DISCUSSION

A.P. DAWID (*The City University, London*):

I have learned to be wary of those who claim that they would like to reconcile the various opposing views on statistical inference. In my experience, the invariable consequence is, rather, a polarisation of attitudes and a great deal of fruitless apoplexy, and Professor Barnard's paper has succeeded in bringing such a reaction. If there is any common attitude that all statisticians might take to Bernoulli's reasoning, it should be that it does not fall within the ambit of *any* of the standard patterns of inference. For example, use of the P-value in the ST model presupposes that the measure of discrepancy is chosen before looking at the data. But if it had happened, say, that the poles of the planetary orbits lay approximately in one plane, rather than being almost coincident, Bernoulli would surely have used a different discrepancy measure, and it seems impossible to correct for this selection effect. This, to me, discredits the ST interpretation of Bernoulli's argument. I do not believe that Bernoulli's reasoning was unsound —it has obvious common-sense appeal— but it is a weakness of *all* modern statistical orthodoxies that they cannot really justify such reasoning.

I don't think it matters much whether or not we can bring about close agreement between proponents of different basic viewpoints. What is important, I believe, is that we should be willing to learn from the insights of our colleagues (both statistical and substantive) of all complexions, and not interpret whatever view we hold so narrowly that we dismiss those insights out of hand. I am happy that this attitude of give-and-take seems to be becoming more common in the statistical community. One area in which I believe it has been fruitful is that of improved estimation in linear models, following Stein's discovery of the inadmissibility of the usual estimator: Hoerl and Kennard (1970), Lindley and Smith (1972), Efron and Morris (1973). More generally, I think that Bayesian ideas will prove extremely valuable to sampling theory statisticians when they come to consider more carefully the modelling process: for example, a Bayesian approach to finite population sampling can be used to justify a superpopulation model (Ericson, 1969). This, to some degree, answers Barnard's questions (ii) and (iii), since such a model would represent the agreed component for a bevy of Bayesians who all shared the view that the elements of the population were exchangeable.

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

In the story about Daniel Bernoulli we have an example of the serious difficulty of trying to make inferences about some particular hypotheses from some data when the hypotheses themselves have been suggested by the data. The null hypothesis of a random distribution on the sphere is tested only after it is noted that the poles of the planetary orbits seem to lie close together. A discrepancy function is chosen after the data have been observed, and then evaluated at these same data points. Under these conditions how are we to interpret the calculated P values?

Every set of data exhibits some peculiarities. It would be very surprising if we could not look over some data and then set up a hypothesis H_0 and a discrepancy

function $D(x)$ that would yield a very small P value based on the same data. But does that mean that H_0 has been discredited? To some extent, perhaps, but not nearly as much as if H_0 and $D(x)$ had been selected *before* the data had been observed. How much must we discount the observed significance because of this double use of the data?

The Bayesian approach suffers from the same dangers. We open the newspaper in the morning and read some data on a topic we had not previously thought about. In order to process the data, we try to think about what our prior distribution would have been before we saw the data so we can calculate our posterior distribution. But we are too late. Who can say what our prior distribution would have been before we saw the data. We lost our virginity when we read the paper.

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

I have three points to make on this thoughtful paper by Professor Barnard.

1. The significance test seems to be a more primitive method than the Bayes factor. Thinking is not free, and thinking about alternatives to the hypothesis under test is often more difficult than thinking up a discrepancy measure, or test statistic, to use. There often seems to be an underlying relation between the discrepancy measure and interesting alternative hypotheses, even when it cannot easily be traced.

2. The arguments I have heard made against significance tests seem to be based either on the misuse of tests or on grounds of ideology. ("Whatever is not overtly Bayesian is useless").

In practice, there seem to be two kinds of hypotheses tested: (a) null hypotheses (no-effect models); and (b) working models subject to diagnostic checks. Sample size considerations can play havoc with tail-area tests in the context of (b); less so in (a). A small tail area should not be relied on as an excuse to consider alternatives to a null hypothesis. But if the tail area is *not* small, one is well advised not to bother to build up elaborate theories to explain an apparent effect, for it could very well have been an accident under the null model. (Practical cases where this latter use appear unreasonable tend to involve a poor choice of test, for example, one in which prior information on the variance is ignored).

