Comment

George A. Barnard

I broadly agree with much of what Good says, although I am sometimes unhappy with his choice of terms. For instance, his reference to maximum likelihood seems to suggest that the method amounts to treating the most likely value of the parameter as if it were certainly the true value. Unfortunately there are some who make this mistake.

It is perhaps unfair, when an author has made quite clear what his topic is to be, to complain that it should have been something else. But I must express regret that Good deferred to a final short paragraph his remarks on the role of statisticians as summarizers of data. Because the objective, efficient summarization of data seems to me to be the most important function of the statistician as such. The function to which Good devotes a great part of his paper, the "cheminement de la pensee," in Emile Meyerson's phrase, or "good thinking," as Good so aptly calls it, is one in which the statistician functions along with many other specialists, and where his role is not the primary one. Besides, as I hope to indicate, the theme of the Bayes/non-Bayes compromise finds an excellent expression in connection with data summarization.

It is of the essence of a summarization of data that it should embody what the data have to say on the topic of interest, omitting only material that is irrelevant, and adding only material accepted as true by all prospective readers of the summary. In the vast majority of cases where continuous measurements are involved, the topic of interest and what is accepted as true by all prospective readers can be expressed in terms of parameters and a specific function $\mathbf{p}(x, \theta)$ of the observations and the parameters, taken to have a known distribution. The summarization procedure consists in transforming $\mathbf{p}$, by a 1–1 transformation, to $(\mathbf{q}, \mathbf{a})$, where $\mathbf{q}$ depends as much as possible on the parameters alone, and $\mathbf{a}$ depends as much as possible on the observations alone. In so far as $\mathbf{a}$ does not involve the parameters its value becomes known when the observations are known. If we then imagine learning the data by first being told the value of $\mathbf{a}$, then, after an interval, being told the values of any functions of the data entering into $\mathbf{q}$, our position just before the interval is similar to what it was at the beginning, with a statistical model now specified by the function $\mathbf{q}$, with its probability function $f(\mathbf{q})$ derived by conditioning on the known value of $\mathbf{a}$. Then learning the values of the data functions involved in $\mathbf{q}$ is like learning the values of the original observations. The data can be efficiently summarized by specifying $\mathbf{q}$, its distribution $f(\mathbf{q})$ and the values of the observational functions entering into $\mathbf{q}$. The value of $\mathbf{a}$ merely tells us the value of a quantity of known distribution, not dependent on any of the parameters; it is irrelevant information.

In some cases a complete separation can be made, so that $\mathbf{q}$ is a function of parameters only, while $\mathbf{a}$ is a function of data only. This is the "full Bayesian case," where the inference consists simply of the posterior distribution of $\mathbf{q}$. The information taken as known in the statistical model has been combined with the data in a fully efficient manner and expressed in the posterior distribution. Cases are, however, rare where a statistical model allowing such treatment can be taken as accepted by all potential readers. More commonly a partial separation is all that can be achieved.

A typical intermediate case is one in which unknown scale and location parameters are involved, with a large number of observations x_i. The function $\mathbf{p}$ then has components $p_i = (x_i - \lambda)/\sigma$, where (λ, σ) are the unknown parameters. If $(\bar{x}, s)$ denotes any convenient location-scale pair of functions of the sample (such, for example, as the sample mean and the sample estimated standard error of the mean), $\mathbf{q}$ may be taken to be $(t, z) = ((\bar{x} - \lambda)/s, (\ln s - \ln \sigma))$, while $\mathbf{a}$ has components $(x_i - \bar{x})/s$. If the conditional density of (t, z) is $f(t, z)$, the data are then objectively and efficiently summarized by specifying (t, z), $f(t, z)$ and the observed values of $(\bar{x}, s)$.

If (t_0, z_0) denotes the result of substituting the observed values of $(\bar{x}, s)$ in (t, z), then $f(t_0, z_0)$ provides the likelihood function of (λ, σ) on the basis of the data. Any reader who can supplement the statistical model with his personal prior for (λ, σ) may combine this with $f(t_0, z_0)$ to derive his posterior for the parameters. If we think of the readers of the summary, some will have their own priors, whereas others may not be prepared to so commit themselves. The specification of (t, z), $f(t, z)$ and the observed values of (x, s) then can be thought of as a Bayes/non-Bayes compromise.

