Selected Works in Probability and Statistics

Selected Works of Debabrata Basu

Springer

Selected Works in Probability and Statistics

For other titles published in this series, go to
www.springer.com/series/8556

Anirban DasGupta
Editor

Selected Works of Debabrata Basu

Springer

Editor
Anirban DasGupta
Department of Statistics and Mathematics
Purdue University
150 N. University Street
West Lafayette, IN 47907, USA
dasgupta@stat.purdue.edu

ISBN 978-1-4419-5824-2 e-ISBN 978-1-4419-5825-9
DOI 10.1007/978-1-4419-5825-9
Springer New York Dordrecht Heidelberg London

Preface to the Series

Springer's Selected Works in Probability and Statistics series offers scientists and scholars the opportunity of assembling and commenting upon major classical works and honors the work of distinguished scholars in probability and statistics. Each volume contains the original papers, original commentary by experts on the subject's papers, and relevant biographies and bibliographies.

Springer is commited to maintaining the volumes in the series with free access on SpringerLink, as well as to the distribution of print volumes. The full text of the volumes is available on SpringerLink with the exception of a small number of articles for which links to their original publisher is included instead. These publishers have graciously agreed to make the articles freely available on their websites. The goal is maximum dissemination of this material.

The subjects of the volumes have been selected by an editorial board consisting of Anirban DasGupta, Peter Hall, Jim Pitman, Michael Sørensen, and Jon Wellner.

Preface and Introduction

When Springer approached me with a proposal for editing a collection of Dev Basu's writings and articles, with original commentaries from experts in the field, I accepted this invitation with a sense of pride, joy, and anxiety. I was a direct student of Dev Basu at the Indian Statistical Institute, and accepted this task with a sense of apprehension. I was initially attracted to Basu because of his clarity of exposition and a dignified and inspiring presence in the classrooms at the ISI. He gave us courses on combinatorics, distribution theory, central limit theorems, and the random walk. To this date, I regard those courses to be the best I have ever taken on any subject in my life. He never brought any notes, never opened the book, and explained and derived all of the material in class with an effortlessness that I have never again experienced in my life.

Then, I got to read some of his work, on sufficiency and ancillarity, survey sampling, the likelihood principle, the meaning of the elusive word *information*, the role of randomization in design and in inference, eliminating nuisance parameters, his mystical and enigmatic counterexamples, and also some of his highly technical work using abstract algebra, techniques of complex and Fourier analysis, and on putting statistical concepts in a rigorous measure theoretic framework. I realized that Basu was also a formidable mathematician. The three signature and abiding qualities of nearly all of Basu's work were clarity of exposition, simplicity, and an unusual originality in thinking and in presenting his arguments. Anyone who reads his paper on randomization analysis (Basu (1980)) will ponder about the use of permutation tests and the role of a statistical model. Anyone who reads his papers on survey data, the likelihood principle, information, ancillarity, and sufficiency will be forced to think about the foundations of statistical practice. The work was fascinating and original, and influenced generations of statisticians across the world. Dennis Lindley has called Basu's writings on foundations "among the most significant contributions of our time to the foundations of statistical inference."

Foundations can be frustrating, and disputes on foundations can indeed be never-ending. Although the problems that we, as statisticians, are solving today are different, the fundamentals of the subject of statistics have not greatly changed. Depending on which particular foundational principle we more believe in, it is still the fundamentals laid out by Fisher, Pearson, Neyman, Wald, and the like, that drive statistical inference. Despite decades and volumes devoted to debates over foundational issues, these issues still remain important. In his commentary on Basu's work on survey sampling in this volume, Alan Welsh says "$\cdots$, and this is characteristic of Basu and one of the reasons (that) his papers are still so valuable; it does challenge the usual way $\cdots$. Statistics has benefitted enormously that Basu made that journey." I could not say it any better. It is with this daunting background that I undertook the work of editing this volume.

This book contains 23 of Basu's most significant articles and writings. These are reprints of the original articles, presented in a chronological order. It also contains eleven commentaries written by some of our most distinguished scholars in the areas of foundations and statistical inference. Each commentary gives an original and contemporary critique of a particular aspect or some particular contribution in Basu's work, and places it in perspective. The commentaries are by George Casella and V. Gopal, Phil Dawid, Tom DiCiccio and Alastair Young, Malay Ghosh, Jay Kadane, Glen Meeden,

Robert Serfling, Jayaram Sethuraman, Terry Speed, and Alan Welsh. I am extremely grateful to each of these discussants for the incredible effort and energy that they have put into writing these commentaries. This book is a much better statistical treasure because of these commentaries.

Terry Speed has eloquently summarized a large portion of Basu's research in his commentary. My comments here serve to complement his. Basu was born on July 5, 1924 in the then undivided Bengal and had a modest upbringing. He obtained an M.A. in pure mathematics from Dacca University in the late forties. His initial interest in mathematics no doubt came from his father, Dr. N. M. Basu, an applied mathematician who worked on the mathematical theory of elasticity under the supervision of Nobel laureate C. V. Raman. Basu told us that shortly after the partition of India and Pakistan, he became a refugee, and crossed the borders to come over to Calcutta. He found a job as an actuary. The work failed to give him any intellectual satisfaction at all. He quit, and at considerable risk, went back to East Pakistan. The adventure did not pay off. He became a refugee for a second time and returned to Calcutta. Here, he came to know of the ISI, and joined the ISI in 1950 as a PhD student under the supervision of C. R. Rao. Basu's PhD dissertation "Contributions to the Theory of Statistical Inference" was nearly exclusively on pure decision theory, minimaxity and admissibility, and risks in testing problems under various loss functions (Basu (1951, 1952a, 1952b)). In Basu (1951), a neat counterexample shows that even the most powerful test in a simple *vs.* simple testing problem can be inconsistent, if the iid assumption for the sequence of sample observations is violated. In Basu (1952a), an example is given to show that if the ordinary squared error loss is just slightly altered, then a best unbiased estimate with respect to the altered loss function would no longer exist, even in the normal mean problem. Basu (1952b) deals with admissible estimation of a variance for permutation invariant joint distributions and for stationary Markov chains with general convex loss functions. It would seem that C. R. Rao was already thinking of characterization problems in probability, and Basu was most probably influenced by C. R. Rao. Basu wrote some early articles on characterizing normal distributions by properties of linear functions, a topic in which Linnik and his students were greatly interested at that time. This was a passing phase.

In 1953, he came to Berkeley by ship as a Fulbright scholar. He met Neyman, and listened to many of his lectures. Basu spoke effusively of his memories of Neyman, Wald, and David Blackwell at many of his lectures. It was during this time that he learned frequentist decision theory and the classical theory of statistics, extensively and deeply. At the end of the Fulbright scholarship, he went back to India, and joined the ISI as a faculty member. He later became the Dean of Studies, a distinguished scientist, and officiating Director of the Research and Training School at the ISI. He pioneered scholastic aptitude tests in India that encourage mathematics students to understand a topic and stop focusing on cramming. The widely popular aptitude test book Basu, Chatterji, and Mukherjee (1972) is an institution in Indian mathematics all by itself. He also had an ingrained love for beautiful things. One special thing that he personally developed was a beautiful flower garden around the central lake facing the entrance of the main residential campus. It used to be known as *Basu's garden*. He also worked for some years as the chief superintendent of the residential student hostels, and when I joined there, I heard stories about his coming by in the wee hours of the morning to check on students, always with his dog. I am glad that he wasn't the superintendent when I joined, because my mischievous friends and I were stealing coconuts from the campus trees around 4:00 am every morning for an early morning carb boost. I can imagine his angst and disbelief and the angry outrage of his puzzled dog at some of his dedicated students hanging from coconut trees at four in the morning, and a few others hiding in the bushes on guard.

The subject of statistics was growing rapidly in the years around the second world war. The most fundamental questions were being raised, and answered. In quick succession, we had the *Neyman-Pearson lemma, Cramér-Rao inequality, Rao-Blackwell theorem, the Lehmann-Scheffé theorem, Neyman's proof of the factorization theorem, Mahalanobis's D^2-statistic, the Wald test, the score test of Rao, and Wald and Le Cam's masterly and all encompassing formulation and development of decision theory as a unified basis for inference*. This was also a golden age for the ISI. Fisher was

spending time there, and so were Kolmogorov, and Haldane. Basu joined the ISI as a PhD student in that fertile and golden era of statistics and the ISI. He was clearly influenced, and deeply so, by Kolmogorov's rigorous measure theoretic development of probability theory, and simultaneously by Fisher's prodigious writings on the core of statistics, maximum likelihood, sufficiency, ancillarity, fiducial distributions, randomization tests, and so on. This influence of Kolmogorov and Fisher is repeatedly seen in much of Basu's later work. The work on foundations is obviously influenced by Fisher's work, and the technical work on sufficiency, ancillarity, and invariance (Basu (1967, 1969)) is very clearly influenced by Kolmogorov. It is not too much of a stretch to call some of Basu's work on sufficiency, ancillarity, and invariance research on abstract measure theory.

Against this backdrop, we see Basu's first transition into raising and answering questions that have something fundamental and original about them. Among the most well known is what everyone knows simply as *Basu's theorem* (Basu (1955a)). It is the only result in statistics and probability that is listed in Wikipedia's *list of Indian inventions and discoveries*, significant scientific inventions and discoveries originating in India in all of recorded history. A few other entries in this list are the hookah, muslin, cotton, pajamas, private toilets, swimming pools, hospitals, plastic surgery, diabetes, jute, diamonds, the number zero, differentials, the Ramanujan theta function, the AKS primality test, Bose-Einstein statistics, and the Raman effect.

The direct part of Basu's theorem says that if $X_1, \cdots, X_n$ have a joint distribution $P_{n,\theta}$, $\theta \in \Theta$, $T(X_1, \cdots, X_n)$ is boundedly complete and sufficient, and $S(X_1, \cdots, X_n)$ is ancillary, then T and S are independent under each $\theta \in \Theta$. The theorem has very pretty applications, and I will mention a few. But, first, I would like to talk a little more of the context of this theorem. He was led to Basu's theorem when he was asked the following question. Consider iid $N(\mu, 1)$ observations $X_1, \cdots, X_n$. Then, every location invariant statistic is ancillary; is the converse true? The converse is not true, and so Basu wanted to characterize the class of all ancillary statistics in this situation. The reverse part of Basu's theorem answers this question; in general, suppose $T(X_1, \cdots, X_n)$ is boundedly complete and sufficient. Then, $S(X_1, \cdots, X_n)$ is ancillary only if it is independent of T under each θ. Typically, in applications, one would take T to be the minimal sufficient statistic, which has the best chance of being also complete. Without completeness, an ancillary statistic need not be independent of T.

Returning to the more well known part of Basu's theorem, namely the direct part, there is an element of sheer beauty about the result. A sufficient statistic is supposed to capture all the information about the parameter that the full data could supply, and an ancillary statistic has none to offer at all. We can think of a rope, with T and S at the two ends of the rope, and θ placed in the middle. T has everything to do with θ, and S has nothing to do with θ whatsoever. They must be independent! The theorem brings together sufficiency, information, ancillarity, completeness, and conditional independence. Terry Speed (Speed (2010), IMS Bulletin), calls it *a beautiful theorem*, which indeed it is. Basu later worked on various other aspects of ancillarity and selection of reference sets; all of these are comprehensively discussed in Phil Dawid's commentary in this volume.

Basu's theorem isn't only pretty. It has also been used by many researchers in diverse areas of statistics and probability. To name a few, the theorem has been used extensively in distribution theory, in deriving Barndorff-Nielsen's magic formula (Barndorff-Nielsen (1983), Small (2010)), in proving theorems on infinitely divisible distributions, in goodness of fit testing (and in particular for finding the mean and the variance of the *WE* statistic for testing exponentiality), and of late in small area estimation. Hogg and Craig (1956), Lehmann (1981), Boos and Hughes-Oliver (1998), and Ghosh (2002) have previously described some of these applications. Larry Brown has provided some very powerful applications of a more general (but less statistically intuitive) version of Basu's theorem in Brown (1986), and Malay Ghosh has indicated applications to empirical Bayes problems in his commentary in this volume. Here are a few of my personal favorite applications of Basu's theorem to probability theory. The attractive part of these examples is that you save on clumsy or boring calculations by using Basu's theorem in a clever way. The final results can be obtained without

using Basu's theorem, but in a pedestrian sort of way. In contrast, by applying Basu's theorem, you do it elegantly.

Example 1 (Independence of Mean and Variance for a Normal Sample) Suppose $X_1, X_2, \cdots, X_n$ are iid $N(\eta, \tau^2)$ for some η, τ. Suppose $\overline{X}$ is the mean and s^2 the variance of the sample values $X_1, X_2, \cdots, X_n$. The goal is to prove that $\overline{X}$ and s^2 are independent, whatever be η and τ. First note the useful reduction that if the result holds for $\eta = 0, \tau = 1$, then it holds for all η, τ. Indeed, fix any η, τ, and write $X_i = \eta + \tau Z_i, 1 \le i \le n$, where $Z_1, \cdots, Z_n$ are iid $N(0, 1)$. Now,

$$\left(\overline{X}, \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right) \stackrel{\mathcal{L}}{=} \left(\eta + \tau\overline{Z}, \tau^2 \sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2\right).$$

Therefore, $\overline{X}$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2$ are independently distributed under (η, τ) if and only if $\overline{Z}$ and $\sum_{i=1}^{n}(Z_i - \overline{Z})^2$ are independently distributed. This is a step in getting rid of the parameters η, τ from consideration. But, now, we will import a parameter! Embed the $N(0, 1)$ distribution into a larger family of $\{N(\mu, 1), \mu \in \mathcal{R}\}$ distributions. Consider now a fictitious sample $Y_1, Y_2, \cdots, Y_n$ from $P_\mu = N(\mu, 1)$. The joint density of $Y = (Y_1, Y_2, \cdots, Y_n)$ is a one parameter Exponential family density with the natural sufficient statistic $T(Y) = \sum_{i=1}^{n} Y_i$. And, of course, $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is ancillary. Since this is an Exponential family, and the parameter space for μ obviously has a nonempty interior, all the conditions of Basu's theorem are satisfied, and therefore, under *each* μ, $\sum_{i=1}^{n} Y_i$ and $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ are independently distributed. In particular, they are independently distributed under $\mu = 0$, i.e., when the samples are iid $N(0, 1)$, which is what we needed to prove. This is surely a very pretty proof of that classic fact in distribution theory.

Example 2 (An Exponential Distribution Result) Suppose $X_1, X_2, \cdots, X_n$ are iid Exponential random variables with mean λ. Then, by transforming $(X_1, X_2, \cdots, X_n)$ to $\left(\frac{X_1}{X_1+\cdots+X_n}, \cdots, \frac{X_{n-1}}{X_1+\cdots+X_n}, X_1 + \cdots + X_n\right)$, one can show by carrying out the necessary Jacobian calculation that $\left(\frac{X_1}{X_1+\cdots+X_n}, \cdots, \frac{X_{n-1}}{X_1+\cdots+X_n}\right)$ is independent of $X_1 + \cdots + X_n$. We can show this without doing any calculations by using Basu's theorem.

For this, once again, by writing $X_i = \lambda Z_i, 1 \le i \le n$, where the Z_i are iid standard Exponentials, first observe that $\left(\frac{X_1}{X_1+\cdots+X_n}, \cdots, \frac{X_{n-1}}{X_1+\cdots+X_n}\right)$ is a (vector) ancillary statistic. Next observe that the joint density of $X = (X_1, X_2, \cdots, X_n)$ is a one parameter Exponential family, with the natural sufficient statistic $T(X) = X_1 + \cdots + X_n$. Since the parameter space $(0, \infty)$ obviously contains a nonempty interior, by Basu's theorem, under each λ, $\left(\frac{X_1}{X_1+\cdots+X_n}, \cdots, \frac{X_{n-1}}{X_1+\cdots+X_n}\right)$ and $X_1 + \cdots + X_n$ are independently distributed.

Example 3 (A Weak Convergence Result Using Basu's Theorem) Suppose $X_1, X_2, \cdots$ are iid random vectors with a uniform distribution in the d-dimensional unit ball. For $n \ge 1$, let $d_n = \min_{1 \le i \le n} ||X_i||$, and $D_n = \max_{1 \le i \le n} ||X_i||$. Thus, d_n measures the distance to the closest data point from the center of the ball, and D_n measures the distance to the farthest data point. We find the limiting distribution of $\rho_n = \frac{d_n}{D_n}$. Although this can be done by using Slutsky's theorem, the Borel-Cantelli lemma, and some direct algebra, we will do so by an application of Basu's theorem.
Toward this, note that for $0 \le u \le 1$,

$$P(d_n > u) = (1 - u^d)^n; \;\; P(D_n > u) = 1 - u^{nd}.$$

As a consequence, for any $k \geq 1$,

$$E[D_n]^k = \int_0^1 ku^{k-1}(1-u^{nd})du = \frac{nd}{nd+k},$$

and,

$$E[d_n]^k = \int_0^1 ku^{k-1}(1-u^d)^n du = \frac{n!\Gamma\left(\frac{k}{d}+1\right)}{\Gamma\left(n+\frac{k}{d}+1\right)}.$$

Now, embed the uniform distribution in the unit ball into the family of uniform distributions in balls of radius θ and centered at the origin. Then, D_n is complete and sufficient, and ρ_n is ancillary. Therefore, once again, by Basu's theorem, D_n and ρ_n are independently distributed under each $\theta > 0$, and so, in particular under $\theta = 1$. Thus, for any $k \geq 1$,

$$E[d_n]^k = E[D_n\rho_n]^k = E[D_n]^k E[\rho_n]^k$$

$$\Rightarrow E[\rho_n]^k = \frac{E[d_n]^k}{E[D_n]^k} = \frac{n!\Gamma\left(\frac{k}{d}+1\right)}{\Gamma\left(n+\frac{k}{d}+1\right)}\frac{nd+k}{nd}$$

$$\sim \frac{\Gamma\left(\frac{k}{d}+1\right)e^{-n}n^{n+1/2}}{e^{-n-k/d}\left(n+\frac{k}{d}\right)^{n+\frac{k}{d}+1/2}}$$

(by using Stirling's approximation)

$$\sim \frac{\Gamma\left(\frac{k}{d}+1\right)}{n^{\frac{k}{d}}}.$$

Thus, for each $k \geq 1$,

$$E\left[n^{1/d}\rho_n\right]^k \to \Gamma\left(\frac{k}{d}+1\right) = E[V]^{k/d} = E[V^{1/d}]^k,$$

where V is a standard Exponential random variable. This implies, because $V^{1/d}$ is uniquely determined by its moment sequence, that

$$n^{1/d}\rho_n \overset{\mathcal{L}}{\Rightarrow} V^{1/d},$$

as $n \to \infty$.

Example 4 (An Application of Basu's Theorem to Quality Control) Herman Rubin kindly suggested that I give this example. He used Basu's theorem to answer a question on statistical quality control while consulting some engineers in the fifties. Here is the problem, which is simple to state.

In a production process, the measurement X of a certain product has to fall between *conformity limits* a, b. For design, as well as for quality monitoring, the management wants to estimate what percentage of items are currently meeting the conformity specifications. That is, we wish to estimate $\theta = P(a < X < b)$. Suppose that for estimation purpose, we have obtained data values $X_1, X_2, \cdots, X_n$, which we assume are iid $N(\mu, \sigma^2)$ for some μ, σ. Then, $\theta = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$. Standard plug-in estimates are asymptotically efficient. But quality control engineers have an inherent preference for the UMVUE. We derive the UMVUE below in closed form by using Basu's theorem and the Lehmann-Scheffé theorem.

By the Lehmann-Scheffé theorem, the UMVUE is the conditional expectation

$$E\left(I_{a\le X_1\le b}\,|\overline{X}, s\right) = P\left(a \le X_1 \le b\,|\overline{X}, s\right)$$

$$= P\left(\frac{a-\overline{X}}{s} \le \frac{X_1-\overline{X}}{s} \le \frac{b-\overline{X}}{s}\,|\overline{X}, s\right)$$

Now, $(\overline{X}, s)$ are jointly sufficient and complete, and $\frac{X_1-\overline{X}}{s}$ is evidently ancillary. Therefore, by Basu's theorem, we get the critical simplification that the conditional distribution of $\frac{X_1-\overline{X}}{s}$ given $(\overline{X}, s)$ is the same as the unconditional distribution (whatever it is) of this ancillary statistic $\frac{X_1-\overline{X}}{s}$. Hence, the UMVUE of θ is

$$= P\left(\frac{a-\overline{X}}{s} \le \frac{X_1-\overline{X}}{s} \le \frac{b-\overline{X}}{s}\right) = F_n\left(\frac{b-\overline{X}}{s}\right) - F_n\left(\frac{a-\overline{X}}{s}\right),$$

where $F_n(t)$ denotes the CDF of the unconditional distribution of $\frac{X_1-\overline{X}}{s}$.

We can, perhaps a little fortunately, compute this in closed form. It is completely obvious that the mean of F_n is zero and that the second moment is $\frac{1}{n}$. With a little calculation, which we will omit, F_n can be shown to be a Beta distribution on $[-1, 1]$ (*in fact, even this fact, which I am not proving here, can be proved by using Basu's theorem*). In other words, F_n has the density

$$f_n(x) = \frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\sqrt{\pi}\,\Gamma(\alpha)}(1-x^2)^{\alpha-1},\ -1 \le x \le 1.$$

The value of α must be $\frac{n-1}{2}$, by virtue of the second moment being $\frac{1}{n}$. Hence, the UMVUE of θ is

$$F_n\left(\frac{b-\overline{X}}{s}\right) - F_n\left(\frac{a-\overline{X}}{s}\right)$$

where

$$F_n(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\int_{-1}^{t}(1-x^2)^{\frac{n-3}{2}}dx$$

$$= \frac{1}{2} + \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\; t\; {}_2F_1\left(\frac{1}{2}, \frac{3-n}{2}; \frac{3}{2}; t^2\right),$$

where ${}_2F_1$ denotes the usual *Gauss Hypergeometric function*. This describes the UMVUE of θ in closed form. Herman Rubin or I have not personally seen this very closed form derivation involving the Hypergeometric function anywhere, but it is another instance where Basu's theorem makes the problem solvable without breaking our back.

Example 5 (A Covariance Calculation) Suppose $X_1, \cdots, X_n$ are iid $N(0, 1)$, and let $\overline{X}$ and M_n denote the mean and the median of the sample set $X_1, \cdots, X_n$. By using our old trick of importing a mean parameter μ, we first observe that the difference statistic $\overline{X} - M_n$ is ancillary. By Basu's theorem, $X_1 + \cdots + X_n$ and $\overline{X} - M_n$ are independent under each μ, which implies

$$\mathrm{Cov}(X_1 + \cdots + X_n, \overline{X} - M_n) = 0 \Rightarrow \mathrm{Cov}(\overline{X}, \overline{X} - M_n) = 0$$

$$\Rightarrow \mathrm{Cov}(\overline{X}, M_n) = \mathrm{Cov}(\overline{X}, \overline{X}) = \mathrm{Var}(\overline{X}) = \frac{1}{n}.$$

We have achieved this result without doing any calculations at all.

Example 6 (Application to Infinite Divisibility) Infinitely divisible distributions are important in both the theoretical aspects of weak convergence of partial sums of triangular arrays, and in real applications. Here is one illustration of the use of Basu's theorem in producing unconventional infinitely divisible distributions. The example is based on the following general theorem (DasGupta (2006)), whose proof uses *both* Basu's theorem and the Goldie-Steutel law (Goldie (1967)).

Theorem Let f be any homogeneous function of two variables, i.e., suppose $f(cx, cy) = c^2 f(x, y) \forall x, y$ and $\forall c > 0$. Let Z_1, Z_2 be iid $N(0, 1)$ random variables and $Z_3, Z_4, \cdots, Z_m$ any other random variables such that $(Z_3, Z_4, \ldots, Z_m)$ is independent of (Z_1, Z_2). Then for any positive integer k, and an arbitrary measurable function g, $f^k(Z_1, Z_2) g(Z_3, Z_4, \cdots, Z_m)$ is infinitely divisible.

A large number of explicit densities can be proved to be densities of infinitely divisible distributions by using this theorem. Here are a few, each supported on the entire real line.

(i) $f(x) = \frac{1}{\pi} K_0(|x|)$, where K_0 denotes the Bessel K_0 function;

(ii) $f(x) = \frac{2\log(|x|)}{\pi^2(x^2 - 1)}$;

(iii) $f(x) = \frac{1}{\sqrt{2\pi}}\left(1 - \sqrt{2\pi}|x|e^{x^2/2}\Phi(-|x|)\right)$;

(iv) $f(x) = \frac{1}{2}\frac{1}{(1 + |x|)^2}$.

Note that the density in (iv) is the so called *GT density*. Of course, we can introduce location and scale parameters into all of these, and make families of infinitely divisible distributions.

Continuing on with some other significant contributions of Basu, his mathematical work as well as his foundational writings on survey sampling helped put sampling theory on a common mathematical footing with the rest of rigorous statistical theory, and even more, in raising and clarifying really fundamental issues. Alan Welsh has discussed the famous example of *Basu's elephants* (Basu (1971)) in his commentary in this volume. This is the solitary example that I know of where a single example has led to the writing of a book with the example in its title (Brewer (2002)). The elephants example must be understood in the context of the theme and also the time. It was written when the Horvitz-Thompson estimator for a finite population total was gaining theoretical popularity, and many were accepting the estimator as an automatic choice. Basu's example reveals a fundamental flaw in the estimator in particular, and in the wisdom of sample space based optimality, more generally. The example is so striking and entertaining, that I cannot but reproduce it here.

Example 7 (Basu's Elephants) "The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back at his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in his herd. Therefore, the owner plans to weigh Sambo and take $50y$ (where y is the present weight of Sambo) as an estimate of the total weight $Y = Y_1 + \ldots + Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive sampling plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So together they work out a compromise sampling plan. With the help of a table of random numbers, they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected, and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why? The estimate ought to be $50y$ of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician. "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz-Thompson estimate in this case?", asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is $\frac{100y}{99}$ and not $50y$." "And how would you have estimated Y," enquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!)."

Some of my other personal favorites are a number of his counterexamples. The examples always used to have something dramatic or penetrating about them. He would take a definition, or an idea, or an accepted notion, and chase it to its core. Then, he would give a remarkable example to reveal a fundamental flaw in the idea and it would be very very difficult to refute it. One example of this is his well known *ticket example* (Basu (1975)). The point of this example was to argue that blind use of the maximum likelihood estimate, even if there is just one parameter, is risky. In the ticket example, Basu shows that the MLE overestimates the parameter by a huge factor with a probability nearly equal to one. The example was constructed to make the likelihood function have a global peak at the wrong place. Basu drives home the point that one must look at the entire likelihood function, and not just where it peaks. Very reasonable, especially these days, when so many of us just throw the data into a computer and get the MLE and feel happy about it. Jay Kadane has discussed Basu's epic paper on likelihood (Basu (1975)) for this volume.

In this very volume, Robert Serfling discusses Basu's definition of asymptotic efficiency through concentration of measures (Basu (1956)) and the counterexample which puts his definition at the extreme opposite pole of the traditional definition of asymptotic efficiency. An important aspect of Basu's definition of asymptotic relative efficiency is that it isn't wedded to asymptotic normality, or $\sqrt{n}$-consistency. You could use it, for example, to compare the mean and the midrange in the normal case, which you cannot do according to the traditional definition of asymptotic relative efficiency.

A third example, but of less conceptual gravity, is his example of an inconsistent MLE (Basu (1955b)). The most famous example in that domain is certainly the Neyman-Scott example (Neyman and Scott (1948)). In the Neyman-Scott example, the inconsistency is caused by a nonvanishing bias, and once the bias is corrected, consistency is retrieved. Basu's example is pathological statistically, but like all his examples, this too makes the point in the most extreme conceivable way. The inconsistent MLE isn't fixable in his example.

One important point about Basu's writings is that it is never clear that he does not like the procedures that classical criteria produce. In numerous writings, he uses a time tested classical procedure. *But he only questions the rationale behind choosing them*. This distinction is important. I feel that in these matters, he is closer to Dennis Lindley, who too, reportedly holds the view that classical statistics generally produces fine procedures, but using the wrong reasons. This seems to be very far from a dogmatic view that all classical procedures deserve to be discarded because of where they came from. But, even when Basu questioned the criteria for selecting a procedure in his writings, and in his seminars, it was always in the best of spirits (George Casella comments in this volume that "the banter between Basu and Kempthorne is fit for a comedy.")

Shortly after he taught us at the ISI, Basu left India and moved to the USA. He joined the faculty of the Florida State University, causing, according to his daughter Monimala, a family rebellion. They would happily settle in Sydney, or Denmark, or Ottawa, or Sheffield, where Basu used to visit frequently, even Timbuktu, but not in a small town in Florida. After he moved to the US, one can see a distinct change of perspective and emphasis in Basu's work. He now started working on more *practical things*; concrete elimination of nuisance parameters, modelling Bayesian bioassays with missing data (Basu and Pereira (1982a)), randomization tests (1980), and Bayesian nonparametrics. His involvement in Bayesian nonparametrics resulted in a beautifully written paper on Dirichlet processes starting from absolute scratch (Basu and Tiwari (1982b)). Jayaram Sethuraman superbly discusses this paper in this volume. In a way, Basu's involvement in Bayesian nonparametrics was perhaps a little surprising. This is not because he could not deal with abstract measure theory; he was an expert on it! But, because, Basu repeatedly expressed his deep rooted skepticism about putting priors on *more than three or four parameters*. He never said what he would do in problems with many parameters. But he would not accept improper priors, or even empirical Bayes. He simply said that he does not know what to do if one has many parameters, because you then just can't write your elicited information into honest priors. In some ways, Basu was a half hearted Bayesian. But, he was very forthcoming.

Basu returned permanently to India 1986. He still taught and lectured at the ISI. I last saw Basu in 1995 at East Lansing. He, his wife Kalyani, and daughter Monimala were all visiting his son Shantanu, who was then a postdoctoral scientist at Michigan State University. I spent a few hours with him, and I told him what I was working on. I asked him if he would like to give a seminar at Purdue. He said that the time came some years ago that he should now only listen, *to young people like you*, and not talk. He said that his time has passed, and he only wants to learn now, and not profess. He often questioned himself. In a nearly spiritual mood, he once wrote (Ghosh (1988)): "What have I got to offer? I am afraid nothing but a set of negative propositions. But with all humility, let me draw the attention of the would be reader to the ancient Vedic (Hindu scripture) saying- "neti, neti, $\cdots$, iti," which means- "not this, not this, $\cdots$, this!"

Basu went back to Calcutta from Lansing, and I received letters from him periodically. In the March of 2001, when I was visiting Larry Brown at the Wharton school, one morning an e-mail came from

B.V. Rao at Calcutta. The e-mail said that he is duty bound to give me the saddest news, that Dr. Basu passed away the night before. I went to Larry Brown's office and gave him the news. Larry looked at me, as if he did not believe what I said, and I saw his eyes glistening up, and he said - "that's just too bad. He was such a good guy." However, the idealism and the inspiration live on, as strongly as in 1973, when he walked into that classroom at the ISI to meet twenty 16 year olds, and fifty minutes later, we were all in love with probability. I know that I speak for numerous people who got to know Basu as closely as we did, that he was a personification of purity, in scholarship and in character. There was an undefinable poetic and ethereal element in the man, his personality, his presence, his writings, and his angelic disposition, that is very very hard to find. He is missed dearly, but he lives in our memory. The legacy remains ever so strong. I do tell my students in my PhD theory course; *read Basu*.

West Lafayette, Indiana, USA Anirban DasGupta

Preface Bibliography

Barndorff-Nielsen, O. (1983). On a formula for the conditional distribution of the maximum likelihood estimator, Biometrika, 70, 343–365.

Basu, D. (1951). A note on the power of the best critical region for increasing sample sizes, Sankhyā, 11, 187–190.

Basu, D. (1952a). An example of non-existence of a minimum variance estimator, Sankhyā, 12, 43–44.

Basu, D. (1952b). On symmetric estimators in point estimation with convex weight functions, Sankhyā, 12, 45–52.

Basu, D. (1955a). On statistics independent of a complete sufficient statistic, Sankhyā, 15, 377–380.

Basu, D. (1955b). An inconsistency of the method of maximum likelihood, Ann. Math. Statist., 26, 144–145.

Basu, D. (1956). The concept of asymptotic efficiency, Sankhyā, 17, 193–196.

Basu, D. (1967). Problems related to the existence of minimal and maximal elements in some families of subfields, Proc. Fifth Berkeley Symp. Math. Statist. and Prob., I, 41–50, Univ. California Press, Berkeley.

Basu, D. (1969). On sufficiency and invariance, in *Essays on Probability and Statistics*, R.C. Bose et al. eds., University of North Carolina Press, Chapel Hill.

Basu, D. (1971). An essay on the logical foundations of survey sampling, with discussions, in *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott eds., Holt, Rinehart, and Winston of Canada, Toronto.

Basu, D. (1975). Statistical information and likelihood, with discussion and correspondence between Barnard and Basu, Sankhyā, Ser. A, 37, 1–71.

Basu, D. (1980). Randomization analysis of experimental data: The Fisher randomization test, with discussions, Jour. Amer. Statist. Assoc., 75, 575–595.

Basu, D., Chatterji, S., and Mukherjee, M. (1972). Aptitude Tests for Mathematics and Statistics, Statistical Publishing Society, Calcutta.

Basu, D. and Pereira, C. (1982a). On the Bayesian analysis of categorical data: The problem of nonresponse, Jour. Stat. Planning Inf., 6, 345–362.

Basu, D. and Tiwari, R. (1982b). A note on the Dirichlet process, in *Essays in Honor of C.R. Rao*, 89–103, G. Kallianpur, P.R. Krishnaiah, and J.K. Ghosh eds., North-Holland, Amsterdam.

Boos, D. and Hughes-Oliver, J. (1998). Applications of Basu's theorem, Amer. Statist., 52, 218–221.

Brewer, K. (2002). Combined Survey Sampling Inference: Weighing Basu's Elephants, Arnold, London.

Brown, L. D. (1986). Fundamentals of Statistical Exponential Families, Institute of Mathematical Statistics, Hayward, California.

DasGupta, A. (2006). Extensions to Basu's theorem, infinite divisibility, and factorizations, Jour. Stat. Planning Inf., 137, 945–952.

Ghosh, J. K. (1988). A Collection of Critical Essays by Dr. D. Basu, Springer-Verlag, New York.

Ghosh, M. (2002). Basu's theorem with applications: A personalistic review, Sankhyā, Ser. A, 64, 509–531.

Goldie, C. (1967). A class of infinitely divisible random variables, Proc. Cambridge Philos. Soc., 63, 1141–1143.

Hogg, R. and Craig, A. T. (1956). Sufficient statistics in elementary distribution theory, Sankhyā, 17, 209.

Lehmann, E. L. (1981). An interpretation of completeness and Basu's theorem, Jour. Amer. Statist. Assoc., 76, 335–340.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations, Econometrica, 16, 1–32.

Small, C. G. (2010). Expansions and asymptotics for statistics, Chapman and Hall, Boca Raton.

Speed, T. (2010). IMS Bulletin, 39, 1, 9–10.

Acknowledgment I am thankful to Peter Hall for his critical help in acquiring the reprint rights of some of Basu's most important articles. Alan Welsh rescued me far too many times from an impending disaster. B.V. Rao and Bill Strawderman went over this preface very very carefully and made numerous suggestions for improving on an earlier draft. My longtime and dear friend Srinivas Bhogle graciously helped me with editing the preface. Amarjit Budhiraja, Shih-Chuan Cheng, Doug Crabill, Ritabrata Dutta, Ingram Olkin, Sastry Pantula, Bhamidi Shankar and Ram Tiwari helped me with acquisition of some of the articles. Shantanu and Monimala Basu gave me important biographical information. Springer's scientific editors Patrick Carr and Marc Strauss were tremendously helpful. Integra Software Services at Pondicherry did a fabulous job of producing this volume. And, John Kimmel was just his usual self, the best supporter of every good cause.

Acknowledgements

This series of selected works is possible only because of the efforts and cooperation of many people, societies, and publishers. The series editors originated the series and directed its development. The volume editors spent a great deal of time compiling the previously published material and the contributors provided comments on the significance of the papers. The societies and publishers who own the copyright to the original material made the volumes possible and affordable by their generous cooperation:

American Institute of Physics
American Mathematical Society
American Statistical Association
Applied Probability Trust
Bernoulli Society
Cambridge University Press
Canadian Mathematical Society
Elsevier
Foundation of the Scandinavian Journal of Statistics
Indian Statistical Institute
Institute of Mathematical Statistics
International Chinese Statistical Association
International Statistical Institute
John Wiley and Sons
l'Institut Fourier
London Mathematical Society
New Zealand Statistical Association
Oxford University Press
Polish Academy of Sciences
Princeton University and the Institute for Advanced Studies
Springer
Statistical Society of Australia
University of California Press
University of Illinois, Department of Mathematics
University of Michigan, Department of Mathematics
University of North Carolina Press

Contents

Contributors

George Casella Distinguished Professor of Statistics, University of Florida, Gainesville, FL 32611, casella@stat.ufl.edu

Philip Dawid Statistical Laboratory, Centre for Mathematical Sciences, CB3 0WB, UK, apd@statslab.cam.ac.uk

Thomas J. DiCiccio ILR School, Cornell University, Ithaca, NY 14853, tjd9@cornell.edu

Malay Ghosh Distinguished Professor of Statistics, University of Florida, Gainesville, FL 32611, ghoshm@stat.ufl.edu

Vikneswaran Gopal Department of Statistics, University of Florida, Gainesville, FL 32611, viknesh@stat.ufl.edu

Joseph B. Kadane, Emeritus Leonard J. Savage University Professor of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, kadane@stat.cmu.edu

Glen Meeden Chairman and Head, School of Theoretical Statistics, University of Minnesota, Minneapolis, MN 55455, glen@stat.umn.edu

Robert J. Serfling Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75083, serfling@utdallas.edu

Jayaram Sethuraman, Emeritus Robert O. Lawton Distinguished Professor of Statistics, Florida State University, Tallahassee, FL 32306, sethu@ani.stat.fsu.edu

T.P. Speed Department of Statistics, University of California, Berkeley, CA 94720, terry@stat.berkeley.edu

A.H. Welsh Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia, alan.welsh@anu.edu.au

G. Alastair Young Department of Mathematics, Imperial College, London, UK, alastair.young@imperial.ac.uk

Author Bibliography

1. D. Basu. Contributions to the Theory of Statistical Inference, PhD Dissertation, Submitted to the University of Calcutta, 1953.
2. D. Basu. A note on the power of the best critical region for increasing sample sizes, Sankhyā, 11, 187–190, 1951.
3. D. Basu. On the limit points of relative frequencies, Sankhyā, 11, 379–382, 1951.
4. D. Basu. On the independence of linear functions of independence chance variables, Bull. Inst. Internat. Statist., 23, II, 83–96, 1951.
5. D. Basu. On the minimax approach to the problem of estimation. Proc. Nat. Inst. Sci. India, 18, 287–299, 1952.
6. D. Basu. An example of non-existence of a minimum variance estimator, Sankhyā, 12, 43–44, 1952.
7. D. Basu. On symmetric estimators in point estimation with convex weight functions, Sankhyā, 12, 45–52, 1952.
8. D. Basu. On a class of admissible estimators of the normal variance, Sankhyā, 12, 57–62, 1952.
9. D. Basu. Choosing between two simple hypotheses and the criterion of consistency, Proc. Nat. Inst. Sci. India, 19, 841–849, 1953.
10. D. Basu and R. G. Laha. On some characterizations of the normal distribution, Sankhyā, 13, 359–362, 1954.
11. D. Basu. On the optimum character of some estimators in multistage sampling problems, Sankhyā, 13, 363–368, 1954.
12. D. Basu. A note on mappings of probability spaces, Vestnik Leningrad Univ., 10, 5, 33–35, 1955.
13. D. Basu. An inconsistency of the method of maximum likelihood, Ann. Math. Statist., 26, 144–145, 1955.
14. D. Basu. A note on the structure of a stochastic model considered by V. M. Dandekar, Sankhyā, 15, 251–253, 1955.
15. D. Basu. On statistics independent of a complete sufficient statistic, Sankhyā, 15, 377–380, 1955.
16. D. Basu. A note on the theory of unbiased estimation, Ann. Math. Statist., 26, 345–348.
17. D. Basu. The concept of asymptotic efficiency, Sankhyā, 17, 193–196, 1956.
18. D. Basu. A note on the multivariate extension of some theorems related to the univariate normal distribution, Sankhyā, 17, 221–224.
19. D. Basu. On statistics independent of sufficient statistics, Sankhyā, 20, 223–226, 1958.
20. D. Basu. On sampling with and without replacement, Sankhyā, 20, 287–294, 1958.
21. D. Basu. The family of ancillary statistics, Sankhyā, 21, 247–256, 1959.
22. D. Basu. Recovery of ancillary information, Sankhyā, 26, 3–16, 1964.
23. D. Basu. Sufficiency and model preserving transformations, Technical Report, Dept. of Statistics, University of North Carolina, Chapel Hill, 1965.

24. D. Basu. Problems related to the existence of minimal and maximal elements in some families of subfields, Proc. Fifth Berkeley Symp. Math. Statist. and Prob., I, 41–50, Univ. California Press, Berkeley, 1967.
25. D. Basu and J. K. Ghosh. Invariant sets for translation parameter families of distributions, Ann. Math. Statist., 40, 162–174, 1969.
26. D. Basu. Role of sufficiency and likelihood principle in sample survey theory, Sankhyā, 31, 441–454, 1969.
27. D. Basu and J. K. Ghosh. Sufficient statistics in sampling from a finite universe, Bull. Inst. Internat. Statist., 42, 850–858, 1969.
28. D. Basu. On sufficiency and invariance, in *Essays on Probability and Statistics*, 61–84, R.C. Bose et al. eds., University of North Carolina Press, 1970.
29. D. Basu. An essay on the logical foundations of survey sampling, with discussions, in *Foundations of Statistical Inference*, 203–242, V. P. Godambe and D. A. Sprott eds., Holt, Rinehart, and Winston of Canada, Toronto, 1971.
30. D. Basu. Discussion of "A note on Basu's examples of anomalous ancillary statistics" by G. A. Barnard and D. A. Sprott, in *Foundations of Statistical Inference*, 163–176, V. P. Godambe and D. A. Sprott eds., Holt, Rinehart, and Winston of Canada, Toronto, 1971.
31. D. Basu, S. Chatterji, and M. Mukherjee. Aptitude Tests for Mathematics and Statistics, Statistical Publishing Society, Calcutta, 1972.
32. D. Basu. Discussion of "The use of the structural model in combining data bearing on the same characteristics" by J. B. Whitney in *Foundations of Statistical Inference*, 393–408, V. P. Godambe and D. A. Sprott eds., Holt, Rinehart, and Winston of Canada, Toronto, 1972.
33. D. Basu. Discussion of "Exact tests, confidence regions, and estimates" by P. Martin-Löf, in *Foundational Questions in Statistical Inference*, 121–138, Memoirs, No. 1, Dept. Theoretical Statist., Inst. Math., Univ. Aarhus, Aarhus, 1974.
34. D. Basu. Discussion of "Conditionality, pivotals, and robust estimation" by G. A. Barnard, in *Foundational Questions in Statistical Inference*, 61–80, Memoirs, No. 1, Dept. Theoretical Statist., Inst. Math., Univ. Aarhus, Aarhus, 1974.
35. D. Basu. Discussion of "Sufficiency, Prediction, and Extreme models,, by S. Lauritzen, in *Foundational Questions in Statistical Inference*, 249–269, Memoirs, No. 1, Dept. Theoretical Statist., Inst. Math., Univ. Aarhus, Aarhus, 1974.
36. Discussion of "Redundancy and its use as a quantitative measure of the deviation between a statistical hypothesis and a set of observational data" by P. Martin-Löf, in *Foundational Questions in Statistical Inference*, 1–42, Memoirs, No. 1, Dept. Theoretical Statist., Inst. Math., Univ. Aarhus, Aarhus, 1974.
37. D. Basu. Statistical information and likelihood, with discussion and correspondence between Barnard and Basu, Sankhyā, Ser. A, 37, 1–71, 1975.
38. D. Basu. On the elimination of nuisance parameters, Jour. Amer. Statist. Assoc., 72, 355–366, 1977.
39. D. Basu. On partial sufficiency: A review, Jour. Stat. Planning Inf., 2, 1–13, 1978.
40. D. Basu. On the relevance of randomization in data analysis, with discussions, *Survey Sampling and Measurement*, 267–339, N. Namboodiri ed., Academic Press, New York, 1978.
41. D. Basu. Discussion of "In dispraise of the exact tes" by Joseph Berkson, Jour. Stat. Planning Inf., 3, 189–192, 1979.
42. D. Basu. Survey Theory, in *Incomplete Data on Sample Surveys*, 3, 407–410, Academic Press, New York, 1979.
43. D. Basu. Randomization analysis of experimental data: The Fisher randomization test, with discussions, Jour. Amer. Statist. Assoc., 75, 575–595, 1980.
44. D. Basu. On ancillary statistics, pivotal quantities, and confidence statements, in *Topics in Applied Statistics*, 1–29, V. P. Chaubey and T. D. Dwivedi eds., Concordia University Publication, 1981.

45. D. Basu. Basu Theorems, Encyclopedia of Statistical Sciences, 1, 193–196, S. Kotz and N. L. Johnson eds., Wiley, New York, 1982.
46. D. Basu and S. C. Cheng. A note on sufficiency in coherent models, Internat. Jour. Math. Math. Sci., 3, 571–582, 1981.
47. D. Basu. Likelihood, Proceedings of National Symposium on Probability and Statistics, Univ. Sao Paulo, Brazil, 1981.
48. D. Basu and C. B. Pereira. On the Bayesian analysis of categorical data: The problem of nonresponse, Jour. Stat. Planning Inf., 6, 345–362, 1982.
49. D. Basu and R. Tiwari. A note on the Dirichlet process, in *Essays in Honor of C.R. Rao*, 89–103, G. Kallianpur, P.R. Krishnaiah, and J.K. Ghosh eds., North-Holland, Amsterdam, 1982.
50. D. Basu and C. B. Pereira. Conditional independence in statistics, Sankhy$\bar{a}$, Ser. A, 45, 324–337, 1983.
51. D. Basu and C. B. Pereira. A note on Blackwell sufficiency and a Skibinsky characterization of distributions, Sankhy$\bar{a}$, Ser. A, 45, 99–104, 1983.
52. D. Basu. Statistical Information and Likelihood: A Collection of Critical Essays, J. K. Ghosh ed., Lecture Notes in Statistics, 45, Springer-Verlag, New York, 1988.
53. D. Basu and C. B. Pereira. Blackwell sufficiency and Bernoulli experiments, Rebrape, 4, 137–145, 1990.
54. D. Basu. Learning statistics from counterexamples, Bayesian Analysis in Statistics and Econometrics, Lecture Notes in Statistics, 75, Springer-Verlag, New York, 1992.

Basu's Work on Randomization and Data Analysis

George Casella and Vikneswaran Gopal

1 Introduction

Sir R. A. Fisher put forward the idea that randomization is a necessary component of any designed experiment. It is accepted without question by most practitioners of statistics. Yet in the two papers

1. Basu, D. (1978) Relevance of randomization in data analysis. *Survey sampling and measurement* 267-292.
2. Basu, D. (1980) Randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association* **75** (371) 575-582.

Basu wonders out loud if randomization is really that important. He argues his case in the context of survey sampling, and when analyzing data using a randomization test.

In [1] Basu covers the survey sampling situation, and the randomization test is the topic of [2]. Although he acknowledges that there is a place for randomization in surveys (see Section 4 of [1]), his belief is the opposite for the randomization test. It is important to note the difference between the randomizations discussed in the two papers. In [1], Basu focuses on *prerandomization* - how to pick a sample from a sampling frame, and how it affects the subsequent analysis of data. In [2], the focus is on the *randomization test*, which was first introduced by Fisher. The two types of randomization are intricately linked, as the first provides a basis for the second. In essence, Basu argues that the absence of prerandomization does not make a dataset worthless, however, because of the total dependence of the randomization test on prerandomization, a randomization test is never valid.

In this commentary, we provide a short summary of Basu's ideas on randomization. That he did not write a great deal more on this topic is, in his own words, "a measure of my diffidence on the important question of the relevance of randomization at the data analysis stage".

G. Casella (✉)
Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grants DMS-0631632 and SES-0631588
e-mail: casella@stat.ufl.edu

V. Gopal (✉)
PhD Candidate, Department of Statistics, University Florida, 102 Griffin-Floyd Hall, Gainesville, FL 32611
e-mail: viknesh@stat.ufl.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_1,

2 Survey Sampling

The main question posed in [1] is about how to analyze the data generated by a survey or experiment. With a series of examples, Basu demonstrates the disadvantages of a frequentist approach, which is closely tied to the exact sampling plan used.

We highlight one of his more striking examples here. Suppose we have a well-defined finite population $\mathscr{P}$, consisting of individually identifiable objects called units. We can perceive of $\mathscr{P}$ as the set $\{1, 2, \ldots, N\}$. Corresponding to each $j \in \mathscr{P}$, there exists an unknown quantity Y_j. The goal of sampling is typically to estimate some function of $(Y_1, Y_2, \ldots, Y_N)$. The method of achieving this is through a sampling plan $\mathscr{S}$, by which we mean a set of rules, following which we can arrive at a subset s of $\mathscr{P}$.

Suppose also that we have a machine that produces $N = 100$ units in a day. However, it is possible for the machine to malfunction at some point, after which *it only produces defective products*. Using the definitions of the previous paragraph, Y_i take on values 1 or 0, depending on whether they are defective or functioning. The aim is to estimate $\theta = \sum Y_i$, the total number of defective products manufactured in a day, by drawing a sample from the N units.