This limited use for significance tests in context (a) is justified by the inequality,

$$\begin{aligned} B(H) &\geq P(T|H)/P(T|H^c) \\ &\geq P(T|H), \end{aligned}$$

where $B(H)$ is the Bayes factor in favour of the null hypothesis H based on the test statistic t, and T is the tail event $\{t \geq t\}$. See Dickey (1977) and references cited therein, and also Good (1950, footnote p. 94). Note that the Bayes factor is approximately the same as the posterior probability for H when it is small and the prior probability is moderate,

$$\begin{aligned} B(H) &= \{P(H|D)/[1-p(H|D)]\}/\{P(H)/[1-P(H)]\} \\ &\doteq \{P(H|D)/1\}/1. \end{aligned}$$

3. It is not generally true that the Bayes factor is a monotonic function of the tail area. In the case of a point hypothesis, H: $\mu = \mu_0$ versus H^c : $\mu \neq \mu_0$, write the Bayes factor in terms of the likelihood function $\ell(\mu)$,

$$B(H) = \ell(\mu_0) / \int \ell(\mu)\, p\,(\mu \,|\, H^c)\, d\mu.$$

Suppose, as is commonly the case, that the tail area decreases to zero as the maximum likelihood estimate $\hat{\mu}$ goes to infinity. If the likelihood has a location form,

$$\ell(\mu) = f(\hat{\mu} - \mu),$$

then it is quite clear that the limiting behaviour of the Bayes factor depends crucially on the relative tail behaviours of f and $p\,(\mu \,|\, H^c)$. For example, if the prior density is supported on a bounded set and the likelihood has a tail like a Student-t density, then the Bayes factor will go to unity (no evidence) instead of zero. Data $\hat{\mu}$ very far away, then, will not distinguish between H and H^c (Dickey, 1977). (It might be said to indicate that neither model is reasonable).

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

I would like to answer Professor Barnard's question concerning Daniel Bernoulli's use of a tail-area probability in an astronomical context. But I have already given a detailed discussion of tail-area probabilities from a Bayesian or rather "Doogian" point of view in Good (1950, pp. 93-94; and 1976a, pp. 162-165). I would be grateful if people interested in this topic would read those few pages. Perhaps the Editor would regard this contribution as too long if I included copies of those pages here. Some slight impression of the nature of those pages may be gleaned from the following footnote from page 94 of Good (1950):

> "There are two independent reason why the factor in favour of H exceeds $P\,(\chi_0^2)$. The first is that to pretend that the result is $\chi \geq \chi_0$ when it is really $\chi = \chi_0$ is unfair to H. The second is that $P\,(\chi \geq \chi_0 | \bar{H}) < 1$, so that the factor from the evidence "$\chi \geq \chi_0$" is
>
> $$P\,(\chi \geq \chi_0 | H)/P\,(\chi \geq \chi_0 | \bar{H}) > P\,(\chi \geq \chi_0 | H) = P\,(\chi_0^2).$$"

After the formal meeting, Professor Barnard drew my attention to Boole (1854, pp.365-368). By using modern terminology Boole's argument can be condensed into the following few lines:

> The final odds of a null hypothesis are equal to the initial odds times the Bayes factor, but we do not usually have physycal knowledge of the initial odds nor of the probability of the observed event given the non-null hypothesis.

Although Boole does not (here) mention Bayes, he is in effect saying that Bayes's theorem cannot be used when the appropriate prior probabilities are *unknown* and Boole could therefore be considered to have somewhat anticipated von Mises (1942). They both ignore the possibility of using partially ordered subjective probabilities.

This possibility is not ignored in the 1950 and 1976 references that I have just mentioned. Those references explain why it makes sense to use tail-area probabilities in many circumstances: they often have a loose relationship to approximate Bayes factors. This relationship forms a part of the Bayes/non-Bayes compromise that I advocate and which Professor Barnard should welcome.

It is worth emphasizing that the Bayesian or Doogian explanation of the use of tail-area probabilities shows very clearly how the sample size is relevant: The larger the sample the more the subjective distribution of the statistic, given that the null hypothesis is false, moves away from the distribution given that is the true. Hence a smaller tail-area probability is required to undermine the null hypothesis. For example, if in Barnard's Bernoulli example there had been a million planets, a tail-area probability of say 1/1000 would have been unconvincing for refuting the null hypothesis (that the normals to the planetary orbits were flat-randomly distributed in all directions).

