(3) It would be of interest to obtain an expansion of the form

$$i^X(\theta) - i^T(\theta) = \alpha + \frac{\beta}{n} + \cdots .$$

We know the expression for α and its geometric interpretation. What about β?

(4) I believe that the choice of a prior distribution is governed by the nature of the parameter and previous knowledge (though vague) about it and should not depend on what experiment is conducted to have further information on it. Jeffreys' invariant prior may have nice properties but it seems to depend on how observations are generated, which may not be acceptable to Bayesians.

ADDITIONAL REFERENCES

RAO, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Cambridge Philos. Soc.* **44** 50–57.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

Comment

N. Reid and D. A. S. Fraser

We congratulate Professor Kass on a very clear and interesting account of the role of differential geometry in asymptotic inference. In particular, his discussion of information loss and recovery through conditioning, and the geometric interpretation of this, adds substantially to the long-standing discussion initiated in Fisher's early work.

The use and implications of conditional analysis are central to the topics in the paper. In this discussion, we expand a little on arguments for and justifications of conditioning, and the use of geometric methods to motivate this.

In the setting discussed in Section 3.1, we can write

$$p_Y(y \mid \theta) = p_{T|A}(t \mid a, \theta)p_A(a) \tag{1}$$

where $Y = (T, A)$ is sufficient, A is ancillary, and the Jacobian has been absorbed into the support differentials. This factorization suggests, as the paper indicates, that inference about θ may be based on the conditional distribution of T given A, without loss of information about θ. Section 3.1.1 gives formal clarity to Fisher's general analysis of information loss and is valuable in giving a precise interpretation of the phrase "without loss of information about θ."

Other arguments can also provide some interpretation of the phrase above. For example the likelihood function obtained from the conditional distribution is the same as the likelihood function from the distribution of the full data Y. Another motivation for conditioning on A when the factorization in (1) holds is that the variability in the outcome that is described by the marginal distribution of A is irrelevant for inference about θ; this is an underlying theme in Fisher's early work expanded in Fisher (1961) and is very clearly presented in the weighing machine example of Cox (1958). Fisher frequently used the term "relevant subset" to refer to the set of sample points having the observed value for the ancillary statistic. However, it seems clear that he attached additional meaning to the term, derived from the physical context from which the statistical problem arose. Indeed, this additional interpretation may well have been primary in Fisher's interpretation of conditioning and the definition of the correct probabilities to use in applications. There does seem to be no fully satisfactory formalization of such "relevant subsets" based on the statistical model alone. The derivation of the Likelihood Principle from the Conditionality Principle discussed in Evans, Fraser and Monette (1986) bears on this.

Most discussions of conditioning are motivated by a few very compelling examples. Subsequent attempts to formalize the operating principle to enable extension to more realistic settings are widely divergent. One development, primarily initiated by Birnbaum (1962, 1972) and Basu (1959, 1964) (see also Buehler, 1982), isolates ancillarity as the essential feature; the discussion of this approach and its relation to Bayesian inference and the likelihood principle is well summarized in Berger and Wolpert (1985).

Another development of conditioning in Fraser (1968, 1979) extends and formalizes one aspect of

N. Reid is Professor in the Department of Statistics at the University of Toronto. Her mailing address is: Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. D. A. S. Fraser is Professor in the Department of Mathematics at York University. His mailing address is: Department of Mathematics, York University, Downsview, Ontario, Canada M3J 1P3.

relevant subsets. Location-scale and other transformation models can be recast in terms of a fixed antecedent error variable, as is indicated for example by $y = X\beta + \sigma e$, where e is a sample from a known distribution. With such models certain components of the variable e are known by calculation once the sample is available. In this sense then, they provide the same kind of information as is provided by the knowledge of which machine was used to take the measurement, in Cox's example, or indeed the same as is provided to the bridge player by his own and the dummy's hand. Unconditional calculation of probabilities is in this framework irrelevant, even incorrect. In all examples of this approach, the conditioning statistic is in fact ancillary; it is simply the motivation for the conditioning that is different. Another advantage of this approach is that it also provides prescriptions for marginal and/or conditional inference for components of vector parameters.