Randomization is injected into the experiment through the choice of the sampling plan. Should we draw a simple random sample? Maybe a stratified sample? Whatever $\mathscr{S}$ we chose, the result of drawing a sample of size 4 would be recorded as, say,

$$Y_{17} = 0, Y_{24} = 0, Y_{40} = 1, Y_{73} = 1.$$

What then would a non-Bayesian statistician do with this data? To apply a randomization analysis, the probability of this sample with respect to the sampling scheme would have to be computed. A complicated enough scheme might even preclude this. A Bayesian, on the other hand, would observe that regardless of the sampling scheme applied, we know that $61 \leq \theta = \sum Y_i \leq 76$, since the first defective occurred in the set $\{25, 26, \ldots, 40\}$. Moreover, the likelihood function would be constant over the set $\{61, 62, \ldots, 76\}$ and we simply base all inference on this. Thus, Basu is invoking the Conditionality and Likelihood Principles to conclude that at the data analysis stage, the exact nature of the sampling plan is not important. He also points out that it in this case a sequential purposive sampling plan would serve our need better.

Notice that the example has been carefully set-up so that the non-Bayesian would be somewhat confused. For example, θ as it is presented here, would not be viewed as a parameter in classical statistics. But Basu, being a Bayesian, does not make a distinction between a random variable and a parameter. The way Basu presents the problem, a Bayesian analysis offers itself as the most natural thing to do. Such an approach avoids the need for obsessive randomization, and extracts information from the sample obtained rather than basing inference on samples that were not drawn.

3 The Test of Randomization

With [2], Basu places the randomization test under his microscope. At the end of his analyses, he concludes that he is unable to justify the use of the randomization test.

In the initial segments of the paper, Basu presents a version of the Fisher randomization test as a precursor to nonparametric tests such as the sign test and the Wilcoxon signed-rank test. Following that, he speculates that Fisher lost interest and belief in the randomization test. The final section of the paper is the most entertaining one. It contains a fictitious conversation between a scientist, a statistician and Basu himself. The three individuals discuss the randomization test introduced by Fisher in Chapter III (Section 21) of [3].

The scientist wishes to test whether a new diet is an improvement over the standard one. 30 animals are divided into 15 homogeneous pairs and from each pair, the scientist selects one subject for the treatment and the other one for the control. The response is the amount of weight gained in a subject after, say, 6 weeks. The data for each pair are recorded as (t_i, c_i). Suppose that for this particular experiment, the scientist records that $t_i - c_i > 0$ for all i, and that $T = \sum_i (t_i - c_i)$ is a large positive number.

H_0 states that the new diet makes no difference to the response. If this null hypothesis were true, it would mean that any difference in response for the i-th pair must have been caused by "nuisance" factors such as subject differences. Under H_0 then, the significance level of the observed statistic would be $\Pr(T' \geq T|H_0) = (1/2)^{15}$, assuming that all treatment assignments were equally likely. Basu takes the position that the randomization test should be applicable even if the randomization were not so. Specifically, he asks why the randomization test yields a different significance level if a biased coin were used to assign treatments within each pair. This apparent breakdown of the methodology is one of the reasons that leads Basu to recommend that the test not be used.

In introducing the article [2] in an earlier collection [4], Basu poses similar questions with regard to the famous tea-tasting experiment, which was also introduced by Fisher in [3].

> A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or tea infusion was first added to the cup. ... Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order.

The subject knows that there are 4 cups of each kind, and her task is to pick out the two groups of cups. Fisher argued that under the (null) hypothesis that the lady does not have the ability to distinguish, if we use the number of matches between the true grouping and the lady's grouping as a statistic, the significance level of a perfect grouping by the lady is given by

$$\Pr(T \geq 8|H_0) = \frac{1}{70}$$

Basu asks a series of questions of this approach:

> Why randomize? Was it because we wanted to keep the Lady in the dark about the actual layout? But then, why did we have to tell the Lady that there were exactly four cups of each kind in the layout and that all the 70 choices were equally likely? Why couldn't we choose just any haphazard looking layout and keep the lady uninformed about the choice? But then, how could we compute the significance level? Instead of randomizing over the full 70 point set, couldn't we randomize over a smaller, say, 10 point set of haphazard arrangements? How can we explain that in that case the same data (x, y) with $T = 8$ will be associated with a significance level of $1/10$? Why are we holding the Lady's response y as fixed and playing this probability game with the ancillary statistic x?

Fisher went some way to explaining some of these questions when he described the purpose of randomization, in Chapter II (Section 10) of [3].

> The element in the experimental procedure which contains the essential safeguard, is that the two modifications of the test beverage are to be prepared "in random order". ... The phrase "random order" itself, however, must be regarded as an incomplete instruction, standing as a kind of shorthand symbol for the full procedure of randomization, by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated.

Fisher says that randomization is what solves the problem of not being able to hold every single factor other than the treatment condition constant. The only solution is to ensure that every treatment allocation has an equal chance of occurring. Any other probability distribution on the treatment assignments could introduce a confounding factor.

For example, suppose that in the diet experiment, a coin that yields a treatment assignment of (t_i, c_i) with probability $1/4$ rather than $1/2$ is used. Then this is clearly against the requirement spelt out by Fisher, because for example, a treatment allocation with 15 (c_i, t_i)'s is more likely than one

with 15 (t_i, c_i)'s. If the animals were kept in a pen divided into 30 cells in a 15×2 arrangement, it is possible that the cells on the left obtained more sunlight and hence caused the animals to gain more weight. This would make the control treatment look good, since more animals on the left would be assigned that treatment.

The validity of the randomization test depends on the prerandomization being carried out properly, which requires that *all* treatment assignments be equally likely. Granted, Fisher never explicitly stated that when he said randomize, he meant for us to impose a uniform distribution on the treatment allocations. However, even if he had made his intentions explicit, would Basu have let him so lightly? We think not. Unless Fisher gave a sound mathematical argument as to why all treatment allocations should be equally likely, Basu's points would still be relevant and fair.

4 Final Thoughts

It is a tremendous joy to read Basu's papers. He presents his view in such a convincing manner that one almost feels ashamed at believing anything to the contrary. However, it is clear from the final sections of [1] that he does not suffer terribly from tunnel vision; he dissects his own arguments and tries to come up with explanations for possible criticisms of his points. It is also evident that he welcomes a good debate. The discussions at the end of [2] provide ample evidence for this. The banter between Basu and Kempthorne in particular, is fit for a comedy (be sure not to miss it!).

(Re-)Reading Basu's papers, which combine an inimitable style of writing with impactful examples, is an educating, enlightening and entertaining experience. At best, we question our assumptions and beliefs, which leads us to gain new insights into classical statistical concepts. At "worst", we embark on a journey to becoming Bayesian.

References

[1] Basu, D. (1978) Relevance of randomization in data analysis. *Survey sampling and measurement* 267–292.
[2] Basu, D. (1980) Randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association* **75** (371) 575–582.
[3] Fisher, R.A. (1966) The design of experiments.
[4] Ghosh, J.K. (1988) Statistical Information and Likelihood, Lecture Notes in Statistics **45**

Basu on Ancillarity

Philip Dawid

1 The origins of ancillarity

The term "ancillary statistic" was introduced by R. A. Fisher (Fisher 1925) in the context of maximum likelihood estimation. Fisher regarded the likelihood function as embodying all the information that the data had to supply about the unknown parameter. At a purely abstract level, this might be regarded as simply an application of the sufficiency principle (SP), since as a function of the data the whole likelihood function (*modulo* a positive constant factor — a gloss we shall henceforth take as read) is minimal sufficient; but that principle says nothing about what we should do with the likelihood function when we have it. Fisher went beyond this stark interpretation, regarding the actual form of the likelihood function as itself somehow embodying the appropriate inference. In some cases, such as full exponential families, the maximum likelihood estimator (MLE) is itself sufficient, fully determining the whole course of the likelihood function; but more generally it is only in many-one correspondence with the likehood function, so that two different sets of data can have associated likelihood functions whose maxima are in same place, but nevertheless differ in shape. Initially, for Fisher, an *ancillary* statistic (from the Latin "ancilla", meaning handmaiden) denoted a quantity calculated from the data which "lent support" to the MLE, by providing additional information about the shape of the likelihood function, over and above the position of its maximum — for example, higher derivatives of the log-likelihood at the MLE. If we regard the spikiness of the likelihood function as telling us something about the (data-dependent) precision of the MLE, we might select a suitable ancillary statistic to quantify this precision: this appears to have been Fisher's original motivation. According to this understanding of an ancillary statistic as describing the shape of the likelihood function, it is necessarily a function of the minimal sufficient statistic. Ideally, the MLE together with it handmaiden would fully determine the likelihood function, the pair then constituting a minimal sufficient statistic.

Fisher (1934) then considered the working out of these general concepts in the special case of a location model, where the MLE fully determines the location of the likelihood function, but is entirely uninformative as to its its shape; while the configuration statistic, *i.e.*, the set of pairwise differences between the observations, constitutes an ancillary statistic, fully determining the shape of the likelihood function, but uninformative about its location. For this model (though, for Fisher's original definition, not necessarily more generally) it is also true that the distribution of this ancillary statistic is entirely independent of the value of the unknown location parameter; furthermore, the conditional distribution of the maximum likelihood estimator, given the configuration, has a density

P. Dawid (✉)
Statistical Laboratory, Centre for Mathematical Sciences, CB3 0WB, UK
e-mail: apd@statslab.cam.ac.uk

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_2,

that has the same shape as the likelihood function. At a certain point, Fisher decided that it was such properties, rather than his original handmaiden conception, that were of crucial general importance, and from that point on the word "ancillary" was used to mean "having a distribution independent of the parameter". Associated with this was the somewhat vague idea of a "conditionality principle" (CP), whereby it is the conditional distribution of the data, given the ancillary statistic, that is regarded as supplying the appropriate "frame of reference" for determining the precision of our estimate. As a simple example lending support to this principle, suppose we first toss a fair coin, and then take 10 observations if it lands heads up, or 100 if it lands tails up. The coin toss does not depend upon the parameter (it is ancillary in the revised sense, although not necessarily in the original sense), and so cannot, of itself, be directly informative about it; but it does determine the precision of the experiment subsequently performed, and it does seem eminently sensible to condition on the number of observations actually taken to obtain a realistic measure of realised precision.

At an abstract level, CP can be phrased as requiring that any inference should be (or behave as if it were) conducted in the frame of reference that conditions on the realised value of an ancillary statistic. One can attempt to draw analogies between this CP and the sufficiency principle, SP, which tells us that our inference should always be (or behave as if it were) based on a sufficient statistic. But is important to note that in either case there may be a choice of statistics of the relevant kind, and we would like to be able to apply the principle simultaneously for all such. Considering first the case of sufficiency, suppose T_1 is sufficient and, in accordance with SP, we are basing our inference on T_1 alone. If now T_1 is a function of T_2, then T_2 is also sufficient: but the property that our inference should be based on T_2 alone is automatically inherited from this property holding for the "smaller" statistic T_1, so we do not need to take any explicit steps to ensure this. In particular, if we can find a *smallest* sufficient statistic T_0, a function of any other sufficient statistic, then basing our inference on T_0 will automatically satisfy SP with respect to *any* choice of sufficient statistic. It is well known that, subject only to mild regularity conditions, such a smallest ("minimal") sufficient statistic can generally be found. Hence it is pretty straightforward to satisfy SP simultaneously with respect to every sufficient statistic: simply base inference on the minimal sufficient statistic.

The case of ancillarity appears very similar, though with the functional ordering reversed. Suppose S_1 is ancillary, and, in accordance with CP, we are basing inference on the conditional distribution of the data, given S_1. If now S_2 is a function of S_1, then S_2 is also ancillary; and the property that inference is conditioned on S_2 is automatically inherited from this property holding for the "larger" statistic S_1. This analysis suggests that — in close analogy with the case of the minimal sufficient statistic — we should aim to identify a *largest* ancillary statistic S_0, of which every ancillary statistic would be a function. Then conditioning on S_0 would automatically satisfy CP, simultaneously with respect to every choice of ancillary statistic.

2 Enter Basu

The above analysis appears unproblematic, and might be thought to make a compelling case for always conditioning on the largest ancillary statistic — an apparently straightforward enterprise. But then along comes Basu, and suddenly things are not so clear!

Basu presented theory and counter-examples to show that in general there is *no* unique largest ancillary statistic, conditioning on which would allow us to apply the conditionality principle unambiguously. Typically there will be a multiplicity of ancillary statistics that are maximal, *i.e.*, cannot be expressed as a non-trivial function of any other ancillary; and in this case no single largest ancillary can exist. Even in what would seem to be the simplest special case, of two independent observations from the normal location model, having $X_i \sim \mathcal{N}(\theta, 1)$ $(i = 1, 2)$, there is no largest ancillary: for any $c \in [-\infty, +\infty]$, the statistic S_c defined as $X_1 - X_2$ if $X_1 + X_2 > c$, $X_2 - X_1$ otherwise, is

ancillary (Basu 1959). But knowing S_c for all c we can recover the full data (X_1, X_2) — which is clearly *not* ancillary. This possibility arises because two statistics can each be marginally ancillary, while not being jointly ancillary. Another example of this phenomenon occurs for the bivariate normal distribution with standard normal marginals and unknown correlation coefficient: the data on either variable singly are ancillary, but this clearly fails when both variables are combined.

Some have argued that such examples merely show that we should not have abandoned Fisher's original conception that an ancillary statistic should itself be a function of the minimal sufficient statistic — which does not hold in the above examples. But Basu has other examples that are not subject to this criticism. One simple example (Basu 1964) involves the outcome X of a single throw of a die, where, for some value of $\theta \in [0, 1]$, the probabilities of obtaining the scores 1–6 are respectively, $(1/12) \times (1 - \theta, 2 - \theta, 3 - \theta, 1 + \theta, 2 + \theta, 3 + \theta)$. Then X itself is minimal sufficient, but there are 6 non-equivalent maximal ancillary functions of X: for example, one such is $Y_1 = 0$ if $X = 1$ or 4, $Y_1 = 2$ if $X = 2$ or 5, $Y_1 = 3$ if $X = 3$ or 6; another is $Y_6 = 0$ if $X = 1$ or 6, $Y_6 = 2$ if $X = 2$ or 5, $Y_6 = 3$ if $X = 3$ or 4. We thus have a choice of ancillaries to condition on. For any such choice, the conditional distribution of X is confined to two possible values, so looking like a biased coin-toss; but the bias is a different function of θ in each case, and there is no clear reason to prefer one of these choices rather than another. Since there is no largest ancillary here, the simple interpretation of the conditionality principle, as enjoining us to make inference in the reference set obtained by conditioning on any ancillary, appears non-operational. Attempts — *e.g.*, Cox (1971), Kalbfleisch (1975) — have been made to restrict CP to apply only to certain ancillaries, but these are either unconvincingly *ad hoc* or fail fully to resolve the difficulty.

In the presence of a choice of maximal ancillaries to condition on, CP could nevertheless be rescued if the conditioning in fact had no effect (or had the same effect in all cases). In another strand of his work, Basu found conditions for this to hold. Thus let T be a *complete* (and hence also minimal) sufficient statistic. Basu (1955) showed that T must be independent of any ancillary statistic, for any value of the parameter. It follows that any inference based only on the marginal distribution of T, which of course respects SP, will also respect CP, with respect to any possible choice of ancillary statistic. This applies, for example, in the example above of two observations from the normal location model, in which the minimal sufficient statistic $\overline{T} = \frac{1}{2}(X_1 + X_2)$ is complete, hence independent of any ancillary (including S_c), so that any inference based on T alone will automatically satisfy CP. However, the general usefulness of this construction is limited, since in many problems (such as the biased die example above) the minimal sufficient statistic is not complete, and its conditional distribution does depend on which ancillary is conditioned on — so that this particular escape route is blocked off.

A related result (though with less direct relevance for CP) is that, under additional conditions, any statistic which is independent of a sufficient statistic is ancillary. In Basu (1955) this was asserted as true without further conditions; a counter-example and corrected version were given in Basu (1958).

3 Ancillarity and likelihood

In the light of these depressing results, one might reach the depressing conclusion that, however appealling CP may seem, there is no general way of satisfying it. And that conclusion is essentially correct if we take a frequentist approach to inference, since we end up with entirely different sample spaces, with entirely different properties, by conditioning on different ancillaries. However, this does not mean that there is no way of making inferences that respect CP. For example, suppose we agree, along with Fisher, that the message in the observation $X = x_0$ is entirely encapsulated in the likelihood function for the parameter θ that this observation generates: $L_0(\theta) \propto \text{Prob}(X = x_0 \mid \theta)$ (we here suppose, purely for simplicity of exposition, that the sample space is discrete). For any ancillary

statistic $S = s(X)$, the conditional probability of $X = x$, given $S = s := s(x)$, is $\text{Prob}(X = x \mid \theta)/\text{Prob}(S = s)$, where the denominator does not involve θ by ancillarity of S; so, for any data x_0 the likelihood function computed in this conditional frame of reference, $L_0^*(\theta) \propto \text{Prob}(X = x_0 \mid S = s_0; \theta)$ (with $s_0 = s(x_0)$), will be identical with the full-data likelihood function, $L_0(\theta) \propto \text{Prob}(X = x_0 \mid \theta)$. It follows that any inference that is based purely on the properties of the observed likelihood function — for instance, the maximum likelihood estimate, the curvature of the log-likelihood at its maximum, a Bayesian posterior based on a fixed prior distribution,... — will be entirely unaffected if we condition on an ancillary statistic, and hence will automatically satisfy CP.

One of the most significant results in this area was Birnbaum's theorem (Birnbaum 1962). This showed that any general method of inference about some parameter θ, applicable across a range of experimental setups, will satisfy both the sufficiency and the conditionality principles if and only if it depends only on the observed likelihood function — *i.e.*, it satisfies the likelihood principle, LP. Basu's investigations led him down this same path, and he did fully accept LP. However, even this was not enough for him, and in many of his works — *e.g.*, Basu (1975); Basu (1977) — he argued that the only sensible way of satisfying LP is to be, or at least act like, a Bayesian with a fixed proper prior distribution. But that is another story.

References

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377–80.

Basu, D. (1958). On statistics independent of sufficient statistics. *Sankhyā*, **20**, 223–6.

Basu, D. (1959). The family of ancillary statistics. *Sankhyā*, **21**, 247–56.

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

Basu, D. (1975). Statistical information and likelihood. *Sankhyā, Series A*, **37**, 1–71.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, **72**, 355–66.

Birnbaum, A. (1962). On the foundations of statistical inference (with Discussion). *Journal of the American Statistical Association*, **57**, 269–326.

Cox, D. R. (1971). The choice between alternative ancillary statistics. *Journal of the Royal Statistical Society, Series B*, **33**, 252–252.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, **22**, 700–25.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London, Series A*, **144**, 285–307.

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika*, **62**, 251–9.

Conditional Inference by Estimation of a Marginal Distribution

Thomas J. DiCiccio and G. Alastair Young

1 Introduction

Conditional inference has been, since the seminal work of Fisher (1934), a fundamental part of the theory of parametric statistics, but is a less established part of statistical practice. Crucial aspects of our understanding of the issues behind conditional inference are revealed by key work of Dev Basu: see, for example, Basu (1959, 1965).

Conditioning has two principal operational objectives: (i) the elimination of nuisance parameters; (ii) ensuring relevance of inference to an observed data sample through the conditionality principle of conditioning on the observed value of an ancillary statistic, when such a statistic exists. The concept of ancillarity here is usually taken to mean distribution constant. The elimination of nuisance parameters is usually associated with conditioning on sufficient statistics, and is most transparently and uncontroversially applied for inference in multiparameter exponential family models. Basu (1977) provides a general and critical discussion of conditioning to eliminate nuisance parameters. The notion of conditioning to ensure relevance, together with the associated problem, which exercised Fisher himself (Fisher, 1935), of recovering information lost when reducing the dimension of a statistical problem (to, say, that of the maximum likelihood estimator, when this estimator is not sufficient), is most transparent in transformation models, such as the location-scale model considered by Fisher (1934).

In some circumstances, issues to do with conditioning are clear cut. Though most often applied as a slick way to establish independence between two statistics, Basu's Theorem (Basu, 1955) shows that a boundedly complete sufficient statistic is independent of every ancillary statistic, which establishes the irrelevance for inference of any ancillary statistic when a boundedly complete sufficient statistic exists.

In many other circumstances, however, through the work of Basu and others, we have come to understand that there are formal difficulties with conditional inference. We list just a few. (1) It is well understood that conflict can emerge between conditioning and conventional measures of repeated sampling optimality, such as power. The most celebrated illustration is due to Cox (1958). (2) Typically there is arbitrariness of what to condition on; in particular, ancillary statistics are often not unique and a maximal ancillary may not exist. See, for instance, Basu (1959, 1965) and McCullagh (1992). (3) We must confront too the awkward mathematical contradiction of Birnbaum (1962), which says that

T.J. DiCiccio (✉)
Department of Social Statistics, ILR School, Cornell University, Ithaca, NY 14853
e-mail: tjd9@cornell.edu

G.A. Young (✉)
Department of Mathematics, Imperial College, London, UK
e-mail: alastair.young@imperial.ac.uk

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_3,

the conditionality principle, taken together with the quite uncontroversial sufficiency principle, imply acceptance of the likelihood principle of statistical inference, which is incompatible with the common methods of inference, such as calculation of p-values or construction of confidence sets, where we are drawn to the notion of conditioning.

Careful, elegant and accessible evaluation of these issues and related core ideas of statistical inference characterise much of the work of Basu, whose analyses had a profound impact on shaping the current prevailing attitude to conditional inference. Calculating a conditional sampling distribution is typically not easy, and such practical difficulties, taken together with the formal difficulties with conditional inference elucidated by Basu and others, have led to much of modern statistical theory being based on notions of inference which automatically accommodate conditioning, at least to some high order of approximation. Of particular focus are methods which respect the conditionality principle without requiring explicit specification of the conditioning ancillary, and which therefore circumvent the difficulties characterised by Basu associated with non-uniqueness of ancillaries.

Much attention in parametric theory now lies, therefore, in inference procedures which are stable, that is, which are based on a statistic that has, to some high order in the available data sample size, the same repeated sampling behaviour both marginally and conditionally given the value of the appropriate conditioning statistic. The notion is that accurate approximation to an exact conditional inference can then be achieved by considering the marginal distribution of the stable statistic, ignoring the relevant conditioning. This idea is elegantly expressed for the ancillary statistic context by, for example, Barndorff-Nielsen & Cox (1994, section 7.2), Pace & Salvan (1997, section 2.8) and Severini (2000, section 6.4). See also Efron & Hinkley (1978) and Cox (1980).

A principal approach to approximation of an intractable exact conditional inference lies in developments in higher-order small-sample likelihood asymptotics, based on saddlepoint and related analytic methods. Book length treatments of this analytic approach are given by Barndorff-Nielsen & Cox (1994) and Severini (2000). Brazzale *et al*. (2007) demonstrate very convincingly how to apply these developments in practice. Methods have been constructed which automatically achieve the elimination of nuisance parameters which is desired in the exponential family setting, though focus has been predominantly on ancillary statistic models. Here, a key development concerns construction of adjusted forms of the signed root likelihood ratio statistic, which require specification of the ancillary statistic, but are distributed, conditionally on the ancillary, as $N(0, 1)$ to third order, $O(n^{-3/2})$, in the data sample size n. Normal approximation to the sampling distribution of the adjusted statistic therefore provides third-order approximation to exact conditional inference: see Barndorff-Nielsen (1986). Approximations which yield second-order conditional accuracy, that is, which approximate the exact conditional inference to error of order $O(n^{-1})$, but which avoid specification of the ancillary statistic, are possible: Severini (2000 section 7.5) reviews such methods.

In the computer age, an attractive alternative approach to approximation of conditional inference uses marginal simulation, or 'parametric bootstrapping', of an appropriately chosen statistic to mimic its conditional distribution. The idea may be applied to approximate the conditioning that is appropriate to eliminate nuisance parameters in the exponential family setting, and can be used in ancillary statistic models, where it certainly avoids specification of the conditioning ancillary statistic.

2 An inference problem

To be concrete in our discussion, we consider the following inference problem. Let $Y = \{Y_1, \ldots, Y_n\}$ be a random sample from an underlying distribution $F(y; \eta)$, indexed by a d-dimensional parameter η, where each Y_i may be a random vector. Let $\theta = g(\eta)$ be a (possibly vector) parameter of interest, of dimension p. Without loss we may assume that $\eta = (\theta, \lambda)$, with θ the p-dimensional interest parameter and λ a q-dimensional nuisance parameter. Suppose we wish to test a null hypothesis of

the form $H_0 : \theta = \theta_0$, with θ_0 specified, or, through the familiar duality between tests of hypotheses and confidence sets, construct a confidence set for the interest parameter θ. If $p = 1$, we may wish to allow one-sided inference; for instance, we may want a test of H_0 against a one-sided alternative of the form $\theta > \theta_0$ or $\theta < \theta_0$, or construction of a one-sided confidence limit. Let $l(\eta) = l(\eta; Y)$ be the log-likelihood for η based on Y. Also, denote by $\hat{\eta} = (\hat{\theta}, \hat{\lambda})$ the overall maximum likelihood estimator of η, and by $\hat{\lambda}_\theta$ the constrained maximum likelihood estimator of λ for a given fixed value of θ. Inference on θ may be based on the likelihood ratio statistic, $W = w(\theta) = 2\{l(\hat{\eta}) - l(\theta, \hat{\lambda}_\theta)\}$. If $p = 1$, one-sided inference uses the signed square root likelihood ratio statistic $R = r(\theta) = \text{sgn}(\hat{\theta} - \theta)w(\theta)^{1/2}$, where $\text{sgn}(x) = -1$ if $x < 0$, $= 0$ if $x = 0$, and $= 1$ if $x > 0$. In a first-order theory of inference, the two key distributional results are that W is distributed as χ^2_p, to error of order $O(n^{-1})$, while R is distributed as $N(0, 1)$, to error of order $O(n^{-1/2})$.

3 Exponential family

Suppose the log-likelihood is of the form $l(\eta) = \theta s_1(Y) + \lambda^T s_2(Y) - k(\theta, \lambda) - d(Y)$, with θ scalar, so that θ is a natural parameter of a multi-parameter exponential family. We wish to test $H_0 : \theta = \theta_0$ against a one-sided alternative, and do so using the signed root statistic R.

Here the conditional distribution of $s_1(Y)$ given $s_2(Y) = s_2$ depends only on θ, so that conditioning on the observed value s_2 is indicated as a means of eliminating the nuisance parameter. So, the appropriate inference on θ is based on the distribution of $s_1(Y)$, given the observed data value of s_2. This conditional inference has the unconditional (repeated sampling) optimality property of yielding a uniformly most powerful unbiased test: see, for example, Young & Smith (2005, section 7.2). The necessary conditional distribution is, in principle, known, since it is completely specified, once θ is fixed. In practice, however, the exact inference may be difficult to construct: the relevant conditional distribution typically requires awkward analytic calculations, numerical integrations, etc., and may even be completely intractable.

DiCiccio & Young (2008) show that in this exponential family context, accurate approximation to the exact conditional inference may be obtained by considering the marginal distribution of the signed root statistic R under the fitted model $F(y; (\theta, \hat{\lambda}_\theta))$, that is, under the model with the nuisance parameter taken as the constrained maximum likelihood estimator, for the given value of θ. This scheme yields inference agreeing with exact conditional inference to relative error of third order, $O(n^{-3/2})$. Specifically, DiCiccio & Young (2008) show that

$$\text{pr}\{R \geq r; (\theta, \hat{\lambda}_\theta)\} = \text{pr}(R \geq r | s_2(Y) = s_2; \theta)\{1 + O(n^{-3/2})\},$$

when r is of order $O(1)$. Their result is shown for both continuous and discrete models. The approach therefore has the same asymptotic properties as saddlepoint methods developed by Skovgaard (1987) and Barndorff-Nielsen (1986) and studied by Jensen (1992). DiCiccio & Young (2008) demonstrate in a number of examples that this approach of estimating the marginal distribution of R gives very accurate approximations to conditional inference even in very small sample sizes. A crucial point of their analysis is that the marginal estimation should fix the nuisance parameter as its constrained maximum likelihood estimator: the same third-order accuracy is not obtained by fixing the nuisance parameter at its global maximum likelihood value $\hat{\lambda}$.

Third-order accuracy can also be achieved, in principle, by estimating the marginal distributions of other asymptotically standard normal pivots, notably Wald and score statistics. However, in numerical investigations, using R is routinely shown to provide more accurate results. A major advantage of using R is its low skewness; consequently, third-order error can be achieved, although not in a relative sense, by merely correcting R for its mean and variance and using a standard normal approximation

to the standardized version of R. Since it is computationally much easier to approximate the mean and variance of R by parametric bootstrapping at $(\theta, \hat{\lambda}_\theta)$ than it is to simulate the entire distribution of R, the use of mean and variance correction offers substantial computational savings, especially for constructing confidence intervals. Although these savings are at the expense of accuracy, numerical work suggests that the loss of accuracy is unacceptable only when the sample size is very small.

4 Ancillary statistic models

In modern convention, ancillarity in the presence of nuisance parameters is generally defined in the following terms. Suppose the minimal sufficient statistic for η may be written as $(\hat{\eta}, A)$, where the statistic A has, at least approximately, a sampling distribution which does not depend on the parameter η. Then A is said to be ancillary and the conditionality principle would argue that inference should be made conditional on the observed value $A = a$.

McCullagh (1984) showed that the conditional and marginal distributions of signed root statistics derived from the likelihood ratio statistic W for a vector interest parameter, but with no nuisance parameter, agree to error of order $O(n^{-1})$, producing very similar p-values whether one conditions on an ancillary statistic or not. Severini (1990) considered similar results in the context of a scalar interest parameter without nuisance parameters; see also Severini (2000, section 6.4.4). Zaretzki *et al.* (2009) show the stability of the signed root statistic R, in the case of a scalar interest parameter and a general nuisance parameter, and their methodology can be readily extended to the case of a vector interest parameter θ to establish stability of signed root statistics derived from W in the presence of nuisance parameters. Stability of the likelihood ratio statistic W is immediate: the marginal and conditional distributions are both χ^2_p to error $O(n^{-1})$.

Since the marginal and conditional distributions of R coincide to error of order $O(n^{-1})$ given $A = a$, it follows that the conditional p-values obtained from R are approximated to the same order of error by the marginal p-values. Moreover, for approximating the marginal p-values, the marginal distribution of R can be approximated to error of order $O(n^{-1})$ by means of the parametric bootstrap; the value of η used in the bootstrap can be either the overall maximum likelihood estimator, $\eta = (\hat{\theta}, \hat{\lambda})$, or the constrained maximum likelihood estimator, $\eta = (\theta, \hat{\lambda}_\theta)$. For testing the null hypothesis $H_0 : \theta = \theta_0$, the latter choice is feasible; however, for constructing confidence intervals, the choice $\eta = (\hat{\theta}, \hat{\lambda})$ is computationally less demanding. DiCiccio *et al.* (2001) and Lee & Young (2005) showed that the p-values obtained by using $\eta = (\theta, \hat{\lambda}_\theta)$ are marginally uniformly distributed to error of order $O(n^{-3/2})$, while those obtained by using $\eta = (\hat{\theta}, \hat{\lambda})$ are uniformly distributed to error of order $O(n^{-1})$ only. Numerical work indicates that using $\eta = (\theta, \hat{\lambda}_\theta)$ improves conditional accuracy as well, although, formally, there is no difference in the orders of error to which conditional p-values are approximated by using the two choices. Though in principle the order of error in approximation of exact conditional inference obtained by considering the marginal distribution of R is larger than the third-order, $O(n^{-3/2})$, error obtained by normal approximation to the sampling distribution of the adjusted signed root statistic, substantial numerical evidence suggests very accurate approximations are obtained in practice. Particular examples are considered by DiCiccio *et al.* (2001), Young & Smith (2005, section 11.5) and Zaretzki *et al.* (2009).

In the case of a vector interest parameter θ, both the marginal and conditional distributions of $W = w(\theta)$ are chi-squared to error $O(n^{-1})$, and hence, using the χ^2_p approximation to the distribution of W achieves conditional inference to error of second order. Here, however, it is known (Barndorff-Nielsen & Cox, 1984) that a simple scale adjustment of the likelihood ratio statistic improves the chi-squared approximation:

$$\frac{p}{\mathrm{E}_{(\theta,\lambda)}\{w(\theta)\}} w(\theta)$$

is distributed as χ^2_p to error of order $O(n^{-2})$. Since $\mathrm{E}_{(\theta,\lambda)}\{w(\theta)\}$ is of the form $p + O(n^{-1})$, it follows that $\mathrm{E}_{(\theta,\hat{\lambda}_\theta)}\{w(\theta)\} = \mathrm{E}_{(\theta,\lambda)}\{w(\theta)\} + O_p(n^{-3/2})$. Thus, estimation of the marginal distribution of W by bootstrapping with $\eta = (\theta, \hat{\lambda}_\theta)$ yields an approximation having error of order $O(n^{-3/2})$; moreover, to error of order $O(n^{-2})$, this approximation is the distribution of a scaled χ^2_p random variable with scaling factor $\mathrm{E}_{(\theta,\hat{\lambda}_\theta)}\{w(\theta)\}/p$. The result of Barndorff-Nielsen & Hall (1988), that

$$\frac{p}{\mathrm{E}_{(\theta,\hat{\lambda}_\theta)}\{w(\theta)\}} w(\theta)$$

is distributed as χ^2_p to error of order $O(n^{-2})$, shows that confidence sets constructed by using the bootstrap approximation to the marginal distribution of W have marginal coverage error of order $O(n^{-2})$.

The preceding results continue to hold under conditioning on the ancillary statistic. In particular,

$$\frac{p}{\mathrm{E}_{(\theta,\lambda)}\{w(\theta)|A = a\}} w(\theta)$$

is, conditional on $A = a$, also χ^2_p to error of order $O(n^{-2})$. The conditional distribution of W is, to error of order $O(n^{-2})$, the distribution of a scaled χ^2_p random variable with scaling factor $\mathrm{E}_{(\theta,\lambda)}\{w(\theta)|A = a\}/p$. Generally, the difference between $\mathrm{E}_{(\theta,\lambda)}\{w(\theta)\}$ and $\mathrm{E}_{(\theta,\lambda)}\{w(\theta)|A = a\}$ is of order $O(n^{-3/2})$ given $A = a$, and using the bootstrap estimate of the marginal distribution of W approximates the conditional distribution to error of order $O(n^{-3/2})$. Thus, confidence sets constructed from the bootstrap approximation have conditional coverage error of order $O(n^{-3/2})$, as well as marginal coverage error of order $O(n^{-2})$.

Bootstrapping the entire distribution of W at $\eta = (\theta, \hat{\lambda}_\theta)$ is computationally expensive, especially when constructing confidence sets, and two avenues for simplification are feasible. First, the order of error in approximation to conditional inference remains of order $O(n^{-3/2})$ even if the marginal distribution of W is estimated by bootstrapping with $\eta = (\hat{\theta}, \hat{\lambda})$, the global maximum likelihood estimator. It is likely that using $\eta = (\theta, \hat{\lambda}_\theta)$ produces greater accuracy; however, this increase in accuracy might not be sufficient to warrant the additional computational demands. Second, instead of bootstrapping the entire distribution of W, the scaled chi-squared approximation could be used, with the scaling factor $\mathrm{E}_{(\theta,\hat{\lambda}_\theta)}\{w(\theta)\}/p$ being estimated by the bootstrap. It could be of interest to examine, by theoretical analysis or numerical examples, which of these two simplifications is preferable. Use of the bootstrap for estimating Bartlett adjustment factors was proposed by Bickel & Ghosh (1990).

References

BARNDORFF-NIELSEN, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307–22.

BARNDORFF-NIELSEN, O. E. & COX, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *J.R. Statist. Soc*. B **46**, 483–95.

BARNDORFF-NIELSEN, O. E. & COX, D. R. (1994). *Inference and Asymptotics*. London: Chapman & Hall.

BARNDORFF-NIELSEN, O. E. & HALL, P. G. (1988). On the level-error after Bartlett adjustment of the likelihood ratio statistic. *Biometrika* **75**, 374–8.

BASU, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya* **15**, 377–80.

BASU, D. (1959). The family of ancillary statistics. *Sankhya* **21**, 247–56.

BASU, D. (1965). Problems relating to the existence of maximal and minimal elements in some families of statistics (subfields). *Proc. Fifth Berk. Symp. Math. Statist. Probab.* **1**, 41-50. Berkeley CA: University of California Press.

BASU, D. (1977). On the elimination of nuisance parameters. *J. Amer. Statist. Assoc.* **72**, 355–66.

BICKEL, J. K. & GHOSH, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction – a Bayesian argument. *Ann. Statist.* **18**, 1070–90.

BIRNBAUM, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc.* **57**, 269–306.

BRAZZALE, A. R., DAVISON, A. C. & REID, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge: Cambridge University Press.

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Stat.* **29**, 357–72.

COX, D. R. (1980). Local ancillarity. *Biometrika* **67**, 279–86.

DICICCIO, T. J., MARTIN, M. A. & STERN, S. E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Can. J. Statist.* **29**, 67–76.

DICICCIO, T. J. & YOUNG, G. A. (2008). Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika* **95**, 747–58.

EFRON, B. & HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with discussion). *Biometrika* **65**, 457–87.

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. R. Soc. Lond. A* **144**, 285–307.

FISHER, R. A. (1935). The logic of inductive inference. *J.R. Statist. Soc.* **98**, 39–82.

JENSEN, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika* **79**, 693–703.

LEE, S. M. S. & YOUNG, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Stat. Prob. Letters* **71**, 143–53.

MCCULLAGH, P. (1984). Local sufficiency. *Biometrika* **71**, 233–44.

MCCULLAGH, P. (1992). Conditional inference and Cauchy models. *Biometrika* **79**, 247–59.

PACE, L. & SALVAN, A. (1997). *Principles of Statistical Inference: from a Neo-Fisherian Perspective*. Singapore: World Scientific.

SKOVGAARD, I. M. (1987). Saddlepoint expansions for conditional distributions. *J. Appl. Probab.* **24**, 875–87.

SEVERINI, T. A. (1990). Conditional properties of likelihood-based significance tests. *Biometrika* **77**, 343–52.

SEVERINI, T. A. (2000). *Likelihood methods in Statistics*. Oxford: Oxford University Press.

YOUNG, G. A. & SMITH, R. L. (2005). *Essentials of Statistical Inference*. Cambridge: Cambridge University Press.

ZARETZKI, R., DICICCIO, T. J. & YOUNG, G. A. (2009). Stability of the signed root likelihood ratio statistic in the presence of nuisance parameters. *Submitted for publication*.

Basu's Theorem

Malay Ghosh

Professor Basu, in his long illustrious career, has made many fundamental contributions to the foundations of statistical inference. Among others, I point out his work on ancillarity, likelihood principle, partial and marginal sufficiency, randomization and foundations of survey sampling.

In spite of all the above contributions, Basu is possibly the most well-known to a vast majority of statisticians for a theorem which bears his name. Basu's Theorem, published in *Sankhya*, 1955, has served several generations of statisticians as a fundamental tool for proving independence of statistics. The theorem itself is beautiful because of its elegance and simplicity, and yet one must acknowledge its underlying depth, as it is built on several fundamental concepts of statistics, such as sufficiency, completeness and ancillarity.

The theorem simply states that if a sufficient statistic T is boundedly complete and a statistic U is ancillary, then T and U are independently distributed. But the theorem is not just useful for what it says. It can be used in a wide range of applications such as in distribution theory, hypothesis testing, theory of estimation, calculation of moments of many complicated statistics, calculation of mean squared errors of empirical Bayes estimators, and even surprisingly, establishing infinite divisibility of certain distributions. The application possibly extends to many other areas of statistics which I have not come across. I strongly believe that even probabilists can benefit by knowing this theorem, since it may provide a handy tool for finding distributions of many complex statistics.

A detailed set of examples showing applications of Basu's Theorem in various branches of statistics is available in Ghosh (2002). I will present only a few of them here. But first I will discuss a few conceptual issues as pointed out in Lehmann (1981) and DasGupta (2006).

Lehmann (1981) pointed out that the properties of minimality and completeness of a sufficient statistic are of a rather different nature. A complete sufficient statistic is minimal sufficient, but the converse is not necessarily true. The existence of a minimal sufficient statistic T, by itself, does not guarantee that there does not exist any function of T which is ancillary. Basu's Theorem tells us that if T is complete in addition to being sufficient, then no ancillary statistic other than the constants can be computed from T. Thus, by Basu's Theorem, completeness of a sufficient statistic T characterizes the success of T in separating the informative part of the data from that part, which by itself, carries no information. The following example illustrates this.

Example 1 Let $X_1, \cdots, X_n$ be independent and identically distributed (iid) with joint probability density function (pdf) belonging to the double exponential family

M. Ghosh (✉)
Distinguished Professor of Statistics, University of Florida, Gainesville, FL 32611
e-mail: ghoshm@stat.ufl.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_4,

$$\mathcal{P} = \left\{ \prod_{i=1}^{n} f_\theta(x_i) : f_\theta(x) = f(x-\theta),\ x \in \mathcal{R}^1, \theta \in \mathcal{R}^1 \right\}, \tag{1}$$

where $f(x) = (1/2)\exp(-|x|)$. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote the ordered X_i's. Then $T = (X_{(1)}, \cdots, X_{(n)})$ is minimal sufficient for $\mathcal{P}$. But clearly T is not boundedly complete as can be verified directly by showing that the expectation of any bounded function of $X_{(n)} - X_{(1)}$ (with the exception of constants) is a nonzero constant not depending on θ, but that the probability that the function is equal to that constant is not 1 (in fact equal to zero). In this example, $X_{(n)} - X_{(1)}$, while ancillary, is not independent of T. On the other hand, if one considers instead the augmented class of all continuous pdf's $\mathcal{P} = \{\prod_{i=1}^{n} f(x_i) : f \text{ continuous}\}$, then T is indeed complete, and Basu's Theorem asserts that there does not exist any non-constant function of T which is ancillary.

Thus, for the double exponential family, sufficiency has not been successful in "squeezing out" all the ancillary material, while for the augmented family, success takes place by virtue of Basu's Theorem.

There are several ways to think of possible converses to Basu's Theorem. One natural question is that if T is boundedly complete and sufficient, and U is distributed independently of T for every $\theta \in \Theta$, then is U ancillary? The answer is no as pointed out by in Koehn and Thomas (1975) in the following example.

Example 2 Let $X \sim$ uniform$[\theta, \theta+1)$, where $\theta \in \Theta = \{0, \pm 1, \pm 2, \cdots, \}$. Then X has pdf $f_\theta(x) = I_{[[x]=\theta]}$, where $[x]$ denotes the integer part of x. It is easy to check that $[X]$ is complete sufficient for $\theta \in \Theta$, and is also distributed independently of X, but clearly X is not ancillary!

The above apparently trivial example brings out several interesting issues. First, since $P_\theta([X] = \theta) = 1$ for all $\theta \in \Theta$, so that $[X]$ is degenerate with probability 1. Indeed, in general, a nontrivial statistic cannot be independent of X, because if this were the case, it would be independent of every function of X, and thus independent of itself! However, this example shows also that if there exists a nonempty proper subset $\mathcal{X}_0$ of $\mathcal{X}$, and a nonempty proper subset Θ_0 of Θ such that

$$\begin{aligned} P_\theta(\mathcal{X}_0) &= 1 \text{ for } \theta \in \Theta_0; \\ &= 0 \text{ for } \theta \in \Theta - \Theta_0, \end{aligned} \tag{2}$$

then the converse to Basu's Theorem may fail to hold. In Example 2, $\mathcal{X} = \mathcal{R}^1$, and Θ is the set of all integers. Taking $\Theta_0 = \{\theta_0\}$ and $\mathcal{X}_0 = [\theta_0, \theta_0 + 1)$, one produces a counterexample to a possible converse to Basu's Theorem.

Basu (1958) gave a sufficient condition for the converse to his theorem. First he defined two probability measures P_θ and $P_{\theta'}$ to be *overlapping* if they do not have disjoint supports. In Example 2, all probability measures $P_\theta, \theta \in \Theta$ are non-overlapping. The family $\mathcal{P}$ is said to be *connected* if for every pair $\{\theta, \theta'\}$, $\theta \in \Theta$, $\theta' \in \Theta$, there exist $\theta_1, \cdots, \theta_k$, each belonging to Θ such that any two members of the sequence overlap. The following theorem is given in Basu (1958).

Theorem 1 *Let $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ be connected, and T be sufficient for $\mathcal{P}$. Then U is ancillary if T and U are independent for every $\theta \in \Theta$.*

It is only the sufficiency and not the completeness of T which plays a role in Theorem 1. An alternative way to think about a possible converse to Basu's Theorem is whether the independence of all ancillary statistics with a sufficient statistic T implies that T is boundedly complete. The answer is again NO as Lehmann (1981) produces the following counterexample.

Example 3 Let X be a discrete random variable assuming values x with probabilities $p(x)$ as given below:

x	-5	-4	-3	-2	-1	1	2	3	4	5
$p(x)$	$\alpha' p^2 q$	$\alpha' pq^2$	$\frac{1}{2}p^3$	$\frac{1}{2}q^3$	$\gamma' pq$	γpq	$\frac{1}{2}q^3$	$\frac{1}{2}p^3$	αpq^2	$\alpha p^2 q$

Here $0 < p = 1 - q < 1$, is the unknown parameter, and $\alpha, \alpha', \gamma, \gamma'$ are known positive constants satisfying $\alpha + \gamma = \alpha' + \gamma' = 3/2$. In this example, $|X|$ is minimal sufficient, $P(X > 0) = 1/2$ so that $U = I_{[X>0]}$ is ancillary. However, if $\alpha \neq \alpha'$, then U is not distributed independently of T.

Lehmann (1981) pointed out that this converse to Basu's Theorem fails to hold because ancillarity is a property of the whole distribution of a statistic, while completeness is a property dealing only with expectations. He showed also that correct versions of the converse could be obtained either by replacing ancillarity with the corresponding first order property or completeness with a condition reflecting the whole distribution.

To this end, define a statistic $V \equiv V(X)$ to be first order ancillary if $E_\theta(V)$ does not depend on $\theta \in \Theta$. Then one has a necessary and sufficient condition for Basu's Theorem.

Theorem 2 *A necessary and sufficient condition for a sufficient statistic T to be boundedly complete is that every bounded first order ancillary V is uncorrelated with every bounded real-valued function of T for every $\theta \in \Theta$.*

An alternative approach to obtain a converse is to modify instead the definition of completeness. Quite generally, a sufficient statistic T is said to be $\mathcal{G}$-complete ($\mathcal{G}$ is a class of functions) if for every $g \in \mathcal{G}$, $E_\theta[g(T)] = 0$ for all $\theta \in \Theta$ implies that $P_\theta[g(T) = 0] = 1$ for all $\theta \in \Theta$. Suppose, in particular, $\mathcal{G} = \mathcal{G}_0$, where $\mathcal{G}_0$ is the class of all two-valued functions. Then Lehmann (1981) proved the following theorem.

Theorem 3 *Suppose T is sufficient and every ancillary statistic U is distributed independently of T. Then T is $\mathcal{G}_0$-complete.*

Basu's Theorem implies the independence of T and U when T is boundedly complete and sufficient, while U is ancillary. This, in turn, implies the $\mathcal{G}_0$-completeness of T. However, the same Example 3 shows that neither of the reverse implications is true. On the other hand, if instead of $\mathcal{G}_0$, one considers $\mathcal{G}_1$ which are conditional expectations of all two-valued functions with respect to a sufficient statistic T, then Lehmann proved the following theorem.

Theorem 4 *A necessary and sufficient condition for a sufficient statistic T to be $\mathcal{G}_1$-complete is that every ancillary statistic U is independent of T (conditionally) for every $\theta \in \Theta$.*

Theorems 2–4, provide conditions under which a sufficient statistic T has some form of completeness (not necessarily bounded completeness) if it is independent of every ancillary U. However, Theorem 1 says that ancillarity of U does not follow even if it is independent of a complete sufficient statistic. As shown in Example 2, $[X]$ is complete sufficient, and hence, by Basu's Theorem, is independent of every ancillary U, but $[X]$ independent of X, but X not ancillary.

Next we provide a few selected examples which show multifaceted applications of Basu's Theorem.

Example 4 (A Distribution Theory Result) Let $\boldsymbol{X}_i = (X_{1i}, X_{2i})^T$ be n iid random variables, each having a bivariate normal distribution with means $\mu_1 (\in \mathcal{R}^1)$ and $\mu_2 (\in \mathcal{R}^1)$, variances $\sigma_1^2 (> 0)$ and $\sigma_2^2 (> 0)$, and correlation $\rho \in (-1, 1)$. Let $\bar{X}_j = n^{-1} \sum_{i=1}^n X_{ji}$, $S_j^2 = \sum_{i=1}^n (X_{ji} - \bar{X}_j)^2$ $(j = 1, 2)$ and $R = \sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)/(S_1 S_2)$. Under the null hypothesis $H_0 : \rho = 0$, $(\bar{X}_1, \bar{X}_2, S_1^2, S_2^2)$ is complete sufficient for $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, while R is ancillary. Thus $(\bar{X}_1, \bar{X}_2, S_1^2, S_2^2)$ is distributed independently of R when $\rho = 0$. Due to the mutual independence of $\bar{X}_1$, $\bar{X}_2$, S_1^2 and S_2^2 when

$\rho = 0$, one gets now the mutual independence of $\bar{X}_1$, $\bar{X}_2$, S_1^2, S_2^2 and R when $\rho = 0$, and the joint pdf of these five statistics is now the product of the marginals. Now to derive the joint pdf $q_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho}(\bar{x}_1, \bar{x}_2, s_1^2, s_2^2, r)$ of these five statistics for an arbitrary $\rho \in (-1, 1)$, by the Factorization Theorem of sufficiency, one gets

$$q_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho}(\bar{x}_1, \bar{x}_2, s_1^2, s_2^2, r) = q_{0,0,1,1,0}(\bar{x}_1, \bar{x}_2, s_1^2, s_2^2, r)\frac{L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)}{L(0, 0, 1, 1, 0)},$$

where $L(\cdot)$ denotes the likelihood function under the specified values of the parameters.