When selecting a significance test criterion we have at least a vague idea of the alternatives to the null hypothesis, and the criterion can be selected as one giving rise to a large expected weight of evidence for distinguishing the non-null hypothesis from the null hypothesis. This weight of evidence (logarithm of the Bayes factor) is based on the test criterion, which does not usually exhaust all the information from the sample.

There are also approximate relationships between Bayes factors based on all the data and tail-area probabilities based on sensible statistics. For example, see Good (1967, 1976b), Good and Crook (1974) and Crook and Good (1980).

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

Professor Barnard inquires as to the scientific value of Daniel Bernoulli's significance test for the hypothesis of uniformity of the planetary orbits on the celestial sphere. Here we are not considering the various ways in which significance tests are routinely misapplied nowadays by even supposedly well-trained statisticians, but rather the significance test in the hands of a master. Although I must be hesitant to critize a Daniel Bernoulli for anything whatsoever, I would still like to question the value of his tests. The only comprehensible purpose of a significance test without specified alternatives is the purpose of deciding when there is a need to search for new and better models. (In Bernoulli's problem there is in fact a natural alternative, namely coplaner orbits, but Professor Barnard wishes us to ignore this). A p value can be used for such a purpose, but so can many other quantities, for example, the surface area of the smallest region of a given shape containing the points, as a percentage of the total surface area. Apart from cases where the p value is an approximation to a posterior probability the p value has no natural interpretation, and so the question "how small is small" for such a surface area corresponds precisely to the question "how small is small" for a p value. Professor Barnard, of course, might not use conventional levels of significance such as .05, .01, but in this case he must tell us how to allow for sample size and choice of the critical region *after seeing the data* in our interpretation of the evidence against the null hypothesis. So I ask what does a p value offer over and above simpler and more direct quantitative measures as a guide in the search for better models? Professor Barnard

suggests (private conversation) that it allows one to compare different problems on a common scale. However, it seems preferable to me to choose whatever feature strikes one's eye in a particular problem. I see no reason to compare different problems on a common scale. Perhaps Professor Barnard could make clear the purpose of such a comparison. Note that with the approach I am suggesting there would be less likelihood of ascribing statistical significance when there is no practical significance, since it rests upon a more direct perception of the striking features of the data. Often Berkson's interocular traumatic test will suffice.

J.B. KADANE (*Carnegie-Mellon University*):

Professor Barnard rightly calls to our attention the question of the reputation of statistics in experimental disciplines. However I disagree with his diagnosis of the problem: he proposes that sharp divisions among us may lose us respect, while I believe that our reputation lies in the quality of statistics we propose.

Significance testing is a critical point in the philosophical discussions surrounding statistics. The basic question is not so much whether Daniel Bernoulli's use of it was felicitous, but whether we are to endorse present day statistical practice which puts great weight on such tests. Several experiences have led me to conclude that significance testing is much less generally useful than its proponents proclaim. Briefly, some of those experiences are:

(1) (testing a new theory). A distinguished colleague had a new theory (of city sizes) he wished to publish in a statistics journal. The journal insisted on a significance test, so he found the *least* powerful test so that his theory would not be rejected, by the test and by the journal. But he never thought that his theory held *exactly*.

(2) (The catastrophy of too much data). In a sociological study of the frequency of contributions to group discussions, there was a theory Kadane, Lewis and Ramage (1969), wanted to compare to the data. After observing significance at less than 10^{-6}, we found ultimately that plotting the data was much more helpful. This was because we had about 10^4 observations.

(3) (The catastrophy of too little data). A governmental wished to know whether a machine extensively tested in the laboratory worked as well in the field. A significance test revealed "no significant difference", although further analysis showed it was working 75% as well, on the basis of 5 observations costing 1 million dollars each.

In each of these cases enhancing the model and estimating a parameter is much more revealing, although often graphical techniques suffice. There may be an extremely limited role for significance tests, in my view, when the following pertain: (i) the null hypothesis is *honestly* believed by some parties and (ii) the alternatives are expensive to figure out and specify prior distributions for. In such cases a significance test may be understood as a (weak) approximation to a proper Bayesian analysis.