An alternative description of the factorization (1) is that T is conditionally sufficient for θ, given A, as in Section 2.2. One advantage of considering the type of factorization provided by (1), rather than the information factorization that follows from it, is that (1) is helpful for clarifying the role of conditioning in inference about a parameter of interest in the presence of nuisance parameters. There seem to be two different separations like (1) that isolate a parameter component,

$$(2) \qquad p_Y(y \mid \psi, \lambda) = p_{T|S}(t \mid s, \psi, \lambda) p_S(s \mid \lambda)$$

and

$$(3) \qquad p_Y(y \mid \psi, \lambda) = p_{T|S}(t \mid s, \psi) p_S(s \mid \psi, \lambda);$$

in these Y would typically be minimal sufficient. In (2) we could regard S as ancillary for the parameter of interest ψ, and thus motivate the use of the conditional distribution of T given S. This is a fairly straightforward extension of the situation in (1) above. In (3), however, S is sufficient for the nuisance parameter λ. Justification for basing inference about ψ on the conditional distribution of T given S is in this setting more pragmatic. Although information about ψ may be available in the marginal distribution of S, it is often assumed that this information cannot be extracted from the marginal distribution, in the absence of knowledge of λ. Investigation of information loss in $p_S(s)$ is one approach to this justification (Amari, 1985, Chapter 8). Another is that similar or unbiased tests of hypotheses can only be generated by such conditioning, if S is complete. Often the situation is clarified by using orthogonal versions of the nuisance parameter (Cox and Reid, 1987). In the very special case that

$$(4) \qquad p_Y(y \mid \psi, \lambda) = p_{T|S}(t \mid s; \psi) p_S(s \mid \lambda)$$

we have that S is ancillary for ψ and sufficient for λ, and both the arguments above can be applied.

Conditional distributions are typically much easier to compute than marginal distributions, especially in high dimensional problems. In some problems, it seems possible to develop conditional techniques from a purely pragmatic point of view: one conditions on some components of the problem of less direct interest, simply for computational or inferential simplicity or convenience. Often the geometry of the sample space can be helpful in determining the most effective direction for conditioning (Fraser and Massam 1985). Work in progress by H.-S. Lee at the University of Toronto shows that often a preferred marginal distribution can be well approximated by a suitably designed conditional distribution.

Kass' paper emphasizes the role of geometry in asymptotics, and it is in this area that the development of the Riemmanian metric and affine connections have proved to be most useful. Geometric arguments can also be useful in exploring the local structure of the sample space. In Fraser and Reid (1988), a one-dimensional conditional distribution for inference about a real parameter ψ in the presence of a nuisance parameter is constructed by differential arguments that examine likelihood change on the sample space in a local manner. An extension to that development that determines a conditional distribution for a pivotal statistic depending on ψ is in progress.

As an example, if we have two samples of size n from exponential distributions with rates λ and $\psi\lambda$ and sample totals y_1 and y_2, the method of local differential analysis in Fraser and Reid (1988) leads to the conditioning equation

$$dy_1 + \hat{\psi} dy_2 = 0.$$

This leads to the conditional distribution given $S = y_1 y_2$. Further details are provided in Fraser and Reid (1987); the conditional distribution gives a large-sample approximation to the usual marginal analysis that refers $y_1/\psi y_2$ to an F distribution on (n, n) degrees of freedom. For the pivotal version of the local analysis the conditioning equation is

$$dy_1 + \psi dy_2 = 0;$$

this leads to conditioning on a ψ-dependent function $S_\psi = y_1 + \psi y_2$. The resulting conditional distribution is the same as the usual marginal distribution. It would be interesting to see clearly the differences between the local analysis just described and the asymptotic methods, which are local in a different manner.

ADDITIONAL REFERENCES

Basu, D. (1959). The family of ancillary statistics. *Sankhyā Ser. A* **21** 247–256.

BASU, D. (1964). Recovery of ancillary information. *Sankhyā Ser. A* **26** 3–16.

BERGER, J. and WOLPERT, R. (1985). *The Likelihood Principle.* IMS, Hayward, Calif.

BIRNBAUM, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* **57** 269–326.