Example 5 This example, taken from Boos and Hughes-Oliver (BH) (1998), is referred to as the *Monte Carlo Swindle*. The latter refers to a simulation technique that ensures statistical accuracy with a smaller number of replications at a level which one would normally expect from a much larger number of replications. Johnstone and Velleman (1985) provide many such examples. One of their examples taken by BH shows that if M denotes a sample median in a random sample of size n from a $N(\mu, \sigma^2)$ distribution, then the Monte Carlo estimate of $V(M)$ requires a much smaller sample size to attain a prescribed accuracy, if instead one finds the Monte Carlo estimate of $V(M - \bar{X})$ and adds the usual estimate of σ^2/n to the same.

We do not provide the detailed arguments of BH to demonstrate this. We point out only the basic identity $V(M) = V(M - \bar{X}) + V(\bar{X})$ as used by these authors. As noticed by BH, this is a simple consequence of Basu's Theorem. As mentioned in Example 2, for fixed σ^2, $\bar{X}$ is complete sufficient for μ, while $M - \bar{X} = \text{med}(X_1 - \mu, \cdots, X_n - \mu) - (\bar{X} - \mu)$ is ancillary. Hence, by Basu's Theorem,

$$V(M) = V(M - \bar{X} + \bar{X}) = V(M - \bar{X}) + V(\bar{X}) = V(M - \bar{X}) + \sigma^2/n.$$

Hogg and Craig (1956) have provided several interesting applications of Basu's Theorem. Among these, there are some hypothesis testing examples where Basu's Theorem aids in the derivation of the exact distribution of $-2\log_e \lambda$ under the null hypothesis H_0, λ being the generalized likelihood ratio test (GLRT) statistic. One common feature in all these problems is that the supports of all the distributions depend on parameters. We discuss one of these examples in its full generality.

Example 6 Let X_{ij} $(j = 1, \cdots, n_i; i = 1, \cdots, k)$ $(k \geq 2)$ be mutually independent, X_{ij} $(j = 1. \cdots, n_i)$ being iid with common pdf

$$f_{\theta_i}(x_i) = [h(x_i)/H(\theta_i)]I_{[0 \leq x_i \leq \theta_i]},\ i = 1, \cdots, k, \tag{3}$$

where $H(u) = \int_0^u h(x)dx$, and $h(x) > 0$ for all $x > 0$. We want to test $H_0 : \theta_1 = \cdots = \theta_k$ against the alternative H_1: not all θ_i are equal. We write $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{in_i})^T$, $i = 1, \cdots, k$ and $\boldsymbol{X} = (\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_k^T)^T$. Also, let $T_i \equiv T_i(\boldsymbol{X}_i) = \max(X_{i1}, \cdots, X_{in_i})$, $i = 1, \cdots, k$, and $T = \max(T_1, \cdots, T_k)$. The unrestricted MLE's of $\theta_1, \cdots, \theta_k$ are $T_1, \cdots, T_k$. Also, under H_0, the MLE of the common θ_i is T. Then the GLRT statistic for testing H_0 against H_1 simplifies to $\lambda(\boldsymbol{X}) = \prod_{i=1}^k H^{n_i}(T_i)/H^n(T)$, where $n = \sum_{i=1}^k n_i$. Hence,

$$\begin{aligned} -2\log_e \lambda(\boldsymbol{X}) = &\sum_{i=1}^k [-2\log_e\{H^{n_i}(T_i)/H^{n_i}(\theta)\}] \\ &- [-2\log_e\{H^n(T)/H^n(\theta)\}], \end{aligned} \tag{4}$$

where θ denotes the common unknown value of the θ_i's under H_0. It follows from (4) that $T_1, \cdots, T_k$ are independent with distribution functions $H^{n_i}(t_i)/H^{n_i}(\theta_i)$, $(i = 1, \cdots, k)$. Hence, under H_0, $H^{n_i}(T_i)/H^{n_i}(\theta)$ are iid uniform(0,1). Accordingly, under H_0,

$$\sum_{i=1}^{k}[-2\log_e\{H^{n_i}(T_i)/H^{n_i}(\theta)\}] \sim \chi^2_{2k}. \tag{5}$$

Also, under H_0, the distribution function of T is $H^n(t)/H^n(\theta)$, and hence, $H^n(T)/H^n(\theta)$ is uniform(0,1) under H_0. Thus, under H_0,

$$-2\log_e[H^n(T)/H^n(\theta)] \sim \chi^2_2. \tag{6}$$

So far, we have not used Basu's Theorem. In order to use it, first we observe that under H_0, T is complete sufficient for θ, while λ is ancillary. Hence, under H_0, T is distributed independently of $-2\log_e\lambda$. Also, from (5),

$$\sum_{i=1}^{k}[-2\log_e\{H^{n_i}(T_i)/H^{n_i}(\theta)\}] = [-2\log_e\lambda] + [-2\log_e\{H^n(T)/H^n(\theta)\}]. \tag{7}$$

The two components in the right hand side of (8) are independent. Now by (6), (7) and the result that if W_1 and W_2 are independent with $W_1 \sim \chi^2_m$ and $W_1 + W_2 \sim \chi^2_{m+n}$, then $W_2 \sim \chi^2_n$, one finds that $-2\log_e\lambda \sim \chi^2_{2k-2}$ under H_0.

The above result should be contrasted to the regular case (when the support of the distribution does not depend on parameters) where under some regularity conditions, $-2\log_e\lambda$ is known to have an asymptotic chisquared distribution. In a similar scenario with n observations and k unknown parameters in general, and 1 under the null, the associated degrees of freedom in the regular case would have been $(n-1)-(n-k) = k-1$ instead of $2(k-1)$.

Empirical Bayes (EB) analysis has, of late, become very popular in statistics, especially when the problem is simultaneous estimation of several parameters. An EB scenario is one in which known relationships among the coordinates of a parameter vector, say, $\theta = (\theta_1, \cdots, \theta_k)^T$ allow use of the data to estimate some features of the prior distribution. For example, one may have reasons to believe that the θ_i are iid from a prior Π_0, where Π_0 is structurally known except possibly for some unknown parameter (possibly vector-valued) λ. A parametric EB procedure is one where λ is estimated from the marginal distribution of the observations.

Often in an EB analysis, one is interested in finding Bayes risks of the EB estimators. Basu's Theorem helps considerably in many such calculations as we demonstrate below.

Example 7 We consider an EB framework as proposed in Morris (1983a, 1983b). Let $X_i|\theta_i$ be independent $N(\theta_i, V)$, where $V(>0)$ is assumed known. Let θ_i be independent $N(z_i^T\boldsymbol{b}, A)$, $i = 1, \cdots, k$. The p-component ($p < k$) design vectors z_i are assumed to be known, and let $\boldsymbol{Z}^T = (z_1, \cdots, z_k)$. We assume rank($\boldsymbol{Z}$)=p. Based on the above likelihood and the prior, the posteriors of the θ_i are independent $N((1-B)X_i + Bz_i^T\boldsymbol{b}, V(1-B))$, where $B = V/(V+A)$. Accordingly, the posterior means, the Bayes estimators of the θ_i are given by

$$\hat{\theta}_i^{BA} = (1-B)X_i + Bz_i^T\boldsymbol{b},\ i = 1, \cdots, k. \tag{8}$$

In an EB set up, $\boldsymbol{b}$ and A are unknown, and need to be estimated from the marginal distributions of the X_i's. Marginally, the X_i's are independent with $X_i \sim N(z_i^T\boldsymbol{b}, V+A)$. Then, writing $\boldsymbol{X} = (X_1, \cdots, X_k)^T$, based on the marginal distribution of $\boldsymbol{X}$, the complete sufficient statistic for $(\boldsymbol{b}, A)$ is $(\hat{\boldsymbol{b}}, S^2)$, where $\hat{\boldsymbol{b}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{X}$ is the least squares estimator or the MLE of $\boldsymbol{b}$, and $S^2 = \sum_{i=1}^{k}(X_i - z_i^T\hat{\boldsymbol{b}})^2$. Also, based on the marginal of $\boldsymbol{X}$, $\hat{\boldsymbol{b}}$ and S^2 are independently distributed with $\hat{\boldsymbol{b}} \sim N(\boldsymbol{b}, (V+A)(\boldsymbol{Z}^T\boldsymbol{Z})^{-1})$, and $S^2 \sim (V+A)\chi^2_{k-p}$. Accordingly $\boldsymbol{b}$ is estimated by $\hat{\boldsymbol{b}}$. The MLE

of B is given by $\min(kV/S^2, 1)$, while its UMVUE is given by $V(k-p-2)/S^2$, where we must assume $k > p+2$ for the latter to be meaningful. If instead, one assigns the prior $\Pi(\boldsymbol{b}, A) \propto 1$ as in Morris (1983a,1983b), then the HB estimator of θ_i is given by $\hat{\theta}_i^{HB} = (1 - B^*(S^2))X_i + B^*(S^2)z_i^T\hat{\boldsymbol{b}}$, where $B^*(S^2) = \int_0^1 B^{\frac{1}{2}(k-p-2)} \exp\left(-\frac{1}{2V}BS^2\right) dB / \int_0^1 B^{\frac{1}{2}(k-p-4)} \exp(-\frac{1}{2V}BS^2) dB$. Thus a general EB estimator of θ_i is of the form

$$\hat{\theta}_i = [1 - \hat{B}(S^2)]X_i + \hat{B}(S^2)\boldsymbol{z}_i^T\hat{\boldsymbol{b}}. \tag{9}$$

We will now demonstrate an application of Basu's Theorem in finding the mean squared error (MSE) matrix $E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T]$, where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)^T$, and expectation is taken over the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{\theta}$. The following theorem provides a general expression for the MSE matrix.

Theorem 5 *With the notations of this section,*

$$\begin{aligned} E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] &= V(1-B)\boldsymbol{I}_k + VB\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T \\ &\quad + E[(\hat{B}(S^2) - B)^2 S^2](k-p)^{-1}(\boldsymbol{I}_k - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T). \end{aligned}$$

Proof Write $\hat{\boldsymbol{\theta}}^{BA} = (1-B)\boldsymbol{X} + B\boldsymbol{Z}\boldsymbol{b}$. Then

$$\begin{aligned} E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] &= E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA} + \hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA} + \hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})^T] \\ &= E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA})^T] + E[(\hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})^T], \end{aligned} \tag{10}$$

since

$$E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA})(\hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})^T] = E[E(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA}|\boldsymbol{X})(\hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})^T] = 0.$$

Now

$$\begin{aligned} E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA})^T] &= E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{BA})^T|\boldsymbol{X}] \\ &= E[Var(\boldsymbol{\theta}|\boldsymbol{X})] = E[V(1-B)\boldsymbol{I}_k] = V(1-B)\boldsymbol{I}_k. \end{aligned} \tag{11}$$

Next after a little algebra, we get

$$\hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}} = (\hat{B}(S^2) - B)(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}}) + B\boldsymbol{Z}(\hat{\boldsymbol{b}} - \boldsymbol{b}).$$

Now by the independence of $\hat{\boldsymbol{b}}$ with $\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}}$, noting $S^2 = ||\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}}||^2$, where $||\cdot||$ denotes the Euclidean norm, and $Var(\hat{\boldsymbol{b}}) = VB^{-1}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$, one gets

$$\begin{aligned} E(\hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{BA} - \hat{\boldsymbol{\theta}})^T &= E[(\hat{B}(S^2) - B)^2(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})^T] \\ &\quad + VB\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T. \end{aligned} \tag{12}$$

Next we observe that

$$(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})^T/S^2 = \frac{\left[(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{b}) - \boldsymbol{Z}\left(\hat{\boldsymbol{b}} - \boldsymbol{b}\right)\right]\left[(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{b}) - \boldsymbol{Z}\left(\hat{\boldsymbol{b}} - \boldsymbol{b}\right)\right]^T (V + A)^{-1}}{||\,(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{b}) - \boldsymbol{Z}\left(\hat{\boldsymbol{b}} - \boldsymbol{b}\right)||^2\,(V + A)^{-1}}$$

is ancillary, and by Basu's Theorem, is independent of S^2, which is a function of the complete sufficient statistic $(\hat{\boldsymbol{b}}, S^2)$. Accordingly,

$$E[(\hat{B}(S) - B)^2(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})^T] = E[(\hat{B}(S) - B)^2 S^2]E[(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})^T / S^2], \quad (13)$$

and then by the formula for moments of ratios,

$$\begin{aligned} E[(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})^T / S^2] &= E[(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})(\boldsymbol{X} - \boldsymbol{Z}\hat{\boldsymbol{b}})^T]/E(S^2) \\ &= (k-p)^{-1}[\boldsymbol{I}_k - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T]. \end{aligned} \quad (14)$$

The theorem follows.

References

BASU, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, **15**, 377–380.

BASU, D. (1958). On statistics independent of a sufficient statistic. *Sankhya*, **20**, 223–226.

BOOS, D.D. and HUGHES-OLIVER, J.M. (1998). Applications of Basu's Theorem. *The American Statistician*, **52**, 218–221.

HOGG, R.V. and CRAIG, A.T. (1956). Sufficient statistics in elementary distribution theory. *Sankhya*, **16**, 209–216.

JOHNSTONE, I. M. and VELLEMAN, P.F. (1985). Efficient scores, variance decomposition, and Monte carlo swindles. *Journal of the American Statistical Association*, **80**, 851–862.

KOEHN, U. and THOMAS, D.L. (1975). On statistics independent of a sufficient statistic: Basu's Lemma. *American Statistician*, **29**, 40–42.

LEHMANN, E.L. (1981). An interpretation of completeness and Basu's Theorem. *Journal of the American Statistical Association*, **76**, 335–340.

MORRIS, C. N. (1983a). Parametric empirical Bayes confidence intervals. In *Scientific Inference, Data Analysis and Robustness*. Eds. G.E.P. Box, T. Leonard and C.F.J. Wu. Academic Press, New York, pp 25–50.

MORRIS, C.N. (1983b) (with discussion). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, **78**, 47–65.

Basu's Work on Likelihood and Information

Joseph B. Kadane

It has been a joy learning from Dev Basu's work on aspects of statistical inference, and especially his deep and often provocative essays on fallacies of common statistical principles. I will limit myself to his epic paper *Statistical Information and Likelihood*.

"Statistical Information and likelihood" is a tour de force in three parts. In the first part, Basu studies the implications of the sufficiency and conditionality principles, and shows that these lead to the likelihood function as the summary of the information in an experiment. His treatment is similar to that of Birnbaum (1962, 1972). His second part reviews non-Bayesian likelihood methods, leaning especially on Fisher's maximum likelihood method (MLE). He criticizes the use of sampling standard errors around the MLE to create confidence intervals in the grounds that they violate the likelihood principle. His third part gives various examples that illuminate what he finds problematic about fiducial arguments, improper Bayesian priors, and simple-null hypothesis testing. Although most of his effort is critical, on the positive side Basu advocates subjective Bayesian analysis with proper priors, and making optimal decisions using a utility (or loss) function.

This essay needs to be understood in the context of its time. It was given in lecture form in 1972, ten years after Fisher's death. Fisher himself vigorously, vociferously, and sometimes with blind fury would attack those who disagreed with him. Basu is speaking from within the Fisherian tradition, and showing, by theorem and by counterexample, that large parts of that tradition simply do not make sense. This took courage and conviction, particularly considering the audience to whom he gave the talk. The discussants, in alphabetical order, were Barnard, Barndorff-Nielsen, Cox, Dempster, Edwards, J.D. Kalbfleisch, Lauritzen, Martin-Lof, and Rasch. Of these, only Dempster had anything supportive to say about Bayesian ideas, and he characterizes himself as a "sometimes Bayesian".

Nonetheless, the discussion is civil and respectful. I am particularly struck by the tone of the exchange of letters between Basu and Barnard. By the end, only a couple of points are still subject to disagreement, and the atmosphere is collegial.

Basu concludes his response to the discussion by writing "The Bayesian and Neyman-Pearson-Wald theories of data analysis are the two poles in current statistical thought. Today I find assembled before me a number of eminent statisticians who are looking for a via media between the two poles. I can only wish you success in an endeavor in which the redoubtable R.A.Fisher failed".

The situation is much the same today. The difficulty lies in what is to be regarded as random and what is to be regarded as fixed. To a classical statistician, the data are random, even after they have been observed, while the parameters are fixed but unknown (whatever that may mean). To a Bayesian, the data, after they are observed, are fixed at the observed values, but the parameters are uncertain,

J.B. Kadane (✉)
University Professor of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
e-mail: kadane@stat.cmu.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_5,

and hence random. There are not convenient middle grounds between these two perspectives. Basu has no hesitation about where he stands, writing "with an experiment already planned and performed, and with the sample x already before us, I do not see any point in speculating about all other samples that might have been." This places him solidly in the Bayesian camp.

Reference

[1] Basu, D. (1975). Statistical information and likelihood, with discussion and correspondence between Barnard and Basu, Sankhy$\bar{a}$, Ser. A, 37, 1–71.

Basu on Survey Sampling

Glen Meeden

Fifty years ago at a large scientific conference a statistician and a probabilist happen to set down together for lunch. In the ensuing small talk the probabilist admitted to knowing nothing about statistics and ask for a brief introduction to the subject. His companion outlined the common scenario of a company receiving a shipment of 1,000 widgets and selecting 20 of them at random to be tested. He then explained how the number of defective widgets in the sample could be used to make inferences about the state of the remaining 980 widgets in the shipment. The probabilist thought about this for a minute and then remarked, "I do not understand how knowledge about the 20 sampled units can tell me anything about the remaining 980 unsampled units." It is easy to forget how nonintuitive it is to claim that learning the observed values of the units in a sample, selected by random sampling, translates to knowledge about the unobserved values of the units remaining in the population.

If $y = (y_1, \ldots, y_N)$ is the vector of unknown population values of the characteristic of interest then given a sample s we denote the observed or seen values by $y(s) = \{y_i : i \in s\}$ and the remaining unobserved or unseen values by $y(s') = \{y_j : j \notin s\}$. For Basu the fundamental question of survey sampling is how can one relate the seen to the unseen. Without some assumption about how these two sets are related knowing $y(s)$ does not tell one anything about $y(s')$. His application of the sufficiency and likelihood principles to survey sampling demonstrated that all we learn from the observed data are the values of the characteristic of interest in the sampled units and that the "true" vector of population values must be consistent with these observed values. Note this fact justifies the probabilist's statement. Moreover, Basu showed that this is true for any sampling plan where, at any stage, the choice of the next population unit to be observed is allowed to depend on the observed values of the characteristic of the previously selected units.

For Basu the Bayesian paradigm was the natural way to relate the unseen to the seen and still follow the likelihood principle. Let $\pi(y)$ be the prior density function or probability function for the Bayesian survey sampler over the parameter space of possible vectors y. The Bayesian selects $\pi(\cdot)$ to represent the prior information and his or her prior beliefs about y. Once the sample has been selected and the seen have been observed inferences are based on the posterior distribution, $\pi(y(s')|y(s))$, of the unseen given the seen and the design plays no role.

When Basu was writing we were, for the most part, restricted to prior distributions whose posterior distributions could only be studied using paper and pencil. With the recent advances in Bayesian computing it is now possible to simulate complete copies of $y(s')$ for many different possible posterior distributions. For such a posterior given $y(s)$ we can form many copies of $y(s')$ and hence many complete copies of the population. Supposes we are interested in estimating the function $\gamma(y)$. For

G. Meeden
Chairman and Head, School of Theoretical Statistics, University of Minnesota, Minneapolis, MN 55455
e-mail: glen@stat.umn.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_6,

each complete simulated copy we can compute the value of γ. Given a large set of such simulated values we can find approximately the corresponding point and interval estimates of γ. The key point in a Bayesian analysis is finding a sensible prior distribution. Once this in hand and the sample has been selected inferences can be made for any function γ of interest.

This is in contrast to the design approach where the sampling design is often an important way to incorporate prior information into a problem. The design along with an unbiased requirement leads to an appropriate estimator. One difficulty with this approach is that each different choice of the function γ requires a different argument. At a more fundamental level this suggests that the design approach does not yield a coherent method of relating the unseen to the seen. Basu never found this approach compelling because it violated the likelihood principle. Furthermore he never had much good to say about unequal probability sampling designs since, again by the likelihood principle, after the sample has been chosen the selection probabilities should play no role at the inferential stage.

Much of survey practice is still design based. It has always seemed curious to me that this one area of statistics where prior information is routinely employed makes use of this information in a way that cannot be justified from the Bayesian perspective. This is especially surprising given Basu's work. It is interesting to speculate why this is so. Part of the reason, I believe, is that it has always been difficult to find sensible and tractable prior distributions for large dimensional problems. This is particularly true in survey sampling which often deals with governmental statistics for which a certain degree of objectivity is expected. The challenge for a Bayesian is to find prior distributions which allow one to make use of the kinds of prior information which are now incorporated into a design. Our ability to now simulate complete copies of a population from more complicated but realistic posterior distributions should help fulfill the promise of Basu's work in the years ahead.

Commentary on Basu (1956)

Robert J. Serfling

Asymptotic relative efficiency of estimators

For statistical estimation problems, it is typical and even desirable that more than one reasonable estimator can arise for consideration. One natural and time-honored approach for choosing an estimator is simply to compare the sample sizes at which the competing estimators meet a given standard of performance. This depends upon the chosen measure of performance and upon the particular population distribution F.

For example, we might compare the sample mean versus the sample median for location estimation. Consider a distribution function F with density function f symmetric about an unknown point θ to be estimated. For $\{X_1, \ldots, X_n\}$ a sample from F, put $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\mathrm{Med}_n = \mathrm{median}\{X_1, \ldots, X_n\}$. Each of $\overline{X}_n$ and Med_n is a *consistent* estimator of θ in the sense of convergence in probability to θ as the sample size $n \to \infty$. To choose between these estimators we need to use further information about their performance. In this regard, one key aspect is *efficiency*, which answers:

Question A *How concentrated about θ is the sampling distribution of $\widehat{\theta}$?*

Criteria for asymptotic relative efficiency

Variance as a measure of performance

A simple and natural criterion relative to the above question is the *variance* of the sampling distribution: the smaller this variance, the more "efficient" is that estimator. In this regard, let us consider "large-sample" sampling distributions. For $\overline{X}_n$, the classical central limit theorem tells us: if F has finite variance σ_F^2, then the sampling distribution of $\overline{X}_n$ is approximately $N(\theta, \sigma_F^2/n)$, i.e., Normal with mean θ and variance σ_F^2/n. For Med_n, a similar classical result [10] tells us: if the density f is continuous and positive at θ, then the sampling distribution of Med_n is approximately $N(\theta, 1/4[f(\theta)]^2 n)$. On this basis, we consider $\overline{X}_n$ and Med_n to perform equivalently at respective sample sizes n_1 and n_2 if

R.J. Serfling (✉)
Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas 75083-0688, USA
e-mail: serfling@utdallas.edu
Website: www.utdallas.edu/~serfling
Support by NSF Grant DMS-0805786 and NSA Grant H98230-08-1-0106 is gratefully acknowledged.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_7,

$$\frac{\sigma_F^2}{n_1} = \frac{1}{4[f(\theta)]^2 n_2}.$$

Keeping in mind that these sampling distributions are only approximations assuming that n_1 and n_2 are "large", we define the *asymptotic relative efficiency (ARE)* of Med to $\overline{X}$ as the *large-sample limit* of the ratio n_1/n_2, i.e.,

$$\text{ARE}(\text{Med}, \overline{X}, F) = 4[f(\theta)]^2 \sigma_F^2. \tag{1}$$

More generally, for any parameter η of a distribution F, and for estimators $\widehat{\eta}^{(1)}$ and $\widehat{\eta}^{(2)}$ which are approximately $N(\eta, V_1(F)/n)$ and $N(\eta, V_2(F)/n)$, respectively, the *ARE of* $\widehat{\eta}^{(2)}$ *to* $\widehat{\eta}^{(1)}$ is given by

$$\text{ARE}(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F) = \frac{V_1(F)}{V_2(F)}. \tag{2}$$

Interpretation: If $\widehat{\eta}^{(2)}$ is used with a sample of size n, the number of observations needed for $\widehat{\eta}^{(1)}$ to perform equivalently is $\text{ARE}(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F) \times n$.

In view of the asymptotic normal distribution underlying the above formulation of ARE in estimation, we may also characterize the ARE given by (2) as the limiting ratio of sample sizes at which the *lengths of associated confidence intervals at approximate level* $100(1-\alpha)\%$,

$$\widehat{\eta}^{(i)} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{V_i(F)}{n_i}}, \ i = 1, 2,$$

converge to 0 at the same rate, when holding fixed the coverage probability $1 - \alpha$. (In practice, of course, consistent estimates of $V_i(F)$, $i = 1, 2$, are used in forming the CI.)

The treatment of ARE for consistent asymptotically normal estimators using the variance criterion had been long well established by the 1950s – see [1] for a string of references.

Probability concentration as a measure

Instead of comparison of asymptotic variance parameters as a criterion, one may quite naturally compare the *probability concentrations* of the estimators in any ε-neighborhood of the target parameter η: $P(|\widehat{\eta}^{(i)} - \eta| > \varepsilon)$, $i = 1, 2$. When we have

$$\frac{\log P(|\widehat{\eta}_n^{(i)} - \eta| > \varepsilon)}{n} \to \gamma^{(i)}(\varepsilon, \eta), \ i = 1, 2,$$

as is typical, then the ratio of sample sizes n_1/n_2 at which these concentration probabilities converge to 0 at the same rate is given by $\gamma^{(1)}(\varepsilon, \eta)/\gamma^{(2)}(\varepsilon, \eta)$, which then represents another ARE measure for the efficiency of estimator $\widehat{\eta}_n^{(2)}$ relative to $\widehat{\eta}_n^{(1)}$. This entails approximation of the sampling distribution in the tails. Accordingly, instead of central limit theory the relevant tool is *large deviation theory*, which is rather more formidable. In the context of hypothesis testing, Chernoff [3] argued that when the sample size approaches infinity it is appropriate to minimize both Type I and Type II error probabilities, rather than minimizing one with the other held fixed. He developed an ARE index essentially based on tail probability approximations. See also [10, **1.15.4**] for general discussion.

How compatible are these two criteria?

Those who have been fortunate enough to observe D. Basu in action, as I was when we were colleagues at Florida State University in the early 1970s, know his talent for inquiring into the boundaries of any good idea. Relative to the present context, when the variance and probability concentration criteria were just becoming established criteria in the 1950s, stemming from somewhat differing orientations, it was Basu who thought of exploring their compatibility. Basu [2] provides an example in which not only do the variance-based and concentration-based measures disagree on which estimator is better. but they do so in the most extreme sense: one ARE is infinite at every choice of F in a given class Ω, while the other ARE is zero for every such F.

Basu's construction is straightforward and worth discussing, so we briefly examine some details. For $X_1, \ldots, X_n$ an i.i.d. sample from $N(\mu, 1)$, put

$$\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i \quad \text{and} \quad S_n = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Basu defines the estimation sequences for μ given by $T = \{t_n\}$ and $T' = \{t'_n\}$, with

$$t_n = (1 - H_n)\overline{X}_n + nH_n \quad \text{and} \quad t'_n = \overline{X}_{[\sqrt{n}]},$$

where $H_n = 1$ if $S_n > a_n$ and 0 otherwise, and a_n satisfies $P(S_n > a_n) = 1/n$. He shows that $\sqrt{n}(t_n - \mu) \xrightarrow{d} N(0, 1)$. Since also $n^{-1/4}(t'_n - \mu) \xrightarrow{d} N(0, 1)$, it follows that the ARE according to (2) is given by

$$\text{ARE}(t_n, t'_n, N(\mu, 1)) = \lim_{n\to\infty} \frac{n^{-1}}{n^{-1/2}} = 0. \tag{3}$$

He also shows that the corresponding ARE based on concentration probabilities for any fixed choice of ε is given by

$$\lim_{n\to\infty} \frac{n^{-1}}{o(n^{-1})} = \infty. \tag{4}$$

An immediate observation about this example is that it is not pathological. Rather, it employs ordinary ingredients characteristic of typical application contexts.

Another important aspect is that the disagreement between the two notions of ARE is as extreme as possible. Not merely differing with respect to whether the ARE is < 1 or > 1, here one version is *infinite* at every choice of F in the class $\Omega = \{N(\mu, 1) : -\infty < \mu < \infty\}$, while the other version is *zero* for every such F.

The details of proof yield the interesting corollary that (4) also gives the *concentration probability* ARE of t_n versus simply $\overline{X}_n$. Thus the estimator which is *optimal* under the *variance* ARE criterion is *infinitely nonoptimal* under the concentration probability ARE criterion.

A slight variation on Basu's $\{t_n\}$ provides an example of *superefficient estimator*, similar to that of J. L. Hodges (see Le Cam, 1953). discussed in Lehmann and Casella (1998). Put

$$t^*_n = A(1 - H_n)\overline{X}_n + nH_n$$

for some constant $A \neq 1$. Then we have that $\sqrt{n}(t^*_n - \mu) \xrightarrow{d} N(0, A^2) + \lim_{n\to\infty} \sqrt{n}\mu(A - 1)$, i.e., $\sqrt{n}(t^*_n - \mu)$ converges to $\pm\infty$ if $\mu \neq 0$ and otherwise converges to $N(0, A^2)$. Therefore, in the

case that $\mu = 0$ and $A < 1$, the estimator t_n^* outperforms the "optimal" estimator. See Lehmann and Casella (1998) for useful discussion of superefficiency.

We see that the content of Basu's example, like all of his contributions to statistical thinking, reflects great ingenuity and insight applied very productively to useful purposes.

Subsequent developments

The impact of Basu [2] thus has been to motivate stronger interest in "large deviation (LD) approaches" to ARE. For example, Bahadur [1] follows up with a deep discussion of this approach along with many constructive ideas. Quite a variety of LD and related moderate deviation approaches are discussed in Serfling [10, Chap. 10]. More recently, Puhalskii and Spokoiny [9] provide an extensive treatment of the LD approach in statistical inference. For convenient elementary overviews on ARE in estimation and testing, see DasGupta [4], Serfling [11], and Nikitin [8], for example.

Acknowledgments Very helpful suggestions of Anirban DasGupta are greatly appreciated and have been used to improve the manuscript. Also, support by NSF Grant DMS-0805786 and NSA Grant H98230-08-1-0106 is gratefully acknowledged.

References

[1] Bahadur, R. R. (1960). Asymptotic relative efficiency of tests and estimates. *Sankhyā* **22** 229–252.
[2] Basu, D. (1956). On the concept of asymptotic relative efficiency. *Sankhyā* **17** 193–196.
[3] Chernoff, H. (1952). A measure of asymptotic relative efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics* **23** 493–507.
[4] DasGupta, A. (1998). Asymptotic relative efficiency. *Encyclopedia of Biostatistics, Vol. I*, 210–215, P. Armitage and T. Colton (Eds.). John Wiley, New York.
[5] Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *University of Calif. Publ. in Statistics* **1** 277–330.
[6] Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd edition. Springer.
[7] Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press.
[8] Nikitin, Y. (2010). Asymptotic relative efficiency in testing. *International Encyclopedia of Statistical Sciences* (M. Lovric, ed.). Springer.
[9] Puhalskii, A. and Spokoiny, V. (1998). On large-deviation efficiency in statistical inference. *Bernoulli* **4** 203–272.
[10] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.
[11] Serfling, R. (2010). Asymptotic relative efficiency in estimation. *International Encyclopedia of Statistical Sciences* (M. Lovric, ed.). Springer.

Commentary on *A Note on the Dirichlet Process*

Jayaram Sethuraman

Consider the standard inference problem where the data $\mathbf{X} = (X_1, \ldots, X_n)$, taking values in $\mathcal{X}^n$, consist of independent observations from a common distribution F or probability measure P on $(\mathcal{X}, \mathcal{B})$. The essentials of Bayesian analysis are well developed for this problem, when the common distribution F is completely specified in terms of a finite number k of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. A prior distribution for this problem is just a suitable distribution on a subset in R^k described by any restrictions the parameters θ must satisfy. The calculation of posterior distribution become routine from this point onwards. Special classes of prior distributions called conjugate families have the property that the posterior distribution is also in the same class. Computations become simpler when the conjugate family is itself parametrized by a finite number of parameters. If the prior distribution in a conjugate family can be described by the parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ then the posterior distribution based on data $\mathbf{X}$ would also be in the same conjugate family with parameters $\boldsymbol{\lambda}^{\mathbf{X}} = (\lambda_1^{\mathbf{X}}, \ldots, \lambda_m^{\mathbf{X}})$; the term "updated" bas been used to describe the parameters of the posterior distribution.

When $\mathcal{X}$ consists of k points, the common distribution P is a discrete distribution and it can be completely described by $\mathbf{p} = (p_1, \ldots, p_k)$ satisfying $p_i \geq 0, i = 1 \ldots, k, \sum_1^k p_i = 1$. In view of the linear dependence among these parameters it is enough to specify a prior distribution just for $(p_1, \ldots, p_{k-1})$; in other words, P is parametrized by $k-1$ parameters satisfying the above conditions. A natural conjugate family is the finite dimensional Dirichlet distribution $\mathcal{D}(\lambda_1, \ldots, \lambda_k)$, which is the distribution of $(\frac{Z_1}{Z}, \ldots, \frac{Z_{k-1}}{Z})$ where $Z = Z_1 + \cdots + Z_k$ and $Z_1, \ldots, Z_k$ are independent Gamma random variables with parameters $\lambda_1, \ldots, \lambda_k$ respectively. The parameters $(\lambda_1, \ldots, \lambda_k)$ will have to satisfy the conditions $\lambda_1 \geq 0, \ldots, \lambda_k \geq 0, \sum_1^k \lambda_i > 0$. When all the λ_i are positive, the distribution $\mathcal{D}(\lambda_1, \ldots, \lambda_k)$ can be defined in a more familiar way by a pdf proportional to $\prod_1^{k-1} p_i^{\lambda_i - 1}(1 - \sum_1^{k-1} p_j)^{\lambda_k - 1}$. The data $\mathbf{X} = (X_1, \ldots, X_n)$ which are i.i.d. P can also be summarized by its empirical distribution function $F_n(j) = \frac{1}{n}\sum_1^n I(X_i = j), j = 1, \ldots, k$. In this case the posterior distribution can easily be shown to also be the finite dimensional Dirichlet distribution $\mathcal{D}(\lambda_1 + nF_n(1), \ldots, \lambda_k + nF_n(k))$.

To perform nonparametric inference when the common distribution P is not restricted to such parametric classes one should study classes of probability distributions for P which varies in the space $\mathcal{P}$ of all probability measures on $(\mathcal{X}, \mathcal{B})$. A natural σ-field in $\mathcal{P}$ is $\sigma(\mathcal{P})$, the smallest σ-field such that sets of the form $\{P : P(B) < r\}$ where $B \in \mathcal{B}$ and $r \in [0, 1]$. One can consider P as $(P(B), B \in \mathcal{B})$ and thus take P as an element of $[0, 1]^\infty$ satisfying the familiar countable additivity assumptions. One can also consider P to be the vectors $(P(B_1), \ldots, P(B_k))$ satisfying

J. Sethuraman (✉)
Distinguished Professor of Statistics, Florida State University, Tallahassee, FL 32306
e-mail: sethu@ani.stat.fsu.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_8,

$P(B_i) \geq 0, i = 1, \ldots, k, \sum_1^k P(B_i) = 1$ over all finite measurable partitions $(B_1, \ldots, B_k)$ of $\mathcal{X}$ and satisfying some other conditions. Ferguson (1973) used the above description to define a probability measure on $\mathcal{P}$ as follows. Let α be a non-zero finite measure on $(\mathcal{X}, \mathcal{B})$. Ferguson (1973) defined $\mathcal{D}_\alpha$, the Dirichlet process with parameter α, to be the probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$ under which $(P(B_1), \ldots, P(B_k))$ has the finite dimensional Dirichlet distribution $\mathcal{D}(\alpha(B_1), \ldots, \alpha(B_k))$, for each measurable partition $(B_1, \ldots, B_k)$ of $\mathcal{X}$.

The paper of Ferguson (1973) revolutionized the subject of nonparametric Bayes methods by showing that when $\mathcal{D}_\alpha$ is used as a prior distribution for P, the posterior distribution becomes the Dirichlet process $\mathcal{D}_{\alpha+nF_n}$ where F_n is the empirical measure of the data $\mathbf{X}$. This means that Dirichlet processes form a conjugate family of priors in the standard nonparametric problem. This paper further showed how to obtain Bayes estimates of several functions of P. It also established that Dirichlet process $\mathcal{D}_\alpha$ is concentrated on the subset of all discrete probability measures in $\mathcal{P}$.

The papers Blackwell (1973) and Blackwell and MacQueen (1973) appeared in the same journal issue where Ferguson's article appeared. These papers gave other ways to define a Dirichlet prior and to establish its properties.

Berk and Savage (1979) also gave an elementary proof of the result that Dirichlet processes concentrate on the collection of discrete measures.

The treatment of measure theoretical issues involved in all this was not satisfactory. For instance it was not clear that the Dirichlet process was well defined as a probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$. Do we know that the set $\mathcal{P}_0 \stackrel{def}{=} \{P : P \text{ is a discrete probability measure}\}$ is in $\sigma(\mathcal{P})$ before asserting that it had probability 1 under a Dirichlet process? How general can the space $\mathcal{X}$ be? These and other such questions remained.

The paper of Basu and Tiwari (1982) is a delightful paper that clears up all these questions. The paper starts out by describing the nature of general Bayes inference. It describes carefully and in detail the properties of finite-dimensional Dirichlet distributions which form a conjugate family in the standard nonparametric problem concerning random variables taking values in $\mathcal{X}$ consisting of a finite number of points. A quick extension when $\mathcal{X}$ is countable is presented next.

The case where $\mathcal{X}$ is a Borel space is the main focus of this paper. (Separable complete metric spaces are examples of Borel spaces). Borrowing from the ideas in Blackwell (1973), Basu and Tiwari establish the existence and properties of a Dirichlet process in this general case, with care and in an elementary way.

The highlight of this paper consists of Sections 6, 7 and 8. A very clear exposition of several measurability issues and the existence of the Dirichlet process are presented in these sections. It is shown that the function $P \rightarrow P_d(\mathcal{X})$ which gives the total sum of the probability masses of the discrete part of P is a measurable function from $(\mathcal{P}, \sigma(\mathcal{P}))$ to $[0, 1]$. Again it is established that the collection, $\mathcal{P}_0$, of all discrete probability measures on $(\mathcal{X}, \mathcal{B})$ is a measurable set in $(\mathcal{P}, \sigma(\mathcal{P}))$. It is only after this has been established, it makes sense to say that Dirichlet processes gives probability 1 to this set and this fact is established next. It is further shown that the collection $\mathcal{P}'$ of all probability measures whose support is $\mathcal{X}$ is also a measurable set in $(\mathcal{P}, \sigma(\mathcal{P}))$. It is only after this has been established, it makes sense to say that $\mathcal{D}_\alpha$ has support $\mathcal{P}'$ if α has support $\mathcal{X}$ and this result is also part of the paper.

David Blackwell's visit to Florida State University in 1978–79 gave Basu, Tiwari and myself inspiration to work on Dirichlet processes. In the course of my lectures at that time I discovered a constructive definition of a Dirichlet process which does not place restrictions of the space $\mathcal{X}$. This result appears in Sethuraman (1994).

References

[1] Basu, D. and Tiwari, R. C. (1982) A note on Dirichlet processes *Statistics and Probability Essays in Honor of C. R. Rao Ed. Kallianpur, G, Krishniah, P. R. and Ghosh, J. K.* North Holland Publ. Co. 82–103.

[2] Berk, R. and Savage, I. R. (1979) Dirichlet processes produce discrete measures: An elementary proof *Contributions to Statistics. Jaroslav Hájek Memorial Vol.* Academia, North Holland, Prague 25–31.
[3] Blackwell, D. (1973) Discreteness of Ferguson selections *Ann. Statist.* **1** 356–358.
[4] Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Pölya schemes *Ann. Statist.* **1** 353–355.
[5] Ferguson, T. S. (1973) A Bayesian analysis if some nonparametric problems *Ann. Statist.* **1** 2009–230.
[6] Sethuraman, J. (1994) A constructive definition of Dirichlet priors (1994) *Statistica Sinica* **4** 639–650.

Commentary on D. Basu's Papers on Sufficiency and Related Topics

T.P. Speed

It is an honour and a pleasure to be able to offer this commentary on some of D. Basu's papers on sufficiency and related topics. In the early 1970s, we thought that we might write a book together on sufficiency. We had many discussions and exchanged a fair amount of material. In particular, we prepared a bibliography on sufficiency which was reasonably comprehensive at the time (Basu and Speed (1975)). But the planned book never came to fruition. I have not worked on this topic since the late 1970s. So the experience of writing this commentary has been a pleasant walk down the memory lane. However, it also means that I may be unaware of some relevant later developments. Accordingly, I begin with an apology in advance to the readers for any such oversight or errors. *Caveat lector!*

General background

I think it is worth setting the scene for Basu's work on sufficiency. A new refugee from the then East Pakistan, he began working towards a PhD in 1950, at the Indian Statistical Institute, under the direction of C. R. Rao. We can assume that during this period, he gained a thorough grounding in the theory and philosophy of statistical inference, in particular, on the work of Fisher. Basu's doctoral dissertation was on estimation and testing in a decision theoretic framework, and to a smaller extent on some characterisation problems for normal distributions. It was not very Fisherian in style, but more mathematical. Undoubtedly, it was influenced by the work of Neyman, Pearson, Wald, notably Rao, and perhaps others. After submitting his PhD thesis (to the Calcutta University) in 1953, he went to the University of California at Berkeley as a Fulbright scholar. It was a long trip by ship. By the end of his time there, if not even before, I believe that he would have thoroughly absorbed the modern version of the Neyman-Pearson-Wald approach to inference, being defined and taught by Erich Lehmann, Charles Stein, Henry Scheffé, Lucien Le Cam, and Jerzy Neyman himself. In fact, some major papers of these statisticians were published in *Sankhyā*, then edited by P. C. Mahalanobis.

Sufficiency background

Fisher introduced sufficiency in his famous 1922 paper (Fisher (1922)) on the mathematical foundations of theoretical statistics. Stigler (1973) is a good source for more background. In this paper, Fisher decreed that if θ is the parameter of concern, and a statistic T contains the whole of the

T.P. Speed (✉)
Department of Statistics, University of California, Berkeley, CA 94720
e-mail: terry@stat.berkeley.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_9,

information that the full sample supplies as to the value of θ, then for any other statistic T_2, and for any θ, the conditional distribution of T_2 given T should be the same under all θ. There and in later papers, Fisher presented several properties of sufficient statistics. In the 1930s, Neyman gave a rigorous proof of the factorisation criterion (Neyman (1935)), while Pitman, Koopman, Darmois, and Brown (Darmois (1935), Koopman (1936), Pitman (1936), Brown (1964)) independently discovered that under suitable regularity conditions, Exponential families were the only classes of distributions to possess nontrivial sufficient statistics at all sample sizes, and minimal sufficient statistics that have the same affine dimension as the parameter space to which θ belongs. Kolmogorov (1942) introduced *Bayes sufficiency*, though Basu was not to be seriously interested in it till the early 1970s. An important development on the practical implications of sufficiency that also came in the 1940s was the independent discovery by C. R. Rao and David Blackwell (Rao (1945), Blackwell (1947)) of the theorem that bears their names. Perhaps more significantly for Basu's work on sufficiency, a 1949 paper by P. R. Halmos and L. J. Savage (Halmos and Savage (1949)) elegantly placed sufficiency within the framework of measure theory, and replaced Fisher's parametric families by a more or less arbitrary family of probabilities, through the consideration of *sufficient σ-fields*. Halmos and Savage obtained their best results under the assumption of a *dominated family of probabilities*, that is, that each member of the family of probabilities possessed a density (*Radon-Nikodym derivative*) with respect to a common σ-finite measure. Their measure-theoretic approach was adopted in nearly all of Basu's papers on sufficiency.

The main themes in Basu's sufficiency papers

The Trinity. Kolmogorov introduced the famous triple $(\Omega, \mathcal{A}, P)$ (Kolmogorov (1933)). In his writings on sufficiency, Basu used the expanded triple $(\Omega, \mathcal{A}, \mathcal{P})$, where $\mathcal{P}$ is a general family of probability measures on $\mathcal{A}$. He used X or $\mathcal{X}$ instead of Ω, and called $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ *the trinity*. The precise nature of parametrization played little or no role in most of his research on sufficiency, the principal exceptions being his discussions of invariance and partial sufficiency, and some of his famous counterexamples.

Null sets. Much of our intuition in statistics is developed for the dominated case, where each of our probability measures in $\mathcal{P}$ has a density with respect to a common σ-finite dominating measure. With just a single probability measure P, we only need to take care with P-null sets. When we work with sufficiency, we need to pay attention to null sets more generally. With a family of probability measures $\mathcal{P}$, the relationships between the P-null sets for different members P of $\mathcal{P}$, and the sets which are P-null for all P in $\mathcal{P}$ (called $\mathcal{P}$-null sets), and the different completions of the underlying measure space all play a critical role. Much of Basu's work on sufficiency is marked by very careful treatment of considerations of null sets.

Completeness and other joint aspects of a statistic and the family of probabilities. In 1950, Lehmann and Scheffé published a landmark paper which highlighted the importance of the notions of *completeness* and *bounded completeness* of a family of distributions (Lehmann and Scheffé (1950)). Later, *weak completeness* came into the picture. Completeness was an essential ingredient of one of Basu's most well known contributions, namely *Basu's theorem*.

Ordering, maximality, and minimality of σ-fields. Another fundamental contribution in the 1950 paper of Lehmann and Scheffé was the introduction of the notion of *minimal sufficiency*. In the measure theoretic framework, this amounted to identification of a minimal sufficient σ-field. On several occasions, Basu studied the ordering of sub-σ-fields with various specific properties, and the questions of existence of maximal or minimal elements among them. As a rule, this was not a simple matter.

The papers (As in Author Bibliography)

Basu's Theorem (paper 15)

Basu's theorem (Basu (1955)) says that a boundedly complete sufficient statistic and any ancillary statistics are independently distributed under all θ. This is Theorem 2 in the paper. This was Basu's first paper on sufficiency, and arguably the most well known. Numerous applications of Basu's theorem, including many in probability theory, are detailed in the commentary of Anirban DasGupta and of Malay Ghosh in this volume; for earlier references on applications of Basu's theorem, see the beautiful exposition in Boos and Hughes-Oliver (1998), and also see DasGupta (2007). Theorem 1 in the paper was a converse, but not correct as stated. In a later paper, Basu gave a correct converse, which describes conditions under which a statistic which is independent of a sufficient statistic under all θ must be ancillary. This has been used in the literature on higher order asymptotics to establish approximate ancillarity of certain P-values; for example, see Lauritzen (2008). It is also worth pointing out that although we regard Basu's theorem purely as a result in statistical inference, it is also a tremendously effective tool in probabilistic calculations. Students of probability would be better equipped if they were trained in applying Basu's theorem to greatly simplify many distributional calculations.

Sufficiency and Finite Population Sampling (papers 29, 27, 26)

In some sense, we can see Basu at his best in his papers on finite population sampling. These papers have several goals, all of which he achieves neatly and eloquently. The first goal involves setting the statistical notion of sampling from a finite universe within the same mathematical framework of all other statistical models, by defining a suitable trinity $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Basu argues that it is natural to take $\mathcal{A}$ as the set of all subsets of the sample space $\mathcal{X}$, and $\mathcal{P}$ as an undominated family of discrete probabilities on $\mathcal{A}$. He then shows that a maximal sufficiency reduction is always at hand. These are also probably the papers in which he shows his Bayesian transition for the first time. One piece of evidence of this is his theorem that once the survey data from the finite universe has been obtained, inference should no longer depend on the sampling design that was actually used. Paper 29 (Basu (1970)) contains the now famous and colorful example of *Basu's elephants*. This example has led many statisticians of subsequent generations to think about the exact role and relevance of sample space based optimality criteria, such as admissibility. The example of Basu's elephants was the subject of an entire book on survey sampling (Brewer (2002)). Paper 27 also had the goal of showing that the counterexamples given by Pitcher (1957) and Burkholder (1961) concerning sufficiency in the undominated case need not discourage statisticians. Indeed, paper 27 shows what a difference an $\mathcal{A}$ makes.

Sufficiency and Invariance (papers 25, 18)

Calling upon invariance (under transformations preserving a statistical model) to select one from competing decision procedures originated in the late 1940s, though undoubtedly there were earlier instances. The approach was widely used in Erich Lehmann's classic text *Testing Statistical Hypotheses*, first published in 1959 (Lehmann (1959)). Invariance is a tool for data reduction, and so is sufficiency. Charles Stein (in some unpublished work), Burkholder (1960) (see Hall, Wijsman, and Ghosh (1965)), Hall, Wijsman, and Ghosh (1965), Berk and Bickel (1968), and Berk (1972) explored the relationship between sufficiency and invariance reductions of the sample data. These papers form

the background for Basu's research in this area, which includes his characteristic search for clear and simple proofs, compelling motivation, a desire to deal very carefully with null sets, and is also of very significant expository value. Let me summarize it briefly. Suppose that $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is a statistical trinity, and $\mathcal{G}$ a group of one-to-one bimeasurable transformations of $(\mathcal{X}, \mathcal{A})$ onto itself which are measure-preserving, i.e., $Pg^{-1} = P$ for all $P \in \mathcal{P}$ and all $g \in \mathcal{G}$. It is not hard to prove that for any $g \in \mathcal{G}$, the sub-σ-field $\mathcal{A}_g = \{A \in \mathcal{A} : g^{-1}A = A\}$ is sufficient for the triple $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ and so interest naturally turns to $\mathcal{A}(\mathcal{G}) = \cap_g \mathcal{A}_g$. The sub-$\sigma$-field $\mathcal{A}(\mathcal{G})$ is the σ-field of *$\mathcal{G}$-invariant sets*. Closely related are the sub-σ-fields of *essentially* and *almost $\mathcal{G}$-invariant* sets (see Basu (1970) for the exact definitions). With these preliminaries, Basu explores conditions under which a minimal sufficient sub-field $\mathcal{T}$ is contained in or coincides with the sub-field of almost $\mathcal{G}$-invariant sets. In the second half of the paper, Basu turns to parameter-preserving transformations, foreshadowing his work on partial sufficiency. But his focus here is firmly on normal models.