But in my statistical practice, (i) is almost never the case (and (ii) is almost always true!). The only exception for me in recent years is an experiment planned with an astrologer who claimed to be able to distinguish drug offenders from others on the basis of birth dates. Here I put some positive probability on the hypothesis of identical frequency of drug offences. In general, however, the null hypothesis has zero prior

probability, and hence zero posterior probability whatever the data. Attempts to rescue even Bayesian versions of hypothesis testing (Dickey (1976) have lead to their abandonment (Kadane and Dickey (1980)).

On Professor Barnard's word that my criteria (i) and (ii) are met in the case of Daniel Bernoulli's application, I do no object to significance testing in this case. But as a general matter, I believe that significance testing threatens the respectability of statistics more than any other single factor.

T. LEONARD (*University of Warwick*):

Professor Barnard has stimulated a general discussion on significance testing on the basis of a practical example with only five observations. Could I simply remark that for larger sample sizes the problem of goodness of fit should no longer be controversial? It is possible to show that we would compare the chisquared statist with the product of the degrees of freedom and the log of the sample size. This approximates the Bayes solution under a very wide range of prior assumptions, and essentially fixes the significance level for any particular sample size. For very large sample sizes it confirms that the standard test is too much ready to reject the null hypothesis.

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

What ought Daniel Bernoulli to have done? Use a Fisher-von Mises distribution on the sphere and look at the posterior distribution of the spread, particularly in relation to the value of the spread corrsponding to a uniform distribution. (This is effectively what Jaynes described modern physicists as doing, in his discussion of Zellner's and Bernardo's papers). The difficulty with a test of a hypothesis using a tail-area, significance level is that there is always something that is sinificant. The introduction of a discrepancy function tacitly introduces the notion of an alternative and hence of the Bayes approach.

A. ZELLNER (*University of Chicago*):

In this interesting contribution, it is indicated that in the *ST* model approach, a discrepancy function $D(x)$ is introduced and no formally stated alternative hypothesis is used. However different choices of the discrepancy function can lead to different results. Could it be that choice of a particular discrepancy function implicitly implies an alternative hypothesis (H_A) which the investigator has in mind?. If so, why not formulate a posterior odds ratio for H_o and the alternative hypothesis, H_A? To be specific, if for a normal mean problem, H_o is the hypothesis that the mean is zero, $\mu = 0$, one might use as a discrepancy function $t^2 = ny^2/s^2$ and compute the *P*-value associated with t^2, i.e., $Pr\{t^2 \geq t_o^2 | H_o\}$ where t_o^2 is the observed value of t^2. The problem here lies in the interpretation of the *P*-value. It is not equal to the posterior probability that the mean is zero, as is well-known. Jeffrey's analysis of $H_o{:}\mu = 0$ vs. $H_A{:}\mu \neq 0$ leads to the following posterior odds ratio, $K_{oA} \doteq \sqrt{\pi\nu/2}/\,(1 + t^2/\nu)^{(\nu-1)/2}$ where ν = n-1, with n the sample size, and involves the 'discrepancy function'' t^2. It is apparent that K_{oA} is a monotonicaly increasing function of the *P*-value and thus, in my opinion, gives a rationalization for the use of *P*-values in this and other problems.

This example illustrates how use of a particular discrepancy function can be rationalized in Bayesian terms. In Bernoulli's problem, with the null hypothesis of a random (uniform) distribution on the unit sphere, it would be interesting to find the alternative hypothesis (or hypotheses) which leads to a posterior odds ratio that is a monotonic function of the particular discrepancy function for the Bernoulli problem mentioned by Barnard and to show how use of various alternative hypotheses affects the form of the discrepancy function. That Bernoulli employed a particular discrepancy function, apparently without justifiying its use should not be interpreted as good statististical practice in general.