BIRNBAUM, A. (1972). More on concepts of statistical evidence. *J. Amer. Statist. Assoc.* **67** 858–886.

BUEHLER, R. J. (1982). Some ancillary statistics and their properties (with discussion). *J. Amer. Statist. Assoc.* **77** 581–594.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

EVANS, M., FRASER, D. A. S. and MONETTE, G. (1986). On principles and arguments to likelihood. *Canad. J. Statist.* **14** 181–194.

FISHER, R. A. (1961). Sampling the reference set. *Sankhyā Ser. A* **23** 3–8.

FRASER, D. A. S. (1968). *The Structure of Inference.* Wiley, New York.

FRASER, D. A. S. (1979). *Inference in Linear Models.* McGraw-Hill, New York.

FRASER, D. A. S. and MASSAM, H. (1985). Conical tests: Observed level of significance and confidence regions. *Statist. Hefte (N.F.)* **26** 1–18.

FRASER, D. A. S. and REID, N. (1987). Fibre analysis and conditional inference. *Proc. 2nd Pacific Area Statistical Conference* 241–248.

FRASER, D. A. S. and REID, N. (1988). On conditional inference for a real parameter: A differential approach on the sample space. *Biometrika* **75** 251–264.

Rejoinder

Robert E. Kass

I am very grateful to the discussants for their comments, which have substantially enriched the material presented here. The remarks of Professors Amari, Barndorff-Nielsen, and Reid and Fraser require no reply. I do, however, wish to answer the specific queries raised by Professors Bernardo and Rao.

With regard to Rao's query (1), concerning characterizations of the information metric, I would refer interested readers to the original work of Centsov (1972) and the newer work of Picard (1989). I am not sure what Rao has in mind in his query (2) about the choice of affine connection. Part of the answer may come from the results of Centsov and Picard, but if Professor Rao is referring to the choice of α in the α-connection, perhaps helpful to the intuition is the observation in Kass (1984) that vanishing of the α-connection coefficients when $\alpha = -1, -1/3, 0, 1$ occurs for the bias-reducing, skewness-reducing, variance-stabilizing, and natural parameterizations, respectively, and when $\alpha = 1/3$, it occurs for the parameterization in which the expected values of the third derivatives of the loglikelihood vanish. These parameterizations were characterized in differential equation form, in the one-parameter case, by Hougaard (1982). There is also a very nice answer to part of Rao's query (3), due to Amari (1985, 1987a). In brief, Amari used higher derivatives of the imbedding $\eta(\cdot)$, defining a curved exponential family, to define both higher-order curvatures and appropriate statistics based on higher-order derivatives of the loglikelihood function. With these he obtained a complete decomposition of the information in the sample as an asymptotic expansion with geometrically-interpretable terms of decreasing order associated with the loglikelihood derivatives.

Finally, Professor Rao's point (4), and Professor Bernardo's request for comments in the rejoinder, concern Jeffreys' general rule for choosing a prior. I have a few things to say about this, though for the sake of brevity I will not try to argue my opinions in detail.

As a preliminary remark, I emphasize that by "reference prior" I mean a prior chosen according to any formal rule that may be applied without detailed consideration of the data-analytic context. Such a prior need not be considered "noninformative" in any well-defined sense. This is an important point, since it is dubious that the concepts of ignorance and lack of information can be given satisfactory definitions. I believe the idea of selecting a prior by *convention*, as a "standard of reference," analogous to choosing a standard of reference in other scientific settings, is due to Jeffreys (1955) page 277. This notion and terminology was adopted by Box and Tiao (1973) page 23. Unfortunately, Bernardo (1979) used the term "reference prior" for a specific rule, rather than the general concept, and this occasionally causes confusion.

There is great convenience in conventional choices, throughout statistics and throughout science. But convenience should not be confused with necessity: one might say that conventions are useful as long as they are not taken too seriously. Thus, I see the convenience in reference priors, just as I recognize the convenience in conventional levels of significance. In applications, however, such conveniences must be questioned. Sometimes they are justifiable timesavers, especially for communicating results, but often they are not. I consider reference priors to be "default" choices, but they are to be used only when their