In paper 25 (Basu and Ghosh (1969)), Basu and Ghosh introduced the concept of *nontrivial weak completeness*. Nontrivial weak completeness means that there are no sets A, not $\mathcal{P}$- equivalent to the empty set or the entire space $\mathcal{X}$, such that $P(A)$ is constant in $P \in \mathcal{P}$. The principal aim of this paper was to explore families $\mathcal{P}$ which are not weakly complete. This was almost entirely restricted to translation parameter families on the real line, circle, or some other compact or locally compact group. Several interesting connections with theorems from harmonic analysis, and as was customary with Basu, a number of interesting examples were described. But no neat general results really came into light.

Partial Sufficiency (paper 39)

When we read Basu's work, it appears that Basu embraced the Bayesian approach to statistical inference because of the failure of the other approaches to deal adequately with inference concerning what he termed as sub-parameters, that is, functions of the global parameter. I think he found sufficiency compelling when inference concerning the entire parameter was the goal, despite some of the problems and paradoxes involving ancillary statistics. In this case, reduction to the likelihood function is the maximal possible reduction, which he probably found appropriate. However, when he turned to ways of carrying out inference for sub-parameters of interest, eliminating nuisance parameters not of interest, and the procession of forms of partial sufficiency, he did not find any solution that stood up to his creative scrutiny (Basu (1978)). Although there has been a lot more work on this topic since Basu's 1978 paper, I don't think that any general satisfactory solution to the problem has emerged. Today, non-Bayesians deal with nuisance parameters on a case-by-case basis, at times aided by special results, such as Barndorff-Nielsen's formula (Barndorff-Nielsen (1983)), or special tools, such as conditional, partial or profile likelihoods (Cox (1975), Severini (1994)). Bayesians would often integrate out all the nuisance parameters, perhaps with some type of a *noninformative prior* (Bernardo and Smith (1994)). Satisfactory general approaches concerning sub-parameters and elimination of nuisance parameters seem as far away today as they did when Basu wrote his probing article in 1978.

Sufficiency and Coherence (paper 46)

In the late 1960s and 1970s, several authors sought to broaden the domain of the nice results due to Halmos and Savage (1949) concerning general, pairwise, and minimal sufficiency, which they proved undeer the assumption that the family of probabilities was a dominated family. Pitcher (1965) defined the notion of *compactness* of a family of probability measures, Mussman (1972) introduced *weak domination*, Hasegawa and Perlman (1974) gave us *coherence*, while Le Cam (1964) explored related

ideas within the framework of vector lattices. Siebert (1979) connects the first three, showing that they are essentially equivalent. While Siebert's publication predates Basu and Cheng (1981), it came after S. C. Cheng's 1978 Florida State PhD thesis written under Basu's direction, from which paper 46 most likely derived. This paper had an expository flavor. But it also gives a useful addendum to the converse part of Basu's theorem. Basu's original converse (Basu (1958)) was in terms of the family of probabilities being *connected*. Koehn and Thomas (1975) strengthened this theorem of Basu to lay down a necessary and sufficient condition for the converse to Basu's theorem to hold. This result of Koehn and Thomas said that non-ancillary statistics independent of a sufficient statistic (under all θ) exist if and only if the family of probabilities admit a *splitting set*. In paper 46, Basu and Cheng show that under the condition of coherence, this necessary and sufficient condition of Koehn and Thomas is exactly the same as Basu's original connectedness condition. Thus, under coherence, the two theorems of Basu precisely characterize the relationship between ancillarity and sufficiency through their independence, a very clean conclusion.

References

Barndorff-Nielsen, O. E. (1983). On a formula for the conditional distribution of the maximum likelihood estimator, Biometrika, 70, 343–365.

Basu, D. (1955). On statistics independent of a complete sufficient statistic, Sankhy$\bar{a}$, 15, 4, 377–380.

Basu, D. (1958). On statistics independent of a sufficient statistic, Sankhy$\bar{a}$, 20, 223–226.

Basu, D. and Ghosh, J. K. (1969). Invariant sets for translation parameter families of measures, Ann. Math. Statist., 40, 162–174.

Basu, D. (1970). An essay on the logical foundations of survey sampling, Foundations of Statistical Inference, 203-242, Holt, Rinehart and Winston, Toronto, Canada.

Basu, D. (1970). On sufficiency and invariance, Essays in Probability and Statistics, 61–84, UNC Press, Chapel Hill.

Basu, D. and Speed, T. P. (1975). Bibliography of sufficiency, Mimeographed Technical Report, Manchester.

Basu, D. (1978). On partial sufficiency: A review, Jour. Statist. Planning Inf., 1, 1–13.

Basu, D. and Cheng, S. C. (1981). A note on sufficiency in coherent models, Internat. J. Math. Sci., 3, 571–582.

Berk, R. and Bickel, P. J. (1968). On invariance and almost invariance, Ann. Math. Statist., 39, 1573–1576.

Berk, R. (1972). A note on sufficiency and invariance, Ann. Math. Statist., 43, 647–650.

Bernardo, J. M. and Smith, A. F. M. (1994). Bayesian Theory, Wiley, New York.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation, Ann. Math. Statist., 18, 105–110.

Boos, D. D. and Hughes-Oliver, J. M. (1998). Applications of Basu's theorem, Amer. Statist., 52, 218–221.

Brewer, K. (2002). Combined Survey Sampling Inference: Weighing of Basu's Elephants, Arnold, London,

Brown, L. D. (1964). Sufficient statistics in the case of independent random variables, Ann. Math. Statist., 35, 1456–1475.

Burkholder, D. (1960). The relation between sufficiency and invariance, I: theory. Invited address at the Central Regional Meeting of the Institute of Mathematical Statistics, Lafayette, Indiana.

Burkholder, D. (1961). Sufficiency in the undominated case, Ann. Math. Statist., 32, 1191–1200.

Cox, D. R. (1975). Partial likelihood, Biometrika, 62, 269–276.

Darmois, G. (1935). Sur les lois de probabilités a estimation exhaustive, C. R. Acad. Sc. Paris, 200, 1265–1266.

DasGupta, A. (2007). Extensions to Basu's theorem, factorizations, and infinite divisibility, Jour. Statist. Planning Inf., 945–952.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, Phil. Trans. Royal Soc. (London), A222, 309–368.

Hall, W. J., Wijsman, R. J., and Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis, Ann. Math. Statist., 36, 575–614.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics, Ann. Math. Statist., 20, 225–241.

Hasegawa, M. and Perlman, M. D. (1974). On the existence of a minimal sufficient sub-field, Ann. Statist., 2, 1049–1055.

Koehn, U. and Thomas, D. L. (1975). On statistics independent of a sufficient statistic: Basu's lemma, Amer. Statist., 29, 1, 40–42.

Kolmogorov, A. N. (1933). Foundations of the Theory of Probability Springer, Berlin.

Kolmogorov, A. N. (1942). Definition of center of dispersion and measure of accuracy from a finite number of observations, Izv. Akad. Nauk Er. Mat., 6, 3–32. (In Russian)
Koopman, B. O. (1936). On distributions admitting a sufficient statistic, Trans. Amer. Math. Soc., 39, 399–409.
Lauritzen, S. (2008). Ancillarity and Conditional Inference, Lecture Notes, Oxford University, UK.
Le Cam, L. (1964). Sufficiency and approximate sufficiency, Ann. Math. Statist., 35, 1419–1455.
Lehmann, E. L. and Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation, Sankhy$\bar{a}$, 10, 305–340.
Lehmann, E. L. (1959). Testing Statistical Hypotheses, Wiley, New York.
Mussman, O. (1972). Vergleich von Experimenten im schwach dominierten Fall, Z. Wahrsch. Verw. Gebiete, 24, 295–308.
Neyman, J. (1935). Su un teorema concernente le cosiddette statistiche sufficienti, Inst. Ital. Atti Giorn, 6, 320–334.
Pitcher, T. S. (1957). Sets of measures not admitting necessary and sufficient statistics or subfields, Ann. Math. Statist., 28, 267–268.
Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy, Proc. Camb. Philos. Soc., 32, 567–579.
Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc., 37, 81–91.
Severini, T. A. (1994). On the approximate elimination of nuisance parameters by conditioning, Biometrika, 81, 649–661.
Siebert, E. (1979). Statistical experiments and their conical measures, Z. Wahrsch. Verw. Gebiete, 45, 247–258.
Stigler, S. (1973). Studies in the history of probability and statistics, Laplace, Fisher, and the discovery of the concept of sufficiency, Biometrika, 60, 439–445.

Basu on Randomization Tests

A.H. Welsh

Basu's paper on randomization tests (Basu, 1980) is a critique of the use of randomization as a basis for inference (i.e. as the source of the variability underlying the application of statistical arguments). The paper focusses on analyzing data obtained from an experiment, making it a direct companion piece to Basu (1978) and Basu (1971) which consider very similar issues in analyzing data obtained from sample surveys. It fits comfortably with Basu's work on sample surveys and more generally on statistical inference, enriching and being enriched by the whole body of work. However, part of the attractiveness and strength of the present paper is that it can also be read alone, without reference to Basu's other work, as a relatively accessible, stimulating illustration of Basu's approach to thinking about statistical inference. It is a classic Basu paper highlighting the hallmarks of his style: it is provocative and challenging, based on simple examples pushed to extremes, and illustrated in an entertaining way by a conversation between three people. Underlying all this of course is deep thinking on serious issues. And, as an additional benefit, the discussion and Basu's rejoinder are insightful and interesting, adding much to the original paper.

The title of the paper puts the focus on randomization tests but in fact Basu discussed both permutation tests (Section 4) and randomization tests (Section 6), treating them both as randomization tests. Permutation tests are operationally similar to randomization tests but different from them because the justification for the test comes from an assumed model rather than from a physical act of randomization. They fit therefore into the standard model-based framework for inference whereas randomization tests fit into the design-based framework. This point was partly acknowledged by Basu at the end of Section 5 and then made strongly by Hinkley, Kempthorne and Rubin in their discussions. In his rejoinder, Basu justified his inclusion of the permutation test by pointing out similarities between it and the randomization test. I think that he was a bit too quick to dismiss the differences but this potentially distracting issue is reduced if we interpret the paper from a more general perspective than the title suggested. As some aspects of the critique apply quite generally to significance tests, the paper can be interpreted usefully as a critique of significance tests which is developed by exploring, as Basu liked to do, particular instances of significance tests.

Basu identified the components of a significance test as a test statistic (which Basu called a test criterion) and a sample space or reference set for determining the tail area probability under the null hypothesis. As we are reminded in the discussion of unequal probability randomization, there is also the distribution under the null hypothesis of the test criterion over the reference set. Basu's critique of the significance testing paradigm is based on the fact that both the choice of the test criterion and the reference set are to some extent arbitrary but important to the outcome. As is pointed out in the

A.H. Welsh (✉)
Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia
e-mail: alan.welsh@anu.edu.au

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_10,

discussions, this emphasizes that significance is not a property of data alone (as is apparently implied by Basu when he uses the phrase "the significance level of the data") but also depends on the test criterion and the reference set. It is always useful to be reminded of these kinds of subtleties.

The fact that in significance testing the choice of the test criterion is arbitrary and different choices lead to different interpretations of the same data is not new to this paper but has been known for a long time. Indeed, such considerations led to the Neyman-Pearson approach of considering alternative hypotheses and using the power of a test to help with the choice of test criterion. Basu's contribution here is to use the simple framework to illustrate the issue very simply and directly by comparing tests based on the mean and median. Even though the conclusion is unsurprising, the illustration is very nice.

For the permutation test, the reference set is determined by what we choose to condition on and Basu pointed out that different choices with different consequences are possible. Typically Basu pushed this to the extreme by constructing a two point reference set which severely limits the significance level. This is an interesting point but even Basu described this choice as "too ridiculous to deserve any serious consideration" and indicated that it is a choice we could choose to avoid. In the case of the randomization test, the physical act of randomization determines the reference set so the choice is made a step earlier by how we choose to do the physical randomization. Basu discussed the impact on the test of extreme choices of randomization showing that randomization inference can be unhelpful (Basu said "founders on the rocks") when we have restricted and/or unequal probability randomization. This point was also made strongly in Basu (1971) and is the main point of this paper: The issues arising in making randomization inferences from data from experiments are the same as those in making design-based inferences from data from sample surveys.

A point which Basu liked to emphasise when criticizing inference which is not fully conditional on the observed data (which means non-Bayesian inference) is the effect on the analysis of changing the information available to the analyst. In this paper, Basu used am imagined three-way conversation between himself (author), a scientist and statistician to illustrate (amongst other things) the effect of changing the nature of the original randomization. This is not as memorable as Basu's famous elephant example (Basu, 1971) but it is written in the same tongue-in-cheek, provocative style with serious intent. No doubt, many statisticians would feel that they can avoid confronting the questions raised by the unwelcome disclosure of additional information but, at least at the level of thinking about statistical inference, we ought to think about the fact that procedures which require less than full conditioning must be changed as more information becomes available to ensure that we keep on using all the information.

Perhaps the deepest and most challenging part of the paper is Basu's discussion of the applicability of the sufficiency principle to the permutation test and the conditionality principle to the randomization test. It is instructive, though perhaps disappointing, that the sufficiency principle does not rule out any of the three (permutation) test statistics Basu considered. In this sense, all three test statistics use all the information in the data. What I find interesting is that this shows how what we mean by "all the information" depends on what we assume. This means that we can adjust the meaning of "all the information", making it an arbitrary concept. In the final section of concluding remarks, Basu pointed out that the outcome of the randomization is an ancillary statistic and then argued that Fisher's conditionality principle means that it should be held fixed in the analysis. That is, randomization should not be used as the basis of inference. This is a real challenge which is difficult to refute without refuting conditional inference in its entirety. Hinkley recognised the strength of the point and put an alternative view in his discussion.

Basu's approach in this paper is to examine randomization and permutation tests in simple cases pushed to extremes to highlight issues with the tests. This is valuable and important but it allows for the temptation to try to minimize the consequences by conceding the points and arguing that the message is that one should try to avoid getting into such extreme cases. If extreme cases are seen as giving one kind of examination, the other side is to ask whether the tests do what they are intended to

do in ideal situations? Basu's discussion of ancillarity is an important contribution to this side of the debate; the issue has equal force in ideal situations (the sample size is large and the the test is based on equiprobable, unrestricted randomization) and in extreme situations and it is not easy to argue around.

There are other points on the "ideal" side of the examination which are not touched on by either Basu or the discussants. For example, in the quotation from Fisher (1960) included by Basu in Section 5, Fisher says that "The utility of such nonparametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly, it is suspected that the simpler tests have been appreciably injured by departures from normality." This was taken up by Hinkley who argued that the point of randomization is to justify the normal-theory analysis. It is worth examining whether it in fact does so. The difficulty is that the randomization (and the permutation analysis) justify every normal theory analysis and we are led into the situation where the fact that everything goes means that nothing goes. Basu might have put it something like this:

Statistician: As I routinely do, after receiving the data, I subjected the data to a standard normal theory analysis. This is justified by the physical act of randomization I asked you to carry out when you designed the experiment. The results show that the treatment has no effect.

Scientist: Thank you for doing that. I should mention that, when collecting the data, I noticed that I recorded an extremely large observation that may be an outlier but then forgot about it. Does this affect your analysis?

Statistician: No, randomization justifies the standard normal theory analysis even when the data are not normally distributed and even if there are outliers in the data. This is one of the great advantages of nonparametric methods.

Author: What if the outlier is generated by a different process from the one that is of interest in the experiment? If I remove the outlier and then apply a standard normal theory analysis, I find that there is a significant treatment effect. Is this conclusion also justified by the randomization?

Scientist: One more thing, I didn't mention that, as is usual in the literature for this kind of data, I gave you the logarithm of the original variables. Does this matter?

Statistician: I would have done the standard normal theory analysis on the original data. Just as a check, I exponentiated the data you provided and redid the analysis; I found no treatment effect.

Author: If all these analyses are equally justified by the randomization, which one should we adopt?

Scientist (utterly flabbergasted): What am I supposed to do?

The justification provided for the normal theory analysis holds for any set of observed data from the experiment so it holds regardless of the scale on which the data are, regardless of the fact that the data may seem to come from a long-tailed distribution and regardless of whether there are extreme outliers in the data or not. This is a kind of extreme robustness: normal-theory analysis is always justified so we do not even need to consider non-normal models. On the other hand, in practice, normal theory analysis is not always justified so the blanket justification ends up undermining itself. Put in a different way, randomization and permutation arguments can justify basing inference on an automatic numerical computation without examining the data at all. This has happened with the design-based analysis of surveys (but arguably much less in the analysis of randomized experiments) and is one reason surveys have become separated from the rest of statistics, something which Basu decried. It has also done robustness no favors and may in part explain the deep resistance to the ideas of robustness in some regions and some areas of statistics.

If a test is valid regardless of the observed values of the data, how should we think about robustness? Robustness theory provides a partial resolution and additional insight by making the distinction between robustness of validity (able to preserve the level) and robustness of efficiency (able to preserve the power). However, it seems difficult to do power calculations for randomization tests (more difficult than for permutation tests), making it difficult to pursue the issues from entirely within a randomization framework. It is interesting that this brings us back to the general issue of choice of test criterion and shows that it is a difficult problem for randomization tests over and above the general difficulty of the problem within the significance testing paradigm.

The value of Basu's work is not ultimately in whether he is right or not, or whether we agree with him or not, but rather in that it confronts us and makes us think deeply about how we analyze data and make statistical inferences. It is not comfortable reading Basu, but it is enriching and there is benefit in rereading the work from time to time to reassess our understanding. This is true in general and very specifically true of his paper on randomization tests.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling, part I. In *Foundations of Statistical Inference*, eds V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, 203–243.

Basu, D. (1978). On the relevance of randomization in data analysis (with discussion). In *Survey Sampling and Measurement*, ed N.K. Namboodiri, New York: Academic Press, 267–339.

Basu, D. (1980). Randomization analysis of experimental data: The Fisher randomization test (with discussion). *J. Amer. Statist. Assoc.* **75**, 575–595.

Fisher, R.A. (1960). *Statistical Methods for Research Workers*, Seventh edition. Edinburgh: Oliver and Boyd.

Basu on Survey Sampling

A.H. Welsh

"A circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants . . .", so begins Example 3 in Basu (1971), the most colorful and striking illustration of Basu's challenges to the design-based analysis of sample survey data. The full story is included in the box for easy reference. The owner decides to take a sample of size $n = 1$ ("As weighing an elephant is a cumbersome process") and is talked out of a non-random sample (select Sambo, the elephant that had the average weight 3 years before) and the model-based estimate $(50y)$ into an unequal probability sample (select Sambo with probability $99/100$ and any of the other elephants with probability $1/4900$) and the Horvitz-Thompson estimator ($100y/99$ if Sambo is selected and $4900y$ if any other elephant is selected). The point of the story is summarised in Figure 1 which shows the log-sampling distributions (i.e. the sampling distributions of the log of the estimators) for samples of size 1 of the model-based estimator and the Horvitz-Thompson estimator for a troupe of 50 elephants. (We plot the log-sampling distributions to improve the visual impact.) On this scale, the model-based estimator is very close to the actual total weight (indicated by an arrow) but, and this is Basu's elegantly made point, the design-unbiased Horvitz-Thompson is far from the actual total weight in every possible sample. The design-based optimality of the Horvitz-Thompson estimator is no consolation to either the circus owner or the "unhappy statistician" who, Basu tells us, "lost his circus job (and perhaps became a teacher of statistics!)".

The elephant story provokes and challenges, delights and frustrates, and ultimately encourages deep thinking on serious issues. Basu argued that the analysis of survey data should be subject to the same general principles as the analysis of other forms of data, and that there should be no special pleading for survey analysis to be treated differently from other statistical analyses. It is not surprising therefore that Basu's elephants illustrate specific points about survey analysis as well as general points about statistical analysis. In the survey context, Basu's elephants illustrate specific difficulties with unbiased estimation, unequal probability samples and design-based analysis. The elephant's bring these together with striking effect but they can also be teased apart and considered separately. One response to the example (Hajek in the discussion) is to suggest a different estimator for θ, the total weight of the elephants: since we know the weights of the elephants from the last time they were weighed, we should use a ratio estimator rather than the Horvitz-Thompson estimator. It is a nice irony that the ratio estimator is slightly design-biased! For another suggestion, see Rao in the discussion of Basu (1978). A different response is to suggest that we use a different design, perhaps with less variable weights, the extreme choice being equal probability sampling with all the weights equal. The

A.H. Welsh (✉)
Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia
e-mail: alan.welsh@anu.edu.au

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_11,

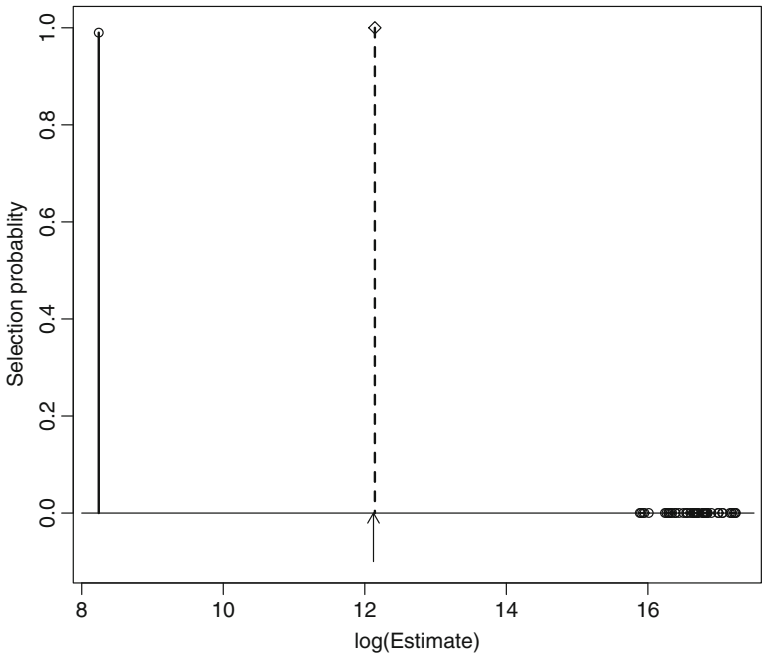

Fig. 1 The design-based log-sampling distributions (i.e. the x-axis is on the log scale) of the model-based and Horvitz-Thompson estimators for a sample of size 1 from a troupe of 50 elephants. The model-based estimator has a degenerate distribution represented by the diamond symbol and the dashed segment. The Horvitz-Thompson estimator has a distribution represented by the circles and solid segments. The true total weight of the troupe is shown by the arrow below the $y = 0$ line

elephants show that if we use unequal probability sampling and the weights do not depend on θ but simply reflect our desire or otherwise to include each unit in the sample, giving the most weight to the observations we want to include least in the sample may not be sensible. Actually, as pointed out in the discussion by Hajek, Godambe and Koop, the sampling design may incorporate prior information about the population and hence depend on θ, but a relationship like this is very difficult to formalize mathematically and so difficult to exploit. Whether for this or for some other reason, Basu did not see value in unequal probability sampling, even though, in simple examples, he did explore some purposive designs for which the selection probabilities are highly unequal. Basu's preferred response, and the motivation for the example, is for us to do a different kind of (non-design-based) analysis which does not depend on the sampling design.

Basu's critique is much broader than unbiased estimation and unequal probability sampling: The fundamental point in his (later) sample survey papers is that the design-based approach contravenes the likelihood principle and hence should not be used for the analysis of survey data. One could argue this from the point of view that there is no likelihood in the design-based framework, although this would open the possible rejoinder that the likelihood principle is then not relevant. Instead, Basu argued that there is a likelihood, the function that equals the probability of selecting the given sample on Ω_x, "the set of parameter points that are consistent with a given sample", and zero otherwise (Basu, 1969). If the sampling design does not depend on the target parameter θ, the design is ancillary and the likelihood is constant on Ω_x. If the ith elephant weighs Y_i, then $\theta = \sum_{i=1}^{N} Y_i$ and, if we sample a single elephant weighing y, the likelihood is constant on the set $\Omega_x = \{\theta \geq y\}$. It is interesting that this likelihood is derived from the sampling design and seems to require a probability sample: If the sample is purposive, there is nothing stochastic in the setup so, although we can simply define the likelihood to be constant on Ω_x and zero elsewhere, this function is not the joint density of the sample

viewed as function of the unknown parameter. Basu felt that we should implement the likelihood principle by doing a Bayesian analysis so he would have specified a prior on θ, thereby introducing a stochastic element, but this still leaves open the question of how to interpret the likelihood which is defined without reference to the prior.

The model-based approach provides another way of introducing probability into survey analysis (by treating $\omega = (Y_1, \ldots, Y_n)$ as a random vector) and hence of obtaining a likelihood (from the distribution of ω). How does this relate to what Basu (1978) called his "neo-Bayesian thesis on sample surveys"? Royall raised the model-based approach in his discussion of Basu (1971) but Basu, although he may have intended to, did not really engage with it. He suggested in response to Royall that superpopulation models are exactly like a Bayesian formulation of the background knowledge. He argued in his (1978) response to the discussion that superpopulation models are not objective and do not even exist, except in the mind, something he presumably also felt of his prior distributions. Since a prior for θ implies a distribution on ω from which the prior can be derived, Basu's neo-Bayesian and the model-based analysis should be able to be put into close numerical agreement by making compatible choices of prior and superpopulation model. These choices are potentially checkable from a census, at least in some cases and at least to the same extent as ordinary statistical models. To this extent at least, they do have an objective existence. It is interesting to explore these issues in more detail in the context of a different Basu example.

In Example 1 from Basu (1978), the population consists of N units, each of which is either defective ($Y_i = 1$) or non-defective ($Y_i = 0$). The units are produced by a mechanical device in such a way that, after the first defective unit, all the rest of the units are defective. The problem is to estimate the number of defective units $\theta = \sum_{i=1}^{N} Y_i$ from the values Y_i observed on a sample s of units. Let v be the largest $i \in s$ such that $Y_i = 0$; if $Y_i = 1$ for all $i \in s$, set $v = 0$. Let w be the smallest $i \in s$ such that $Y_i = 1$; if $Y_i = 0$ for all $i \in s$, set $w = N + 1$. Basu pointed out that, with probability one, $\theta \in \Omega_x = [N - w + 1, N - v]$ and this implies that some samples are much more informative than others: the best sample has $w = v + 1$ because then we know θ exactly. Although it is not our primary concern here, the design-based analysis of the example is interesting too because it explicitly permits us, if we so choose, to ignore the structure of the population. For example, if we select the sample by simple random sampling without replacement, we can use the expansion estimator $\hat{\theta}_E = (N/n) \sum_{i \in s} Y_i$ to produce an optimal, design-unbiased estimator of θ which is often outside the interval Ω_x. Of course, we expect to do better by incorporating the structure of the population into the sampling design. For example, we can select the first unit at random; at the $(k + 1)$th step, select the next unit at random in $[v^{(k)} + 1, w^{(k)} - 1]$, where $v^{(k)}$ and $w^{(k)}$ are the values of v and v from the first k observations; and continue until $w^{(k)} = v^{(k)} + 1$ or $k = n$. This is a stochastic version of the purposive design discussed by Basu which might be used in a design-based analysis, provided the design-based analyst can work out the sample inclusion probabilities needed to construct an estimator of θ. However, Basu would still have criticised this analysis as being in conflict with the likelihood principle.

Basu did not present his own analysis for this example, but we can construct an analysis he might have agreed to. For a sampling design which does not depend on θ, the likelihood is constant on the interval $[N - w + 1, N - v]$ so, if the prior density of θ is $q(\theta)$, the posterior density is

$$q(\theta|x) = \frac{q(\theta)}{\sum_{t=N-w+1}^{N-v} q(t)}, \qquad \theta = N - w + 1, N - w + 2, \ldots, N - v,$$

(Basu, 1969). When we are interested in a point estimate of θ, we use the posterior mean

$$\hat{\theta}_q = \frac{\sum_{t=N-w+1}^{N-v} t q(t)}{\sum_{t=N-w+1}^{N-v} q(t)}.$$

In the model-based approach, we model the distribution of ω. For this particular population, it is completely equivalent to model θ or the label of the first defective unit $M = N - \theta + 1$. The optimal mean squared error predictor of θ is given by $\hat{\theta}_p = N + 1 - E(M|s) = N - w + 1 + E(w - M|s)$, where the second expression is written in the familiar form of a sample contribution plus a non-sample contribution. Now we know that $Y_i = 0$ for $i \leq v$ and $Y_i = 1$ for $i \geq w$ so the sample information is that $m \in [v + 1, w]$. It follows that

$$P(M = m|s) = P(M = m|v + 1 \leq M \leq w) = \frac{P(M = m)}{\sum_{k=v+1}^{w} P(M = k)} \quad m = v + 1, v + 2, \ldots, w,$$

from which we can compute $E(M|s)$ and hence

$$\hat{\theta}_p = N + 1 - \frac{\sum_{k=v+1}^{w} kP(M = k)}{\sum_{k=v+1}^{w} P(M = k)}.$$

Algebraically, $\hat{\theta}_p$ equals the posterior mean $\hat{\theta}_q$ when $q(t) = P(M = N + 1 - t)$ (i.e. q is the distribution of $\theta = N + 1 - M$), supporting Basu's (1971) response to Royall. Adopting a prior for θ is implicitly adopting a distribution for ω and, at least numerically, the consequences can be made to match. For example, if we model M as having a uniform distribution on $\{0, 1, \ldots, N\}$, then

$$\hat{\theta}_p = N - w + 1 + (w - v)/2$$

and this equals the posterior mean $\hat{\theta}_q$ when the prior for θ is uniform on $\{0, 1, \ldots, N\}$. If instead we model M as having a geometric distribution with parameter π, then we can show that

$$\hat{\theta}_p = N - w + 1 + \left[w - \frac{1 + v\pi - (1 - \pi)^{w-v}(1 + w\pi)}{\pi\{1 - (1 - \pi)^{w-v}\}} \right].$$

and this equals the posterior mean $\hat{\theta}_q$ when the prior for θ is the distribution of $\theta = N + 1 - M$ and M has a geometric distribution with parameter π. Importantly, algebraic equality does not mean that $\hat{\theta}_q$ and $\hat{\theta}_p$ have the same content and meaning. From Basu's Bayesian perspective, π is a hyperparameter which we are free to specify: as $\pi \to 0$, $\hat{\theta}_q \to N - w + 1 + (w - v - 1)/2$ and as $\pi \to 1$, $\hat{\theta}_q \to N - v$ so that $\hat{\theta}_q \in \Omega_x \backslash [N - w + 1, N - w + 1 + (w - v - 1)/2]$. The data tells us that $\theta \in \Omega_x$ and the prior selects a particular point in Ω_x to resolve the arbitrariness of where, but the choice is entirely driven by us through the prior rather than the data. (If the prior has no support on Ω_x, the sample represents an event with prior probability zero and there is no usable posterior distribution. Thus, even if we have strong beliefs about the value of θ, the prior should still put some probability on every possible value of θ.) From the model-based perspective, π is an unknown parameter which we need to estimate. An often attractive way to do this is to use the maximum (model-based) likelihood estimator. The model-based likelihood is obtained from the population density of M (in this case, the geometric distribution) by treating ω as the complete data, the sample as incomplete data with the non-sample data missing (in a way determined by the design) and summing the complete data likelihood over the unobserved data. If we assume the sampling design is uninformative (i.e. sample selection does not depend on ω), then we can proceed straightforwardly. The model-based likelihood is not the same as the design-based likelihood used by Basu. Nonetheless it is also a likelihood, so we can impose a prior distribution on π and then do a Bayesian analysis which can be viewed as a hierarchical version of Basu's analysis with a hyperprior on π. Alternatively, we can substitute $\hat{\pi}$ into $\hat{\theta}_q$ as we do in the model-based analysis and view the result as an empirical Bayes predictor. Basu only considered enumerative inference about finite-population parameters like θ and had no interest

in analytic inference about hyperparameters like π; the model-based approach allows us to make analytic inferences about parameters like π as well as enumerative inference about finite-population parameters like θ. If we were to pursue the analysis beyond point estimation to inference, Basu would have objected to our using the sampling distribution in the model-based analysis and insisted on using his posterior distribution.

It is a challenge to describe how Basu would have analysed specific examples because he wrote more about what he would not do than what he would do and, when asked specifically, declined to provide more than general comments. In his very brief response to the discussion to Basu (1978), he wrote that analysing survey (probably meaning any) data is more an art than a science and he could say no more than that the analysis should be Bayesian (in the sense of fixing the sample and speculating about the parameters). Basu (1978) was clear that we need to know how the data were collected in order to analyse them - but, other than explicitly rejecting the design-based approach to doing this, he did not explain how to incorporate the data collection process into the analysis. It is natural for a Bayesian to include it in the prior specification, although this may be very difficult to achieve, particularly with purposive sampling. One possible role for probability sampling then is to simplify the way the data were collected and hence the prior specification. Basu (1969, 1978) also argued that except in simple populations, purposive sampling is too hard to justify (although, as Rao pointed out in the discussion to Basu (1978), this is not the case with Royall's purposive designs) and probability sampling can help a statistician defend his or her integrity. Basu's views on the role of randomisation are close to those of Royall (1976) and Rubin (1978).

Basu's papers on survey sampling should be read by everyone with an interest in survey sampling, indeed in statistics. The discussion papers are the most stimulating: they can be read starting with Basu (1971), referring to Basu (1969) for technical support, and then Basu (1978) or the other way round. The discussions and responses from Basu enhance the papers, stimulating much further thought. The paper by Basu and Ghosh (1967) on sufficiency is written in a very different, much more technical style. Basu (1958) is a traditional design-based paper, written before Basu became a Bayesian. It does not challenge the basic design-based framework in the same way as the later papers but, and this is characteristic of Basu and one of the reason his papers are still so valuable, it does challenge the usual method of analysing samples collected with replacement. With hindsight, it is tempting to see hints in Basu (1958) of what was to come, but it is a long way from there to the elephants, the circus and the "unhappy statistician". Statistics has benefitted enormously from the fact that Basu made that journey, questioning each step of the way.

References

Basu, D. (1958). Sampling with and without replacements. *Sankhya* **20**, 287–294.

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhya* **31**, 441–454.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part I (with discussion). In *Foundations of Statistical Inference*, eds V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, 203–243.

Basu, D. (1978). On the relevance of randomization in data analysis (with discussion). In *Survey Sampling and Measurement*, ed N.K. Namboodiri, New York: Academic Press, 267–339.

Basu, D. and Ghosh, J.K. (1967). Sufficient statistics in sampling from a finite universe. *Bull. Int. Statist. Inst.* **42**, 850–859.

Royall, R.M. (1976). Current advances in sampling theory: Implications for human observational studies (with discussion). *Amer. J. Epidem.* **104** 463–477.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomisation. *Ann. Statist.*, **6** 34–58.

ON SYMMETRIC ESTIMATORS IN POINT ESTIMATION WITH CONVEX WEIGHT FUNCTIONS

By D. BASU

(Research Fellow of the National Institute of Sciences)
Statistical Laboratory, Calcutta

1. Introduction

Let $X_1, X_2, \ldots X_n$ be a set of chance variables whose joint cumulative distribution function $F(x_1, x_2, \ldots x_n)$ is known to belong to some sub-space Ω of the space of all possible distribution functions F. As for instance it may be known that the X's are independently and indentically distributed so that Ω is the sub-space of all d.f.'s of the form

$$F = G(x_1)G(x_2)\ldots G(x_n), \qquad \ldots \quad (1.1)$$

where $G(x)$ is some one dimensional distribution function.

In point-estimation the problem is to estimate some population characteristic $\theta = \mu(F)$, where $\mu(F)$ is a real valued functional defined for all $F \epsilon \Omega$, with the help of an estimator $t = t(x_1, x_2, \ldots x_n)$ where x_i is a random observation on the chance variable X_i.

Let $W(t, F)$, for any fixed $R \epsilon \Omega$, denote the different weights that the statistician attaches to the different values of t as estimates of $\mu(F)$ and let

$$r(F \mid t) = \int_R W(t,F)dF, \qquad \ldots \quad (1.2)$$

where R is the n-dimensional sample-space, be the risk function associated with the estimator t. We assume that there exist estimators t for which the integral (1.2) is convergent for all $F \epsilon \Omega$.

If $r(F\, t_1) \leqslant r(F/t_2)$ for all $F \epsilon \Omega$ with the sign of inequality holding for at least one F then t_1 is said to be uniformly more powerful than t_2.

The estimator t_0 will be called admissible if there exists no estimator t uniformly more powerful than t_0,

In this paper we restrict ourselves to only such weight functions as are convex (downwards) functions of t for every $F \epsilon \Omega$. That is

$$W\left(\frac{t_1+t_2}{2}, F\right) \leqslant \tfrac{1}{2}W(t_1,F) + \tfrac{1}{2}W(t_2,F) \qquad \ldots \quad (1.3)$$

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_12, © Springer Science+Business Media, LLC 2011

for all t_1 and t_2. If the sign of equality holds only when $t_1=t_2$ then the function will be called strictly convex. As for example the following functions are all strictly convex.

$$\text{(i)} \quad t-\theta|^{\mu}, \mu>1, \qquad \text{(ii)} \quad e^{|t-\theta|}-1$$
$$\text{(iii)} \quad a|t-\theta|+b(t-\theta)^2, \qquad a\geqslant 0, \quad b>0$$

The function $|t-\theta|$ is convex but not strictly so.

2. ADMISSIBILITY AND SYMMETRY*

We prove the following:

Theorem 1: *If every $F\epsilon\Omega$ is symmetric in x_i and x_j and if the weight function $W(t, F)$ be strictly convex then every admissible estimator must be essentially symmetric in x_i, x_j.*

Proof: Let $t=t(x_1, x_2 \ldots x_n)$ be any admissible estimator and let t_1 be obtained from t by interchanging x_i with x_j.

From the symmetry of F in x_i, x_j it follows that t and t_1 are identically distributed for all $F\epsilon\Omega$.

$$\therefore \qquad r(F|t)=r(F|t_1) \text{ for all } F\epsilon\Omega.$$

Now if we define $t_0=\frac{1}{2}(t+t_1)$ then from the convexity of $W(t, F)$ we have

$$r(F|t_0)\leqslant \tfrac{1}{2}r(F|t)+\tfrac{1}{2}r(F|t_1)$$
$$=r(F|t). \qquad \ldots \quad (2.1)$$

Since t is admissible the sign of equality must hold everywhere in (2.1). From the strict convexity of $W(t, F)$ if follows that the set of points where $t\neq t_1$ must be of F-measure zero for all $F\epsilon\Omega$. Thus if the weight function be strictly convex all admissible estimators must by essentially symmetric in x_i, x_j. If the weight function be convex but not strictly so then there may exist admissible estimators which are not essentially symmetric in x_i, x_j. But since corresponding to any such unsymmetric estimator there always exist an estimator symmetric in x_i, x_j and generating the same risk function it follows that we need not go beyond estimators that are symmetric in x_i x_j.

Corollary: *If every $F\epsilon\Omega$ is symmetric in all the x's then for the purpose of estimation with a convex weight function we need restrict ourselves to only symmetric functions of the x's.*

If however we want to restrict our choice of t to a particular class of estimators then, for a particular weight function, the above results can be true without F being completely symmetric in x_i, x_j. For example suppose that $W(t, F)=(t-\theta)^2$ and that we want to restrict our choice of t only to linear functions of the x's. We note that if the first two moments of x_i and its product moments with the other x's are the same as the corresponding moments of x_j then for any linear estimator t

$$r(F|t)=r(F|t_1) \text{ for all } F\epsilon\Omega.$$

*The attention of the auther has been drawn to a paper by Paul R. Halmos entitled "The Theory of Unbiased Estimation" in the *Annals of Mathematical Statistics*, Vol. 17 (1946), where the results proved in this section were partially anticipated.

where t_1 is obtained from t by interchanging x_i with x_j and so the proof of the theorem applies.

We now show how the above considerations of symmetry lead to simple proofs of results which are otherwise difficult to obtain.

3. Sample from a finite population

Consider a finite population sith N values $a_1, a_2, \ldots a_N$ and let

$$\alpha = \frac{1}{N}\Sigma a_i \text{ and } \sigma^2 = \frac{1}{N}\Sigma(a_i - \alpha)^2.$$

be the population mean and variance.

Let a random sample $x_1, x_2, \ldots x_n$ of size n be drawn without replacement from the population (x_i is the ith sample drawn). It is clear that the probability

$$P(x_1 = a_{i_1}\ x_2 = a_{i_2} \ldots x_n = a_{i_n}) = \frac{(N-n)!}{N!}$$

and is the same for all the $(N)/(N-n)!$ possible choices of the indices $i_1, i_2, \ldots i_n$ from the set $1, 2, \ldots N$. Thus it follows that the joint distribution of the chance vector $(x_1, x_2, \ldots x_n)$ is symmetric in all the x's.

Hence in the class of all estimators of the form

$$t = c_1x_1 + c_2x_2 + \ldots + c_nx_n \qquad \ldots \quad (3.1)$$

we need consider only those for which all the c's are equal *i,e.* $t = c\bar{x}$. Let θ be the population characteristic we want to estimate and let $W(t,F) = (t-\theta)^2$
Then

$$r(F \mid c\bar{x}) = E(c\bar{x} - \theta)^2$$
$$= c^2V(\bar{x}) + (c\alpha - \theta)^2$$

where

$$V(\bar{x}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}.$$

If in particular $\theta = \alpha$ then

$$r(F \mid c\bar{x}) = c^2V(\bar{x}) + (c-1)^2\alpha^2$$

and we observe that $c\bar{x}$ cannot be admissible unless $0 \leqslant c \leqslant 1$. For corresponding to any c_1 outside the interval $0 \leqslant c \leqslant 1$ we can always find another c_0 in the interval such that $c_0\bar{x}$ is uniformly more powerful than $c_1\bar{x}$. It is conjectured that for the weight function $(t-\theta)^2$ every $c\bar{x}$ $(0 \leqslant c \leqslant 1)$ is an admissible estimator in the entire class of all possible estimators.

Again in the class of all quadratic estimators we need consider only symmetric estimators of the form

$$t = a\Sigma x_i^2 + b \sum_{i \neq j} x_ix_j + c\,\Sigma\, x_i + d. \qquad \ldots \quad (3.2)$$

If σ^2 be the population characteristic we want to estimate and if we add the further criterion of unbiasedness then from

$$E(t)=an(\sigma^2+\alpha^2)+bn(n-1)\left(-\frac{1}{N-1}\sigma^2+\alpha^2\right)+cn\alpha+d$$

$$\equiv\sigma^2 \text{ for all } \alpha \text{ and } \sigma^2.$$

We have

$$an-b\frac{n(n-1)}{N-1}=1$$

$$an+bn(n-1)=0$$

$$c=d=0.$$

Solving for a and b and substituting in (3.2) we have that in the class of all unbiased quadratic estimators of σ^2 the estimator

$$t=\frac{N-1}{N}\cdot\frac{\Sigma(x_i-\bar{x})^2}{n-1} \qquad \ldots \quad (3.3)$$

is uniformly the best estimator provided the weight function $W(t,F)$ is convex (downwards). It is believed that the estimator (3.3) is admissible in the unrestricted class of all estimators.

4. THE MARKOFF SET-UP

Consider the familiar Markoff set-up where $x_1, x_2,\ldots x_n$ is a set of chance variables with equal variances σ^2 and expected values

$$Ex_i=a_{i1}\tau_1+\ldots+a_{im}\tau_m \qquad i=1,2,\ldots n,\ m<n$$

where the a_{ij}'s are known constants and the τ's are unknown parameters. Without loss of generality we may assume that the rank of the matrix $(a_{ij})=m$

The problem is to extimate the τ's and σ^2.

At first let us assume that the x's are independently and normally distributed. We shall later on see how far this assumption can be relaxed.

We can always find a unitary orthogonal transformation

$$(z\vdots\ y)=x(B\vdots\ C) \qquad \ldots \quad 4.1)$$

where $B'=(b_{ij}) \quad i=1,2,\ldots n-m,\ j=1,2,\ldots n$

and $C'=(c_{ij}) \quad i=1,2,\ldots m,\ j=1,2,\ldots n$

such that $Ez_i=0 \quad i=1,2,\ldots n-m$

and $Ey_i=\xi_i\neq 0 \quad i=1,2,\ldots m$

where the ξ's are independent linear functions of the τ's.

From the independence and normality of the x's it follows that the z's and the y's are independent normal variables with the same variance.

Thus if we want to set-up a linear unbiassed estimator of $g_1\xi_1+\ldots+g_m\xi_m$ then from the symmetry of the z's and the condition of unbiassedness it follows that we must choose from

$$t=a(z_1+z_2+\ldots+z_{n-m})+g_1y_1+\ldots+g_my_m. \qquad \ldots \quad (4.2)$$

If further we take our weight function as the square of the error then from the fact that $z_1+z_2+\ldots+z_{n-m}$ is uncorrelated with $g_1y_1+\ldots+g_my_m$ it follows that 'a' must be zero in (4.2).

We now consider the problem of estimating σ^2. In the class of quadratic estimators of σ^2 we must, because of the symmetry of the z's choose from the class

$$t=a\Sigma z_i^2+b\sum_{i\neq j} z_iz_j+\Sigma c_jz_iy_j+\Sigma d_{ij}y_iy_j+e\Sigma z_i+\Sigma f_iy_i+g.$$

From the condition of unbiasedness we have

$$t=\frac{1}{n-m}\Sigma z_i^2+\left\{b\Sigma z_iz_j+\Sigma c_jz_iy_j+e\Sigma z_i\right\}$$

$$=S_0+Q \qquad \ldots \quad (4.3)$$

If we take the weight function as square of the error then from the fact that S_0 is uncorrelated with Q it follows that the minimum variance of t will be attained when $V(Q)=0$, *i.e.* when $Q\equiv 0$. Thus S_0 is the minimum variance unbiased quadratic estimator of σ^2. Following the technique of Rao (1952) it can be shown that in the class of all unbiased estimators S_0 is the minimum variance estimator of σ^2 and also that any linear function of the y's is the minimum variance unbiased estimator of its expected value. For proving this the assumption of independence and normality of the x's play an essential role.

If, however, we want to restrict our choice of t to only quadratic functions and take the square of the error as the weight function then it is apparent from the remarks at the end of §2 that the above proof will hold even in the less restricted situation where the moments and the product moments of the z's and the y's up to order four are symmetrical in the z's and further where S_0 be uncorrelated with Q. This will be so if the z's and the y's be mutually uncorrelated up to order four and if the third moments of the of the z's be zeros.

Definition: A set of chance variables $x_1, x_2,\ldots x_n$ are said to be mutually uncorrelated up to order p if

$$Ex_1^{p_1}x_2^{p_2}\ldots x_n^{p_n}=Ex_1^{p_1}\cdot Ex_2^{p_2}\ldots Ex_n^{p_n}$$

for all non-negative integers $p_1,p_2,\ldots p_n$ such that $p_1+p_2+\ldots+p_n=p$.

Let $x_1, x_2, \ldots x_n$ be mutually uncorrelated up to order p and let

$$y_i = a_{i1}x_1 + \ldots + a_{in}x_n \qquad \ldots (4.4)$$
$$i = 1,2,\ldots n$$

be a linear transformation of the x's.

Under what conditions the y's also will be mutually uncorrelated up to order p?

Since the chance variables $\{a_i x_i + b_i\}$ $i=1, 2,\ldots n$ will also be mutually uncorrelated up to order p it follows that we can, without any loss of generally, assume that the x's have zero means and unit variances. By adjusting the scales of the y's we can then have that the y's also have zero means and unit variances.

$$\therefore \qquad 1 = V(y_i) = \sum_{r=1}^{n} a^2_{ir} \qquad i=1, 2,\ldots n$$

and
$$0 = Ey_i Ey_j = Ey_i y_j$$
$$= \sum_{1}^{n} a_{ir}a_{jr} \; (i \neq j, \; i,j = 1,2,\ldots n)$$

$\therefore$ the transformation (4.4) must be a unitary orthogonal transformation.

Let $c_i(t)$ and $k_i(t)$ be the cumulant generating functions of x_i and y_i respectively znd let $c(t_i, t_2,\ldots t_n)$ and $k(t_1, t_2,\ldots t_n)$ be the joint cumulant generating functions of the x's and the y's respectively. Also let c_{im} and k_{im} be the mth cumulants of x_i and y_i respectively ($i=1,2,\ldots n$, $m=1,2,\ldots p$). We know that

$$c_{i1} = k_{i1} = 0 \text{ and } c_{i2} = k_{i2} = 1 \qquad i=1,2,\ldots n.$$

Now since the x's are uncorrelated up to order p it follows that

$$c(t_1,t_2,\ldots t_n) = c_1(t_1) + \ldots + c_n(t_n) \qquad \ldots (4.5)$$

up to terms with power $\leqslant p$.

Hence
$$k_i(t) = c(a_{i1}t, a_{i2}t,\ldots a_{in}t)$$
$$= c_1(a_{i1}t) + \ldots + c_n(a_{in}t) \qquad \ldots (4.6)$$

up to terms with power $\leqslant p$.

Also
$$k(t_1,t_2,\ldots t_n) = c(\Sigma a_{ij}t_i, \ldots \Sigma a_{in}t_i) \qquad \ldots (4.7)$$
$$= c_1(\Sigma a_{i1}t_i) + \ldots + c_n(\Sigma a_{in}t_i)$$

upto terms with power $\leqslant p$.

A necessary and sufficient condition for the y's to be uncorrelated up to order p is that

$$k(t_1,t_2,\ldots t_n) = k_1/t_1) + \ldots + k_n/t_n) \qquad \ldots (4.8)$$

upto terms with power $\leqslant p$.

Now it is easily seen that if

$$c_{im}=0 \text{ for } m=3, 4,...p \qquad i=1, 2,...n$$

then (4.8) will be satisfied and then

$$k_{im}=0 \text{ for } m=3, 4,...p, \qquad i=1,2,...n.$$

Now if (4.8) be true then we have

$$\begin{aligned} c_i(t)&=k(a_{1i}t, a_{2i}t,...a_{ni}t) \\ &=k_1(a_{1i}t)+...+k_n(a_{ni}t) \end{aligned} \qquad \ldots \quad (4.9)$$

up to terms with power $\leqslant p$.