Other compromises will be needed in practice. It will be unwise to suppose the distribution of $\mathbf{p}$ to be exactly known, so a "model adjustment parameter"

George A. Barnard is Professor Emeritus at the University of Essex. His mailing address is: Mill House, 54 Hurst Green, Brightlingsea, Colchester, Essex C07 0EH, England.

should be introduced to allow for this. It is a comforting fact—responsible for the fact that dubious normality assumptions have not ruined the reputation of statisticians—that $f(t, z)$ often turns out to be affected very little by changes in the model adjustment parameter. Again, location-scale problems, with their generalizations which take up the bulk of texts on applied statistical inference, are special in that a large **a** can be found which is wholly free from dependence on parameters. In other cases, compromises have to be made between the "size" of **a** and the extent to which it does not change with changes in the parameters. Much of the art of statistical inference consists in judicious choice of statistical model and of the ancillary **a** so as to produce a useful summary which is broadly acceptable while remaining highly efficient.

I sketch these ideas here to invite Good to comment on how they appear to him to fit with his ideas of Bayes/non-Bayes compromise. One specific aspect may be picked out in connection with the location scale problem sketched above. A reader who is unprepared to feed in his own personal prior for (λ, σ) might argue as follows: If S denotes any set in the space of (t, z), determined without reference to the observed values of (x, s), and if I guess that (t, z) belongs to S, the probability P that I guess right is given by integrating f over S. But, knowing the observed $(\bar{x}, s)$, and nothing else about (λ, σ), to guess that (t, s) falls in S is the same as to guess that (λ, σ) falls in the set $\{(\lambda, \sigma): (t_0, z_0), \in S\}$, so the probability that this guess is right is also P. This is a version of the fiducial argument. Fisher came to recognize, near the end of his life, that he had perhaps been mistaken over this argument (see Barnard, 1987). I think his mistake consisted in his assumption that any quantity concerning which a probability statement can be made must be a random variable in the sense of Kolmogoroff. I find Good's suggestion, that the man who had the perception to isolate the property of sufficiency was misled by faulty notation, implausible.

Before leaving the subject of summarization I would make the point that science is a cumulative endeavor, built upon continual repetition of repeatable experiments (shown to be repeatable just by being repeated). The emphasis that Fisher laid on the use of internal estimates of error has led, in the softer sciences, to undue emphasis on the interpretation of single, isolated experiments, with their p-values, as if a single experiment could ever establish the existence of a natural phenomenon. If the evidence from a number of experiments is to be combined, this must be done before any prior judgments are inserted—independent likelihoods can be multiplied together, but not independent posteriors.

To return to the major part of Good's paper, I see his major contribution to our thinking in the stress he has laid on the multiple nature of measurable uncertainty as probability, credibility and so on, and in his taxonomy of these concepts. Since being "cured" of philosophy by Wittgenstein in 1933, I have come to see taxonomy as the principle valuable activity of good philosophers. But such activity carries its own risks. It is all too easy to suppose, wrongly, that a word or phrase in current use must possess a generally definable meaning. The extent to which an event F caused another event E may, for purely practical reasons, require assessment within legal systems which employ the dubious concept of damages. But as a general concept of potential value in science, or even in morality, I see no use for it. I have similar doubts about the word "cause."

"Weight of evidence" is a term which can be given a precise meaning in reasonably well-defined circumstances—namely when it is possible to calculate the probability of an event E from an hypothesis H and also from the negation of H. I agree with Good that in such cases "weight of evidence" has a precise and useful meaning. (And I may add that according to my possibly faulty recollection C. S. Peirce said the same, in his book *Chance, Love and Logic*, without adding the condition that H and not-H must be equally probable *a priori*.)

"Induction" is another word which perhaps generates more confusion than it is worth. It so easily tempts us toward the notion of a "law of nature," thought of as necessarily universal.

Good's criticisms of mechanical interpretations of p-values are well taken. Fisher's emphasis on their "exactness" arose, I think, because he wrote his *Statistical Methods* before the heat of his controversy with Karl Pearson had cooled—accuracy of p-values being a central issue in that debate. Although he made far more changes in successive editions than he is given credit for, he was loth to rewrite sections which may have given a false impression without actually being wrong. But I have doubts about Good's proposal to standardize P's to sample size 100. Pitman's point (1965) is too little known—that from the strict Neyman-Pearson point of view, to minimize the long run frequency of errors of the second kind, subject to an upper bound on the frequency of errors of the first kind, we *must* adjust our critical p-values to the sensitivity of our tests. Because the sensitivity of a test depends on factors other than just the sample size, standardizing the latter could mislead.

