technical justification, that certain popular nonparametric techniques apply in situations where it seems meaningless even to talk of a likelihood function, but I have learned to expect that each of these techniques either has a Bayesian

* Professor Savage was invited to conclude the discussion.

validation or will be found to have only illusory value as a method of inference.

To illustrate the question of an alternative method with the topic of interval estimation, the theory of subjective probability often justifies a rather sharply determined belief that an unknown parameter lies in a given interval, as I explained in the part of my main talk dealing with precise estimation. If circumstances are not favourable, as, for instance, when only one or two degrees of freedom are available for estimating a variance, the theory of subjective probability will not allow us to conclude at all sharply what probability ought to be associated with a given interval. To put it differently, a very crude measurement does not overwhelm the differences to be expected between personal opinions. Formally, however, the theory of confidence intervals (and the theory of fiducal intervals also) does not hesitate to base 95 per cent intervals for a variance on one degree of freedom. To be sure, we all know clearly what a 95 per cent interval for variance based on one degree of freedom is. It is a mechanical process so associating with each sample an interval that, no matter what the actual variance is, the probability that the variance will be covered by the random interval is 95 per cent. As we all know, this does not mean that whenever variance is measured with one degree of freedom you would be willing to bet 19 to 1 after seeing the measurement that the particular confidence interval associated with it includes the true parameter. Imagine, for example, that two Meccans carefully drawn at random differ from each other in height by only 0·01 mm. Would you offer 19 to 1 odds that the standard deviation of the height of Meccans is less than 1·13 mm? That is the 95 per cent upper confidence limit computed from chi-squared with one degree of freedom. No, I think you would not have even enough confidence in that limit to offer odds of 1 to 1. The only use I know for a confidence interval is to have confidence in it. When such confidence is not justifiable, is it not empty to say that the confidence interval procedure solves a problem at which the subjective theory throws up its hands?

What am I supposed to publish?

Sometimes listeners to an exposition of Bayesian statistics get the misimpression that they are being urged to publish their own opinions

as their analysis of an empirical study. For example, van Dantzig had the impression on reading my *Foundation of Statistics* (Savage, 1954) that I was urging statisticians to write their own opinions into the scientific publications of their clients. Because of this misconception, van Dantzig (1957) called his review of my book 'Statistical priestcraft'.

Incidentally, I do not ordinarily refer to the relation between the statistician and his client in questions of theoretical statistics, for I regard the separation between statistician and client as an accidental detail of real life that we should try to overcome. If the client had sufficient time, energy, and talent he could be his own statistician, and it seems to me that the first object for theoretical study is such a statistically well endowed investigator. In practice, I conceive that the consulting statistician should, to the best of his ability, lend his mind to his client, or make himself one with his client. There are of course great practical difficulties in bringing about the desired unity and understanding, and the importance of discussing such problems is not to be overlooked, but I have not been discussing them here.

Now, when we Bayesians emphasize that all opinions are but opinion, we do not mean that a scientist publishing the results of his investigation has said the last word when he tells the world what his opinion is. Quite the contrary, the first thing that he ought to tell is what he has observed. In principle, he should do this so well that his peers will know what happened as well as if they had done the experiment themselves. This is in idealization quite unachievable in practice, but approximation to it is the core of a serious research report. In particular, numerical data should be reported as fully as is practical. The most excusable abbreviations are perhaps attempts to condense the data by means of sufficient statistics, but even these are often detrimental, because the sufficient statistics are sufficient only for some model that is nominally accepted, but that might justifiably be rejected in view of some details lost in condensing the data to a sufficient statistic.

Not in principle as an essential but as a courtesy and perhaps as a practical necessity, the scientist may present an opinion that he hopes will be more or less public. His argument would be of the following form, though some parts of it might be left tacit: 'I suppose that you

all, like me, will agree on such and such aspects of our prior opinions and on such and such a model of the experiment. According to Bayes's theorem, we now all have approximately such and such opinions in common until one of us has more data on the basis of which to revise opinions.' A simple example occurs implicitly, I think, whenever someone reports that he made, say, five measurements on a heretofore ill-determined physical constant and gives their mean $\bar{x}$ and standard deviation s. No one will have a sharp prior opinion about the constant or the precision of the experimental method, so, if the possibility of bias in the measurements is neglected, everyone concerned should have, and will be content to have, for his posterior distribution nearly a t-distribution on four degrees freedom about $\bar{x}$ and scaled by $s/\sqrt{5}$. (Of course, the possibility of bias is actually of great practical importance in such an experiment, and gives the best of them a tentative quality not often reflected in textbook discussion.)