From (4.6) and (4.9) we have

$$k_{im}=a_{i1}^m c_{1m}+a_{i2}^m c_{2m}+ \ldots+a_{in}^m c_{nm}$$

and $\qquad \ldots \quad (4.10)$

$$c_{im}=a_{1i}^m k_{1m}+a_{2i}^m k_{2m}+...+a_{ni}^m k_{nm}$$

$$i=1,2,...n, \ m=1,2,...p.$$

Hence from a Lemma proved earlier (Basu 1951) it follows that if no $a_{ij}=\pm 1$ then (4.10) can be satisfied if and only if

$$c_{im}=k_{im}=0 \qquad i=1,2...n, \ m=3,4,...p. \qquad \ldots \quad (4.11)$$

If some $a_{ij}=\pm 1$ then it means that y_i is a function of x_j alone and that no other y involves x_j. Then the uncorrelatedness (up to order p) of y_i with the other y's will follow from the uncorrelatedness (up to order p) of the x's and no further restriction on the cumulants of x_j need be imposed. Thus ignoring such trivial cases we have the following:

Theorem 2: *If $x_1, x_2,...x_n$ be uncorrelated up to order p then a necessary and sufficient condition that there exist non-trivial linear transformations of the x's into $y_1, y_2,...y_n$ such that the y's also are uncorrelated up to order p is that the x's(and therefore the y's) are normal up to order p i.e. $c_{im}=0$, $m=3, 4,...p$, $i=1, 2,...n$.*

Also if the x's have the same variance then any orthogonal transformation will make the y's uncorrelated up to order p.

Thus if in the Markoff set-up we assume that the x's are mutually uncorrelated up to order four and that $\beta_1=0$ and $\beta_2=3$ for all the x's (i.e. the x's are normal up to order four) then the transformation (4.1) will make the z's and the y's uncorrelated and normal up to order four and so in the same way as before we prove that

$$S_0=\frac{1}{n-m}\Sigma z_i^2 \qquad \ldots \quad (4.12)$$

is the minimum variance estimator of σ^2 in the class of all unbiased polynomial estimators of degree not exceeding two.

Hsu (1938) and Rao (1952) proved the same result over a less restricted distribution space Ω but had, therefore, to restrict the scope of the choices for the estimator t.

Hsu, for instance, restricts the choice of t to the class of unbiased quadratic forms xAx' for which $V(xAx')$ is independent of the unknown parameters $\tau_1,\tau_2,\ldots\tau_m$. Rao considers only definite quadratic forms.

REFERENCES

BASU, D. (1951): On the independence of linear functions of independent chance variables. *Proc. Int. Stat. Conf.*, India, 1951 (in press).

HSU, P. L. (1938): On the best unbiased quadratic estimate of the variance. *Stat. Res. Mem.*, **2**.

RAO, C. R. (1952): Some theorems on minimum variance estimation. *Sankhyā* **12**, Parts 1 & 2.

WALD, A. (1950): *Statistical Decision Functions*. John Wiley & Sons, New York.

AN INCONSISTENCY OF THE METHOD OF MAXIMUM LIKELIHOOD

By D. Basu

University of California, Berkeley

An example was given by Neyman and Scott [2] to show that there are situations where the method of maximum likelihood leads to inconsistent estimators. In their example considered, the observations were supposed to be drawn from an infinite sequence of distinct populations involving an infinite sequence of nuisance parameters.

An example is given here to demonstrate that even in simple situations where all the observations are independently and identically distributed and involve only one unknown parameter, the method of maximum likelihood may lead us astray. The example typifies the situations where the correct method of setting up a point estimate should begin with a test of hypothesis between two composite alternatives.

Let A be the set of all rational numbers in the closed interval (0, 1) and B any countable set of irrational numbers in the same interval. Let X be a random variable that takes the two values 0 and 1 with

$$P(X = 1) = \begin{cases} \theta & \text{if } \theta \varepsilon A, \\ 1 - \theta & \text{if } \theta \varepsilon B. \end{cases}$$

If r is the total number of 1's in a set of n random observations on X, then from the rationality of r/n it follows at once that the maximum likelihood estimator of θ is r/n. But r/n converges (in probability) to θ or $1 - \theta$ according as $\theta \varepsilon A$ or $\theta \varepsilon B$.

Now, since A and B are both countable sets, it follows [1] that there exists a consistent test for the composite hypothesis $\theta \varepsilon A$ against the composite alterna-

Received March 19, 1954.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_13,

tive $\theta \varepsilon B$. In other words, there exists a function φ_n of the first n observations such that φ_n takes only the two values 0 and 1 and such that

$$P(\varphi_n = 1 | \theta) \to \begin{cases} 0 & \text{if } \theta \varepsilon A, \\ 1 & \text{if } \theta \varepsilon B, \end{cases} \qquad n \to \infty.$$

It then follows readily that the estimator

$$t_n = (1 - \varphi_n) r/n + \varphi_n (1 - r/n)$$

is a consistent estimator of θ. Thus, though there exist consistent estimators for θ, the maximum likelihood estimator is not consistent.

For a simple construction for the function φ_n, let $\{\alpha_i\}$ and $\{\beta_i\}$, for $i = 1, 2, \cdots$, be two enumerations of the sets A and B, respectively, and let δ_k be the distance between the two sets $(\alpha_1, \cdots, \alpha_k)$ and $(1 - \beta_1, \cdots, 1 - \beta_k)$. Let $k(n)$ be the largest integer k such that $\delta_k > n^{-1/4}$. Note that $k(n)$ increases monotonically to infinity. Let I_{kn} and J_{kn} be the open intervals of length $n^{-1/4}$ centered around α_k and $1 - \beta_k$ respectively and let

$$R_n = \bigcup_{k \leqq k(n)} I_{kn}, \qquad S_n = \bigcup_{k \leqq k(n)} J_{kn}.$$

For every n, the sets R_n and S_n are clearly disjoint. Now consider a fixed k. For all n for which $k(n) \geqq k$ we have

$$\begin{aligned} P(r/n \varepsilon S_n \mid \theta = \beta_k) &\geqq P(r/n \varepsilon J_{kn} \mid \theta = \beta_k) \\ &= P(\, |r/n - [1 - \beta_k]| < -n^{-1/4} \mid \theta = \beta_k) \\ &\to 1 \text{ as } n \to \infty, \end{aligned}$$

because r/n is asymptotically normal with mean $1 - \beta_k$ and asymptotic s.d. $\sqrt{\beta_k(1 - \beta_k)/n}$. By the same argument we have

$$\begin{aligned} P(r/n \varepsilon S_n \mid \theta = \alpha_k) &\leqq 1 - P(r/n \varepsilon R_n \mid \theta = \alpha_k) \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

Now, if we define

$$\varphi_n = \begin{cases} 1 & \text{if } r/n \varepsilon S_n \\ 0 & \text{otherwise,} \end{cases}$$

then φ_n clearly has the required property.

Acknowledgement. I wish to thank Mr. J. L. Hodges, Jr., for his help in the writing of this note.

REFERENCES

[1] CHARLES KRAFT, "On the problem of consistent and uniformly consistent statistical procedures", Unpublished Ph.D. thesis submitted to the University of California (1954).

[2] J. NEYMAN AND E. L. SCOTT, "Consistent estimates based on partially consistent observations," *Econometrica*, Vol. 16 (1948), pp. 1–32.

ON STATISTICS INDEPENDENT OF A COMPLETE SUFFICIENT STATISTIC

By D. BASU

Indian Statistical Institute, Calcutta

1. Introduction

If $\{P_\theta\}$, $\theta\epsilon\Omega$, be a family of probability measures on an abstract sample space $\mathfrak{S}$ and T be a sufficient statistic for θ then for a statistic T_1 to be stochastically independent of T it is necessary that the probability distribution of T_1 be independent of θ. The condition is also sufficient if T be a boundedly complete sufficient statistic. Certain well-known results of distribution theory follow immediately from the above considerations. For instance, if $x_1, x_2, \ldots, x_n$, are independent $N(\mu, \sigma)$'s then the sample mean $\bar{x}$ and the sample variance s^2 are mutually independent and are jointly independent of any statistic f (real or vector valued) that is independent of change of scale and origin. It is also deduced that if $x_1, x_2, \ldots, x_n$, are independent random variables such that their joint distribution involves an unknown location parameter θ then there can exist a linear boundedly complete sufficient statistic for θ only if the x's are all normal. Similar characterizations for the Gamma distribution also are indicated.

2. Definitions

Let $(\mathfrak{S}, \mathcal{A})$ be an arbitrary measurable space (the sample space) and let $\{P_\theta\}$, $\theta\epsilon\Omega$, be a family of probability measures on $\mathcal{A}$.

Definition 1: Any measurable transformation T of the sample space $(\mathfrak{S}, \mathcal{A})$ onto a measurable space $(\mathcal{T}, \mathcal{B})$ is called a statistic. The probability measures on $\mathcal{B}$ induced by the statistic T are denoted by $\{P_\theta^T\}, \theta\epsilon\Omega$.

For every $\theta\epsilon\Omega$ and $A\epsilon\mathcal{A}$ there exists an essentially unique real valued $\mathcal{B}$-measurable function $f_\theta(A \mid t)$ on $\mathcal{T}$ such that the equation

$$P_\theta(A \cap T^{-1}B) = \int_B f_\theta(A \mid t)\, dP_\theta^T \qquad \ldots \quad (1)$$

holds for every $B\epsilon\mathcal{B}$. The set of points t for which $f_\theta(A \mid t)$ falls outside the closed interval (0, 1) is of P_θ^T-measure zero for every $\theta\epsilon\Omega$. We call $f_\theta(A \mid t)$ the conditional probability of A given that $T = t$ and that θ is the true parameter point.

Definition 2: A statistic T is said to be independent of the parameter θ if, for every $B\epsilon\mathcal{B}$, $P_\theta^T(B)$ is the same for all $\theta\epsilon\Omega$.

Definition 3: The two statistics T and T_1, with associated measurable spaces $(\mathcal{T}, \mathcal{B})$ and $(\mathcal{T}_1, \mathcal{B}_1)$ respectively, are said to be stochastically independent of each other if, for every $B\epsilon\mathcal{B}$ and $B_1\epsilon\mathcal{B}_1$

$$P_\theta(T^{-1}B \cap T_1^{-1}B_1) \equiv P_\theta(T^{-1}B) P_\theta(T_1^{-1}B_1)$$

for all $\theta\epsilon\Omega$.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_14,

Now,

$$P_\theta(T^{-1}B \cap T_1^{-1}B_1) \equiv \int_B f_\theta(T_1^{-1}B_1 \mid t)\, dP_\theta^T.$$

It follows, therefore, that a necessary and sufficient condition in order that T and T_1 are stochastically independent is that the integrand above is essentially independent of t, i.e.

$$f_\theta(T_1^{-1}B_1 \mid t) = P_\theta(T_1^{-1}B_1) = P_\theta^{T_1}(B_1)$$

for all $t \epsilon T$ excepting possibly for a set of P_θ^T-measure zero.

Definition 4: The statistic T is called a sufficient statistic (Halmos and Savage, 1949) if for every $A \epsilon \mathcal{A}$ there exists a function $f(A \mid t)$ which is independent of θ and which satisfies equation (1) for every $\theta \epsilon \Omega$.

Let G be the class of all real valued, essentially bounded, and $\mathcal{B}$-measurable functions on $\mathcal{T}$.

Definition 5: The family of probability measures $\{P_\theta^T\}$ is said to be boundedly complete (Lehmann and Scheffé, 1950) if for any $g \epsilon G$ the identity

$$\int_T g(t) dP_\theta^T \equiv 0 \quad \text{for all } \theta \epsilon \Omega \qquad \ldots \quad (2)$$

implies that $g(t) = 0$ excepting possibly for a set of P_θ^T-measure zero for all θ. $\{P_\theta^T\}$ is called complete if the condition of essential boundedness is not imposed on the integrand in (2). The statistic T is called complete (boundedly complete) if the corresponding family of measures $\{P_\theta^T\}$ is so.

3. SUFFICIENCY AND INDEPENDENCE

For any two statistics T_1 and T we have for any $B_1 \epsilon \mathcal{B}_1$

$$P_\theta^{T_1}(B_1) = P_\theta(T_1^{-1}B_1) = \int_{\mathcal{T}} f_\theta(T_1^{-1}B_1 \mid t) dP_\theta^T. \qquad \ldots \quad (3)$$

Now if T be a sufficient statistic then the integrand is independent of θ and if, moreover, T_1 is stochastically independent of T then the integrand is essentially independent of t also. Thus, the right hand side of (3) is independent of θ and so we have

Theorem 1: *Any statistic T_1 stochastically independent of a sufficient statistic T is independent of the parameter θ.*

That the direct converse of the above result is not true will be immediately apparent if we take for the sufficient statistic T the identity mapping of $(\mathfrak{S}, \mathcal{A})$ into itself. No statistic T_1 independent of θ will then be stochastically independent of T excepting in the trival situation where T_1 is essentially equal to a constant. We, however, have the following weaker but important converse.

Theorem 2: *If T be a boundedly complete sufficient statistic then any statistic T_1 which is independent of θ is stochastically independent of T.*

Proof: Since T is sufficient the integrand in (3) is independent of θ. It is also essentially bounded. Now the left hand side of (3) is independent of θ since T_1 is independent of θ. Hence, from bounded completeness of $\{P_\theta^T\}$ it follows that the integrand in (3) is essentially independent of t as well. That is, T_1 is stochastically independent of T.

In the next section we demonstrate how the above theorem may be used to get a few interesting results in distribution theory.

4. Some characterizations of distributions with location and scale parameters

Let $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$ be a random variable in an n-dimensional Euclidean space whose probability distribution involves an unknown location parameter μ and a scale parameter $\sigma > 0$. Then any measurable function $f(x_1, x_2, \ldots, x_n)$ which is independent of change of origin and scale, i.e.

$$f\left(\frac{x_1-a}{b}, \ldots, \frac{x_n-a}{b}\right) \equiv f(x_1, \ldots, x_n)$$

for all a and $b>0$ is independent of the unknown parameter (μ, σ). Now, if there exists a boundedly complete sufficient statistic T for (μ, σ) then f must be stochastically independent of T. For example, if $x_1, x_2, \ldots, x_n$, are independent observations on a normal variable with mean μ and s.d. σ then it is well known that $T = (\bar{x}, s)$ is a sufficient statistic ($\bar{x}$ is the sample mean and s the sample s.d.). The completeness of T follows from the unicity property of the bivariate Laplace transform. It then follows from Theorem 2 that any measurable function $g(\bar{x}, s)$ of $\bar{x}$ and s is stochastically independent of any measurable function $f(x_1, x_2, \ldots, x_n)$ of the observations that is independent of change of origin and scale. The functions g and f need not be real valued. For instance, we may have

$$g = (\Sigma x_i^2, \sum_{i \neq j} x_i x_j)$$

and

$$f = \left(\frac{\Sigma(x_i-\bar{x})^3}{s^3}, \frac{\Sigma(x_i-\bar{x})^4}{s^4}, \ldots\right).$$

Again the stochastic independence of $\bar{x}$ and s follows from the fact that, for any fixed σ, the statistic $\bar{x}$ is a complete sufficient statistic for μ and that s, by virtue of its being independent of change of origin, is independent of the location parameter μ.

Now let $x_1, x_2, \ldots, x_n$, be independent random variables with joint d.f. $F_1(x_1-\theta), F_2(x_2-\theta), \ldots, F_n(x_n-\theta)$.* Since θ is a location parameter it follows that any linear function $\Sigma a_i x_i$ with $\Sigma a_i = 0$ is independent of θ. If $\Sigma b_i x_i$ is a boundedly complete sufficient statistic for θ then from Theorem 2 it follows that $\Sigma a_i x_i$ is independent of $\Sigma b_i x_i$.

* For the sake of notational convenience, we make no distinction between random variables and the values that they may assume.

Now, since $\Sigma b_i x_i$ is a sufficient statistic it follows that every $b_i \neq 0$. For, if possible, let $b_j = 0$. Then x_j is stochastically independent of $\Sigma b_i x_i$ and so from Theorem 1 x_j is independent of the parameter θ which contradicts the assumption that the d.f. of x_j is $F_j(x_j - \theta)$. Again, we can take all the a_i's different from zeros. Thus, the two linear functions $\Sigma a_i x_i$ and $\Sigma b_i x_i$ (with non-zero coefficients) of the independent random variables $x_1, x_2, \ldots, x_n$, are stochastically independent. Therefore,† all the x_i's must be normal variables. We thus have the following:

Theorem 3: *If $x_1, x_2, \ldots, x_n$, are independent random variables such that their joint d.f. involves an unknown location parameter θ then a necessary and sufficient condition in order that $\Sigma b_i x_i$ is a boundedly complete sufficient statistic for θ is that $b_i > 0$ and that x_i is a normal variable with mean θ and variance b_i^{-1} $(i = 1, 2, \ldots, n)$.*

Let us now turn to the case of the Gamma variables. Let $x_1, x_2, \ldots, x_n$, be independent Gamma variables with the same scale parameter $\theta > 0$, i.e., the density function of x_i is

$$f_i(x)dx = \frac{1}{\Gamma(m_i)\theta^{m_i}} x^{m_i - 1} e^{-x/\theta} dx \qquad (x \geqslant 0, \quad \theta > 0, \quad m_i > 0).$$

It is clear then that Σx_i is a sufficient statistic for θ and its completeness follows from the unicity property of the Laplace transform. Thus, we at once have the well known result that Σx_i is stochastically independent of any function $f(x_1, x_2, \ldots, x_n)$ that is independent of change of scale (i.e. independent of θ).

Recently it has been proved by R. G. Laha that if $x_1, \ldots, x_n$, are independent and identically distributed chance variables and if Σx_i is independent of $\Sigma a_{ij} x_i x_j / (\Sigma x_i)^2$ then (under some further assumptions) all the x_i's must be Gamma variables. Using this result we can immediately get a characterization of the Gamma distribution analogous to Theorem 3.

REFERENCES

BASU, D. (1951) : On the independence of linear functions of independent chance variables. *Bull. Int. Stat. Inst.*, **33**, Pt. 2, 83-96.

DARMOIS, G. (1951) : Sur diverses proprietes characteristiques de la loi de probabilite de Laplace-Gauss, *Bull. Int. Stat. Inst.*, **33**, Pt. 2, 79-82.

——— (1953) : Analyse generale des liaisons stochastiques—Etude particuliere de l'analyse factorielle lineaire. *Rev. Inst. Internat. Stochastique*, **21**, 2–8.

HALMOS, P. R. AND SAVAGE, L. I. (1949) : Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Stat.*, **20**, 225-241.

LEHMANN, E. L. AND SCHEFFÉ, H. (1950) : Completeness, similar regions and unbiased estimation. *Sankhyā*, **10**, 305-340.

LAHA, R. G. (1954) : On a characterization of the Gamma distribution. *Ann. Math. Stat.*, **25**, 784-787.

† This result was first conjectured (and proved under certain assumptions) by the author in 1951. The proof without any assumption is due to G. Darmois (1953).

MISCELLANEOUS

THE CONCEPT OF ASYMPTOTIC EFFICIENCY

By D. BASU

Indian Statistical Institute, Calcutta

1. Summary

Partly of an expository nature this note brings out the fact that an estimator, though asymptotically much less efficient (in the classical sense) than another, may yet have much greater probability concentration (as defined in this article) than the latter.

2. Definitions

Let $\{X_i\}$, $i = 1, 2, \ldots$ be an infinite sequence of independent and identically distributed random variables whose common distribution function F is known to belong to a family Ω of one dimensional distribution functions. Let $\mu = \mu(F)$ be a real valued functional defined on Ω. By an estimator $T = \{t_n\}$ of μ we mean a sequence of real valued measurable functions of $\{X_i\}$, where t_n is a function of $X_1, X_2, \ldots, X_n$ only $(n = 1, 2, \ldots)$. The estimator T is said to be an asymptotically normal estimator of μ if there exists a sequence $\{\sigma_n(F)\}$ of positive numbers such that as $n \to \infty$

$$\{t_n - \mu(F)/\sigma_n(F)\} \Longrightarrow N(0, 1) \quad \text{for all } F \in \Omega$$

where $\Longrightarrow$ stands for convergence in law and $N(0, 1)$ for the standard normal variable. The sequence $\{\sigma_n(F)\}$ is called the asymptotic standard deviation of T. A necessary and sufficient condition in order that $\{\sigma_n(F)\}$ and $\{\sigma_n'(F)\}$ may both be called the asymptotic standard deviation of T is

$$\lim_{n \to \infty} \{\sigma_n(F)/\sigma_n'(F)\} \equiv 1 \quad \text{for all } F \in \Omega .$$

A necessary and sufficient condition in order that the asymptotically normal estimator T is also consistent is

$$\lim_{n \to \infty} \sigma_n(F) \equiv 0 \quad \text{for all } F \in \Omega .$$

Let $\mathcal{T}$ be the family of all consistent asymptotically normal estimators of μ. We consider only the space $\mathcal{T}$.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_15,

3. THE PARTIAL ORDER OF EFFICIENCY

Two elements T and T' of $\mathcal{T}$ are said to be equally efficient (or equivalent) if they have the same asymptotic s.d.s, *i.e.* if

$$\lim_{n \to \infty} \{\sigma_n(F)/\sigma'_n(F)\} \equiv 1 \quad \text{for all } F \in \Omega \qquad \ldots \quad (3.1)$$

where $\{\sigma_n(F)\}$ and $\{\sigma'_n(F)\}$ are the corresponding asymptotic s. d.'s.

It is easily verified that the above equivalence relation is reflexive, symmetric, and transitive.

If $$\limsup_{n \to \infty} \{\sigma_n(F)/\sigma'_n(F)\} \leqslant 1 \quad \text{for all } F \in \Omega$$

and $$\liminf_{n \to \infty} \{\sigma_n(F)/\sigma'_n(F)\} < 1 \quad \text{for some } F \in \Omega$$

then we say that T is more efficient than T' and write $T \supset T'$. It is easily seen that the relation $\supset$ induces a partial order on $\mathcal{T}$.

It is known that there do not exist a maximal element in $\mathcal{T}$ with respect to the partial order $\supset$, *i.e.* there do not exist any element $T \in \mathcal{T}$ which is either equivalent to or more efficient than any alternative $T' \in \mathcal{T}$. As a matter of fact it has been demonstrated (LeCam, 1953) how given any $T \in \mathcal{T}$ we can always find a $T' \in \mathcal{T}$ such that $T' \supset T$.

4. THE PARTIAL ORDER OF CONCENTRATION

The estimator $T = \{t_n\}$ of μ is consistent if for all $\epsilon > 0$ and $F \in \Omega$

$$p_n(\epsilon, F) = P\{|t_n - \mu| > \epsilon \,|\, F\} \to 0 \text{ as } n \to \infty .$$

If we work with the simple loss function that is zero or one according as the error in the estimate is $\leqslant \epsilon$ or $> \epsilon$ then $p_n(\epsilon, F)$ is the risk (or expected loss) when the estimator is used with observations on $X_1, X_2, \ldots, X_n$ only.

The rapidity with which $p_n(\epsilon, F) \to 0$ may be considered to be a measure of the asymptotic accuracy or concentration of T. This motivates the following definition of a partial order on $\mathcal{T}$ (and as a matter of fact on the wider family of all consistent estimators of μ).

Definition : The estimator T with the associated sequences of risk functions $p_n(\epsilon, F)$ is said to have greater concentration than T' with the associated sequences $p'_n(\epsilon, F)$ if, for all $\epsilon > 0$ and $F \in \Omega$,

$$\limsup_{n \to \infty} \{p_n(\epsilon, F)/p'_n(\epsilon, F)\} \leqslant 1$$

with the limit inferior being < 1 for some $\epsilon > 0$ and some $F \in \Omega$. We then write $T > T'$.

Intuitively it may seem reasonable to expect that $T \supset T'$ implies $T > T'$. That this is not so is demonstrated in the next section. An example is given where

$$\lim_{n \to \infty} \frac{\sigma_n(F)}{\sigma'_n(F)} \equiv 0 \qquad \text{for all } F \in \Omega, \qquad \ldots \quad (4.1)$$

whereas

$$\lim_{n \to \infty} \frac{p_n(\epsilon, F)}{p'_n(\epsilon, F)} \equiv \infty \qquad \text{for all } \epsilon > 0 \text{ and } F \in \Omega. \qquad \ldots \quad (4.2)$$

5. An example

Let each of the X_i's be $N(\mu, 1)$, the problem being to estimate μ.

Let $$\bar{X}_n = \sum_1^n X_i/n \text{ and } S_n = \sum_1^n (X_i - \bar{X}_n)^2.$$

Then $\bar{X}_n$ and S_n are mutually independent random variables and the distribution of S_n is independent of μ. Let a_n be the upper $100/n$ % point of S_n and let

$$H_n = \begin{cases} 0 & \text{if} \quad S_n \leqslant a_n, \\ 1 & \text{if} \quad S_n > a_n. \end{cases}$$

Now let $$T = \{t_n\}$$

where $$t_n = (1 - H_n)\bar{X}_n + n\, H_n$$

and $$T' = \{t'_n\}$$

where $$t'_n = \bar{X}_{[\sqrt{n}]}.$$

(By $[x]$ we mean the largest integer not exceeding x.)

Since $$P(H_n = 0) = 1 - \frac{1}{n} \to 1,$$

it follows (vide Cramér, p. 254) that $\sqrt{n}(t_n - \mu) = \sqrt{n}(\bar{X}_n - \mu) + \sqrt{n}\, H_n(n - \bar{X}_n)$

$$\Longrightarrow N(0, 1)$$

when μ is the true mean.

Hence, $T \in \mathcal{T}$ with asymptotic s.d. $= \{n^{-1/2}\}$. Also $T' \in \mathcal{T}$ with asymptotic s.d. $= \{n^{-1/4}\}$.

Therefore (4.1) is satisfied. Again, since $\bar{X}_n$ is independent of H_n it follows that, for every $n > \mu+\varepsilon$,

$$P(|t_n-\mu| > \epsilon|\mu) = P(H_n = 0)\, P(|\bar{X}_n-\mu| > \epsilon|\mu) + P(H_n = 1)$$

$$= \frac{1}{n} + o\left(\frac{1}{n}\right)$$

because $P(|\bar{X}_n-\mu| > \epsilon|\mu) = o\left(\frac{1}{n}\right)$, as may be easily verified.

Whereas $$P(|t'_n-\mu| > \epsilon|\mu) = o\left(\frac{1}{n}\right)$$

Therefore (4.2) also is satisfied.

It may be noted that in the example given the s.d. of t_n is not asymptotically equal to the asymptotic s.d. of T. But this can be easily arranged to be true by, say, taking a_n for the upper $100/n^4$ % point of S_n.

REFERENCES

CRAMÉR, H. (1946) : *Methematical Methods of Statistics*, Princeton University Press.

LECAM, L. (1953) : *On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates*, University of California Press.

Paper received : January, 1955.

ON STATISTICS INDEPENDENT OF SUFFICIENT STATISTICS

By D. BASU

Indian Statistical Institute, Calcutta

SUMMARY. In an earlier paper (1955) the author stated that any statistic independent of a sufficient statistic must have the same distribution for all values of the unknown parameter. An example is given here to show that the proposition is not true in the generality stated above. Conditions under which the proposition is true are discussed.

1. Introduction

Let X be a random variable (sample) taking values in an arbitrary sample space $\mathcal{X}$ with the associated σ-field of measurable sets $\mathcal{A}$ and the family of probability measures $\{P_\theta\}$, $\theta\epsilon\Omega$. By a statistic $T = T(X)$ we mean a measurable characteristic of the sample X, i.e., T is an $\mathcal{A}$-$\mathcal{B}$ measurable transformation of the measurable space $(\mathcal{X}, \mathcal{A})$ into some measurable space $(\mathcal{T}, \mathcal{B})$. The family of induced (by the mapping T) probability measures on $\mathcal{T}$ is denoted by $\{P_\theta T^{-1}\}$, $\theta\epsilon\Omega$.

If $P_\theta T^{-1}$ is the same for all $\theta\epsilon\,\Omega$ then it is clear that an observation on the random variable T will be of no use for making any inference about the parameter θ. In this case we may say that the statistic T contains no information about the parameter θ. On the other hand if T be a sufficient statistic then we may say that T contains the whole of the information about θ that is contained in the sample X. Barring these two extreme situations it is not possible to make a general assessment of how much (or what proportion) of information is contained in a particular statistic. The author feels that the question 'How much information is contained in T ?' should be rephrased as 'How effective an observation on T is for making a particular inference about θ ?' Clearly the answer will depend on the kind of inference (tests of hypotheses, point or interval estimation etc.) that we wish to make and also on our idea (or criterion) of effectivity. An element of arbitrariness is bound to enter into any attempted definition of the amount of information in a statistic.

One interesting feature of Fisher's definition (1921) of the amount of information is that it is additive for independent statistics. That is, if T_1 and T_2 are any

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_16,

two statistics that are independent for every $\theta \epsilon\ \Omega$ then the amount of information in (T_1, T_2) is equal to the sum of the informations contained in T_1 and T_2 separately. This, however, does not appear to the author to be a necessary requirement for a satisfactory definition of information. It is possible to think of situations where T_1 and T_2 are equally informative (identically distributed for example) and are independent of one another but still, when an observation on T_1 is given, very little extra information will be supplied by an observation on T_2. For example, suppose we have a population whose distribution we know to be either $N(0, 1)$ or $N(5, 1)$. A single observation from the population will identify the true distribution with a great measure of certainty. Given one observation from the population very little extra information will be obtained from a second observation from the population. Surely the total information contained in two independent observations from the population is much less than twice that contained in a single observation. The following is a more extreme example.

Suppose it is known that a bag contains 10 identical balls numbered $\theta+1$, $\theta+2, \ldots, \theta+10$ where the unknown parameter θ takes any one of the values 0, 10, 20, 30, ... Suppose two balls are drawn one by one with replacement and let T_i be the number on the i-th ball drawn ($i = 1, 2$). Here T_1 and T_2 are identically distributed independent statistics and each is sufficient for θ. Given an observation on T_1 the distribution of T_2 gets completely specified and hence T_1 contains as much information as is contained in (T_1, T_2).

Independence of statistics is sometimes loosely interpreted as follows :— 'If the statistic T_2 is independent of T_1 then knowing what the realization of T_1 has been in a particular trial gives us no information about the possible realization of T_2 in the same trial.' When the probability measure on the sample space is only partially known the above interpretation of independence is no longer true. The example in the previous paragraph very forcefully brings this point out.

2. STATISTICS INDEPENDENT OF A SUFFICIENT STATISTIC

In the previous section we have given an example to show that a statistic can be independent of a sufficient statistic and still contain a great deal of information about the parameter. The example demonstrates that Theorem 1 in (1955) is not true in the generality stated there.[1] Under some mild restrictions, however, the theorem remains true.

[1] Mr. R. H. Farrell of the University of Illinois independently arrived at the conclusions contained in this section. In a letter to the author Mr. Farrell discussed the mistake in the proof of the theorem and gave an example very similar to the one considered above.

Let T be a sufficient statistic and let $A \epsilon \mathcal{A}$ be any fixed event that is independent of T. From the sufficiency of T it follows that there exists a $\mathcal{B}$-measurable real valued function $f(A \mid t)$ on $\mathcal{T}$(called the conditional probability of A given $T = t$) such that for any $B \epsilon \mathcal{B}$

$$P(A \bigcap T^{-1}B) = \int_B f(A \mid t) d\, P_\theta T^{-1} \text{ for all } \theta \epsilon \Omega.$$

From the independence of the event A and the statistic T it follows that for any $\theta \epsilon \Omega$,

$$f(A \mid t) = P_\theta(A) \qquad \ldots \quad (2.1)$$

almost everywhere $[P_\theta T^{-1}]$ in t.

From (2.1) we cannot conclude that $P_\theta(A)$ is the same for all $\theta \epsilon \Omega$.

If the two measures $P_{\theta_1} T^{-1}$ and $P_{\theta_2} T^{-1}$ on $(\mathcal{T}, \mathcal{B})$ overlap, [i.e., for any set $B \epsilon \mathcal{B}$, $P_{\theta_1} T^{-1}(B) = 1$ implies that $P_{\theta_2} T^{-1}(B)$ is positive], then it is very easy to see that (2.1) implies the equality of $P_{\theta_1}(A)$ and $P_{\theta_2}(A)$.

Let us write $\theta_1 \Longleftrightarrow \theta_2$ if $P_{\theta_1} T^{-1}$ and $P_{\theta_2} T^{-1}$ overlap. The equality of $P_\theta(A)$ and $P_{\theta'}(A)$ can be deduced if there exists a finite number of parameter points $\theta_1, \theta_2, \ldots \theta_k$ such that

$$\theta \Longleftrightarrow \theta_1 \Longleftrightarrow \theta_2 \ldots \Longleftrightarrow \theta_k \Longleftrightarrow \theta'. \qquad \ldots \quad (2.2)$$

We say θ and θ' are connected (by the statistic T) if there exists $\theta_1, \theta_2, \ldots \theta_k$ satisfying (2.2).

Thus, we have the

Theorem : *If T be a sufficient statistic and if every pair of θ's in Ω are connected (by T), then any event A independent of T has the same probability for all $\theta \epsilon \Omega$.*

As a corollary we at once have that under the conditions of the above theorem any statistic T_1 independent of the sufficient statistic T contains no information about the parameter.

REFERENCES

BASU, D. (1955) : On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377.

FISHER, R. A. (1921) : The mathematical foundations of theoretical statistics. *Phil. Trans. Royal Soc.* A, **222**, 309.

Paper received : May, 1958.

MISCELLANEOUS

ON SAMPLING WITH AND WITHOUT REPLACEMENT

By D. BASU

Indian Statistical Institute, Calcutta

SUMMARY. Certain aspects of sampling with or without replacement, with equal or unequal probabilities, are considered here in some details. Some comparisons have been made between the with and without replacement sampling schemes. When we are sampling with replacement the estimate should not depend on the number of times that any particular unit may appear in the sample. Thus, certain estimation procedures in current use are shown to be inefficient.

1. Introduction

Suppose a given population has N units. Let Y_j be some real valued characteristic of the j-th population unit ($j = 1, 2, \ldots, N$). Consider the problem of estimating the population mean

$$\bar{Y} = N^{-1} \Sigma \; Y_j.$$

Let

$$\sigma^2 = N^{-1} \; \Sigma \, (Y_j - \bar{Y})^2$$

be the population variance.

If we draw a sample of size n from the population with equal probabilities and with replacement then the variance of the sample mean is $n^{-1}\sigma^2$. If, on the other hand, we draw a sample of the same size n but this time without replacement then the variance of the sample mean is $n^{-1}\sigma^2(N-n)(N-1)^{-1}$. Thus, it is usually claimed that sampling without replacement is better as it leads to an estimator of $\bar{Y}$ with a smaller variance. A little reflection, however, will show that this comparison between the two methods of sampling is not usually quite fair. Let us take a simple example. Suppose the units are villages and Y_j the number of households in the j-th village. Here the cost of selecting the sample villages from a given frame is negligible compared to the cost of travelling to the selected villages and ascertaining the exact number of households in the selected villages. Generally speaking, we need only consider the cost of measuring the Y-characteristics of the selected units—the small cost involved in selecting the units from a frame may be taken to be a part of the large overhead cost. In sampling with replacement it is then the number ν of distinct units appearing in the sample (and not the sample size n) that roughly determines the cost.

2. The distribution of ν

If ν be the number of distinct units appearing in a sample of size n drawn with equal probabilities and with replacements from a population with N units then it is clear that the distribution of ν depends only on n and N. It is not difficult to show (Feller p. 92) that

$$P(\nu = s) = N^{-n} \binom{N}{s}[1 - \binom{s}{1}(s-1)^n + \binom{s}{2}(s-2)^n \; \ldots]$$

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_17, © Springer Science+Business Media, LLC 2011

where s runs from 1 to the smaller of n and N.

In terms of the 'differences of zeros' we may write the above in the more elegant form

$$P(\nu = s) = N^{-n} \binom{N}{s} \Delta^s 0^n \qquad \ldots \quad (2.1)$$

where Δ is the usual difference operator with unit increments and $\Delta^s\ 0^n$ is to be interpreted as $\Delta^s\ x^n$ at $x = 0$.

From (2.1) we have the probability generating function of ν as

$$P_\nu(t) = E\ t^\nu = N^{-n} \sum_{s=0}^{N} \left[\binom{N}{s} t^s \Delta^s 0^n\right] = N^{-n}(1+t\Delta)^N\ 0^n. \qquad \ldots \quad (2.2)$$

(Note that $\Delta^s\ 0^n = 0$ for $s = 0$ and $s > n$).

Writing $1+t$ for t in the probability generating function we have the factorial moment generating function of ν as

$$F_\nu(t) = N^{-n}(\mathfrak{E}+t\Delta)^N 0^n \qquad \ldots \quad (2.3)$$

where $\mathfrak{E} = 1+\Delta =$ the usual increment operator with unit increments.

$$\therefore \quad E(\nu) = N^{-n} \binom{N}{1} \mathfrak{E}^{N-1} \Delta\ 0^n = N^{-n} \binom{N}{1} (\mathfrak{E}^N - \mathfrak{E}^{N-1})\ 0^n = N\left[1 - \left(\frac{N-1}{N}\right)^n\right]. \qquad \ldots \quad (2.4)$$

Also $E\nu(\nu-1) = N^{-n} \binom{N}{2} 2\mathfrak{E}^{N-2} \Delta^2\ 0^n = N^{-n}\ N(N-1)(\mathfrak{E}^N - 2\mathfrak{E}^{N-1} + \mathfrak{E}^{N-2})\ 0^n$

$$= N(N-1)\left[1 - 2\left(\frac{N-1}{N}\right)^n + \left(\frac{N-2}{N}\right)^n\right]$$

$$\therefore \quad V(\nu) = N\left(\frac{N-1}{N}\right)^n - N^2\left(\frac{N-1}{N}\right)^{2n} + N(N-1)\left(\frac{N-2}{N}\right)^n. \qquad \ldots \quad (2.5)$$

3. SAMPLING COST CONSIDERATIONS

If we assume that the variable part of the cost of sampling is proportional to the number of distinct units in the sample then we may compare the two methods of estimating $\bar{Y}$, as described in § 1, in the following manner. The expected sampling cost for a sample of size n with replacement is equal to the sampling cost for a sample of size $E\nu = N\left[1 - \left(\frac{N-1}{N}\right)^n\right]$ without replacement (let us conveniently forget the fact that $E\nu$ is not necessarily an integer). The variances for the sample means for the two cases are then $n^{-1}\sigma^2$ and $(E\nu)^{-1}\sigma^2(N-E\nu)(N-1)^{-1}$ respectively. A little computation will show that the former is larger. Thus, from this comparison, sampling with replacement appears to be worse than sampling without replacement. This comparison between the two methods heavily depends on the assumption of linearity of the cost function and as such is not very satisfactory. For a different cost function sampling with replacement may appear to fare better than sampling without replacement. The issue that is raised in the next section is perhaps more pertinent to the problem.

4. Two estimators from a with replacement sample

If we draw a sample of size n with replacements and with equal probabilities and if ν be the number of distinct units appearing in the sample then the average Y-characteristic of the ν distinct units is also an unbiased estimator of $\bar{Y}$. We may enquire whether this estimator is better or worse than the average over all the n units. The variance of the former is

$$E\left(\frac{N-\nu}{N-1}\cdot\frac{\sigma^2}{\nu}\right) \qquad \ldots \quad (4.1)$$

whereas that of the latter is $n^{-1}\sigma^2$.

It is not possible to give a simple expression for (4.1). In the next section we shall give an indirect proof of the inequality.

$$E\left(\frac{N-\nu}{N-1}\cdot\frac{\sigma^2}{\nu}\right) < \frac{\sigma^2}{n} \quad (\text{if } n > 1) \qquad \ldots \quad (4.2)$$

Thus, the average Y-characteristics of the ν distinct units in the sample is a better estimator than the average over all the n units. This result may at first appear to be a little unfamiliar, even surprising. Let us take, for example, the familiar Binomial model where we draw n balls at random one by one and with replacements from an urn containing N identical balls Np of which are white, the rest being black. Here the sample observation consists of a sequence of n white or black balls. The number r of white balls in the sample then constitutes a complete[1] sufficient statistic for the unknown parameter p. (Note that in this situation the distribution of the sample depends only on p and not on N.) Hence r/n is the uniformly minimum variance unbiased estimator of p. Now suppose that the N balls are distinguishable from one another (as for example when the balls are villages) or suppose we put distinguishing marks on the balls drawn before they are replaced. The sample observation then is a sequence of n balls and there are N^n possible sample observations each having the same probability. (In the previous case the probability of getting a particular sample observation was $p^r(1-p)^{n-r}$ where r is the number of white balls in the sample). Here the sample observation is more detailed than in the previous case and actually contains more information about the parameter p. Now r is no longer a sufficient statistic. The ν distinct balls that came in the sample is a sufficient statistic and nothing less than this can be sufficient. Consider now a third kind of sample observation where for each of the n balls that are drawn we note down only its colour and the fact whether this particular ball was drawn before or not. Here the sample can be represented as a sequence of n whites and blacks with cross marks at ν places (ν a variable) to indicate at which draws we had the distinct balls. The sample observation now is more detailed than the first case and less so than the second. If ρ be the number of distinct white balls then the statistic (ρ, ν) is sufficient. The conditional expectation of r/n for fixed values of (ρ, ν) is ρ/ν and so by the Rao-Blackwell theorem ρ/ν is better than r/n. Here the statistic (ρ, ν) though sufficient is not complete. This is obvious from the fact that the distribution of ν is independent of the parameter p. Thus

[1]The distribution of r is complete if there are at least $n+1$ admissible values for p. Thus if N be smaller than n then the distribution of r will not be complete.

we are unable[1] to prove that ρ/ν is the best unbiased estimator of p. The proof sketched above only demonstrators that, with the additional information of which are the distinct units, the standard estimator r/n is no longer the best estimator and that it is in fact worse than ρ/ν. In the next section we give a proof of inequality (4.2) in the general case.

5. PROOF OF (4.2.)

Let there be N population units and let Y_j be the Y-characteristic of the j-th population unit ($j = 1, 2, \ldots, N$). A sample $\mathbf{S}$ of size n is drawn one by one, with equal probabilities and with replacements. Let y_i be the observed Y-characteristic of the i-th sample unit ($i = 1, 2, \ldots, n$). For each sample unit suppose we also note down its unit index (if a particular sample unit happens to be the j-th population unit then its unit-index is j). Let u_i be the unit-index of the i-th sample unit and let $\mathbf{x}_i = (y_i, u_i)$. We can then record the sample observation as

$$\mathbf{S} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

where the $\mathbf{x}_i$'s are independently and identically distributed random vectors.

Let ν be the number of distinct sample units and let $u_{(1)} < u_{(2)} < \ldots < u_{(\nu)}$ be their unit-indices written in an ascending order. Let $y_{(i)}$ be the Y-characteristic of the sample unit with unit-index $u_{(i)}$ and let $\mathbf{x}_{(i)} = (y_{(i)}, u_{(i)})$ $i = 1, 2, \ldots, \nu$.

Consider now the set of 'order-statistics'

$$\mathbf{T} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)} \ldots, \mathbf{x}_{(\nu)}).$$

The usual estimator of $\bar{Y}$ (the population mean) is based on all the n observations and is

$$\bar{y} = \bar{y}(\mathbf{S}) = n^{-1} \sum_1^n y_i$$

whereas the estimator based on the ν distinct units is

$$\bar{y}^* = \bar{y}^*(\mathbf{T}) = \nu^{-1} \sum_1^\nu y_{(i)}.$$

Now, for fixed $\mathbf{T}$, the conditional distribution of $\mathbf{x}_i$ is concentrated at the ν points $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(\nu)}$ with equal probabilities at each of these points.

$$\therefore \quad E(y_i | \mathbf{T}) = \bar{y}^* \quad (i = 1, 2, \ldots, n)$$

and hence

$$E(\bar{y} | \mathbf{T}) = \bar{y}^*.$$

Since $\bar{y}$ is an unbiased estimator of $\bar{Y}$, it follows at once that $\bar{y}^*$ is also unbiased. It also follows that, for any convex (downwards) loss function, $\bar{y}^*$ has a uniformly better risk function than $\bar{y}$. In particular

$$V(\bar{y}^*) \leqslant V(\bar{y})$$

the sign of equality holding only when $n = 1$.

Thus the inequality (4.2) is proved. We may note in passing that $\mathbf{T}$ is a sufficient statistic here though not a complete one. No uniformly best unbiased estimator of $\bar{Y}$ exists.

---

[1] If the parameter N is also unknown then it is easily demonstrated that the distribution of ν is complete. In this situtation we believe that ρ/ν is the best unbiased estimator of p.

6. The case when ν is fixed in advance

In the previous sections we fixed n and had ν as a random variable. Here we consider the situtation where the number ν of distinct units in the sample is fixed in advance. We go on drawing samples one by one, with equal probabilities, and with replacements until we get n distinct units. The probability distribution of the number n of samples drawn may be obtained as follows. The event $n = k$ means that in the first $k-1$ draws there are exactly $\nu-1$ distinct units and that the k-th unit drawn is different from the $\nu-1$ distinct units that appeared in the first $k-1$ cases. Thus from (2.1) it follows that

$$P(n = k) = [N^{-(k-1)} \binom{N}{\nu-1} \Delta^{\nu-1} O^{k-1}] \left(1 - \frac{\nu-1}{N}\right) = \binom{N-1}{\nu-1} \Delta^{\nu-1} \left(\frac{x}{N}\right)^{k-1}\Big|_{x=0} \quad \ldots \quad (6.1)$$

where $k = \nu, \nu+1, \ldots$ ad inf.

From (6.1) it follows that the probability generating function of n is

$$P_n(t) = E\, t^n = \sum_{k=\nu}^{\infty} t^k P(n = k) = \binom{N-1}{\nu-1} t \sum_{k=\nu}^{\infty} \Delta^{\nu-1} \left(\frac{xt}{N}\right)^{k-1}\Big|_{x=0}$$

where Δ operates on x.

Since $\Delta^{\nu-1} x^r = 0$ for $r < \nu-1$ we have

$$P_n(t) = \binom{N-1}{\nu-1} t\Delta^{\nu-1} \left[\sum_{k=1}^{\infty} \left(\frac{xt}{N}\right)^{k-1}\right]_{x=0} = \binom{N-1}{\nu-1} t\Delta^{\nu-1} \left(1 - \frac{xt}{N}\right)^{-1}\Big|_{x=0} \quad \ldots \quad (6.2)$$

In a like manner we have

$$E(n) = \binom{N-1}{\nu-1} \Delta^{\nu-1} \left(1 - \frac{x}{N}\right)^{-2}\Big|_{x=0} \quad \ldots \quad (6.3)$$

If $\bar{y}$ be the average Y-characteristic of all the n observations then $\bar{y}$ is an unbiased estimator of $\bar{Y}$ with variance

$$V(\bar{y}) = E\frac{\sigma^2}{n} = \sigma^2 \sum_{k=1}^{\infty} \frac{1}{k} P(n = k)$$

$$= \sigma^2 \binom{N-1}{\nu-1} \Delta^{\nu-1} \sum_{k=1}^{\infty} \frac{1}{k} \left(\frac{x}{N}\right)^{k-1}\Big|_{x=0}$$

$$= \sigma^2 \binom{N-1}{\nu-1} \Delta^{\nu-1} \left[\frac{N}{x} \log \frac{N}{N-x}\right]_{x=0} \quad \ldots \quad (6.4)$$

where, for $x = 0$, $\frac{N}{x} \log \frac{N}{N-x}$ is to be interpreted as 1.

If $\bar{y}^*$ be the average Y-characteristics of the ν distinct units then

$$V(\bar{y}^*) = \frac{N-\nu}{N-1} \frac{\sigma^2}{\nu}. \quad \ldots \quad (6.5)$$

That (6.5) is smaller than (6.4) may be proved in precisely the same way as we proved a similar result in §5.

7. SAMPLING WITH UNEQUAL PROBABILITIES

We consider now the more general situation where sampling is done with different probabilities attached to the different population units. Let P_j be the probability associated with the j-th population unit ($\Sigma P_j=1$). Suppose a sample of size n is taken with replacement and with the P_j's as the probabilities. Let us suppose that for the i-th sample unit we record its Y-characteristic y_i, its probability of selection p_i, and its unit-index $u_i (i = 1, 2, ..., n)$. Thus, the sample is

$$\mathbf{S} = (\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_n)$$

where

$$\mathbf{x}_i = (y_i, p_i, u_i) \quad (i = 1, 2, ..., n).$$

Clearly the $\mathbf{x}_i$'s are independently and identically distributed random vectors.

As in §5 let us define ν as the number of distinct x_i's and as before let $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots \mathbf{x}_{(\nu)}$ be an arrangement of the ν distinct $\mathbf{x}_i$'s in ascending order of their unit-indices. Let

$$\mathbf{T} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, ..., \mathbf{x}_{(\nu)}).$$

It is easily seen that in this case also $\mathbf{T}$ is a sufficient statistic. Given $\mathbf{T}$, there are only

$$n!\ \Sigma'\ (\alpha_1!\ \ \alpha_2!\ \ \ldots.\ \ \alpha_\nu!)^{-1}$$

(where the summation is taken over all positive integer α_i's such that $\alpha_1+ \ldots +\alpha_\nu = n$) values that $\mathbf{S}$ may take and the probability for each of these can be computed from the information $\mathbf{T}$ alone. Thus, any admissible estimator of $\bar{Y}$ must necessarily be a function of the statistic $\mathbf{T}$ alone. The usual estimator $\bar{y}$ of $\bar{Y}$ is based on all the n observations and is

$$\bar{y} = \bar{y}(\mathbf{S}) = \frac{1}{n} \sum_1^n \left(\frac{y_i}{Np_i} \right). \qquad \ldots \quad (7.1)$$

From the sufficiency of $\mathbf{T}$ it follows that

$$\bar{y}^* = E(\bar{y} \mid \mathbf{T}) = E \left(\frac{y_1}{Np_1} \middle| \mathbf{T} \right) \qquad \ldots \quad (7.2)$$

is a better unbiased estimator of $\bar{Y}$.