Good's use of the past tense in relation to the idea that subjective probability might be "the most basic kind" leaves one in doubt whether he would still so argue. What is meant here by "basic" is not clear to me; but taking it in one sense I would object to the idea by pointing to the use of "probability" in quantum theory where probability density is equated to the squared modulus of the psi function. The theory amounts to a doctrine that there exist "systems" whose "state" can be described by a psi function satisfying certain rules of combination and of evolution in time. These "systems" relate to objectively describable repeatable experimental set-ups; and the theory is related to such set-ups mainly by interpreting the squared modulus of the psi function as a "long run frequency probability" over repetitions of such set-ups. No subjective element enters into this, although in relation to a single such set-up an observer may associate the quantum-theoretical probability with a subjective probability of the same magnitude. There are many fascinating puzzles here, well described by David Mermin in the April 1985 issue of *Physics Today*.

Like Good I see the future of the foundations of statistical inference in Bayes/non-Bayes compromises involving hierarchical models, objective data summarizations and in other directions. It is a pleasure to have been invited to discuss.

ADDITIONAL REFERENCES

BARNARD, G. A. (1987). R. A. Fisher—a true Bayesian? *Internat. Statist. Rev.* **55** 183–189.

PITMAN, E. J. G. (1965). Some remarks on statistical inference. In *Bernoulli, Bayes, Laplace* (J. Neyman and L. Le Cam, eds.) 215–216. Springer, New York.

Comment

James O. Berger

I recall being surprised upon first encountering the considerable interest of many philosophers in probability and statistics, interest at an often detailed technical level. Perhaps even more unusual is a serious professional interest in philosophy from a statistician or probabilist. Jack Good has had such a professional interest, virtually from the beginning of his career, and it is indeed a pleasure to view the world of "probabilistic philosophy" through his eyes.

One of the cornerstones of probabilistic philosophy was the development of the Bayesian and expected utility paradigms for processing information and making decisions. The paradigms were, however, an incomplete representation of reality, until Good incorporated the concept of partially ordered probabilities into their structures. I have written, in some depth, about this aspect of Good's work in Berger (1987), and so will refrain from further comments here.

I found Good's comment, that "... the future of statistics ... will be a compromise between hierarchical Bayesian methods and methods that seem superficially to be non-Bayesian," quite interesting. It is true that hierarchical Bayesian methods (including their empirical Bayes approximations) often have no workable classical analogues, and hence will be indispensable to the future of statistics; was more than this intended by the comment?

Isn't the left hand side of (2) often called a "weighted likelihood ratio"? I have several times been cynically amused that some statisticians will have no qualms about basing a decision on a weighted likelihood ratio with rather arbitrarily chosen weight functions, but will cry out in horror at the thought of using a Bayes factor with a prior that is actually thought about!

Another way of trying to understand the type of correction to a p-value given in (4), is to observe that, as long as N is at least moderately large,

$$\frac{p\text{-value}}{\text{Bayes factor}} \cong \frac{2\sigma g(\theta_0)}{\sqrt{N}[z + (.75)z^{-1}]};$$

here σ is the standard deviation of an observation, $g(\theta_0)$ is the value of the prior density as it approaches the null model θ_0 and z is the standardized (normal) test statistic $z = \sqrt{N}(\bar{x} - \theta_0)/\sigma$. Thus a p-value will behave roughly like a Bayes factor if it is multiplied by $\sqrt{N}$. (The above formula further suggests that multiplying p by $[z + (.75)z^{-1}]$ might be a beneficial standardization, but this is a comparatively minor additional correction.)

The idea of choosing a (perhaps crude) Bayes factor to be the significance test criterion certainly should be beneficial to classical testing. What, however, is the value of this to a Bayesian, who feels that all tail

James O. Berger is The Richard M. Brumfield Distinguished Professor of Statistics, Department of Statistics, Purdue University, West Lafayette, Indiana 47907.