Finally, and quite incidentally, the investigator may choose to tell his peers some of the things that he feels in his bones without having any public grounds for conviction. This is frequently done, and of course serves some practical purposes, but it is an utter misconception to imagine that Bayesian statistics attaches central theoretical importance to the experimenter's publication of his personal opinion. Rather, we hope he will so publish that each reader can best form his own personal opinion.

How do statistical inferences differ from inferences generally?

Professor Bartlett has stressed that statistical inferences are special and that mathematical theory can be applied to them in a manner that is impossible, or at least unusual, outside of statistics. I concur in this, as I have said in my talk, but Professor Bartlett and I may not be in perfect agreement as to where this difference resides.

As nearly as I can make out, the most characteristic thing about problems in mathematical statistics is the role in each of some specific model, that is, a specific function $\Pr(x|\lambda)$. The model reflects what is taken to be public agreement about the probability of the datum x as a function of the parameter λ. This is the structure even of so-called nonparametric problems.

Though I call such a well-defined public model characteristic, I am

not really sure to what extent it is essential to statistical practice and to what extent it is induced by habit or convention. For one thing, in so far as you accept the likelihood principle you will agree that one really needs only the likelihood of x as a function of λ, not probability of x as a function of λ. Still more, since usually no one really takes the model seriously as anything but a tentative approximation, we may some day learn how to express ourselves more accurately and fully.

The public character of practical models sometimes has to do with large numbers and the statistical order that can come out of chaos. But symmetry also can give rise to public agreement, without large numbers.

So far as the preceding remarks are concerned, the problem of inferring something about the bias of a penny from three tosses would seem to be a problem in statistical inference. When the ensign of small-sample statistics flew high, it would hardly have been questioned that this problem or the problem of estimating a variance from one or two degrees of freedom was a statistical problem. But perhaps today some of you will feel with me that problems based on excessively small samples, though they must necessarily merge gradually with those based on adequate samples, do not quite belong to the main line of statistics. At any rate, the problems that I have called precise measurement have an important property that can hardly be overemphasized. For such problems lead, in practice, to posterior opinions that are nearly the same from person to person. In testing problems, there can also be public agreement, but not of quite so subtle a kind as in precise estimation; a test may produce overwhelming practical evidence in favour of, or against, a hypothesis, but it does not leave everybody with nearly the same posterior odds. The precision by means of which some experiments induce practical public agreement also often has its source in the law of large numbers and the like. Perhaps, in the last analysis, there is no other source of such precision, but it seems important to mention that, in principle, a single measurement with an instrument of known high accuracy nearly induces the same normal posterior distribution for everyone.

Science or business? Inference or decision?

Some recent discussions of the foundations of statistics have been complicated by assertions that some statistical theory may be valid

for business but not for science, and often confused with that distinction there has been another to the effect that problems of inference are very different from problems of decision.

It does not seem to me that any evidence has ever been brought forward that a statistical theory philosophically sound for practical affairs is inappropriate for science or vice versa. Indeed, it seems unlikely that such a thing should occur at a philosophic level, for many kinds of business considerations can, and properly do, enter the loftiest laboratories – how to allocate the time and money of the laboratory to various problems, for example.

The distinction between inference and decision does seem meaningful to a Bayesian. *Inference* is for us the art of arriving at posterior probabilities; *decision* is concerned directly with action. But, from the Bayesian point of view, the two concepts are not in disharmony with one another. Inference is useful in decision, and the posterior probabilities that figure in inference are, like all probabilities, defined in principle in terms of potential decisions.

What kinds of probability are there?

To me personal, or subjective, probability is the only kind that makes reasonably rigorous sense, and it answers all my needs for a probability concept. So far as this conference is concerned, however, I do not urge so extreme a position on anyone else. If I can leave you thinking that personal probability is interesting and potentially valuable for statistics, my main point will have been made, whether you continue to believe that other concepts of probability are valid or not.

There has been no perceptible defence of the symmetry, or necessary concept of probability here, and I do not really think that concept is tenable. However, as time will show, Sir Harold Jeffreys, a defender of the necessary concept, has made a great and lasting contribution to statistics that has been too little studied.

For some of you, it seems to fly in the face of common sense to deny the existence of frequency probability. But right philosophy sometimes is counter to common sense, and de Finetti has carefully worked out a subjectivistic analysis of the situation in which we ordinarily talk about frequencies, as I have mentioned [on p. 69]. From our point of view, the truth behind the frequency concept of probability is thus

a phenomenon clearly explicable in terms of subjective probability. Similarly, we subjectivists believe that personal probability gives good insight into the truth behind the quest for a necessary definition. The capacity to understand, and to take advantage of, other attempts to formulate a probability concept contributes to the evidence, for me, that the subjective theory is on the right track.