It is not difficult to show that

$$P(\mathbf{x}_1 = \mathbf{x}_{(i)} \mid \mathbf{T}) = \frac{p_{(i)} \sum'' \frac{(n-1)!}{\alpha_1!\ \alpha_2! \ldots \alpha_\nu!}\ p_{(1)}^{\alpha_1}\ p_{(2)}^{\alpha_2} \cdots p_{(\nu)}^{\alpha_\nu}}{\sum' \frac{n!}{\alpha_1! \alpha_2! \ldots \alpha_\nu!}\ p_{(1)}^{\alpha_1}\ p_{(2)}^{\alpha_2} \cdots p_{(\nu)}^{\alpha_\nu}}$$

$$= c_i (\text{say}) \quad (i = 1, 2, ..., \nu)$$

where Σ' means summation over all integral α's such that

$$\alpha_1+\alpha_2+\ldots+\alpha_\nu = n \text{ and } \alpha_k > 0 \text{ for } k = 1, 2, ..., \nu$$

and Σ'' means summation over all integral α's such that

$$\alpha_1+\alpha_2+\ldots+\alpha_\nu = n-1, \alpha_i \geqslant 0 \text{ and } \alpha_k > 0 \text{ for } k \neq i.$$

Thus

$$\bar{y}^* = E\left\{ \frac{y_1}{Np_1} \middle| \mathbf{T} \right\} = \sum_i c_i \frac{y_{(i)}}{Np_{(i)}}. \qquad \ldots \quad (7.3)$$

Unfortunately, it is rather troublesome to compute the c_i's. In the particular case where $n = 3$ and $\nu = 2$ we have

$$c_1 = (2p_{(1)}+p_{(2)})/3(p_{(1)}+p_{(2)})$$

and

$$c_2 = (p_{(1)}+2p_{(2)})/3(p_{(1)}+p_{(2)}).$$

The estimator $\bar{y}^*$, though demonstrated to be superior to the usual estimator $\bar{y}$, cannot be of much use for large scale sample surveys. It is even more troublesome to estimate the variance of $\bar{y}^*$. An unbiased estimator for the variance of $\bar{y}$ is

$$\frac{1}{n(n-1)} \sum_{1}^{n} \left(\frac{y_i}{Np_i} - \bar{y} \right)^2. \qquad \ldots \quad (7.4)$$

The above will over-estimate the variance of $\bar{y}^*$ and so it will be on the safe side to take (7.4) as an estimator of the variance of $\bar{y}^*$.

8. The mean and variance of ν

In § 2 we have given the distribution of ν for the particular case where sampling is done with equal probabilities. In the unequal probability set-up the distribution of ν becomes very messy. Here we give expressions for the mean and variance of ν.

Let z_j be the characteristic function of the event that the sample of n units includes the j-th population unit.

Clearly

$$P(z_j = 1) = 1 - Q_j^n \quad (j = 1, 2, \ldots, N)$$

where

$$Q_j = 1 - P_j.$$

Since

$$\nu = z_1 + z_2 + \ldots + z_N \qquad \ldots \quad (8.1)$$

we have

$$E(\nu) = \sum_{1}^{N} (1 - Q_j^n). \qquad \ldots \quad (8.2)$$

Also

$$V(\nu) = \Sigma\, V(z_j) + \sum_{i \neq j} \operatorname{cov}(z_i, z_j). \qquad \ldots \quad (8.3)$$

Now,

$$V(z_j) = Q_j^n (1 - Q_j^n)$$

and

$$\begin{aligned} \operatorname{cov}(z_i, z_j) &= P(z_i = z_j = 1) - P(z_i = 1) P(z_j = 1) \\ &= (1 - Q_i^n - Q_j^n + Q_{ij}^n) - (1 - Q_i^n)(1 - Q_j^n) \\ &= -(Q_i^n\ Q_j^n - Q_{ij}^n) \end{aligned}$$

where

$$Q_{ij} = 1 - P_i - P_j$$

$$\therefore \quad V(\nu) = \Sigma\, Q^n(1 - Q^n) - \sum_{i \neq j} (Q^n Q^n - Q_{ij}^n) = \Sigma\, Q_i^n - (\Sigma\, Q_i^n)^2 + \sum_{i \neq j} Q_{ij}^n. \qquad \ldots \quad (8.4)$$

9. Unequal probabilities and without replacement

Now let us consider the case of sampling without replacement and with different probabilities. As in §7 let P_j be the probability attached to the j-th population unit. Let y_i and p_i be the Y-characteristic and the probability corresponding to the i-th sample unit ($i = 1, 2, \ldots, n$). Writing $\mathbf{x}_i = (y_i, p_i)$ we can record the sample observation[1] as

$$\mathbf{S} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n). \qquad \ldots \quad (9.1)$$

[1]Here we need not record the unit indices of the sample units as we know that they must be all different. Unless we have some additional information about the population units it appears that it is impossible to utilise any information about the sample unit-indices to improve on any estimator of $\bar{Y}$.

Now, let us order the $\mathbf{x}_i$'s by some method, say, in ascending order of the y_i's and for $\mathbf{x}_i$'s with equal y_i's, in ascending order of their p_i's. Let $\mathbf{x}_{(i)}$ be the i-th order statistic and let

$$\mathbf{T} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}) \qquad \ldots \quad (9.2)$$

be the set of order statistics.

For given $\mathbf{T}$, there are $n!$ or less (in the case where some of the $\mathbf{x}_{(i)}$'s are the same) values that $\mathbf{S}$ may take and the conditional probability for each of these may be computed from the information $\mathbf{T}$ alone. Thus, in this case also $\mathbf{T}$ is a sufficient statistic. Hence no estimator that is not a function of $\mathbf{T}$ alone can be admissible. Any estimator that makes use of the order in which the sample was drawn can be uniformly improved upon by its conditional expectation given $\mathbf{T}$.

For example, consider the particular case of $n = 2$. We may estimate $\bar{Y}$ from the first component of $\mathbf{S}$, i.e., from $\mathbf{x}_1$. This estimator is, of course, $y_1 | Np_1$.

Now
$$P(\mathbf{x}_1 = \mathbf{x}_{(1)} | \mathbf{T}) = \frac{p_{(1)}\, p_{(2)}/(1-p_{(1)})}{p_{(1)}\, p_{(2)}/(1-p_{(1)}) + p_{(2)}\, p_{(1)}/(1-p_{(2)})} = \frac{1-p_{(2)}}{2-p_{(1)}-p_{(2)}} \qquad \ldots \quad (9.3)$$

and similarly

$$P(\mathbf{x}_1 = \mathbf{x}_{(2)} | \mathbf{T}) = \frac{1-p_{(1)}}{2-p_{(1)}-p_{(2)}} \qquad \ldots \quad (9.4)$$

$$\therefore \quad E\left(\frac{y_1}{Np_1} \,\middle|\, \mathbf{T}\right) = \left[(1-p_{(2)})\ \frac{y_{(1)}}{Np_{(1)}} + (1-p_{(1)})\frac{y_{(2)}}{Np_{(2)}}\right] \times$$

$$\times (2-p_{(1)}-p_{(2)})^{-1}. \qquad \ldots \quad (9.5)$$

The estimator (9.5) is uniformly better than $y_1 | Np_1$.

Since $\mathbf{T}$ is not a complete sufficient statistic, we cannot prove that (9.5) is the uniformly best estimator. For further discussion about the problem dealt with in this section one may refer to (Murthy, 1957).

ACKNOWLEDGEMENT

The author wishes to thank Mr. S. Raja Rao who suggested the problem in 1952 and then collaborated with the author in obtaining the results contained in § 2 to § 6.

REFERENCES

FELLER, W. (1957) : *An introduction to probability theory and its applications*, John Wiley & Sons New York.

FRASER, D. A. S. (1957) : *Non parametric methods in statistics*, John Wiley & Sons New York.

MURTHY, M. N. (1957) : Ordered and unordered estimators in sampling without replacement. *Sankhyā*, **18**, 379.

Paper received : January, 1958.

THE FAMILY OF ANCILLARY STATISTICS

By D. BASU

Indian Statistical Institute, Calcutta

SUMMARY. Though the marginal distributions of the ancillary statistics are independent of the parameter they are not useless or informationless. A set of ancillaries may sometimes summarise the whole of the information contained in the sample. A classification of the ancillaries in terms of the partial order of their information content is attempted here. In general there are many maximal ancillaries. Among the minimal ancillaries there exists a unique largest one. When there exists a complete sufficient statistic, the problem of tracking down the maximal and minimal ancillaries becomes greatly simplified.

1. Introduction

An ancillary[1] statistic is one whose distribution is the same for all possible values of the unknown parameter. A statistic that is not ancillary may be called 'informative'. The classical example of an ancillary statistic is the following :

Example (a) : Let X and Y be two positive valued random variables with the joint density function

$$f(x, y) = e^{-\theta x - \frac{y}{\theta}}, \; x > 0, \; y > 0, \; \theta > 0.$$

Here $F = XY$ is an ancillary statistic. The maximum likelihood estimator $T = \sqrt{Y/X}$ of θ is not a sufficient statistic. However, the pair (F, T) is jointly sufficient.

The above example shows that though an ancillary statistic, by itself, fails to provide any information about the parameter, yet in conjunction with another statistic—which, as we shall presently see, need not be informative—may supply valuable information[2] about the parameter. In the following example we have given a family of ancillary statistics that are jointly equivalent to the whole sample.

Example (b) : Let X and Y be independent normal variables with unknown means θ and unit standard deviations. Here $X-Y$ is an ancillary statistic. It is commonly believed that every ancillary statistic (in this situation) is necessarily a function of $X-Y$. That, however, is not true.

Let

$$F_c = F_c(X, \; Y) = \begin{cases} X-Y & \text{if} \quad X+Y < c \\ Y-X & \text{if} \quad X+Y \geqslant c \end{cases}$$

where c is a fixed constant.

[1] The name 'ancillary' is due to Fisher (1925). The name 'distribution-free' is also in use and perhaps would have been more appropriate in the present context.

[2] See Fisher (1956) for a discussion of how the ancillary information may (according to Fisher) be recovered.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_18, © Springer Science+Business Media, LLC 2011

Since $X-Y$ and $Y-X$ are identically distributed and each is independent of $X+Y$ it at once follows that F_c is independent of $X+Y$ and has the same distribution as that of $X-Y$. Thus, F_c is ancillary for each c. Consider now the family $\{F_c\}$, $-\infty < c < \infty$, of ancillary statistics. For fixed X and Y the different values of F_c (for varying c) are either $X-Y$ or $Y-X$. The value c_0 of c where F_c changes sign (F_c does not change sign only if $X-Y = 0$ and that is a null event) is the value of $X+Y$. Thus, given $F_c(X, Y)$ for all c we can find $X+Y$ and $X-Y$. Hence, the family $\{F_c\}$ of ancillary statistics is equivalent to the whole sample (X, Y). The countable family $\{F_c\}$ where c runs through the set of rational numbers is easily seen to be also equivalent to (X, Y).

The author (Basu ; 1955, 1958) has shown that, under very mild restrictions, any statistic independent of a sufficient statistic is ancillary and that the converse proposition is also true, provided the sufficient statistic is complete.

In Example (b) the statistic $T = X+Y$ is a complete sufficient statistic. A statistic F can, therefore, be ancillary if and only if F is independent of T. The following is a general method for constructing statistics independent of T. Start with any ancillary statistic F. In general, there will be many measure-preserving transformations of F (i.e. a mapping φ of the range space of F into itself such that $\varphi(F)$ and F are identically distributed). For each real t, define a measure-preserving transformation φ_t of F. Then, take the statistic $\varphi_T(F)$. Subject to some measurability restrictions, $\varphi_T(F)$ will be independent of T and hence will be ancillary. In Example (b) we took $F = X-Y$ and $\varphi_t(F) = F$ or $-F$ according as $t < c$ or $\geqslant c$.

If a statistic F is ancillary then every (measurable) function of F is also ancillary. The statistic F_2 is said to include (or be more informative than) the statistic F_1 if F_1 can be expressed as a function of F_2. In this case we write $F_2 \supset F_1$ or $F_1 \subset F_2$. Two statistics are said to be equivalent if each can be expressed as a function of the other.

Example (c) : Let $X_1, X_2, \ldots X_n$ be n independent observations on a normal variable with mean θ and s.d. unity. Then each of the $n-1$ statistics

$$F_1 = X_1-X_2,\ F_2 = (X_1-X_2,\ X_1-X_3),\ \ldots\ F_{n-1} = (X_1-X_2, X_1-X_3, \ldots X_1-X_n)$$

is ancillary and

$$F_1 \subset F_2 \subset \cdots \subset F_{n-1}$$

The two ancillary statistics F_{n-1} and $F = (X_2-X_1,\ X_2-X_3, \ldots X_2-X_n)$ are easily seen to be equivalent.

From Example (b) it is obvious that F_{n-1} does not include all ancillary statistics.

An ancillary statistic M is said to be 'maximal' if there exists no non-equivalent ancillary M^* such that $M \subset M^*$. Thus, given any ancillary F, either it is maximal or there exists an ancillary $F^* \supset F$. Given any ancillary F_0, there exists (Theorem 2)

a maximal ancillary $M \supset F_0$. In general there exists many non-equivalent maximal ancillaries. A typical property (Cor. to Theorem 4) of a maximal ancillary M is that, for any ancillary F not included in M, the pair (M, F) is informative.

A minimal ancillary is one that is included in every maximal ancillary. Among the class of minimal ancillaries there exists (Theorem 5) a unique largest one G_0. In the absence of a better name we prefer to call G_0 the laminal ancillary. G_0 includes every minimal ancillary and is included in every maximal ancillary. A typical property (Theorem 6) of a minimal ancillary G is that, for any ancillary F, the pair (G, F) is ancillary.

If there exists a complete sufficient statistic G, then, any ancillary statistic F, such that the pair (G, F) is essentially equivalent to the whole sample, is shown (Theorem 7) to be essentially maximal. Under some further restrictions, the laminal ancillary is shown (Theorem 8) to be essentially equivalent to a constant.

In the following sections we elaborate on the above sketch of the family-tree of ancillary statistics. For the sake of elegance and brevity of exposition we use the language of sub σ-fields. Reference may be made to Bahadur (1954, 1955) for excellent expositions of the sub σ-field approach.

2. Definitions

Let $(\mathcal{X}, \mathcal{B})$ be an arbitrary measurable space and let $\{P_\theta\}$, $\theta \in \Omega$ be a family of probability measures on $\mathcal{B}$. Any statistic T induces a sub σ-field $\mathcal{B}_T \subset \mathcal{B}$. Instead of dealing with statistics it is more convenient (in the present context) to deal with the corresponding sub σ-fields.

Definition 1: The event $A \in \mathcal{B}$ is said to be ancillary if $P_\theta(A)$ is the same for all $\theta \in \Omega$. The family of all ancillary events is denoted by $\mathcal{A}$.

It is easy to check that the family $\mathcal{A}$ is closed for complementation and countable disjoint unions. However, in general $\mathcal{A}$ is not closed for intersection (i.e. $\mathcal{A}$ is not a σ-field).

In order to show that the family $\mathcal{A}$ in Example (b) do not constitute a σ-field, we have only to check that

$$P_\theta\,[X-Y>0 \text{ and } F_c\,(X, Y)>0] = P_\theta\,(X-Y>0 \text{ and } X+Y<c)$$

$$= \tfrac{1}{2} \int_{-\infty}^{c} \frac{1}{\sqrt{4\pi}}\, e^{-\frac{1}{4}(x-2\theta)^2}\, dx$$

which varies with θ.

In Example (b) the Borel-extension of $\mathcal{A}$ is $\mathcal{B}$.

Example (d) : Let $\mathcal{X}$ consist of the three points a, b and c and let the corresponding probability measures be $\frac{1}{4}-\theta$, $\frac{1}{2}$, and $\frac{1}{4}+\theta$ respectively, where $0 < \theta < \frac{1}{4}$. Here $\mathcal{A}$ consists of the four sets ϕ, $[b]$, $[a, c]$ and $\mathcal{X}$ and so $\mathcal{A}$ is a sub σ-field of $\mathcal{B}$.

Definition 2 : A σ-field $\mathcal{F}$ is said to be ancillary if $\mathcal{F} \subset \mathcal{A}$. A σ-field that is not ancillary is called informative.

A statistic is ancillary or informative according as the corresponding σ-field is so.

Definition 3: Two ancillary sets A and B are said to conform if AB is also ancillary. If A conforms to B then we write $A \sim B$. Since $P_\theta(AB)+P_\theta(AB') = P_\theta(A)$ it follows that $A \sim B$ if and only if $A \sim B'$.

If A conforms to every one of a sequence of disjoint sets $B_1, B_2 \ldots$ then it is easy to check that $A \sim \bigcup B_i$.

Definition 4 : Let Γ_0 be the family of all ancillary sets B such that $B \sim A$ for all $A \in \mathcal{A}$.

Clearly ϕ and $\mathcal{X}$ belong to Γ_0. From what we have said before it follows that Γ_0 is closed for complementation and countable disjoint unions.

Theorem 1 : *The family* Γ_0 *is a* σ*-field.*

Proof : It is enough to show that Γ_0 is closed for intersection. Let B_1 and B_2 both belong to Γ_0 and let $A \in \mathcal{A}$. From $B_2 \in \Gamma_0$ it follows that $B_2A \in \mathcal{A}$. From $B_1 \in \Gamma_0$ it then follows that $B_1B_2A \in \mathcal{A}$. Since A is an arbitrary ancillary set, it follows that $B_1B_2 \in \Gamma_0$.

We shall later on see that the ancillary σ-field Γ_0 corresponds to the laminal ancillary G_0 that we have referred to in §1.

The family $\mathcal{A}$ of ancillary sets is a σ-field if and only if every pair of ancillary sets conform to one another, i.e. if $\mathcal{A} = \Gamma_0$.

Example (e) : Let $\mathcal{X}$ consist of the five points a, b, c, d and e with the corresponding probabilities $\frac{1}{2}$, θ, θ, $\frac{1}{4}-\theta$ and $\frac{1}{4}-\theta$ respectively, where $0 < \theta < \frac{1}{4}$. In this case Γ_0 consists of the four sets ϕ, [a], $[b, c, d, e]$, and $\mathcal{X}$. The two sets $[b, d]$ and $[b, e]$ are both ancillary but they do not conform. Here $\mathcal{A}$ is wider than Γ_0 and is not a σ-field.

Definition 5 : The ancillary σ-field $\mathcal{F}_2$ is said to include the ancillary σ-field $\mathcal{F}_1$ (in symbols $\mathcal{F}_2 \supset \mathcal{F}_1$ or $\mathcal{F}_1 \subset \mathcal{F}_2$) if every element of $\mathcal{F}_1$ is an element of $\mathcal{F}_2$.

The above partial order on ancillary σ-fields corresponds to the inclusion relationship for ancillary statistics.

Definition 6 : The ancillary σ-field $\mathcal{M}$ is said to be maximal if there exists no other ancillary σ-field $\mathcal{M}^*$ such that $\mathcal{M}^* \supset \mathcal{M}$.

Definition 7 : The intersection of all the maximal ancillary σ-fields is called the laminal ancillary.

The laminal ancillary is the largest ancillary that is included in all maximal ancillaries.

3. Existence and characterizations of maximal and laminal ancillaries

The following theorem is fundamental.

Theorem 2 : *Given any ancillary σ-field $\mathcal{F}_0$ there exists a maximal ancillary σ-field $\mathcal{M} \supset \mathcal{F}_0$.*

Proof : We first prove that given any family $\{\mathcal{F}_j\}$, $j \in J$ of ancillary σ-fields that are linearly ordered (by the inclusion relationship), the Borel-extension $\mathcal{F}$ of $\bigcup \mathcal{F}_j$ is also ancillary.

Clearly, $\bigcup \mathcal{F}_j$ contains ϕ and $\mathcal{X}$ and is closed for complementation. Since $\{\mathcal{F}_j\}$ is linearly ordered it follows that $\bigcup \mathcal{F}_j$ is also closed for finite unions. That is, $\bigcup \mathcal{F}_j$ is a field of sets.

Since each $\mathcal{F}_j$ is ancillary, the restriction of P_θ to $\bigcup \mathcal{F}_j$ is a measure Q that does not depend on θ. From the fundamental Extension Theorem of measures (Kolmogorov, 1933) we know that the extension of Q to $\mathcal{F}$ is unique.

It follows at once that the restriction of P_θ to $\mathcal{F}$ is the same for all θ, i.e. $\mathcal{F}$ is an ancillary σ-field.

Now let $\mathcal{C}$ be the family of all ancillary σ-fields that include $\mathcal{F}_0$. Since corresponding to any linearly ordered sub-family of $\mathcal{C}$ there exists an ancillary σ-field that includes every member of the sub-family it follows from Zorn's Lemma that $\mathcal{C}$ has a maximal element.

Let $\{\mathcal{M}_i\}$, $i \in I$ be the family of all maximal ancillary σ-fields. We at once have the

Theorem 3 : $\mathcal{A} = \bigcup \mathcal{M}_i$

Proof : We have only to note that corresponding to any element A of $\mathcal{A}$ there exists an ancillary σ-field that contains A as an element and then apply Theorem 2.

Corollary : *If $\{\mathcal{M}_i\}$ consists of only one σ-field $\mathcal{M}_0$ then $\mathcal{A} = \mathcal{M}_0 = \Gamma_0$.*

Thus, in any situation where there are non-conforming ancillary sets, the family $\{\mathcal{M}_i\}$ has at least two members.

In Example (d) there is a unique maximal ancillary. In Example (e) there are exactly two maximal ancillaries namely :

$\mathcal{M}_1$ = the σ-field spanned by $[a]$ and $[b, d]$

and $\mathcal{M}_2$ = the σ-field spanned by $[a]$ and $[b, e]$.

Theorem 4 : *If the ancillary set A does not belong to the maximal ancillary $\mathcal{M}$ then A does not conform to at least one element of $\mathcal{M}$.*

Proof: Suppose on the contrary that A cnforms to every element of $\mathcal{M}$. Consider the family $\mathcal{M}^*$ of sets $AX \bigcup A'Y$ where X and Y are arbitrary elements of $\mathcal{M}$. Clearly $\mathcal{M} \subset \mathcal{M}^*$ but not conversely.

Since $$(AX \bigcup A'Y)' = AX' \bigcup A'Y',$$

and $$\bigcup (AX_i \bigcup A'Y_i) = A(\bigcup X_i) \bigcup A'(\bigcup Y_i)$$

and $$P_\theta (AX \bigcup A'Y) = P_\theta (AX) + P_\theta (A'Y),$$

it follows that $\mathcal{M}^*$ is also an ancillary σ-field.

This, however, contradicts the maximality of $\mathcal{M}$.

Corollary : *If $\mathcal{M}$ be any maximal ancillary and if the ancillary σ-field $\mathcal{F}$ is not included in $\mathcal{M}$ then the smallest σ-field containing both $\mathcal{M}$ and $\mathcal{F}$ is informative.*

Theorem 5 : $\bigcap \mathcal{M}_i = \Gamma_0$

Proof: Since every element of Γ_0 conforms (by definition) to every ancillary event, if follows from Theorem 4 that $\Gamma_0 \subset \mathcal{M}_i$ for all i, i.e. $\Gamma_0 \subset (\bigcap \mathcal{M}_i)$.

Now let $B \in \bigcap \mathcal{M}_i$ and A be an arbitrary ancillary set. From Theorem 3 it follows that $A \in \mathcal{M}_i$ for some i.

Hence B and A are together as elements of some $\mathcal{M}_i$ and so $B \sim A$.

Since A is arbitrary it follows that $B \in \Gamma_0$.

$\therefore$ $(\bigcap \mathcal{M}_i) \subset \Gamma_0$ and so the equality is proved.

Theorem 6 : *For any ancillary σ-field $\mathcal{F}$ the smallest σ-field containing both $\mathcal{F}$ and Γ_0 is also ancillary.*

Proof: Consider the family of 'rectangular' sets $X \bigcap Y$ where $X \in \mathcal{F}$ and $Y \in \Gamma_0$. From the definition of Γ_0 it follows that all such sets are ancillary and that they conform to one another. The family of sets that may be formed by finite unions of rectangular sets form a field of sets and each of them is ancillary. The rest follows from the Extension Theorem of Measures.

4. WHEN A COMPLETE SUFFICIENT STATISTIC EXISTS

In general there exist many maximal ancillaries. For instance, in Example (b) there are uncountably many maximal ancillaries. In order to see this, let us consider the family $\{A_c\}$ of ancillary events where $A_c = \{(X, Y) \mid F_c(X, Y) > 0\}$. If $c < d$, then

$$P_\theta (A_c A_d) = P_\theta(X - Y > 0 \text{ and } X + Y < c) + P_\theta(Y - X > 0 \text{ and } X + Y \geqslant d)$$

$$= \tfrac{1}{2} \left[1 - \int_c^d \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{4}(x-2\theta)^2} dx \right]$$

which varies with θ.

Thus, the members of the family $\{A_c\}$ of ancillary sets are mutually non-conforming. Hence the maximal ancillaries including the different members of the family are all different.

Though there may exist many maximal ancillaries, it is not, in general, easy to prove the maximality of a particular ancillary. However, in the situations where we have a complete sufficient statistic, it is rather easy to demonstrate the maximality of a large class of ancillaries.

The following property of complete sufficient statistics is useful.[1] Here we state and prove the result in terms of σ-fields.

Lemma (Basu, 1955) : *If $\mathcal{G} \subset \mathcal{B}$ be a boundedly complete sufficient σ-field and A any ancillary event, then A is independent of $\mathcal{G}$.*

Proof : Let $\varphi = P(A \mid \mathcal{G})$ be the conditional probability of A given $\mathcal{G}$. That is, φ is a $\mathcal{G}$-measurable function such that

$$P_\theta(AG) = \int_G \varphi dP_\theta \text{ for all } \theta \epsilon\, \Omega \text{ and } G \epsilon\, \mathcal{G}.$$

Since $\mathcal{G}$ is sufficient, it follows that φ may be chosen to be independent of θ. Also the set of x's for which $\varphi(x)$ lies outside the interval (0, 1) is of zero-measure for each $\theta \epsilon\, \Omega$.

Taking $G = \mathcal{X}$ we have

$$P_\theta\,(A) = \int_{\mathcal{X}} \varphi\, dP_\theta \quad \text{for all } \theta \epsilon\, \Omega.$$

Since $P_\theta\,(A)$ is independent of θ and φ is $\mathcal{G}$-measurable, it follows from the bounded completeness of $\mathcal{G}$ that $\varphi = P_\theta\,(A)$ almost surely for all $\theta \epsilon\, \Omega$.

$$\therefore \quad P_\theta\,(AG) = \int_G \varphi dP_\theta$$

$$= P_\theta\,(A) P_\theta\,(G) \text{ for all } \theta \epsilon\, \Omega \text{ and } G \epsilon \mathcal{G}.$$

That is, A is independent of all $G \epsilon\, \mathcal{G}$

Before proceeding further we need a slightly wider definition of maximality for an ancillary σ-field.

Definition 8 : The two $\mathcal{B}$-measurable sets A and B are said to be essentially equal if

$$P_\theta\,(A \Delta B) \equiv P_\theta\,(AB' \bigcup A'B)$$

$$= 0 \quad \text{for all } \theta \epsilon\, \Omega.$$

Definition 9 : Two sub σ-fields $\mathcal{F}_1$ and $\mathcal{F}_2$ are said to be essentially equivalent if corresponding to any set belonging to one of them there exists an essentially equal set belonging to the other.

Definition 10 : Any ancillary σ-field that is essentially equivalent to a maximal ancillary is called essentially maximal.

Theorem 7 : *If $\mathcal{G}$ be a boundedly complete sufficient σ-field then any ancillary $\mathcal{F}$ such that the Borel-extension of $\mathcal{G} \bigcup \mathcal{F}$ is essentially equivalent to $\mathcal{B}$, is essentially maximal.*

[1] See Basu (1955) and Hogg and Craig (1956) for some other interesting applications.

Proof : Let $\mathcal{M}$ be a maximal ancillary including $\mathcal{F}$ and let M be an arbitrary element of $\mathcal{M}$. For proving the essential maximality of $\mathcal{F}$ we have to establish the existence of an $F_0 \in \mathcal{F}$ such that F_0 is essentially equal to M.

Let $\mathcal{B}^*$ be the Borel extension of $\mathcal{F} \bigcup \mathcal{G}$. Since $\mathcal{B}^*$ is essentially equivalent to $\mathcal{B}$, there exists an $M^* \in \mathcal{B}^*$ such that M^* is essentially equal to M.

Since $M \in \mathcal{M} \supset \mathcal{F}$ and M^* is essentially equal to M, it follows that M^* is an ancillary set conforming to every $F \in \mathcal{F}$. Clearly, the two measures P and Q on $\mathcal{F}$, defined by the relations $P(F) = P_\theta(F)$ and $Q(F) = P_\theta\ (M^*F)$, are both independent of θ.

Therefore, the conditional probability function

$$\varphi = P_\theta(M^* | \mathcal{F}) = \frac{dQ}{dP}$$

is independent of θ.

Thus, φ is an $\mathcal{F}$-measurable function on $\mathcal{X}$ such that

$$P_\theta\ (M^*F) = \int_F \varphi dP_\theta \quad \text{for all} \quad \theta \in \Omega \text{ and } F \in \mathcal{F}.$$

Let F and G be typical elements of $\mathcal{F}$ and $\mathcal{G}$ respectively. Since $\mathcal{F}$ is ancillary and $\mathcal{G}$ is boundedly complete sufficient, it follows (from the Lemma) that $\mathcal{F}$ and $\mathcal{G}$ are independent.

$$\therefore \quad \int_{FG} \varphi dP_\theta = \int_{\mathcal{X}} (\varphi\ \mathcal{X}_F)\ \mathcal{X}_G\ dP_\theta \quad (\mathcal{X}_F \text{ and } \mathcal{X}_G \text{ are characteristic functions of } F \text{ and } G)$$

$$= \int_{\mathcal{X}} (\varphi\ \mathcal{X}_F)\ dP_\theta \int_{\mathcal{X}} \mathcal{X}_G\ dP_\theta \quad (\because \mathcal{F} \text{ and } \mathcal{G} \text{ are independent})$$

$$= P_\theta\ (M^*F)\ P_\theta\ (G) \qquad \ldots\ (\alpha)$$

Again, since $M^* \sim F$ it follows (from the Lemma) that M^*F is independent of G.

$$\therefore \quad \int_{FG} \mathcal{X}_{M^*} dP_\theta = P_\theta\ (M^*FG) = P_\theta\ (M^*F)\ P_\theta\ (G). \qquad \ldots\ (\beta)$$

From (α) and (β) we have

$$\int_{FG} (\varphi - \mathcal{X}_{M^*})\ dP_\theta = 0 \text{ for all } F \in \mathcal{F} \text{ and } G \in \mathcal{G}.$$

Since $\varphi - \mathcal{X}_{M^*}$ is $\mathcal{B}^*$-measurable it at once follows that

$$\int_B (\varphi - \mathcal{X}_{M^*})\ dP_\theta = 0 \text{ for all } B \in \mathcal{B}^*.$$

Therefore, for each $\theta \in \Omega$, $\varphi(x) - \mathcal{X}_{M^*}(x) = 0$ for almost all x.

Let $F_0 = \{x \mid \varphi(x) = 1\}$. Clearly $F_0 \in \mathcal{F}$ and is essentially equal to M^*.

Since M^* is essentially equal to M the Theorem is proved.

In Example (b), $X+Y$ is a complete sufficient statistic. Also for any fixed c, the pair $(X+Y, F_c)$ is equivalent to the sample (X, Y). Hence it follows that every F_c is an essentially maximal ancillary. In Example (c), the ancillary F_{n-1} together with the complete sufficient statistic $X_1+X_2+\ldots+X_n$ is equivalent to the whole sample and, therefore, is essentially maximal. A large number of similar situations are covered by Theorem 7.

Having partially settled the question of maximal ancillaries let us turn our attention to the laminal ancillary.

The laminal ancillary is the largest ancillary σ-field that is included in all maximal ancillaries. From Theorem 5 we have that the class Γ_0 of ancillary sets C that conform to every ancillary set is the laminal ancillary.

Let Λ be the family of sets that are essentially equal to either the empty set ϕ or the whole space $\mathcal{X}$. That is, Λ is the family of all sets E such that $P_\theta(E)$ is either $\equiv 0$ or $\equiv 1$ for all $\theta \in \Omega$. It is easy to check that Λ is a σ-field and that $\Lambda \subset \Gamma_0$. The following theorem covers a number of important cases.

Theorem 8 : *If the following conditions are satisfied then* $\Gamma_0 = \Lambda$.

i) $\mathcal{F}$ *is an essentially maximal ancillary.*

ii) *There exists an informative set* G *which is independent of* $\mathcal{F}$.

iii) *For every* $F \in \mathcal{F}$ *such that* $0 < P_\theta(F) < 1$ *there exists* $F^* \in \mathcal{F}$ *such that* $P_\theta(F^*) = P_\theta(F)$ and $P_\theta(FF^*) < P_\theta(F)$.

Proof : Let C be an arbitrary element of Γ_0. We have to prove that $P_\theta(C) = 0$ or 1. If possible let $0 < P_\theta(C) < 1$.

Now, $\mathcal{F}$ is essentially equivalent to a maximal ancillary and C belongs to every maximal ancillary. Hence, there exists $F \in \mathcal{F}$ which is essentially equal to C. Thus, F conforms to every ancillary set and $0 < P_\theta(F) < 1$.

Let G and F^* satisfy conditions (ii) and (iii) respectively and let $A = GF \cup G'F^*$. Since G is independent of $\mathcal{F}$, we have

$$\begin{aligned} P_\theta(A) &= P_\theta(G)P_\theta(F)+P_\theta(G')P_\theta(F^*) \\ &= P_\theta(F)[P_\theta(G)+P_\theta(G')] \\ &= P_\theta(F). \end{aligned}$$

That is, A is an ancillary set.

Now

$$AF = GF \cup G'(FF^*)$$

and, therefore,

$$\begin{aligned} P_\theta(AF) &= P_\theta(G)P_\theta(F)+P_\theta(G')P_\theta(FF^*) \\ &= P_\theta(FF^*)+P_\theta(G)[P_\theta(F)-P_\theta(FF^*)]. \end{aligned}$$

Let us note that $P_\theta(FF^*)$ and $P_\theta(F)-P_\theta(FF^*)$ are both independent of θ and that the latter is not zero. Again since G is informative $P_\theta(G)$ is not independent of θ. Hence AF is informative, which is a contradiction. Therefore, $P_\theta(C) = 0$ or 1, i.e. $C \in \Lambda$, which proves the theorem.

If the conditions of Theorem 7 are satisfied then $\mathcal{F}$ and any informative $G \in \mathcal{G}$ satisfies conditions (i) and (ii) of Theorem 8. We have then only to check whether condition (iii) is satisfied or not. If the restriction of P_θ to $\mathcal{F}$ be non-atomic then it is very easy to see that condition (iii) is also satisfied.

In Examples (b) and (c) the (essentially) maximal ancillaries have continuous (non-atomic) distributions and so Theorem 8 holds. Most of the familiar cases where a complete sufficient statistic exists fall under the above category.

Example (f) : Let X be a single observation on a normal variable with mean zero and standard deviation σ. Here X^2 is a complete sufficient statistic.

Let $$Y = \begin{cases} -1 & \text{of } X < 0 \\ 1 & \text{if } X \geqslant 0 \end{cases}$$

Here the pair (Y, X^2) is equivalent to the whole sample X.

$\therefore$ Y is an essentially maximal ancillary.

The sub σ-field generated by Y consists of the four sets ϕ, $(-\infty, 0)$, $[0, \infty)$ and $\mathcal{F}$. Condition (iii) of Theorem 8 is clearly satisfied. Therefore, the laminal ancillary Γ_0 is the same as Λ.

ACKNOWLEDGEMENT

I wish to thank Dr. R. R. Bahadur for some useful discussions.

REFERENCES

BAHADUR, R. R. (1954) : Sufficiency and statistical decision functions. *Ann. Math. Stat.*, **25**, 423.

——— (1955) : Statistics and subfields. *Ann. Math. Stat.*, **26**, 490.

BASU, D. (1955) : On statistics independent of a complete sufficient statistics. *Sankhyā*, **15**, 377.

——— (1958) : On statistics independent of a sufficient statistic. *Sankhyā*, **20**, 223.

FISHER, R. A. (1925) : Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700.

——— (1956) : *Statistical Methods and Scientific Inference*, Oliver and Boyd, London.

HOGG, R. V. and CRAIG, A. T. (1956) : Sufficient statistics in elementary distribution theory. *Sankhyā*, 17, 209.

KOLMOGOROV, A. N. (1933) : *Foundations of The Theory of Probability*, Chelsea Publishing Company, New York.

Paper received : July, 1959.

RECOVERY OF ANCILLARY INFORMATION*

By D. BASU

Indian Statistical Institute

1. Introduction

The main upsurge of late Professor R. A. Fisher's theory of Statistical Inference took place within a brief span of about 10 years (1920-30) after the first world war. It was during this period that Fisher came out with the brilliant and now famous notions of (a) likelihood, (b) fiducial probability, (c) information and intrinsic accuracy, (d) sufficiency and (e) ancillary statistics and recovery of information — concepts around which the superstructure of the theory is built.

Many eminent statisticians and mathematicians have made detailed localised studies of some particular aspect of Fisher's theory and some of these studies gave rise to important streams of fundamental research in statistical theory. The author (Basu, 1959) made a very much localised study of the notion of ancillary statistics from the purely mathematical point of view. This note is a follow up study of the earlier paper (Basu, 1959) from the statistical angle. Here we discuss the very controversial subject matter of 'recovery of ancillary information' through proper choice of 'reference sets.' For the purpose of pinpointing our attention to the basic issues raised, we restrict ourselves to the one-parameter set-up only.

In the one parameter set-up Fisher defines an ancillary statistic as one whose probability (sampling) distribution is free of the parameter and which, in conjunction with the maximum likelihood estimator of θ (the parameter), is sufficient. The use of ancillary statistics has been recommended in two different inference situations namely the point estimation problems and the testing of hypotheses problems.

In point estimation problems the use of a suitably chosen ancillary statistic is recommended in the situations where the maximum likelihood estimator T of θ is not a sufficient statistic. The use of T as an estimator of θ will then entail a certain loss of information which, according to Fisher, may be meaningfully (at least in the large sample case) measured, in some situations, as follows. The information contained in the whole sample X is defined as

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta) \,\middle|\, \theta \right]$$

where $f(x \mid \theta)$ is the frequency or density function for X. Similarly the information contained in a statistic T (which may be vector-valued) is measured by the function

$$J(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} \log g(T \mid \theta) \,\middle|\, \theta \right]$$

* The paper has been included in *Contributions to Statistics*, Presented to Professor P. C. Mahalanobis on the occasion of his 70th birthday.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_19,

where $g(t \mid \theta)$ is the frequency or density function for the estimator T. The difference

$$\lambda(\theta) = I(\theta) - J(\theta)$$

may then be taken as a measure of the information lost. Under certain regularity conditions the following results may be proved :

(1) $\lambda(\theta) \geqslant 0$ for all values of θ;

(2) $\lambda(\theta) \equiv 0$ if and only if T is a sufficient statistic;

(3) if T together with the statistic Y be sufficient for θ then the information contained in the pair (T, Y) is $I(\theta)$;

(4) if Y be ancillary and the pair (T, Y) is sufficient then

$$I(\theta) = E[J(\theta \mid Y) \mid \theta]$$

where $J(\theta \mid y)$ is the conditional amount of information contained in T under the condition that Y takes the value y, i.e.

$$J(\theta \mid y) = E\Big[-\frac{\partial^2}{\partial \theta^2} \log f(T \mid y, \theta) \mid Y = y, \theta \Big]$$

where $f(t \mid y, \theta)$ is the conditional frequency (density) function for T under the condition $Y = y$. The relation (4) follows directly from the observation that the joint frequency (density) function $h(t, y \mid \theta)$ of (T, Y) may be factorized as

$$h(t, y \mid \theta) = g(y) f(t \mid y, \theta)$$

where $g(y)$ is the θ-free (Y being ancillary) marginal frequency (density) function for Y.

Now, consider a situation where the maximum likelihood estimator T is not sufficient but where we are able to find another statistic Y whose marginal distribution is θ-free and which complements T in the sense that the pair (T, Y) is jointly sufficient. The statistic Y by itself contains no information about θ. But in a sense it 'summarises in itself' the quantum of information $\lambda(\theta)$ that is lost in the use of T as an estimator of θ. The problem is how to recover this apparent loss of information. According to Fisher it is a mistake to calculate the information content of T with reference to the whole sample space, i.e. with reference to the marginal distribution of T. The appropriate measure of the information content of T is $J(\theta \mid Y)$ not $J(\theta)$. Having observed T and Y we should consider the conditional distribution of T with the observed value of Y as the condition. We take as our 'reference set' not the whole sample space but the sub-set of those sample points that could give rise to the observed value of the ancillary statistic Y.

RECOVERY OF ANCILLARY INFORMATION

The following two quotations from Fisher's writings emphasise the analogy that he has repeatedly drawn between the sample size and the ancillary statistic.

"Having obtained a criterion for judging the merits of an estimate in the real case of finite samples, the important fact emerges that, though sometimes the best estimate we can make exhausts the information in the sample, and is equivalent for all future purposes to the original data, yet sometimes it fails to do so, but leaves a measurable amount of the information unutilized. How can we supplement our estimate so as to utilize these too? It is shown that some, or sometimes all of the lost information may be recovered by calculating what I call ancillary statistics, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it. Their function is, in fact, analogous to the part which the size of our sample is always expected to play, in telling us what reliance to place on the result." (Fisher, 1935).

"When sufficient estimation is possible, there is no problem, but the exhaustive treatment of the cases in which no sufficient estimate exists is now seen to be an urgent requirement. This at present is in the interesting stage of being possible sometimes, though, so far as we know, not always. I have spoken of the sufficient estimates as containing in themselves the whole of the information provided by the data. This is not strictly accurate. There is always one piece of additional, or ancillary, information which we require, in conjunction with even a sufficient estimate, before this can be utilized. That piece of information is the size of the sample or, in general, the extent of the observational record. We always need to know this in order to know how reliable our estimate is. Instead of taking the size of the sample for granted, and saying that the peculiarity of the cases where sufficient estimation is possible lies in the fact that the estimate then contains all the further informations required, we might equally well have inverted our statement, and, taking the estimate of maximum likelihood for granted, have said that the peculiarity of these cases was that, in addition, nothing more than the size of the sample was needed for its complete interpretation. This reversed aspect of the problem is the more fruitful of the two, once we have satisfied ourselves that, when information is lost, this loss is minimised by using the estimate of maximum likelihood. The cases in which sufficient estimation is impossible are those in which, in utilizing this estimate, other ancillary information is required from the sample beyond the mere number of observations which compose it. The function which this ancillary information is required to perform is to distinguish among samples of the same size those from which more or less accurate estimates can be made; or, in general, to distinguish among samples having different likelihood functions, even though they may be maximised at the same value. Ancillary information never modifies the value of our estimate; it determines its precision." (Fisher, 1936).

Often the 'extent of observational record' is planned in advance and is taken for granted in the subsequent analysis of the data. If we take n independent observations on a normal variable with unknown mean θ and known standard deviation,

we see no need to bother about any characteristic of the sample other than the sample mean $\bar{x}$; but yet the fact remains that without some knowledge about n the maximum likelihood estimator $\bar{x}$ of θ will be hardly of any use to any statistician. Along with the information that the sample is drawn from a normal population and the observed value of $\bar{x}$, we need to know the value of the sample size n. The 'reliability' of the estimator $\bar{x}$ is interpreted in terms of its average performance in repeated sampling with the fixed sample size n.

What happens if 'chance' plays (or is allowed to) a part in the determination of n? Suppose we toss a true coin and, depending on whether the outcome is a 'head' or a 'tail', we draw a sample of size 10 or 100. It is easily verified that the sample mean $\bar{x}$ is still the maximum likelihood estimator of θ but that it no longer is a sufficient statistic. It is the pair $(\bar{x}, n)$, where n is the (variable) sample size, that is sufficient for θ. Here n is an ancillary statistic taking the two values 10 and 100 with equal probabilities. Now, having drawn a sample of size n (which is either 10 or 100) and having estimated θ by the sample mean $\bar{x}$, how does the statistician report the 'reliability', the precision, the information content of the estimate? There is no gainsaying the fact that a sample of size 10 will lead to a less reliable estimate than a sample of size 100. Having drawn a sample of size 10 should the statistician turn a blind eye to the actual smallness of the sample size and try to figure out the long run performance of his estimation procedure in a hypothetical series of experimentations in which 50% of the cases he draws sample of size 10 and the other 50% of the cases the sample size is 100? What should be the reference set for judging the performance characteristic of the estimator—the 10 dimensional Euclidean space R_{10} or the union of R_{10} and R_{100}? The author agrees with Fisher that, having drawn a sample with the ancillary statistic $n = 10$, the statistician should judge (if at all he must) the performance of the maximum likelihood estimator $\bar{x}$ in the conditional sample space (restricted reference set) R_{10}. However, the author feels that Fisher, in his writings on ancillary statistics and choice of reference sets, has pushed the above analogy with the sample size a little too far, thereby giving rise to some logical difficulties the real nature of which will be discussed later.

In problems of testing Fisher uses ancillary statistics for the determination of the 'true' level of significance. Having selected the test criterion—a measure of the extent to which the observed sample departs from the expected one under the null hypothesis—the level of significance is the probability (under the null hypothesis) of getting a sample with a larger criterion score than the one actually obtained. In the presence of a suitable ancillary statistic Y, Fisher recommends that the level of significance of a test should be computed by referring to the conditional sample space determined by the set of all possible samples for which the value of Y is the one presently observed.

The following example worked out in Fisher (1956, pp. 163-69) is reproduced here with quotations with the idea of bringing out the essential features of the method

of 'Recovery of Information' (as envisaged by Fisher in the context of point estimation) through proper choice of 'reference sets'.

Example: Let us suppose that we have N pairs of independent observations on the pair (X, Y) of positive random variables with joint probability density function

$$p(x, y \mid \theta) = e^{-\left(\theta x + \frac{y}{\theta}\right)}, \quad x > 0, y > 0, \theta > 0.$$

Let (X_i, Y_i), $i = 1, 2, \ldots, N$, be the N pairs of observations and

let $$T = \sqrt{\Sigma Y_i / \Sigma X_i} \text{ and } U = \sqrt{(\Sigma X_i)(\Sigma Y_i)}.$$

It is easy to check that

(i) T is the maximum likelihood estimator of θ,

(ii) T is not sufficient for θ,

(iii) the pair (T, U) is jointly sufficient for θ,

(iv) U is an ancillary statistic, i.e. the marginal distribution of U is θ-free.

"Since the likelihood cannot be expressed in terms of only θ and T, there will be no sufficient estimate, and some information will be lost if the sample is replaced by the estimate T only."

The amount of information supplied by the whole sample (of N pairs of observations) is

$$I(\theta) = 2N/\theta^2$$

while the amount of information contained in the statistic T is

$$J(\theta) = \frac{2N}{\theta^2} \frac{2N}{2N+1}.$$

"The loss of information is less than half the value of a single pair of observations, and never exceeds one third of the total available. Nevertheless its recovery does exemplify very well the mathematical processes required to complete the logical inference."

Here, U is the ancillary statistic and so we have to consider the conditional distribution of T given U. From this conditional distribution the conditional information content of T works out as

$$J(\theta \mid U) = \frac{2N}{\theta^2} \frac{K_1(2U)}{K_0(2U)}$$

where K_0 and K_1 are Bessel functions.

Let us note that the information content $J(\theta \mid U)$ "depends upon the value of U actually available, but has an average value, when variations of U are taken into account, of

$$2N/\theta^2,$$

the total amount expected on the average from N observations, none is now lost. The information is recovered and inference completed by replacing the distribution

of T for given size of sample N, by the distribution of T given U, which indeed happens not to involve N at all. In fact, U has completely replaced N as a means of specifying the precision to be ascribed to the estimate. In both cases the estimate T is the same, the calculation of U enables us to see exactly how precise it is, not on the average, but for the particular value of U supplied by the sample."

2. THE SAMPLE SIZE ANALOGY

This section is devoted to a detailed discussion on the oft-drawn analogy between the sample size and an ancillary statistic. While drawing the analogy Fisher seems to be always thinking of the sample size n as the only determinant of the sampling experiment. The reliability of an estimate is assessed in terms of the average performance of the experimental procedure in a hypothetical series of experimentations with the sample size n fixed at the level actually obtained in the sample at hand. Fisher always interpreted the reliability (information content, variance etc.) of an estimate in terms of the average performance of some well-defined experimental (estimation) procedure in some hypothetical sequence of experimentations. When Fisher talks of the reliability of an estimate the adjective 'reliability' is used only as a transferred epithet and is actually meant to be attached to the estimation procedure that has given rise to the estimate.

In the example on page 56 the statistician (for some unknown reasons) decided to choose between a sample of size 10 or one of size 100 on the basis of the flip of a coin. Here a random choice is being made between two sampling experiments $\mathcal{E}_{10}$ and $\mathcal{E}_{100}$ with the associated sample spaces R_{10} and R_{100} and the corresponding probability distributions. More generally, suppose $\mathcal{E}_\nu$ is the sampling experiment corresponding to a sample of size ν and suppose the observed sample size n is determined by a θ-free (parameter-free) chance mechanism. Once we recognize that the estimate T of θ is generated by the random choice $\mathcal{E}_n$ from the family $\{\mathcal{E}_\nu\}$ of available experimental procedures we ought to transfer the reliability index of the chosen experiment $\mathcal{E}_n$ to the estimate T. When the statistician is forced to make a selection from a family of available experimental procedures he should report the reliability of the procedure actually selected by him. The following example is somewhat more realistic than the one considered on page 56.