In addition, it is the case that Bayes' factor (BF), the ratio of the posterior odds ratio to the prior odds ratio, can be interpreted as an "inverse" discrepancy function. For large sample size in many problems, $-2\ell nBF \doteq \chi^2_q - q\,\ell n\nu$ or $BF \doteq \nu^{q/2}\exp\{-\chi^2_q/2\}$, where $-2\ell nLR \doteq \chi^2_q$, with LR = the likelihood ratio, q = the number of restrictions under the null hypothesis, and ν = degrees of freedom. For this large sample approximation, χ^2_q can be interpreted as a discrepancy function in Barnard's sense but is not as satisfactory as BF which has a direct interpretation and involves a dependence on χ^2_q *and* the quantities q and ν.

REPLY TO THE DISCUSSION

G.A. BARNARD (*University of Waterloo*):

Since discussion concentrated on the first part of my paper I will confine my reply to this. I hope the issues raised in the second part may be discussed more fully at another Conference as pleasant and stimulating as this one.

I agree entirely with Joe Kadane and with Morris DeGroot. In their day to day work statisticians are almost always concerned with estimation rather than with hypothesis testing. But the importance of an issue cannot be judged entirely on the basis of its frequency of occurrence. The need for significance tests, such as Daniel Bernouilli's arises at the growing points of science, when a new departure, involving concepts not yet thought of, is required. Such occasions are rare, but their importance cannot be over-estimated. And before undertaking the arduous task of thinking up new concepts we would normally insist on P values much lower than the fossilised numbers 0.05, 0.01, or even 0.001; this, at least partly, because we need to make allowance for selection, though the size of this allowance cannot be determined with any precision.

All the other discussants seem to assume that it is just as easy to compute $\Pr(E|\text{ not-}H)$ as it is to compute $\Pr(E|H)$. Only if this is so can we convert the measure of *relative* plausibility given by Bayes Theorem:

$$\Pr(H|E) \,/\, \Pr(H'|E) = (\Pr(E|H)/\Pr(E|H')).\ \Pr(H)/\Pr(H')$$

into an absolute measure by setting H' = not-H. But this is, almost by definition, impossible when not-H involves concepts not yet thought of.

To give just one illustration, in the paper to which Good refers in his contribution, he assumes that it is known that the observations are independent; but such an assumption would often be false in real life. I would suggest that the many and strange

forms of dependence that could arise would defeat the possibility of computing $\Pr(E \mid \text{not-}H)$ in this case.

In practice we ussually can think of not-H as consisting of the disjunction of a mixture of well specified alternatives (such as Lindley's suggestion, in Daniel Bernoulli's case) with an ill-specified 'something else'. For the well specified alternatives we should quote the likelihood ratio versus H, while for the 'something else' we can have not alternative to the P-value. I look forward to the day when in situations such as those we are considering we will specify, not only H and P, but also a specific (and reasonable) H', with its associated, the likelihood ratio. But we should not pretend to the omniscience involved in assuming that (H or H') exhaust the range of possibilities.

REFERENCES IN THE DISCUSSION

BOOLE, G. (1854). *An Investigation in the Laws of Thought*. New York: Dover.

CROOK, J.F. and GOOD, I.J. (1980). On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part. II. *Ann. Statist.* (in press).

DICKEY, J.B. (1976). Approximate posterior distributions. *J. Amer. Statist. Assoc.* **71**, 680-689.

DICKEY, J.M. (1977). Is the tail area useful as an approximate Bayes factor?. *J. Amer. Statist. Assoc.* **72**, 138-142.

EFRON, B. and MORRIS, C. (1973). Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc. B* **35**, 379-421.

ERICSON, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. B* **31**, 195-233.

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

— (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. B* **29**, 399-431. (with discussion). Corrigendum **36** (1974), 109.

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

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

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

HOERL, A.E. and KENNARD, R.W. (1970). Ridge regression: biased estimation for non-orthogonal problems. *Technometrics* **12**, 55-67.

KADANE, J.B., LEWIS, G.H. and RAMAGE, J.G. (1969). Horvarth's theory of participation in group discussion. *Sociometry* **32**, 348-361.

KADANE, J.B. and DICKEY, J.M. (1980). Bayesian decision theory and the simplification of models. *Evaluation of Econometric Models.* (J. Kmenta and J. Ramseyu, eds) New York: Academic Press.

LINDLEY, D.V. and SMITH, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. B* **34**, 1-41.

VON MISES, R. (1942). On the correct use of Bayes' formula. *Ann. Math. Statist,* **13**, 156-165.