Example : Suppose, from a finite population of N units, we draw a sample of size s with replacements. Let $X_1, X_2, \ldots, X_N$ be the population values and $x_1, x_2, \ldots, x_s$ the s sample values (arranged in order of their appearances). The problem is to estimate the population mean $\overline{X} = (X_1+\cdots+X_N)/N$ and the obvious estimate is the sample mean $\bar{x} = (x_1+\ldots+x_s)/s$ which has a standard deviation of $\sigma/\sqrt{s}$ where σ^2 is the population variance. But will it be correct to recognize s as the true sample size? Since the sample was drawn with replacements, it is plausible that some of the population units came repeatedly in the sample. In realistic sample survey situations, we can usually distinguish between the population units. Consider the extreme situation where all the s sample units happen to be the same (population unit).

Confronted with a situation like this the statistician would surely hesitate to report the reliability of his estimate as $\sigma/\sqrt{s}$. In this extreme case the honest statistician will have to admit that he had drawn an unlucky sample whose effective size is only 1 (and not s) and report the reliability of his estimate as σ (and not $\sigma/\sqrt{s}$). More generally, consider the situation where n is the number of distinct units in the sample and $x_1^*, x_2^*, \ldots, x_n^*$ are the corresponding sample values (arranged say in an increasing order of their population unit-indices). It is easy to see that the probability distribution of n involves only N and s, and since they are known constants, the statistic n is ancillary. If we define $\bar{x}^*$ as

$$\bar{x}^* = (x_1^* + \ldots + x_n^*)/n$$

then it is well recognized [see Basu (1958) for a detailed discussion on this] that $(x_1^*, x_2^*, \ldots, x_n^*)$ is a sufficient statistic and that $\bar{x}^* = E(\bar{x} \mid x_1^*, \ldots, x_n^*)$. Hence, by the Rao-Blackwell theorem, $\bar{x}^*$ is better than $\bar{x}$ as an unbiased estimator of $\bar{X}$. All right, we agree to estimate $\bar{X}$ by $\bar{x}^*$ and forget all about $\bar{x}$. But our troubles are not yet over. What is the standard deviation of $\bar{x}^*$? For a fixed n, the conditional distribution of $(x_1^*, \ldots x_n^*)$ is the same as that of a simple random sample of size n from the population of N values. We see then that the ancillary statistic n is really a sample size. When we draw a sample of size s with replacements and agree to take note of only the sufficient statistic $(x_1^*, x_2^*, \ldots, x_n^*)$ we are effectively drawing a simple random sample of variable size n. If we denote by $\mathcal{E}_i$ $(i = 1, 2, \ldots, s)$ the experimental procedure of drawing a simple random sample of size i from the population of N units and by p_i the parameter-free probability that $n = i$, then the above experimental procedure is the same as that of selecting one of the experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_s$ with probabilities $p_1, p_2, \ldots, p_s$ (and then carrying out the experiment $\mathcal{E}_n$ so selected). From what we have said earlier, it then follows that we should assess the reliability of $\bar{x}^*$ in terms of that of the experiment $\mathcal{E}_n$ actually selected, i.e. $V(\bar{x}^*)$ should be reported as

$$\frac{N-n}{N-1}\,\frac{\sigma^2}{n} \qquad \ldots \quad (2.1)$$

and not as

$$E\left(\frac{N-n}{N-1}\,\frac{\sigma^2}{n}\right). \qquad \ldots \quad (2.2)$$

Let us repeat once again that when reporting the reliability of an estimate we are actually saying something about the long term average performance of some well-defined estimation procedure. Both (2.1) and (2.2) are reliability indices—(2.1) for the experimental procedure $\mathcal{E}_n$ with a fixed n and (2.2) for the experimental procedure where n is allowed to vary (in the parameter-free manner described earlier) from trial to trial. We may briefly summarise the basic Fisherian point of view (in the present context) as follows :

(a) Suppose a whole family $\{\mathcal{E}_y\}$, where y runs over an index set $\mathcal{Y}$, of statistical experiments is available any one of which may be meaningfully performed for the purpose of making a scientific inference about some physical quantity θ.

(b) And suppose that the experiment that the statistician actually performs is recognized to be equivalent to the two-stage experiment of first selecting at random a point Y in $\mathcal{Y}$ and then performing the experiment $\mathcal{E}_Y$.

(c) Suppose, further, that the probability distribution of Y in $\mathcal{Y}$ is θ-free.

Under the above conditions, the true 'reference set' for the statistician is the experiment $\mathcal{E}_Y$ (with its associated sample space and probability distributions) and not the two-stage experiment described in (b) above. The reliability (information content, standard deviation, significance level etc.) of the inference made about θ should be assessed in terms of the average performance characteristic of the inference procedure in a long hypothetical sequence of independent repetitions under identical conditions of the experiment $\mathcal{E}_Y$ (where Y is supposed to be held fixed at the particular point that obtains in the present instance).

Under the above circumstances Fisher would like the statistician to say something to the following effect : "It was rather silly of me to let chance have a hand in the determination of the experiment $\mathcal{E}_Y$ for me. But I now recognize that I have performed the experiment $\mathcal{E}_Y$ and see no reason whatsoever to fuss about the other experiments in the family $\{\mathcal{E}_y\}$ that might have been handed down to me by chance. The inference that I make about θ is the most appropriate one for the experiment $\mathcal{E}_Y$, and the assessment of the reliability of my inference is made with reference to the experiment $\mathcal{E}_Y$ alone."

Now, let us consider a general inference situation and see whether the above arguments hold in the presence of an ancillary statistic. Let $\mathcal{E}$ be an arbitrary statistical experiment performed with a view to elicit some information about a physical quantity θ. From the mathematical standpoint we are then concerned with the trio $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ where $\mathcal{X} = \{x\}$ is the sample space and $\mathcal{A} = \{A\}$ is the σ-field of events on which the family $\mathcal{P} = \{P_\theta\}$ of probability measures is defined. For the sake of simplicity, we ignore the possibility of nuisance parameters and we assume that different possible values of θ are associated with different probability distributions. Now, let Y be an ancillary statistic taking values in the space $\mathcal{Y}$. Corresponding to each point y in $\mathcal{Y}$ we then have (under some regularity conditions) a trio $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$, where $\mathcal{X}_y$ is the sub-set of points of $\mathcal{X}$ for which $Y = y$ and $\mathcal{P}_y = \{P_\theta^y\}$ is the family of conditional probability distributions on a σ- field $\mathcal{A}_y$ of sub-sets of $\mathcal{X}_y$.

We have only to imagine a conceptual experiment $\mathcal{E}_y$ that gives rise to the trio $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$ and the analogy that we have been trying to drive home is complete. The statistical experiment $\mathcal{E}$ is then equivalent to the two-stage experiment of first observing the random variable Y (whose distribution is θ-free) and then performing the conceptual experiment $\mathcal{E}_Y$ leading ultimately to a point X in $\mathcal{X}$.

Why the insistence on Y being an ancillary statistic? The sample X that we arrive at through the experiment $\mathcal{E}$ (or the equivalent two-stage breakdown $Y-\mathcal{E}_Y$), is the only source of our information about θ. Now, according to Fisher, the likelihood function $L(\theta)$ for the sample X is the sole basis for making any judgement about θ.

Nothing else need be taken cognizance of. Let us observe that the likelihood function $L(\theta)$ is the same (excepting for a θ-free multiplicative constant) whether we consider X to be generated by the experiment $\mathcal{E}$ or the conceptual experiment $\mathcal{E}_Y$. This, according to the author, is the main explanation as to why in the above Y-decomposition of the probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ into the family $\{\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y\}$ the statistic Y has to have a θ-free distribution. If Y be ancillary then the choice of the 'reference set'$(\mathcal{X}_Y, \mathcal{A}_Y, \mathcal{P}_Y)$ does not affect the likelihood scale.

3. A LOGICAL DIFFICULTY

The decomposition of the probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ into the family of probability structures $\{\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y\}$ depends on the ancillary statistic Y. Which ancillary statistic Y to work with ? The author made a rather comprehensive study (Basu, 1959) of the family of ancillary statistics. It was noted that each of the two statistics Y_1 and Y_2 may individually be ancillary but jointly not so. Thus, in case of a controversy as to which one of the two ancillaries Y_1 and Y_2 should determine the 'reference set' one cannot solve the dilemma by referring to the conditional probability structure conditioned by the observed values of both Y_1 and Y_2. Consider the following example :

Example : Let (X, Y) have a bivariate normal distribution with zero means, unit standard deviations, and unknown correlation coefficient θ. If (X_i, Y_i), $i = 1, 2, \ldots, n$, be n pairs of independent observations on (X, Y) then we see at once that the set of n observations $(X_1, X_2, \ldots, X_n)$ on X is an ancillary statistic. Similarly $(Y_1, Y_2, \ldots, Y_n)$ also is ancillary. But the two ancillary statistics together is the whole data and is obviously sufficient. In regression studies the statistician often ignores the sampling variations in the observed X-values and treats them as preselected experimental constants. If, in the above situation, he be justified in similarly treating the observed Y-values also, then how do we define the sampling error for the estimate of θ?

Fisher recommended the choice of the 'reference set' with the help of an ancillary Y that complements the maximum likelihood estimator T in the sense that T is not sufficient but the pair (T, Y) is. This, however, is not a sufficient specification for Y. The two statistics Y_1 and Y_2 may each be ancillary complements to the maximum likelihood estimator T and lead to different 'reference sets' and different reliability indices for the estimator T. The following very simple example clearly brings out the above possibility.

Example : Suppose we have a biased die about which we have enough information to assume the following probability distribution :

scores :	1	2	3	4	5	6
probability distributions	$\frac{1-\theta}{12}$	$\frac{2-\theta}{12}$	$\frac{3-\theta}{12}$	$\frac{1+\theta}{12}$	$\frac{2+\theta}{12}$	$\frac{3+\theta}{12}$

where the parameter θ can take any value in the closed interval [0, 1]. Let $\mathcal{E}$ stand

for the experiment of rolling the die only once leading to the observed score X. It is easily seen that the maximum likelihood estimator T is defined as follows :

X :	1	2	3	4	5	6
$T(X)$:	0	0	0	1	1	1

and leads to the partition (1, 2, 3), (4, 5, 6) of the sample space. Here X is the minimal sufficient statistic and T is not sufficient. In this example there are six non-equivalent[1] ancillary complements to T. They may be listed as follows :

X	1	2	3	4	5	6
$Y_1(X)$	0	1	2	0	1	2
$Y_2(X)$	0	1	2	0	2	1
$Y_3(X)$	0	1	2	1	0	2
$Y_4(X)$	0	1	2	2	0	1
$Y_5(X)$	0	1	2	1	2	0
$Y_6(X)$	0	1	2	2	1	0

Each of the six statistics $Y_1, Y_2, \ldots, Y_6$ is a maximal[2] ancillary.

The statistic T induces the partition (1, 2, 3) and (4, 5, 6) whereas Y_1 induces the partition (1, 4), (2, 5) and (3, 6). Since each Y_1-partition intersects each T-partition in a one-point set, it follows that the pair (T, Y_1) is equivalent to X which is the minimal sufficient statistic. Thus, Y_1 is an ancillary complement to T. In like manner, we prove that each of the other Y_i's is an ancillary complement to T. The joint probability distribution for T and Y_1 is described in the following table :

T \ Y_1	0	1	2	total
0	$\frac{1-\theta}{12}$	$\frac{2-\theta}{12}$	$\frac{3-\theta}{12}$	$\frac{2-\theta}{4}$
1	$\frac{1+\theta}{12}$	$\frac{2+\theta}{12}$	$\frac{3+\theta}{12}$	$\frac{2+\theta}{4}$
total	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	1

The Y_1-decomposition of the experiment of rolling the biased die once is as follows :

"Choose one of the three pairs (1, 4), (2, 5) and (3, 6) with probabilities 1/6, 1/3, and 1/2 respectively. Then select one from the chosen pair with probabilities

$\frac{1-\theta}{2}$ and $\frac{1+\theta}{2}$, if the chosen set is (1, 4),

or, $\frac{2-\theta}{4}$ and $\frac{2+\theta}{4}$, if the chosen set is (2, 5),

or, $\frac{3-\theta}{6}$ and $\frac{3+\theta}{6}$, if the chosen set is (3,6)".

[1]Two statistics are said to be equivalent if they lead to the same partition of the sample space.

[2]See Basu (1959) for the definition of a maximal ancillary.

The physical experiment of rolling the biased die once may then be imagined to be equivalent to the two-stage (conceptual) experiment of first choosing (in a θ-free manner) one of three biased coins and then tossing the selected coin once.

Let us observe that we have six different decompositions of $\mathcal{E}$ corresponding to the six different ancillary complements to T. How do we recover the information lost in T ?

Let us suppose that the observed value of X is 5. The corresponding values of T, Y_1, Y_2 and Y_3 are respectively 1, 1, 2 and 0. The conditional distributions of the maximum likelihood estimator T under the three conditions $Y_1 = 0$, $Y_2 = 2$, and $Y_3 = 0$ are as follows :

range of T :		0	1
conditional probability distribution of T under the condition	$Y_1 = 1$	$\frac{(2-\theta)}{4}$	$\frac{(2+\theta)}{4}$
	$Y_2 = 2$	$\frac{(3-\theta)}{5}$	$\frac{(2+\theta)}{5}$
	$Y_3 = 0$	$\frac{(1-\theta)}{3}$	$\frac{(2+\theta)}{3}$

Thus, in this situation we find that different choices of ancillary statistics lead to different 'reference sets' and different reliability indices for the estimator T. There exists no unique way of recovering the ancillary information.

4. Conceptual statistical experiments

The author believes that the real trouble lies in our failure to recognize the difference between a real (performable) and a conceptual (non-performable) statistical experiment.[1] Every real experiment gives rise to a probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ but the converse is not true. On page 60 we saw how the probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, generated by the experiment $\mathcal{E}$, may be decomposed into the family $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$, $y \in \mathcal{Y}$, of probability structures and there we conceived of an experiment $\mathcal{E}_y$ corresponding to $(\mathcal{X}_y, \mathcal{A}_y, \mathcal{P}_y)$. In general, the experiments $\mathcal{E}_y$ are non-performable. If (as in page 56) the statistician selects (on the flip of a coin) between a sample of size 10 and one of size 100, he is making a random choice between two performable statistical experiments. But in the example considered on page 61 the statistician can only conceive (in six different ways) of a breakdown of the experiment of once rolling the biased die into a two-stage experiment of first making a (θ-free) random selection between three biased coins and then tossing the selected coin once. In this example the statistician has a die to experiment with; but where are the coins ?[2]

[1] The author does not think it necessary to enter into a lengthy discussion on the reality or performability of a statistical experiment.

[2] Of course, the experiment of rolling the die repeatedly until, say, either 2 or 5 appears (and then observing only the final score) is essentially equivalent to tossing once a biased coin with probabilities $(2-\theta)/4$ and $(2+\theta)/4$. But who is interested in such a wasteful experiment ? The author would classify such experiments under the conceptual (non-performable) category.

When the sample size n is determined in a θ-free manner the statistician may be justified in regarding n as a pre-fixed experimental constant and in contemplating the long term performance characteristic of the experimental procedure[1] with the sample size fixed at the level actually obtained. More generally, when the statistician is presented with a θ-free choice between a family $\{\mathcal{E}_y\}$ of performable experimental procedures then it would be correct to treat the Y actually obtained as a pre-determined experimental constant. His sole concern should be the experiment $\mathcal{E}_Y$ he is actually presented with and none of the other members of the family $\{\mathcal{E}_y\}$. Difficulty arises in the attempted generalization to non-performable meaningless experiments. The kind of situations where an experiment $\mathcal{E}$ may be decomposed into a family $\{\mathcal{E}_y\}$ of real experiments are very rare indeed. The author is not aware of any example where such a decomposition (into real experiments) may be effected in more than one way. The elementary example considered on page 61 establishes the possibility of a multiplicity of ancillary decompositions into conceptual experiments. The ancillary argument of Fisher cannot be extended to such cases. The sample size analogy for the ancillary statistic appears to be a false one. We end this discourse with a very elementary example where there exists an essentially unique maximal[2] ancillary decomposition of the experiment but yet the ancillary argument leads us to a rather curious and totally unacceptable 'reference set'.

Example : Let X be an observable random variable with uniform probability distribution over the interval $[\theta, \theta+1)$ where $0 \leqslant \theta < \infty$. For the sake of simplicity we consider the case of a single observation on X. The sample space $\mathcal{X}$ is the half line $[0, \infty)$. The likelihood function, for the observation X, is

$$L(\theta) = \begin{cases} 1 & \text{if } X-1 < \theta \leqslant X \\ 0 & \text{otherwise} \end{cases}$$

Thus, the integer part $[X]$ of X has as good a claim to be considered a maximum likelihood estimator for θ as any other point in the interval $(X-1, X]$. It is easy to check that $[X]$ is not a sufficient statistic. Do there exist an ancillary complement to $[X]$?

Consider the fractional part $\varphi(X) = X-[X]$ of X. It is not difficult to show that $\varphi(X)$ is an ancillary statistic with uniform probability distribution over the unit interval $[0, 1)$. Indeed, it is possible to show that $\varphi(X)$ is an essentially maximal ancillary in the sense that every ancillary statistic is essentially a function of $\varphi(X)$. As $X = [X]+\varphi(X)$, it follows at once that the pair $([X], \varphi(X))$ is equivalent to X and hence $\varphi(X)$ is the ancillary complement to $[X]$.

For a given $\theta = [\theta]+\varphi(\theta)$, the observation $X = [X]+\varphi(X)$ lies in the interval $[\theta, \theta+1)$, i.e., with probability one,

$$\theta = [\theta]+\varphi(\theta) \leqslant [X]+\varphi(X) < [\theta+1]+\varphi(\theta) = \theta+1.$$

[1]Throughout this paper we are regarding the inference procedure as a well-defined part of the experimental procedure.

[2]That is, a decomposition with respect to a maximal ancillary as defined in Basu (1959).

From the above, it follows that

$$[X] = \begin{cases} [\theta] & \text{if } \varphi(X) \geqslant \varphi(\theta) \\ [\theta+1] & \text{if } \varphi(X) < \varphi(\theta) \end{cases} \qquad \ldots \quad (4.1)$$

Since $\varphi(X)$ has a uniform distribution over $[0, 1)$, it follows that the marginal distribution of $[X]$ is concentrated at the two points $[\theta]$ and $[\theta+1] = [\theta]+1$ with probabilities $1-\varphi(\theta)$ and $\varphi(\theta)$ respectively. Hence

$$\begin{aligned} E([X]\,|\,\theta) &= [\theta](1-\varphi(\theta))+([\theta]+1)\,\varphi(\theta) \\ &= [\theta]+\varphi(\theta) \\ &= \theta \end{aligned}$$

i.e. $[X]$ is an unbiased estimator of θ. And

$$V([X]\,|\,\theta) = \varphi(\theta)\ (1-\varphi(\theta)).$$

Now, since $\varphi(X)$ is the ancillary complement to $[X]$, let us see what 'reference set' it leads us to. Given $\varphi(X)$, the sample $X = [X]+\varphi(X)$ can vary over the restricted set

$$\varphi(X),\ 1+\varphi(X),\ 2+\varphi(X), \ldots\ldots\ldots$$

From the relation (4.1) it is now clear that, for any fixed θ, the conditional distribution of $X = [X]+\varphi(X)$, given $\varphi(X)$, is degenerate at the point

$$[\theta]+\varphi(X), \qquad \text{if} \quad \varphi(\theta) \leqslant \varphi(X),$$

or, at the point $\qquad [\theta+1]+\varphi(X), \text{ if} \quad \varphi(\theta) > \varphi(X).$

Thus, the 'reference set', corresponding to an observed value of the ancillary statistic $\varphi(X)$, is a one-point, degenerate probability structure. The conditional distribution of the maximum likelihood estimator $[X]$, given $\varphi(X)$ and θ, is degenerate at the point $[\theta]$ or $[\theta+1]$ depending on whether $\varphi(\theta) \leqslant \varphi(X)$ or $\varphi(\theta) > \varphi(X)$. Acceptance of this 'reference set' will alter the status of $[X]$ from a statistical variable to an unknown constant.

Writing $Y = \varphi(X)$, the two-stage Y-decomposition of the experiment $\mathscr{E}$ of making a single observation on X will then be as follows :

(i) Select a number Y at random (with uniform probability distribution) in the unit interval $[0, 1)$

(ii) Determine the value of $[\theta]$ and whether

(a) $\varphi(\theta) \leqslant Y$ or

(b) $\varphi(\theta) > Y$

and then write $$X = \begin{cases} [\theta]+Y & \text{in case of (a)} \\ [\theta]+1+Y & \text{in case of (b).} \end{cases}$$

The second stage of the Y-decomposition is clearly non-performable.

REFERENCES

BASU, D. (1958) : On sampling with and without replacements. *Sankhyā*, **20**, 287.

——— (1959) : The family of ancillary statistics. *Sankhyā*, **21**, 247.

COX, D. R. (1958) : Some problems connected with statistical inference. *Ann. Math. Stat.*, **29**, 357.

FISHER, R. A. (1925) : Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700.

——— (1934) : Two new properties of mathematical likelihood. *Proc. Royal Soc.*, **144**A, 285.

——— (1935) : The logic of inductive inference. *J. Roy. Stat. Soc.*, **98**, 39.

——— (1936) : Uncertain inference. *Proc. American Academy of Arts and Sciences*, **71**, 245.

——— (1956) : *Statistical Methods and Scientific Inference*, Oliver and Boyd, London.

OWEN, A. R. G. (1948) : Ancillary statistics and fiducial distribution. *Sankhyā*, **9**, 1.

RAO, C. R. (1952) : Minimum variance estimation in distributions admitting ancillary statistics. *Sankhyā*, **12**, 53.

PROBLEMS RELATING TO THE EXISTENCE OF MAXIMAL AND MINIMAL ELEMENTS IN SOME FAMILIES OF STATISTICS (SUBFIELDS)

D. BASU
UNIVERSITY OF NORTH CAROLINA
and
INDIAN STATISTICAL INSTITUTE

1. Summary

In statistical theory one comes across various families of statistics (subfields). For each such family, it is of some interest to ask oneself as to whether the family has maximal and/or minimal elements. The author proves here the existence of such elements in a number of cases and leaves the question unsolved in a number of other cases. A number of problems of an allied nature are also discussed.

2. Introduction

Let $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$ be a given probability structure (or statistical model). A statistic is a measurable transformation of $(\mathfrak{X}, \mathcal{A})$ to some other measurable space. Each such statistic induces, in a natural manner, a subfield (abbreviation for sub-σ-field) of $\mathcal{A}$ and is, indeed, identifiable with the induced subfield.

Between subfields of $\mathcal{A}$ there exists the following natural partial ordering.

DEFINITION 1. *The subfield $\mathcal{A}_1$ is said to be larger than the subfield $\mathcal{A}_2$ if every member of $\mathcal{A}_2$ is also a member of $\mathcal{A}_1$.*

A slightly weaker version of the above partial order is the following.

DEFINITION 2. *The subfield $\mathcal{A}_1$ is said to be essentially larger than the subfield $\mathcal{A}_2$ if every member of $\mathcal{A}_2$ is $\mathcal{P}$-equivalent to some member of $\mathcal{A}_1$.*

As usual, two measurable sets A and B are said to be $\mathcal{P}$-equivalent if their symmetric difference $A \Delta B$ is P-null for each $P \in \mathcal{P}$.

Given a family $\mathcal{F}$ of subfields (statistics), one naturally inquires as to whether $\mathcal{F}$ has a largest and/or least element in the sense of definition 1. In the absence of such elements in $\mathcal{F}$, one may inquire about the possible existence of maximal and/or minimal elements. An element $\mathcal{A}_0$ of $\mathcal{F}$ is a maximal (minimal) element of $\mathcal{F}$, if there exists no other element $\mathcal{A}_1$ in $\mathcal{F}$ such that $\mathcal{A}_1$ is larger (smaller) than $\mathcal{A}_0$. In the absence of maximal (minimal) elements in $\mathcal{F}$, one may look for elements

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_20,

that are essentially largest (least) or are essentially maximal (minimal) in the sense of the weaker partial order of definition 2.

The particular case in which $\mathfrak{F}$ is the family of all sufficient subfields has received considerable attention. The largest element of $\mathfrak{F}$ is clearly the total subfield $\mathcal{A}$ itself. If $\mathcal{P}$ is a dominated family of measures, then it is well known that $\mathfrak{F}$ has an essentially least element in terms of the weaker partial order of definition 2. In general, $\mathfrak{F}$ does not have even essentially minimal elements. If, however, an essentially minimal element exists, then it must be essentially unique, and thus, the essentially least element of $\mathfrak{F}$ (see corollary 3 to theorem 4 in [3]).

In [1] the author considers the family $\mathfrak{F}$ of ancillary subfields. A subfield $\mathcal{A}_0$ is said to be ancillary if the restriction to $\mathcal{A}_0$ of the class $\mathcal{P}$ of probability measures shrinks the class down to a single probability measure. The least ancillary subfield is clearly the trivial subfield, consisting of only the empty set $\varnothing$ and the whole space $\mathfrak{X}$. The existence of maximal elements in the family of ancillary subfields is demonstrated in [1]. In general, there exists a multiplicity of maximal ancillary subfields.

In sections 3 to 6 we list four problems that are similar to the problem of ancillary subfields. In section 7 we develop a general method to demonstrate the existence of maximal elements in these four cases. In section 8 we discuss some related questions, and in section 9 we list a number of other problems.

3. The family $\mathfrak{F}_1$ of $\mathcal{B}$-independent subfields

Let $\mathcal{B}$ be a fixed subfield. A subfield is said to be $\mathcal{B}$-independent (independent of $\mathcal{B}$) if $P(BC) \equiv P(B)P(C)$ for all $B \in \mathcal{B}$, $C \in \mathcal{C}$ and $P \in \mathcal{P}$.

Let $\mathfrak{F}_1$ be the family of all $\mathcal{B}$-independent subfields. Clearly, the least element of $\mathfrak{F}_1$ is the trivial subfield. Even in very simple situations, $\mathfrak{F}_1$ has no largest, or essentially largest, element. In section 7 we shall show that $\mathfrak{F}_1$ always has maximal elements. Consider the two examples.

EXAMPLE 1(a). Let $\mathfrak{X}$ consist of the four points a, b, c, and d, and let $\mathcal{P}$ consist of only one probability measure—the one that allots equal probabilities to the four points. Let $\mathcal{B}$ consist of the four sets $\varnothing$, $\mathfrak{X}$, $[a, b]$, and $[c, d]$. Then the two subfields $\mathcal{C}_1$ and $\mathcal{C}_2$, consisting respectively of

$$\begin{aligned} &\mathcal{C}_1\colon \varnothing, \mathfrak{X}, [a, c] \quad \text{and} \quad [b, d], \\ &\mathcal{C}_2\colon \varnothing, \mathfrak{X}, [a, d] \quad \text{and} \quad [b, c], \end{aligned} \tag{3.1}$$

are both maximal $\mathcal{B}$-independent subfields. Incidentally, in this case, $\mathcal{C}_1$ and $\mathcal{C}_2$ happen to be independent of each other.

EXAMPLE 1(b). Let $x_1, x_2, \cdots, x_n$ be n independent normal variables with equal unknown means φ and equal unknown standard deviations θ. Let $\mathcal{B}$ be the subfield induced by the statistic

$$\bar{x} = (x_1 + x_2 + \cdots + x_n)/n, \tag{3.2}$$

and let $\mathcal{C}$ be induced by the set of differences

$$D = (x_1 - x_n, x_2 - x_n, \cdots, x_{n-1} - x_n). \tag{3.3}$$

Here $\mathcal{C}$ is $\mathcal{B}$-independent, but it is not the largest $\mathcal{B}$-independent subfield. Indeed, in this situation there are infinitely many maximal elements in $\mathfrak{F}_1$ (see example 1 in [1]). However, it is possible to show that $\mathcal{C}$ is an essentially maximal element in $\mathfrak{F}_1$. In the above example, one may reverse the role of $\bar{x}$ and D and ask oneself as to whether $\bar{x}$ is a maximal D-independent statistic. It is of some interest to speculate about the truth or falsity of the following general proposition.

PROPOSITION 1. *If $\mathcal{C}$ is a maximal (or essentially maximal) $\mathcal{B}$-independent subfield, then $\mathcal{B}$ is a maximal (or essentially maximal) $\mathcal{C}$-independent subfield.*

4. The family $\mathfrak{F}_2$ of φ-free subfields

Let us suppose that the members of the class $\mathcal{P}$ are indexed by two independent parameters θ and φ; that is,

$$\mathcal{P} = \{P_{\theta,\varphi} | \theta \in \Theta, \varphi \in \Phi\}, \tag{4.1}$$

the parameter space being the Cartesian product $\Theta \times \Phi$.

A subfield $\mathcal{C}$ is called φ-free if the restriction of $\mathcal{P}$ to $\mathcal{C}$ leads to a class of probability measures that may be indexed by θ alone; that is for all $C \in \mathcal{C}$ the probability $P_{\theta,\varphi}(C)$ is a function of θ only. Let $\mathfrak{F}_2$ be the family of all φ-free subfields. Evidently, the concept of φ-free subfields is a direct generalization of the concept of ancillary subfields.

The trivial subfield is again the least element of $\mathfrak{F}_2$. That $\mathfrak{F}_2$ always has maximal elements will be demonstrated later. In general, $\mathfrak{F}_2$ has a plurality of maximal elements.

EXAMPLE 2(a). Let $\mathfrak{X}$ consist of the five points a, b, c, d, and e, and let $\mathcal{P} = \{\mathcal{P}_{\theta,\varphi}\}$ consist of the probability measures

(4.2)

x	a	b	c	d	e
$P_{\theta,\varphi}(x)$	$1 - \theta$	$\theta\varphi$	$\theta\varphi$	$\theta(\frac{1}{2} - \varphi)$	$\theta(\frac{1}{2} - \varphi)$

where $0 < \theta < 1$ and $0 < \varphi < \frac{1}{2}$.

There are exactly 12 subsets of $\mathfrak{X}$ whose probability measure is φ-free, and they are $\mathfrak{X}$, $[a]$, $[b, d]$, $[b, e]$, $[c, d]$, $[c, e]$, and their complements. As these 12 sets do not constitute a subfield, it is clear that there cannot exist a largest element in $\mathfrak{F}_2$. The two subfields $\mathcal{C}_1$ and $\mathcal{C}_2$ consisting respectively of

(4.3)

$\mathcal{C}_1$: $\mathfrak{X}$, $[a]$, $[b, d]$, $[c, e]$ and their complements,

$\mathcal{C}_2$: $\mathfrak{X}$, $[a]$, $[b, e]$, $[c, d]$ and their complements,

are the two maximal elements of $\mathfrak{F}_2$.

EXAMPLE 2(b). Let $x_1, x_2, \cdots, x_n$ be n independent and identically distributed variables with a cumulative distribution function (cdf) of the type $F(x - \varphi/\theta)$, $-\infty < \varphi < \infty$, $0 < \theta < \infty$, where the function F is known and φ, θ are the so-called location and scale parameters.

The subfield $\mathcal{C}$ generated by the $n - 1$ dimensional statistic,

$$D = (x_1 - x_n, x_2 - x_n, \cdots, x_{n-1} - x_n), \tag{4.4}$$

is φ-free in the sense defined before. In general, it is not true that $\mathcal{C}$ is the largest element of the family $\mathfrak{F}_2$ of φ-free subfields. In the particular case where F is the cdf of a normal variable, the subfield $\mathcal{C}$ may be shown to be an essentially maximal element of $\mathfrak{F}_2$. Let us observe that in this particular case, $\mathfrak{F}_2$ is the same as $\mathfrak{F}_1$ of example 1(b). The following proposition may well be true.

PROPOSITION 2. *Whatever may be F, the subfield $\mathcal{C}$ (as defined above) is an essentially maximal element of the family $\mathfrak{F}_2$ of φ-free (φ being the location parameter) subfields.*

Suppose in example 2(b) we reverse the role of φ and θ and concern ourselves with the family $\mathfrak{F}_2^*$ of θ-free subfields, that is, with subfields every member of which has a probability measure that does not involve the scale parameter θ. The author believes that the following proposition is generally true.

PROPOSITION 3. *Every θ-free subfield is also φ-free, that is, $\mathfrak{F}_2^* \subset \mathfrak{F}_2$.*

In the particular case where F is the *cdf* of a normal variable, the truth of proposition 3 has been established in [4].

5. The family $\mathfrak{F}_3$ of $\mathcal{G}$-similar subfields

Let $\mathcal{G} = \{g\}$ be an arbitrary but fixed class of measurable transformations of $(\mathfrak{X}, \mathfrak{A})$ into itself. For each $P \in \mathcal{P}$, the transformation $g \in \mathcal{G}$ induces a probability measure Pg^{-1} on $(\mathfrak{X}, \mathfrak{A})$. A subfield $\mathcal{C}$ will be called $\mathcal{G}$-similar if, for each $g \in \mathcal{G}$ and $P \in \mathcal{P}$, the restriction of the two measures P and Pg^{-1} to $\mathcal{C}$ are identical. In other words, $\mathcal{C}$ is $\mathcal{G}$-similar if for all $\mathcal{C} \in \mathcal{C}$,

$$Pg^{-1}(C) \equiv P(C) \qquad \text{for all} \quad P \in \mathcal{P} \quad \text{and} \quad g \in \mathcal{G}. \tag{5.1}$$

Let $\mathfrak{F}_3$ be the family of all $\mathcal{G}$-similar subfields. One may look upon $\mathfrak{F}_3$ as the family of subfields that are induced by statistics $T(x)$ such that $T(x)$ and $T(gx)$ are identically distributed for each $P \in \mathcal{P}$ and $g \in \mathcal{G}$. The least element of $\mathfrak{F}_3$ is, of course, the trivial subfield. As we shall see later, $\mathfrak{F}_3$ always has maximal elements and, in general, a plurality of them.

EXAMPLE 3(a). Let $\mathfrak{X}$ be the real line and $\mathcal{P} = \{P_\theta | -\infty < \theta < \infty\}$, where P_θ is the uniform distribution over the interval $(\theta, \theta + 1)$. Let $\mathcal{G}$ consist of the single transformation g defined as $gx =$ the fractional part of x. It is easy to check that for all θ in $(-\infty, \infty)$, $P_\theta g^{-1} = P_0$.

In this example, the subfield $\mathcal{C}$ is $\mathcal{G}$-similar if and only if each member of $\mathcal{C}$ has a probability that is θ-free. Thus, the family $\mathfrak{F}_3$ of $\mathcal{G}$-similar subfields is the same as the family of ancillary subfields. Here, $\mathfrak{F}_3$ has a largest element, and that is the subfield of all Borel sets A such that the two sets A and $A + 1$ are essentially equal with respect to the Lebesgue measure.

EXAMPLE 3(b). Let $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ be as in example 2(b) where F is known and φ, θ are the location and scale parameters. Define the shift transformation g_a as

$$g_a(x_1, x_2, \cdots, x_n) = (x_1 + a, x_2 + a, \cdots, x_n + a), \tag{5.2}$$

where a is a fixed real number. Let $\mathcal{G} = \{g_a | -\infty < a < \infty\}$ be the class of all shift transformations.

Denoting the joint distribution of $(x_1, x_2, \cdots, x_n)$ by $P_{\varphi,\theta}$, we note at once that

$$P_{\varphi,\theta}\, g_a^{-1} = P_{\varphi+a,\theta}. \tag{5.3}$$

In this example, the family $\mathfrak{F}_3$ of $\mathcal{G}$-similar subfields is the same as the family $\mathfrak{F}_2$ of φ-free subfields.

Let us call the set A $\mathcal{G}$-invariant if $A \in \mathfrak{A}$ and $g^{-1}A = A$ for all $g \in \mathcal{G}$. Likewise, let us call A almost $\mathcal{G}$-invariant if the two sets $g^{-1}A$ and A are $\mathcal{P}$-equivalent for all $g \in \mathcal{G}$. Let $\mathfrak{B}_i$ and $\mathfrak{B}_a$ be respectively the class of $\mathcal{G}$-invariant and almost $\mathcal{G}$-invariant sets. It is easy to check that $\mathfrak{B}_i$ and $\mathfrak{B}_a$ are members of the family $\mathfrak{F}_3$ of $\mathcal{G}$-similar subfields. The following proposition should be provable under some conditions.

PROPOSITION 4. *The subfield $\mathfrak{B}_a$ of almost $\mathcal{G}$-invariant sets is a maximal $\mathcal{G}$-similar subfield.*

Under some general conditions it should also be true that the subfield $\mathfrak{B}_i$ of $\mathcal{G}$-invariant sets is an essentially maximal element of $\mathfrak{F}_3$. This is so in the case of example 3(b) where F is the cdf of a normal variable.

6. The family $\mathfrak{F}_4$ of $\mathfrak{B}$-linked subfields

Let $\mathfrak{B}$ be a fixed subfield of $\mathfrak{A}$. A subfield $\mathfrak{C}$ will be called $\mathfrak{B}$-linked if $\mathfrak{B}$ is sufficient for $(\mathfrak{C}, \mathcal{P})$; that is, for every $C \in \mathfrak{C}$, there exists a $\mathfrak{B}$-measurable mapping $Q(C, \cdot)$ of $\mathfrak{X}$ into the unit interval such that, for all $B \in \mathfrak{B}$ and $P \in \mathcal{P}$,

$$P(BC) = \int_B Q(C, \cdot)\, dP(\cdot). \tag{6.1}$$

Let $\mathfrak{F}_4$ be the family of all $\mathfrak{B}$-linked subfields. The trivial subfield is again the least element of $\mathfrak{F}_4$. We shall presently see that $\mathfrak{F}_4$ always has maximal elements.

EXAMPLE 4(a). (i) Let $\mathfrak{B}$ be the trivial subfield. It is easy to see, in this instance, that $\mathfrak{F}_4$ is the same as the family of all ancillary subfields.

(ii) Let us suppose that $\mathcal{P}$ is indexed by the parameters φ and θ. Let $\mathfrak{B}$ be a fixed φ-free subfield, that is, a member of $\mathfrak{F}_2$ as defined in section 4. In this instance, every $\mathfrak{B}$-linked subfield is also φ-free.

(iii) Let $\mathfrak{B}$ be a sufficient subfield. In this case $\mathfrak{F}_4$ is the family of all subfields.

EXAMPLE 4(b). Let $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ be as in example 1(b), and let $\mathfrak{B}_0$ be the subfield induced by the sample variance $\Sigma(x_i - \bar{x})^2/n$. If $\mathfrak{C}$ is the subfield induced by

$$D = (x_1 - x_n, x_2 - x_n, \cdots, x_{n-1} - x_n), \tag{6.2}$$

then it is easy to check that $\mathfrak{C}$ is $\mathfrak{B}_0$-linked. Since $\mathfrak{B}_0$ is φ-free, it follows that every $\mathfrak{B}_0$-linked subfield is also φ-free. It is possible to show that $\mathfrak{C}$ is an essentially maximal $\mathfrak{B}_0$-linked subfield. The truth of the following proposition is worth investigating.

PROPOSITION 5. *If $\mathcal{B}_0$ and $\mathcal{C}$ are as in example* 4(b), *then $\mathcal{C}$ is an essentially largest element of the family $\mathfrak{F}_4$ of the $\mathcal{B}_0$-linked subfields.*

7. Existence of maximal elements

In this section we develop some general methods to prove the existence of maximal elements in the families $\mathfrak{F}_1$, $\mathfrak{F}_2$, $\mathfrak{F}_3$, and $\mathfrak{F}_4$. Let us first note a common feature of the four families of subfields. Each $\mathfrak{F}_i$ $(i = 1, 2, 3, 4)$ is the totality of all subfields that can be embedded in a certain class $\mathcal{E}_i$ of measurable sets. This will be clear once we define the four classes $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$, and $\mathcal{E}_4$ of measurable sets.

DEFINITIONS. (i) *Let $\mathcal{E}_1$ be the class of all $\mathcal{B}$-independent (see section 3) sets*; $\mathcal{E}_1 = \{A | P(AB) = P(A)P(B), \text{ for all } P \in \mathcal{P}, B \in \mathcal{B}\}$.

(ii) *Let $\mathcal{E}_2$ be the class of all φ-free (see section 4) sets*; $\mathcal{E}_2 = \{A | P_{\varphi,\theta}(A) \text{ does not involve } \varphi\}$.

(iii) *Let $\mathcal{E}_3$ be the class of all $\mathcal{G}$-similar (see section 5) sets*; $\mathcal{E}_3 = \{A | P(g^{-1}A) = P(A) \text{ for all } P \in \mathcal{P}, g \in \mathcal{G}\}$.

(iv) *Let $\mathcal{E}_4$ be the class of all $\mathcal{B}$-linked (see section 6) sets*; $A \in \mathcal{E}_4$ *if and only if there exists a $\mathcal{B}$-measurable mapping $Q(A, \cdot)$ of $\mathfrak{X}$ into the unit interval such that $P(AB) = \int_B Q(A, \cdot)\, dP(\cdot)$ for all $P \in \mathcal{P}$ and $B \in \mathcal{B}$.*

It is now clear that, for $i = 1, 2, 3, 4$,

$$\mathfrak{F}_i = \{\mathcal{C} | \mathcal{C} \text{ is a subfield and } \mathcal{C} \subset \mathcal{E}_i\}, \tag{7.1}$$

that is, $\mathfrak{F}_i$ is the family of all subfields that can be embedded in the class $\mathcal{E}_i$ of measurable sets.

Our first general result is the following.

THEOREM 1. *Each $\mathcal{E}_i$, $(i = 1, 2, 3, 4)$ has the following properties:*

(a) $\varnothing \in \mathcal{E}_i$, $\mathfrak{X} \in \mathcal{E}_i$;

(b) $A \in \mathcal{E}_i$, $B \in \mathcal{E}_i$, $A \subset B \Rightarrow B - A \in \mathcal{E}_i$;

(c) *$\mathcal{E}_i$ is closed for countable disjoint unions.*

The proof of theorem 1 is routine and hence omitted. An immediate consequence of theorem 1 is the following.

COROLLARY. *Each $\mathcal{E}_i$, $(i = 1, 2, 3, 4)$ is a monotone class of sets.*

The following are our fundamental existence theorems.

THEOREM 2. *If $\mathcal{E}$ is a given monotone class of sets, and $\mathfrak{F}$ is the family of all Borel fields that could be embedded in $\mathcal{E}$, then corresponding to each element $\mathcal{C}$ of $\mathfrak{F}$, there exists a maximal element $\tilde{\mathcal{C}}$ of $\mathfrak{F}$ such that $\mathcal{C} \subset \tilde{\mathcal{C}}$.*

PROOF. Let $\{\mathcal{C}_t | t \in T\}$ be an arbitrary subfamily of $\mathfrak{F}$, which is linearly ordered with respect to the partial order of inclusion relationship, and let $\mathcal{C}_0 = \bigcup_{t \in T} \mathcal{C}_t$.

Since $\{\mathcal{C}_t\}$ is linearly ordered, it follows that $\mathcal{C}_0$ is a field of sets. The monotone extension of $\mathcal{C}_0$ is then the same as the Borel extension $\mathcal{C}_1$ of $\mathcal{C}_0$. Since $\mathcal{E}$ is monotone and $\mathcal{C}_0 \subset \mathcal{E}$, it follows that $\mathcal{C}_1 \subset \mathcal{E}$ and hence $\mathcal{C}_1 \in \mathfrak{F}$. Thus, every linearly ordered subfamily of $\mathfrak{F}$ has an upper bound in $\mathfrak{F}$. Theorem 2 is then a consequence of Zorn's Lemma.

An immediate consequence of theorems 1 and 2 is theorem 3.

THEOREM 3. *For each* $\mathcal{C} \in \mathfrak{F}_i$ *there exists a maximal element* $\tilde{\mathcal{C}}$ *in* $\mathfrak{F}_i$ *such that* $\mathcal{C} \subset \tilde{\mathcal{C}}$, $(i = 1, 2, 3, 4)$.

8. Some general results

Let $\mathcal{E}$ be a class of measurable sets having the same characteristics as those of the classes $\mathcal{E}_i$ in theorem 1. That is,

(a) $\varnothing \in \mathcal{E}$, $\mathfrak{X} \in \mathcal{E}$;

(b) $A \in \mathcal{E}, B \in \mathcal{E}, A \subset B \Rightarrow B - A \in \mathcal{E}$;

(c) $\mathcal{E}$ is closed for countable disjoint unions.

Let $\mathfrak{F}$ be the family of all the subfields that may be embedded in $\mathcal{E}$, and let $\mathfrak{F}_0$ be the subfamily of all the maximal elements in $\mathfrak{F}$. That $\mathfrak{F}_0$ is not vacuous has been established in theorem 2.

Two members A and B of $\mathcal{E}$ are said to 'conform' if $AB \in \mathcal{E}$. The set $A \in \mathcal{E}$ is said to be 'conforming' if $AB \in \mathcal{E}$ for all $B \in \mathcal{E}$. If every member of $\mathcal{E}$ is conforming, then $\mathcal{E}$ must itself be a Borel field; hence, there is no problem since $\mathfrak{F}_0$ consists of a single member, namely $\mathcal{E}$ itself. A subfield is 'conforming' if every one of its members is so.

THEOREM 4. *Let* $\mathfrak{D}$ *be the class of all the conforming sets in* $\mathcal{E}$, *that is,*

$$A \in \mathfrak{D} \Leftrightarrow A \in \mathcal{E} \quad \text{and} \quad AB \in \mathcal{E} \quad \text{for all} \quad B \in \mathcal{E}. \tag{8.1}$$

Let $\mathfrak{M}$ *stand for a typical element of* $\mathfrak{F}_0$; *that is,* $\mathfrak{M}$ *is a maximal element of* $\mathfrak{F}$:

(i) $\mathfrak{M}$ *is a maximal element of* $\mathfrak{F}$ *if and only if* $A \in \mathcal{E} - \mathfrak{M}$ *implies that* A *does not conform to at least one member of* $\mathfrak{M}$;

(ii) $\mathfrak{D}$ *is a subfield and is equal to the intersection of all the maximal elements in* $\mathfrak{F}$. *It is the largest conforming subfield;*

(iii) $\mathcal{C}$ *is a conforming subfield if and only if for* $\mathcal{B} \in \mathfrak{F}$ *it is true that* $\mathcal{C} \vee \mathcal{B} \in \mathfrak{F}$, *where* $\mathcal{C} \vee \mathcal{B}$ *stands for the least subfield containing both* $\mathcal{C}$ *and* $\mathcal{B}$.

PROOF. Let $\mathfrak{M} \in \mathfrak{F}_0$, and let A be a fixed member of $\mathcal{E} - \mathfrak{M}$. If possible, let A conform to all the members of $\mathfrak{M}$. Consider the class $\mathfrak{M}^*$ of sets of the type $AM_1 \cup A'M_2$, where M_1 and M_2 are arbitrary members of $\mathfrak{M}$. It is easy to check that $\mathfrak{M}^* \in \mathfrak{F}$ and that $A \in \mathfrak{M}^*$ and $\mathfrak{M} \subset \mathfrak{M}^*$. This violates the supposition that $\mathfrak{M}$ is a maximal element of $\mathfrak{F}$. Thus, the 'only if' part of (i) is proved. To prove the 'if' part we have only to observe that if $\mathfrak{M}$ is not maximal, then there exists a larger subfield $\mathfrak{M}^* \subset \mathcal{E}$ and this implies the existence of an $A \in \mathcal{E} - \mathfrak{M}$ that conforms to every member of $\mathfrak{M}$.

Since every member of $\mathfrak{D}$ conforms by definition to every member of $\mathcal{E}$, it is an immediate consequence of (i) that $\mathfrak{D} \subset \mathfrak{M}$ for each $\mathfrak{M} \in \mathfrak{F}_0$, that is, $\mathfrak{D} \subset \bigcap \mathfrak{M}$.

Now let M and E be typical members of $\bigcap\mathfrak{M}$ and $\mathcal{E}$ respectively. From theorem 2 there exists a maximal element $\mathfrak{M}_0$ in $\mathfrak{F}$ which contains the subfield consisting of $\varnothing$, E, E', and $\mathfrak{X}$. Thus, M and E are together in the subfield $\mathfrak{M}_0$, and hence they must conform. Since E is arbitrary, it follows that $M \in \mathfrak{D}$. We have thus proved the equality of $\mathfrak{D}$ and $\bigcap\mathfrak{M}$, and have incidentally proved the equality of $\mathcal{E}$ and $\bigcup\mathfrak{M}$. Since each $\mathfrak{M}$ is a subfield, it is now clear that $\mathfrak{D} = \bigcap\mathfrak{M}$

is also a subfield. That it is the largest conforming subfield follows from its definition.

Now let $\mathcal{C}$ be an arbitrary conforming subfield; that is, let $\mathcal{C}$ be a subfield of $\mathcal{D}$. For each $\mathcal{B} \in \mathcal{F}$ there exists (theorem 2) a maximal element $\mathcal{M}$ of $\mathcal{F}$ such that $\mathcal{B} \subset \mathcal{M}$. But $\mathcal{C} \subset \mathcal{D} \subset \mathcal{M}$. Therefore, $\mathcal{C} \vee \mathcal{B} \subset \mathcal{M} \subset \mathcal{E}$, that is $\mathcal{C} \vee \mathcal{B} \in \mathcal{F}$. This proves the 'only if' part of (iii). The 'if' part is trivial.

For example, let $\mathcal{E}$ be the class of all $\mathcal{B}$-linked sets (see sections 6 and 7) in the probability structure $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{B}$ is a fixed subfield of $\mathcal{A}$. If the set A is $\mathcal{B}$-linked, that is, if there exists a $\mathcal{B}$-measurable function $Q(A, \cdot)$ satisfying definition (iv) of section 7, then it is easily seen that AB is $\mathcal{B}$-linked for every $B \in \mathcal{B}$. We have only to define $Q(AB, \cdot)$ as $Q(A, \cdot)\, I(B, \cdot)$, where $I(B, \cdot)$ is the indicator of B.

In this case, $\mathcal{B}$ is a conforming subfield. Theorem 4(iii) then asserts that for every $\mathcal{B}$-linked subfield $\mathcal{C}$, the subfield $\mathcal{B} \vee \mathcal{C}$ is also $\mathcal{B}$-linked. It will be of some interest to find out conditions under which $\mathcal{B}$ is the largest conforming subfield, that is, $\mathcal{B} = \mathcal{D}$.

9. Some further problems

In this section we list four problems that are mostly unsolved.

(A) *Separating subfields.* Let $\mathcal{P}$ be a class of 'distinct' probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$. That is, for each pair P_1, P_2 of members of $\mathcal{P}$ there exists a measurable set $A \in \mathcal{A}$ such that $P_1(A) \neq P_2(A)$. A subfield $\mathcal{B}$ will be called 'separating' if the restriction of $\mathcal{P}$ to $\mathcal{B}$ gives rise to a class of distinct measures. For example, every sufficient subfield is separating. No ancillary or φ-free (see section 4) subfield is separating.

Let $\mathcal{F}_5$ be the family of all separating subfields. By definition, $\mathcal{A}$ is the largest element of $\mathcal{F}_5$. What can we say about the existence of minimal elements in $\mathcal{F}_5$? A variant of this problem has recently received some attention in the USSR ([6], [8]). A partition Π of $\mathcal{X}$ into a class of disjoint measurable sets $\{A_t\}$ will be called 'separating' if, for each pair P_1, P_2 of member of $\mathcal{P}$, there exists a member A_t of the partition Π such that $P_1(A_t) \neq P_2(A_t)$. A separating partition is called minimal if there exists no other separating partition with a smaller number of parts. Let $\nu(\mathcal{P})$ stand for the number, possibly infinite, of parts in a minimal separating partition. What can we say about $\nu(\mathcal{P})$?

EXAMPLE 5(a). Consider the class $\mathcal{P}$ of all normal distributions on the real line with unit variances. Here $\nu(\mathcal{P}) = 2$. Any partition of the real line into two half lines is clearly separating and, of course, minimal. The corresponding subfield is a minimal element of $\mathcal{F}_5$.

EXAMPLE 5(b). Let $\mathcal{P}$ be the family of uniform distributions on $[0, \theta]$, $0 < \theta < 1$. In this case $\nu(\mathcal{P}) = 3$ (see [6]).

EXAMPLE 5(c). If $\mathcal{P}$ consists of a finite number of measures $P_1, P_2, \cdots, P_n$, then $\nu(\mathcal{P}) \leq n$. If $\mathcal{P}$ consists of a countable number of continuous measures, then $\nu(\mathcal{P}) = 2$ (see [6], [8]).

(B) *Partially sufficient subfields.* The notion of partial sufficiency, as introduced by Fraser [5], is as follows.

Let $\mathcal{P} = \{P_{\varphi,\theta}\}$, $\varphi \in \Phi$, $\theta \in \Theta$, be a family of probability measures indexed by the two independent parameters φ and θ. A subfield $\mathcal{B} \subset \mathcal{A}$ will be called θ-sufficient for $\mathcal{A}$ (or simply θ-sufficient) if

(i) $\mathcal{B}$ is φ-free in the sense of section 4, and

(ii) for each $A \in \mathcal{A}$ there exists a choice of the conditional probability (function) of A given $\mathcal{B}$ that does not depend on θ; that is, for each $\varphi \in \Phi$, there exists a $\mathcal{B}$-measurable function $Q_\varphi(A, \cdot)$ that maps $\mathfrak{X}$ to the unit interval in such a manner that

$$P_{\varphi,\theta}(AB) \equiv \int_B Q_\varphi(A, \cdot)\, dP_{\varphi,\theta}(\cdot) \tag{9.1}$$

for all $B \in \mathcal{B}$ and $\theta \in \Theta$.

Let $\mathfrak{F}_6$ be the family of all θ-sufficient subfields. Under what conditions can we prove that $\mathfrak{F}_6$ is not vacuous? What about the minimal and maximal elements in $\mathfrak{F}_6$?

(C) *Complete subfields.* Given a probability structure $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$, we call a subfield $\mathcal{B}$ 'complete' if for a $\mathcal{B}$-measurable, $\mathcal{P}$-integrable function f, the integral $\int_{\mathfrak{X}} f\, dP \equiv 0$, for all $P \in \mathcal{P}$, when, and only when, f is $\mathcal{P}$-equivalent to zero. Let $\mathfrak{F}_7$ be the family of all complete subfields. What can we say about the existence of maximal and minimal elements in $\mathfrak{F}_7$?

Let us terminate this list of problems with a final one.

(D) *Complementary subfield.* Let $(\mathfrak{X}, \mathcal{A})$ be a given measurable space and let $\mathcal{B}$ be a fixed subfield of $\mathcal{A}$. A subfield $\mathcal{C}$ will be called a complement to $\mathcal{B}$ if $\mathcal{B} \vee \mathcal{C} = \mathcal{A}$, that is, if $\mathcal{A}$ is the least Borel field that contains both $\mathcal{B}$ and $\mathcal{C}$.

Let $\mathfrak{F}_8$ be the family of all subfields that are complements to $\mathcal{B}$. For example, if $\mathcal{B}$ is the trivial subfield, then $\mathfrak{F}_8$ consists of a single element, namely $\mathcal{A}$ itself. If $\mathcal{B} = \mathcal{A}$, then $\mathfrak{F}_8$ consists of all subfields of $\mathcal{A}$.

Of course, $\mathcal{A}$ is the largest element of $\mathfrak{F}_8$. It is easy to construct examples where $\mathfrak{F}_8$ has a multiplicity of minimal elements. Whether $\mathfrak{F}_8$ always has a minimal element is not known.

10. An addendum

Of the several speculatory statements made (and listed as propositions) in this paper, E. L. Lehmann has recently proved proposition 3 under some conditions on F. Counterexamples to propositions 1 and 2 have been obtained by J. K. Ghosh.

REFERENCES

[1] D. Basu, "The family of ancillary statistics," *Sankhyā*, Vol. 21 (1959), pp. 247–256.

[2] ———, "On maximal and minimal sub-fields of certain types," *Institute of Statistics Mimeo Series, No. 422*, University of North Carolina, 1965.

[3] D. L. BURKHOLDER, "Sufficiency in the undominated case," *Ann. Math. Statist.*, Vol. 32 (1961), pp. 1191–1200.
[4] G. B. DANTZIG, "On the non-existence of tests of Student's hypothesis having power functions independent of σ," *Ann. Math. Statist.*, Vol. 11 (1941), pp. 186–191.
[5] D. A. S. FRASER, "Sufficient statistics with nuisance parameters," *Ann. Math. Statist.*, Vol. 27 (1956), pp. 838–842.
[6] A. M. KAGAN and V. N. SUDAKOV, "Separating partitions of certain families of measures," Vestnik Leningrad University, No. 13 (1964), pp. 147–150. (In Russian.)
[7] T. S. PITCHER, "Sets of measures not admitting necessary and sufficiency statistics or sub-fields," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 267–268.
[8] S. M. VISIK, A. A. COBRINKSII, and A. L. ROSENTHAL, "A separating partition for a finite family of measures," *Teor. Verojatnost. i Primenen.*, Vol. 9 (1964), pp. 165–167. (In Russian.)

The Annals of Mathematical Statistics
1969, Vol. 40, No. 1, 162-174

INVARIANT SETS FOR TRANSLATION-PARAMETER FAMILIES OF MEASURES

BY D. BASU[1] AND J. K. GHOSH[2]

Indiana Statistical Institute

1. Introduction. In this paper we discuss a number of problems which have their origin in statistics but whose main interest is measure-theoretical. It is to the statistician interested in abstract harmonic analysis and to the harmonic analyst interested in statistics that the paper is addressed.

Let $\mathcal{P} = \{P\}$ be a family of probability measures on an arbitrary measurable space $(X, \mathcal{A})$. The set $A \in \mathcal{A}$ is called '$\mathcal{P}$-invariant' (in preference to the more familiar expression 'similar region') if $P(A)$ is a constant in P. The class $\mathcal{A}(\mathcal{P})$ of $\mathcal{P}$-invariant sets contains all sets that are $\mathcal{P}$-equivalent to the empty set or the whole space, and is closed for complementation and countable disjoint unions. In general, $\mathcal{A}(\mathcal{P})$ is not a sub-σ-field of $\mathcal{A}$.

The set $A \in \mathcal{A}(\mathcal{P})$ is 'non-trivial' if A is not $\mathcal{P}$-equivalent to the empty set or the whole space. If every member of $\mathcal{A}(\mathcal{P})$ is 'trivial' then we call the family 'weakly complete'. The name is suggested by the fact that 'completeness' $\Rightarrow$ 'bounded completeness' $\Rightarrow$ 'weak completeness'.[3] That weak completeness does not imply bounded completeness is seen from the example where X consists of only three points with a probability distribution θ, θ and $1 - 2\theta$, where $0 < \theta < \frac{1}{2}$.

If $\mathcal{P}$ is not weakly complete, i.e., if there exist non-trivial $\mathcal{P}$-invariant sets, then we call the family of measures 'weakly incomplete.' If $\mathcal{P}$ consists of a finite number of non-atomic measures, then its weak incompleteness is an immediate consequence of a well-known result due to Liapunov [12]. In this situation the class $\mathcal{A}(\mathcal{P})$ is very wide and contains sets of all 'sizes.'

Our main concern is with the weak incompleteness of families of probability measures. Here, we restrict ourselves almost exclusively to the particular situation where $\mathcal{P}$ is a translation parameter family of probability measures. At the risk of some repetitions, this paper brings together a number of results some of which have been noted elsewhere.

2. Notations and a few basic propositions. For the sake of simplicity of exposition we consider the case where X is the additive group of real numbers.

Received 8 August 1966; revised 15 December 1967.

[1] Research was done mainly at the University of Chicago and supported in part by Research Grant No. NSF GP 3707 from the Division of Mathematical, Physical and Engineering Sciences of the National Science Foundation, and in part by the Statistics Branch, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] Research was done mainly at the University of Illinois and supported partly by NSF Grant GP-3814.

[3] $\mathcal{P}$ is complete if $\int f\,dP = 0$ for all $P \in \mathcal{P}$ implies $f = 0$ a.e. $\mathcal{P}$. $\mathcal{P}$ is boundedly complete if f is bounded and $\int f\,dP = 0$ for all $P \in \mathcal{P}$ implies $f = 0$ a.e. $\mathcal{P}$.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_21, © Springer Science+Business Media, LLC 2011

But most of the results of this section are true for arbitrary locally compact commutative groups. We shall consider this general setting in a later section.

We take $\mathfrak{A}$ to be the σ-field of Borel sets and we denote the Lebesgue measure by λ. Though our main interest is in probability measures it is necessary for certain results to consider the class M of all bounded signed measures. M will be equipped with a topology in the usual way by setting norm of $\mu \in M$ as $\sup_A |\mu|(A)$. With this norm and convolution (denoted by $*$) as multiplication, M becomes a Banach algebra. We denote by M_p the class of probability measures.

For each $\mu \in M$ we have the translation parameter family $[\mu] = \{\mu_\theta \mid \theta \in X\}$ where $\mu_\theta(A) = \mu(A - \theta)$ and $A - \theta = \{x - \theta \mid x \in A\}$. Let $\mathfrak{A}[\mu]$ stand for the class of $[\mu]$-invariant sets, i.e., the class of sets A with the property $\mu_\theta(A) = \mu(A)$ for all θ. We shall write $\check{f}(x)$ to denote $f(-x)$.

A class of sets $\mathfrak{B}$ is translation invariant if $B \in \mathfrak{B} \Rightarrow B + t \in \mathfrak{B}$. We state below a number of results with a sketch of proof where necessary.

LEMMA 1. *$\mathfrak{A}(\mu)$ is a translation invariant monotone class which is closed with respect to complements and countable disjoint unions and contains the empty set and the whole space X.*

LEMMA 2. *If $A \in \mathfrak{A}$ and $\phi(\theta) = \mu_\theta(A)$ then $\lambda(A)\mu(X) = \int \phi(\theta)\, d\lambda(\theta)$.*

Proof of Lemma 2 is an easy application of Fubini's theorem. Alternatively one may state the above lemma as: "For all unitary measures μ the convolution $\lambda * \mu = \lambda$." If $\lambda(A) = \infty$ but $\mu(X) = 0$ we take $\lambda(A) \cdot \mu(X)$ as zero.

LEMMA 3. *If $A \in \mathfrak{A}[\mu]$ and $\mu(X) \neq 0$ then $\lambda(A) = 0$ or infinity according as $\mu(A) = 0$ or $\neq 0$.*

This lemma implies that for a probability measure μ a non-trivial $[\mu]$-invariant set has infinite Lebesgue measure.

LEMMA 4. *If $A \in \mathfrak{A}[\mu]$ then $A \in \mathfrak{A}[\mu * \nu]$ and $(\mu * \nu)(A) = \mu(A)\nu(X)$.*

COROLLARY. *If μ and ν are unitary and A belongs to both $\mathfrak{A}[\mu]$ and $\mathfrak{A}[\nu]$ then $\mu(A) = \nu(A)$.*

PROOF. $\mu(A) = (\mu * \nu)(A) = (\nu * \mu)(A) = \nu(A)$ by Lemma 4.

For each measurable set A let $M(A)$ be the family of measures μ such that A is $[\mu]$-invariant. $M(A)$ always contains the null measure. The relation between A and $M(A)$ is a dual of the relation between μ and $\mathfrak{A}[\mu]$. The following theorem is essentially a restatement of the preceding results.

THEOREM 1. *Let A be a fixed set in $\mathfrak{A}$.*

(i) *$M(A)$ is a closed ideal of the Banach algebra, M,* i.e., *$M(A)$ is a closed linear subspace of M which is closed with respect to convolution with every $\nu \in M$.*

(ii) *If M_p is the class of probability measures, then $M(A) \cap M_p$ is a (possibly empty) closed convex subset of M and $\mu(A)$ is a constant for all μ in $M(A) \cap M_p$ and this common value is zero or positive according as $\lambda(A) = 0$ or ∞.*

The following two lemmas prove that we are not working in a vacuum and that weakly incomplete translation parameter families do exist.

LEMMA 5. *If μ be the (normalized) restriction of the Lebesgue measure λ to the interval $J = [a, a + \delta)$ then*

(i) *$A \in \mathfrak{A}[\mu]$ if and only if A is an essentially periodic set with period δ,* i.e.,

$A \triangle (A + \delta)$ *is a Lebesgue null set.*

(ii) $\mathcal{A}[\mu]$ *contains sets of all sizes.*

(iii) $\mathcal{A}[\mu]$ *is a sub-σ-field.*

PROOF. If $A \in \mathcal{A}[\mu]$ and I_A be the indicator of A then, from the fact that

$$\int_\theta^{\theta+\delta} I_A(x)\, d\lambda(x) \text{ is a constant in } \theta,$$

it follows that $I_A(\theta + \delta) - I_A(\theta) = 0$ for almost all (a.e. $[\lambda]$) values of θ. And this in turn implies the essential periodicity of A with period δ.

To prove the 'if' part of (i) let us first note that if A be essentially periodic with period δ, then, for each integer n, A is essentially equal to $A + n\delta$. Also note that, for each $\theta \in X$, the sequence of sets $\{J + \theta + n\delta\}$, $n = 0, \pm 1, \pm 2, \cdots$ is a partition of X. The $[\mu]$-invariance of A then follows from the following chain of equalities.

$$\begin{aligned}
\delta\mu(A - \theta) &= \lambda[(A - \theta) \cap J] = \lambda[A \cap (J + \theta)] \\
&= \textstyle\sum_n \lambda[A \cap (J + \theta) \cap (J + n\delta)] \\
&= \textstyle\sum_n \lambda[(A - n\delta) \cap (J + \theta - n\delta) \cap J] \\
&= \textstyle\sum_n \lambda[A \cap J \cap (J + \theta - n\delta)] \\
&= \lambda(A \cap J) = \delta\mu(A).
\end{aligned}$$

Now, if $0 < \alpha < 1$ and A_0 be a sub-set of J such that $\lambda(A_0) = \alpha\delta$ and $A = \mathsf{U}_n(A_0 + n\delta)$ then it is clear that $A \in \mathcal{A}[\mu]$ and that $\mu(A) = \alpha$. This proves (ii). The proof of (iii) is elementary and hence omitted. In a later section we shall indicate how to generalize the 'if' part of (i) to general topological groups. It is easy to see how the lemma may be generalized to an arbitrary Euclidean space.

Let us contrast Lemma 5 with the following:

LEMMA 6. *If μ be the uniform discrete distribution over the two points a and $a + \delta$ then*

(i) *the empty set and the whole space are the only trivial $[\mu]$-invariant sets;*

(ii) *A is a non-trivial member of $\mathcal{A}[\mu]$ if and only if A and $A + \delta$ are complements of each other; all such sets are of size $\frac{1}{2}$;*

(iii) $\mathcal{A}[\mu]$ *is not a sub-σ-field.*

A detailed proof of the above lemma is perhaps unnecessary. Only observe that A is a non-trivial $[\mu]$-invariant set if and only if, for every x it is true that A contains exactly one of the two elements x and $x + \delta$. Now, for each x let $S_x = \{x + n\delta \mid n = 0, \pm 1, \cdots\}$ and let B be a set that has exactly one point in common with each S_x. For instance, we may take B to be the interval $[a, a + \delta)$. It is now easily seen that the set $A = \mathsf{u}(B + m\delta)$ where m runs through the set of even integers is a typical non-trivial member of $\mathcal{A}[\mu]$.

Lemma 6 tells us that the translation parameter family $[\mu]$ generated by any uniform two-point discrete distribution μ is weakly incomplete. It is of some interest to note that if μ is not uniform then the family $[\mu]$ is boundedly (hence

weakly) complete but is incomplete. The generalization of Lemma 6 to the case where μ is a uniform discrete distribution on a finite number, say n, of points in an arithmetical progression is almost immediate. In this case we have non-trivial $[\mu]$-invariant sets of sizes i/n $(i = 1, 2, \cdots, n - 1)$.

We thus see that weakly incomplete translation-parameter families do exist and that if $[\mu]$ is weakly incomplete then so also is $[\nu]$ if μ is a 'factor' of ν (i.e., if $\nu = \mu * \sigma$).

In the next section we consider the case where $[\mu]$ is a dominated family of measures.

3. The dominated case.

3.1. *Miscellaneous results.*

LEMMA 7. *If the translation-parameter family* $[\mu]$ *is dominated by a* σ*-finite measure* σ *then it is also dominated by the Lebesgue measure* λ.

The proof is given by Ferguson (Lemma 2 in [8]).

In view of the above lemma in the dominated case we may take λ as the dominating measure.

LEMMA 8. *If* μ *be a probability measure dominated by* λ *then given* $\epsilon > 0$ *there exists* $\delta > 0$ *such that* $\sup_A |\mu(A - \theta) - \mu(A - \theta')| < \epsilon$ *whenever* $|\theta - \theta'| < \delta$.

The above lemma holds for an arbitrary bounded signed measure by splitting it up into its positive and negative parts. This result is well-known, Rudin ([18], p. 3). This lemma was also proved by Ferguson [7] but there is a gap in his proof; he wrongly asserts that $\lambda[A \triangle (A - \theta)] \to 0$ as $\theta \to 0$ which is true only if $\lambda(A) < \infty$. It is possible to construct an alternative proof by constructing a sequence of continuous probability densities $p_n \to d\mu/d\lambda$ a.e. and applying Scheffé's theorem [19].

The above lemma gives us an insight into how to construct a set that is approximately $[\mu]$-invariant. If A be a periodic set with period δ then note that so also is the function $\phi(\theta) = \mu(A - \theta)$. If we choose δ sufficiently small then A would be approximately $[\mu]$-invariant. The above considerations lead to the following generalization.

THEOREM 2. *Let* $\{f(x, \theta)\}$ *be a family of probability density functions with respect to the Lebesgue measure* λ *on the real line* X *and let* θ *be real valued. If, for each* $x \in X$, *the function* $f(x, \theta)$ *is continuous in* θ, *then given any* $0 < \alpha < 1$, $\epsilon > 0$ *and* $0 < K < \infty$ *there exists a set* A *such that* $|P_\theta(A) - \alpha| < \epsilon$ *for* $|\theta| < K$.

In the case of a dominated location-parameter family we may take $K = \infty$ and drop the assumption of continuity for f.

The first part of the above proposition follows from Liapunov's theorem and Scheffé's convergence theorem. It is implicitly stated in a paper of Dvoretzky, Wald and Wolfowitz [7]. The second part follows from the remarks preceding the theorem.

The following characterization of bounded completeness, vide [9], is a reformulation of a famous Tauberian theorem of Wiener.

THEOREM 3. (Wiener) *Let* μ *be dominated by the Lebesgue measure. Then* $[\mu]$

is boundedly complete iff *the Fourier transform* $\int_{-\infty}^{\infty} e^{-itx}\, d\mu(x) = \hat{\mu}(t)$ *does not vanish anywhere.*

Wiener's theorem leads to an interesting necessary condition for weak incompleteness.

THEOREM 4. *If* μ *is dominated by the Lebesgue measure* λ *then a necessary condition that* $[\mu]$ *be weakly incomplete is that* $\hat{\mu}(t)$ *vanishes at an infinite number of points.*

PROOF. Let $d\mu = f\, d\lambda$, $f \in L_1(\lambda)$. We shall show that if $\hat{\mu}(t) = 0$ only at a finite number of points, then $[\mu]$ is weakly complete. Let $\hat{\mu}(t) = 0$ iff $t = t_1, \cdots, t_m$ and define

$$M = \{\psi; \quad \psi \in L_\infty(\lambda), \quad \psi(x) = \sum a_j e^{it_j x} \text{ a.e. } (\lambda)\}.$$

Let N be the closed ideal in $L_1(\lambda)$ generated by f. By Wiener's theorem ([18], 7.2.4), $N = \{\phi; \phi \in L_1(\lambda), \hat{\phi}(t) = 0 \text{ if } t = t_1, \cdots, t_m\}$. Suppose $\int_A f(x - \theta)\, d\lambda = c$. Then $\int_{-\infty}^{\infty} (I_A(x) - c)\phi(x)\, d\lambda = 0$ if $\phi \in N$. Since M is finite dimensional it follows that $I_A - c \in M$, i.e., $I_A(x) = c + \sum a_j e^{it_j x}$ a.e. (λ). Since every open set has positive λ-measure, the continuous function $c + \sum a_j e^{it_j x}$ can only take the values 0 or 1. Finally, R being connected, this implies that either $I_A(x) = 0$ a.e. (λ) or $I_A(x) = 1$ a.e. (λ). This completes the proof.

Theorems 3 and 4 allow us to construct easily a μ such that $[\mu]$ is weakly complete but not boundedly complete. For example let $d\mu = x^2(2\pi)^{-\frac{1}{2}} e^{-x^2/2}\, d\lambda$. Then $\hat{\mu}(t) = (1 - t^2)e^{-t^2/2}$ which vanishes iff $t = \pm 1$.

3.2 *Sufficiency and Neyman structure.* We continue to consider the class of dominated measures. Equivalently we consider the class of Lebesgue integrable functions $L_1(X)$.

Let μ be a given dominated probablility measure with $d\mu = f\, d\lambda$. A natural thing to look for is a $\{\mu\}$-invariant set with Neyman structure. A set A has Neyman structure if there exists a sufficient sub-σ-field $\mathcal{A}_1$ for $\{\mu\}$ and $P_\mu(A \mid \mathcal{A}_1) =$ constant. As the following theorem, Theorem 4 shows there exist no such non-trivial sets.

Let $\mathcal{A}_s$ be the minimal sufficient σ-field which exists by Bahadur's theorem [1].

THEOREM 5. $\mathcal{A}_s = \mathcal{A}$ wrt λ, i.e., *for any* $A \in \mathcal{A}$ *there exists* $B \in \mathcal{A}_s$ *such that* $\lambda(A \triangle B) = 0$.

The above theorem follows from a more general result of Pitcher [17]. This theorem may be interpreted as saying the minimal sufficient sub σ-field is the maximal invariant sub σ-field under the group of transformations (in fact identity only) which leave each P_θ invariant. In this form the theorem is true for any separable, locally compact abelian group G in place of R, with $m_0 = \lambda$ taken as the Haar measure.

3.3 *Characterisation of some weakly incomplete families.* We confine attention to all μ dominated by λ and identify μ_f with f where $f \in L_1(\lambda)$ and $d\mu = f\, d\lambda$.

Two of our main problems are to characterise $M(A)$ where $M(A)$ is the class of all μ_f with $A \in \mathcal{A}[\mu_f]$ (at least for some sets A) and to give conditions under which $\mathcal{A}[\mu]$ is a σ-field. We shall throw some light on these questions by solving a related problem.

Consider the following problem: given a translation invariant sub-σ-field $\mathfrak{A}_1$ of $\mathfrak{A}$ characterize all $f \in L_1(\lambda)$ such that $\mathfrak{A}_1 \subset \mathfrak{A}[\mu]$ where $d\mu = f\,d\lambda$.

Obviously unless $\mathfrak{A}_1$ is a proper sub-σ-field no such f can exist.

Before this problem can be solved it is necessary to know the translation invariant proper sub-σ-fields. We do not know if the following method is the only way of generating them.

Take any countable closed subgroup X_1 of the reals i.e., $X_1 = \{\pm n\delta;\ n = 0, 1, 2, \cdots\}$, $\delta > 0$, and consider all members $A \in \mathfrak{A}$ which are invariant under translations by elements of X_1. Any X_1-invariant Borel set is obtained by taking a Borel set A in $[0, \delta)$ and then forming the union $\bigcup_{n=0,\pm1,\ldots}(A + n\delta)$. The class of such X_1-invariant sets forms a translation invariant proper sub-σ-field of $\mathfrak{A}$, isomorphic to the class of Borel sets in $[0, \delta)$. We shall denote this class by $\mathfrak{A}_{X_1}$. We can prove the following:

THEOREM 6. *If $\mathfrak{A}_1$ is a countably generated translation invariant proper sub-σ-field of $\mathfrak{A}$, then either $\mathfrak{A}_1$ is the trivial σ-field or there is a subgroup $\{X_1 = \pm n\delta;\ n = 0, 1, 2, \cdots,\}\delta > 0$ such that $\mathfrak{A}_1$ is the class of all X_1-invariant sets of $\mathfrak{A}$.*

PROOF. Let A denote an atom of $\mathfrak{A}_1$. Let $X_A = \{x;\ x \in X,\ x + A = A\}$. Then it is easy to check that X_A is a subgroup of X not depending on A and that in fact X_A is the atom containing the origin. By [3], Theorem 3, $\mathfrak{A}_1$ contains all X_A-invariant measurable sets. Hence by [16], Theorem 7.2, X_A is closed. So $X_A = X$ or $\{0\}$ or $\{n\delta;\ n = 0, \pm1, \cdots\}$, $\delta > 0$. The theorem follows from this.

In the following we write $\mathfrak{A}_1$ for $\mathfrak{A}_{X_1}$.

Unfortunately, though $\mathfrak{A}$ itself is countably generated its sub-σ-fields need not have the same property. It is true that to any sub-σ-field of $\mathfrak{A}$ there corresponds countably generated sub-σ-field equivalent with respect to λ but we are unable to prove the conjecture this suggests, namely, that if $\mathfrak{B}$ is any translation invariant sub-σ-field of $\mathfrak{A}$ then $\mathfrak{B} = \mathfrak{A}$ or $\mathfrak{A}_1$ or the trivial σ-field wrt λ. It may be interesting to observe that if $\mathfrak{B}$ has in addition the property that there exists a set $B \in \mathfrak{B}$ with $0 < \lambda(B) < \infty$, then $\mathfrak{B} = \mathfrak{A}$ wrt λ; for define $d\mu = (I_B(x)\,d\lambda)/\lambda(B)$ and apply Theorem 5. (Alternatively this result can be directly proved and Theorem 5 derived from it.) Also see Theorem 13 in this connection.

Let us notice that the $[\mu]$-invariant sub-σ-field of Lemma 6 is equivalent to $\mathfrak{A}_1$ wrt λ.

In the following we write μ_f for μ to indicate $d\mu = f\,d\lambda$ but if there is no fear of confusion we shall use μ instead of μ_f.

THEOREM 7. *Let $X_1 = \{\pm n\delta;\ \delta > 0,\ n = 0, 1, 2, \cdots\}$ and $\mathfrak{A}_1$ the sub-σ-field of all X_1-invariant sets of $\mathfrak{A}$. Then $\mathfrak{A}_1 \subset \mathfrak{A}[\mu_f]$ iff f lies in the closed ideal in $L_1(\lambda)$ generated by f_δ = characteristic (= indicator) function of the interval $[0, \delta)$.*

PROOF. Without loss of generality we take $\delta = 1$. That $\mathfrak{A}_1 \subset \mathfrak{A}[\mu_f]$ if f lies in the closed ideal generated by f_1 follows from Theorem 1 and Lemma 6. We now prove the converse also holds. Suppose $\mathfrak{A}_1 \subset \mathfrak{A}[\mu_f]$. Then

$$\begin{aligned}\hat{f}(2n\pi) &= \int_{-\infty}^{\infty} e^{-2n\pi ix}\,d\mu \\ &= \int_{-\infty}^{\infty} E_\mu(e^{-2n\pi ix} \mid \mathfrak{A}_1)\,d\mu \\ &= \mu(X)\int_{-\infty}^{\infty} e^{-2n\pi ix} f_1\,d\lambda.\end{aligned}$$

Since $E_\mu(e^{-2n\pi ix} \mid \mathcal{A}_1) = e^{-2n\pi ix}$ and by Lemma 5

$$\begin{aligned}\mu(A)\nu(X) &= \nu(A)\mu(X) \quad \text{where} \quad d\nu = f_1\, d\lambda,\\ &= \mu(X)\textstyle\int_0^1 e^{-2n\pi ix}\, d\lambda\\ &= 0 \quad \text{if} \quad n \text{ is any non-zero integer.}\end{aligned}$$

Since $\hat{f}_1(t) = 0$ iff $t = 2n\pi$ where n is any non-zero integer, f lies in the closed ideal generated by f_1 according to Theorem 7.2.4 of Rudin [18].

THEOREM 8. *Let $\mathcal{A}_1$ be as in Theorem 7. Then*

$$\mathcal{A}_1 \equiv \mathcal{A}[\mu_f] \text{ (up to } \lambda\text{-null sets) if } \hat{f}(t) = 0 \text{ for } t = 2n\pi/\delta, \quad n = \pm 1, \pm 2, \cdots,$$
$$\neq 0 \text{ otherwise.}$$

PROOF. Suppose $\hat{f}(t)$ satisfies the given condition. Then f_δ and f generate the same closed ideal, where f_δ is defined in Theorem 7. Hence $\mathcal{A}[\mu_f] = \mathcal{A}[\mu_{f_\delta}] = \mathcal{A}_1$ upto λ-null sets.

A referee has pointed out that the converse is not true, and conjectures that the theorem remains true if $\hat{f}(t) = 0$ at $t = 2n\pi/\delta$, $n = \pm 1, \pm 2$, and at a finite number of additional points.

THEOREM 9. *Let $0 < \delta' < \delta$ and δ'/δ be an irrational number, $A = \{x + n\delta;\ 0 \leqq x < \delta',\ n = 0, \pm 1, \pm 2 \cdots\}$. Then the class $M(A)$ of all μ_A such that $A \in \mathcal{A}[\mu]$ is the closed ideal generated by f_δ, where f_δ is the indicator function of the interval $[0, \delta)$.*

PROOF. It is well-known that the set of all numbers of form $m + n\xi$, m, n integers ξ irrational, is dense in X. Using this and Lemma 1 one can show that if $\mu \in M(A)$ then $\mathcal{A}[\mu]$ contains the field generated by sets

$$A = \{x + n\delta;\ 0 < \delta'' < x < \delta''' < \delta,\ n = 0, \pm 1, \pm 2, \cdots\},$$

and hence being monotone $\mathcal{A}[\mu] \supset \mathcal{A}_1$. The result now follows from Theorem 7.

It will be noticed that Theorem 8 gives sufficient conditions for $\mathcal{A}[\mu]$ to be a given sub-σ-field (of certain structure) and Theorem 9 characterizes $M(A)$ for a particular type of A. An f satisfying the conditions of Theorem 8 is $f = \phi * f_1$ where f_1 is the indicator function of $[0, 1)$ and ϕ is the standardized normal density. Both f and f_1 generate the same ideal in $L_1(\lambda)$.

4. A more general formulation. Let us consider these problems in a somewhat more general setting. Instead of X we may work with any locally compact group. But it will be sufficient for our purposes to consider a separable locally compact abelian group G. In particular this implies G is σ-compact and hence the Haar measure, to be denoted by m_0, is σ-finite.

The method of constructing weakly incomplete families in the preceding section can be generalized as follows. Let G_1 be a countable subgroup. Suppose there exists a measurable subset G_0 of G such that it has one and only one element from each coset of G_1. Note that $m_0(G_0) > 0$. For $G = \bigcup_{g \in G_1}(G_0 + g)$ is the countable union of disjoint sets of equal measure $m_0(G_0)$. We further assume $m_0(G_0) < \infty$ and let $f_0 = I_{G_0}/m_0(G_0)$ where I stands for indicator func-

tion. Then it is easily checked that if A is any G_1-invariant set then $I_A * \tilde{f}_0 =$ constant. As before we can construct new weakly incomplete families starting with f_0. But in this general setting it can be shown that not all weakly incomplete families are obtainable in this way; see for example the next section on the circle group. [Also Lemma 7 shows that if the real line X is given the discrete topology then there exist weakly incomplete families but none obtainable by what we have called above the method of 3.3. However in this case the group is not even σ-compact.] The known structure of locally compact abelian groups should make it quite easy to give necessary and sufficient conditions under which we can construct a weakly incomplete family as in Lemma 6.

We conclude this section with a simple result.

THEOREM 10. *If a measurable G_0 exists then G_1 is closed and $[\mu]$ is constant on all G_1-invariant sets; then $(G, \mathcal{A}_1, [\mu])$ is isomorphic to $(G/G_1, \mathcal{B}, m_1)$ where $\mathcal{A}_1$ is the class of G_1-invariant sets of $\mathcal{A}$, $\mathcal{B}$ is the class of Borel sets of the quotient G/G_1 and m_1 is the normalized Haar measure on G/G_1.*

PROOF. The first part follows from Theorems 5.2, 7.2 of Mackey [16]. The second part follows from Lemma 5 by taking $d\nu = f_0\, dm_0$ where f_0 is as in the discussion preceding this theorem.

This result shows translations on $(G, \mathcal{A}_1, [\mu])$ do not provide really new examples of measure preserving transformations.

5. Circle group. It is natural that we should try to study convolutions by using Fourier transforms. The difficulty here is that if G is non-compact then the indicator function of a non-trivial $[\mu]$-invariant set does not belong to $L_1(m_0)$. In order to overcome this difficulty we may work with a compact group. In this section we work with the circle group.

Let G be the group of complex numbers of modulus one with multiplication as the group operation or equivalently the additive group of reals modulo 2π; we follow the second interpretation henceforth.

Since G is compact the closed countable subgroups are the finite groups, in this case the cyclic groups of the form

$$G_1 = \{2\pi K/n_0\,; \qquad K = 0, 1, \cdots, n_0 - 1\}.$$

A typical G_1-invariant set is

$$A = \{x + K2\pi/n_0\,; \qquad 0 < x < a, \qquad K = 0, 1, \cdots, n_0 - 1\}$$

where $0 < a < 2\pi/n_0$.

Let $G_0 = \{x;\ 0 \leqq x < 2\pi/n_0\}$ and $f_0 = I_{G_0}/(2\pi/n_0)$. As before we assume $d\mu = f(x)\, dm_0$.

In this case we can offer a different proof of Theorem 9 (and hence of Theorem 7) which is illuminating.

If A is $[\mu]$-invariant, then

$$\hat{I}_A(n)\hat{\bar{f}}(n) = c \quad \text{if} \quad n = 0$$
$$= 0 \quad \text{if} \quad n \neq 0$$

where $\hat{\phi}(n)$ for $\phi \varepsilon L_1(m_0)$ denotes the Fourier transform of ϕ at n, i.e., $\int_0^{2\pi} \phi(x)e^{-inx}\,dx/2\pi$ and $\bar{\phi}$ denotes the complex conjugate of ϕ.

$$\therefore \quad \hat{f}(n) = 0 \quad \text{if} \quad \hat{I}_A(n) \neq 0 \quad \text{and} \quad n \neq 0.$$

Take $A = \{x + 2\pi K/n_0\,;\ 0 < x < a,\ K = 0, 1, \cdots, n_0 - 1\}$ and $0 < a < 2\pi/n_0$ with a/π and hence $an_0/2\pi$ an irrational number. Then

$$\hat{I}_A(n) \neq 0 \quad \text{if} \quad n = \pm Kn_0\,, \qquad K = 1, 2, \cdots.$$

Therefore if $[\mu]$ is constant on A, then

$$(1) \qquad \hat{f}(n) = 0 \quad \text{if} \quad n = \pm Kn_0\,, \qquad K = 1, 2, \cdots.$$

Also, it is easily checked that

$$(2) \qquad \hat{f}_0(n) = 0 \quad \text{iff} \quad n = \pm Kn_0\,, \qquad K = 1, 2, \cdots.$$

The desired conclusion namely that f lies in the closed ideal generated by I_A follows from (1) and (2) by Theorem 7.2.4 [18].

To show that the above result about $M(A)$ is not true for all $A \varepsilon \mathfrak{A}$ we consider the following example suggested by the preceding arguments. Let $X = \{x + K2\pi/n_0\,;\ 0 < x < a,\ K = 0, 1, \cdots, n_0 - 1\}$ where $0 < a < 2\pi/n_0$ and a/π is irrational. Let $f_1 = I_A/m_0(A)$. Then $[\mu_1]$ generated by f_1 is invariant on $A = [0, 2\pi/n_0)$. But since $\hat{f}_1(n) \neq 0$ if $n = \pm Kn_0$, $K = 1, 2, \cdots$, it follows that f_1 cannot be obtained by convolution with a uniform distribution on $[0, 2\pi/n_1)$ for any n_1. Incidentally by Pitcher's theorem [17] the minimal sufficient sub σ-field for $[\mu_1]$ is properly contained in $\mathfrak{A}$ and A is not merely $[\mu_1]$-invariant but also has Neyman structure.

Theorem 3 holds for any locally compact abelian group and Theorem 4 for any locally compact connected abelian group. In particular we get

THEOREM 11. *If $[\mu_f]$ is weakly incomplete then $\hat{f}(n) = 0$ at an infinite number of points.*

In this case one can give an elementary proof not using Wiener's theorem.

PROOF. Suppose if possible $\hat{f}(n)$ vanishes at a finite number of points. Hence if $I_A * \tilde{f} =$ constant, then $\hat{I}_A(n) \neq 0$ only at a finite number of points. Hence inversion is possible and we get

$$I_A(x) = \sum e^{inx}\hat{I}_A(n)$$

which is a continuous function of x. Hence

$$A = \{x: I_A(x) = 1\} = \{x: I_A(x) \neq 0\}$$

is both open and closed. Hence $A = G$ or ϕ i.e., $[\mu_f]$ is not weakly complete.

6. Compact groups. This section contains some remarks on other compact separable groups.

For compact groups it is easy to characterize all translation invariant sub-σ-fields. However we do not know to what extent Theorems 7, 8 and 9 hold.

THEOREM 12. *If G is separable and compact and $\mathfrak{A}_1$ is a translation invariant*

sub-σ-field then there exists a compact subgroup G_1 such that $\mathcal{A}_1$ differs from the class of G_1-invariant sets by m_0-null sets.

PROOF. For $f \in L_1(m_0)$ let $\pi f = E_{m_0}(f \mid \mathcal{A}_1)$. Then it is easily seen that π commutes with translations. Hence by Theorem 3.8.3 of [18] $\pi f = f * \mu_0$ where $\hat{\mu}_0(\gamma) = 1$ or 0, $\gamma \in \Gamma$ and Γ is the dual of G. Let Γ_1 be the set of all γ such that $\hat{\mu}_0(\gamma) = 1$. If $\gamma \in \Gamma_1$, $\pi\gamma = \gamma * \mu_0 = \gamma$. Hence since π is a conditional expectation operator, $\pi(f\gamma) = \gamma\pi f$ if $\gamma \in \Gamma_1$. Since this holds for all $f \in L_1(m_0)$ this means if $\gamma \in \Gamma_1$, $\gamma(x) = 1$ a.e. μ_0. Since Γ is countable and hence Γ_1, we can choose a μ_0-null set N_0 such that $\gamma(x) = 1$ if $x \notin N_0$ and $\gamma \in \Gamma_1$. From this it easily follows that Γ_1 is a subgroup of Γ. Hence if G_1 is the anihilater of Γ_1 then the uniqueness of Fourier transforms implies μ_0 is the normalized Haar measure corresponding to G_1. The required result now follows since $E_{m_0}(f \mid \mathcal{A}_1) = f * \mu_0 = E_{m_0}(f \mid \mathcal{A}_2)$ where $\mathcal{A}_2$ is the class of G_1-invariant sets of $\mathcal{A}$.

It will be noticed that the main part of the above proof consists in giving a different and much simpler proof of Theorem 1.1 of Pitcher [17] for the compact abelian case. A proof for compact monothetic groups can be found in [1], pp. 356, 357.

If f has Fourier transform that vanishes at a finite number of points, then the problem of deciding whether $[\mu_f]$ is weakly incomplete can be reduced to a question on finite groups. We shall not give the details since the answer is not known even for finite groups. Instead we content ourselves with stating the following which essentially shows how the reduction can be effected.

Let Γ be the dual of G.

THEOREM 13. *Suppose $\hat{f}$ vanishes at a finite number of points. If $[\mu]$ is weakly incomplete then*

(i) *the smallest subgroup Δ of Γ containing zeros of $\hat{f}$ is finite,*

(ii) *G/H is finite where H is the anihilator of Δ,*

(iii) *there exists a set A consisting of some of the cosets of H such that $I_A * \check{f} = C$, $0 < C < 1$.*

The proof is omitted. It depends on Theorem 3.3.2 of Rudin [18].

We conclude this section with two equivalent formulations of the problem of characterizing all weakly incomplete families of probability measures.

FIRST FORMULATION. Which subsets N of Γ have the property (W) that $N = \{\gamma; I_A(\gamma) = 0\}$ for some set A with $0 < m_0(A) < 1$? $[\mu]$ is a weakly incomplete family of probability measure iff $N_0 = \{\gamma; \hat{f}(\gamma) \neq 0, \gamma \neq 0\}$, where 0 stands for the identity element of Γ, is a subset of some N with property (W).

SECOND FORMULATION. Given $N_1 = \{\gamma; \hat{f}(\gamma) = 0 \text{ or } \gamma = 0\}$ where 0 is the identity element of Γ, can one construct $\phi \in L_2(\Gamma)$ such that $\phi = 0$ outside N_1 and $\phi(\gamma) = 1$ or 0 for all γ but not equal to the indicator of G or the empty set? If one can, then $A = \{x; \hat{\phi}(x) = 1\}$ is non-trivially $[\mu]$-invariant. In the case of finite groups the problem thus becomes one of characterizing the support of idempotent measures.

It seems that even for finite groups, worse still even for finite cyclic groups, the problem of characterizing all weakly incomplete families is extremely hard.

It is interesting to note that if G is a cyclic group of order p where p is a prime number, then the Haar measure is the unique weakly incomplete family.

7. A maximal property of the maximal invariant sub σ-field. In this section we consider an application of the concept of weak completeness to solve a problem raised by Basu [6].

Suppose $Y_1, \cdots, Y_n$ are real valued iid with density $f(x - \theta)$ (wrt Lebesgue measure), $-\infty < \theta < \infty$; let $(X^{(n)}, \mathcal{A}^{(n)}, \mathcal{P}^{(n)})$ denote their joint distribution and, without loss of generality, the basic space on which they are defined. Thus $dP_\theta^{(n)} = \prod f(x_i - \theta)\, d\lambda^{(n)}$ where $\lambda^{(n)}$ is the n-dimensional Lebesgue measure. Clearly if $\mathcal{A}_I$ is the maximal invariant sub-σ-field induced by $Y_2 - Y_1, \cdots, Y_n - Y_1$, then $\mathcal{A}_I$ is $\mathcal{P}^{(n)}$-invariant. But even in the case of normal density with unknown mean θ and known variance it can be shown that $\mathcal{A}_I$ does not contain all $\mathcal{P}^{(n)}$-invariant sets [14], p. 227. Basu [6] has raised the question whether $\mathcal{A}_I$ is a maximal $\mathcal{P}^{(n)}$-invariant sub-σ-field, i.e., whether $\mathcal{A}_I$ is not contained in any $\mathcal{P}^{(n)}$-invariant sub-σ-field $\mathcal{B}$ which has at least one set B such that $P_\theta^{(n)}(A \triangle B) = P_0^{(n)}(A \triangle B) > 0$ for all $A \in \mathcal{A}_I$. Our theorem in this section gives a necessary and sufficient condition for maximality. It leads to an example showing that in general $\mathcal{A}_I$ is not even maximal. Also as a corollary we have an easily verifiable sufficient condition.

Let the conditional density of Y_1 wrt $\lambda^{(1)} = \lambda$ be $f_0^{Y_1}(x_1 \mid Y_i - Y_1 = \lambda_i, i = 2, \cdots, n)$. We shall write it as $f_0(x_1 - \theta; \lambda_2, \cdots, \lambda_n)$. In the following $\mathcal{A}^{(1)} = \mathcal{A}$ the class of linear Borel sets.

THEOREM 14. *$\mathcal{A}_I$ is a maximal $\mathcal{P}^{(n)}$-invariant sub-σ-field* iff *there does not exist any family of sets $A(\lambda_2, \cdots, \lambda_n) \in \mathcal{A}$ for all $(n-1)$-tuples $(\lambda_2, \cdots, \lambda_n)$ satisfying*

(i) $\int_{A(\lambda_2, \cdots, \lambda_n)} f_0(x_1 - \theta; \lambda_2, \cdots, \lambda_n)\, d\lambda = c(\lambda_2 \cdots, \lambda_n)$

where $c(\lambda_2, \cdots, \lambda_n)$ is free of θ and $0 < c < 1$ on a set of $(\lambda_2, \cdots, \lambda_n)$ having positive measure under $P_0^{Y_2 - Y_1, \cdots, Y_n - Y_1}$.

(ii) $A = \{(x_1, \cdots, x_n); x_1 \in A(x_2 - x_1, \cdots, x_n - x_1)\}$ *is a $\mathcal{A}^{(n)}$-measurable set.*

The first condition says the family of conditional densities is weakly incomplete for a set of $(\lambda_2, \cdots, \lambda_n)$ of positive probability. Condition (ii) implies these sets can be combined in a "measurable" way.

PROOF. Suppose (i) and (ii) hold. Then $P_\theta(A \mid \mathcal{A}_I)$ is free of θ; hence $P_\theta(B)$ is free of θ for $B \in \mathfrak{F}$ where $\mathfrak{F}$ is the class of finite disjoint unions of sets of form $D_1 \cap D_2$, $D_1 = A, A^c$ or R and $D_2 \in \mathcal{A}_I$. Let $\mathcal{B}$ be the smallest σ-field containing $\mathfrak{F}$. Then $\mathcal{B}$ is $\mathcal{P}^{(n)}$-invariant. Also by (i) $A \notin \mathcal{A}_I$ and so $\mathcal{A}$ is a proper sub-σ-field of $\mathcal{B}$.

Conversely, suppose $\mathcal{A}_I$ is contained in a $\mathcal{P}^{(n)}$-invariant sub-σ-field $\mathcal{B}$ with a set $B \in \mathcal{B}$ such that $P_\theta^{(n)}(A \triangle B) = P_0^{(n)}(A \triangle B) > 0$ for all $A \in \mathcal{A}_I$. Then

$$P_0^{(n)}(A \mid \mathcal{A}_I) = C(Y_2 - Y_1, \cdots, Y_n - Y_1) \tag{1}$$

is strictly between 0 and 1 with positive probability under $P_0^{(n)}$. Then for

$$(2) \qquad \int_{A(\lambda_2,\cdots,\lambda_n)} f_0(x_1 - \theta; \lambda_2, \cdots, \lambda_n)\, d\lambda = C(\lambda_2, \cdots, \lambda_n)$$

a.e. $P_0^{Y_2-Y_1,\cdots,Y_n-Y_1}$ for $\mathfrak{B}$ being $\mathcal{P}^{(n)}$-invariant we may take $P_\theta^{(n)}(A \mid \mathfrak{A}_I) = P_0^{(n)}(A \mid \mathfrak{A}_I)$. Hence there exists a set N of measure zero under $P_0^{Y_2-Y_1,\cdots,Y_n-Y_1}$ such that if $(\lambda_2, \cdots, \lambda_n)\, \varepsilon'\, N$ then (2) holds for all rational θ; by Lemma 9. (2) holds for all θ if $(\lambda_2, \cdots, \lambda_n)\, \varepsilon\, N$. This completes the proof.

COROLLARY. *$\mathfrak{A}_I$ is a maximal $\mathcal{P}^{(n)}$-invariant sub-σ-field if*

$$\int_{-\infty}^{\infty} e^{-itx_1} f_0(x_1 ; \lambda_2, \cdots, \lambda_n)\, d\lambda \not\equiv 0$$

for any t for almost all $(\lambda_2, \cdots, \lambda_n)$ under $P_0^{Y_2-Y_1,\cdots,Y_n-Y_1}$.

The sufficient condition for maximality given in the corollary holds for the case of normally distributed random variables with mean θ and variance unity.

To show that $\mathfrak{A}_I$ is not always maximal consider the following example: Let $f_0(x) = 1$ if $0 \leqq x \leqq 1$. Then

$$f_0(x_1 ; \lambda_2, \cdots, \lambda_n) = 1/(a_2 - a_1) \quad \text{if} \quad a_1 < x_1 < a_2$$

where a_1, a_2 are functions of $(\lambda_2, \cdots, \lambda_n)$ and $a_1 < a_2$ with probability one under $P_0^{Y_2-Y_1,\cdots,Y_n-Y_1}$. Let the sample space of Y_1 be divided up into left closed right open intervals $[i\lambda, (i + 1)\lambda)$ where $2\lambda = a_2 - a_1$, $i = 0, \pm 1, \pm 2, \cdots$. Let $A(\lambda_2, \cdots, \lambda_n)$ be the union of those intervals that have an even i in their left end point. Then $A(\lambda_2, \cdots, \lambda_n)$ satisfies (i) and (ii) of Theorem 15 with $C(\lambda_2, \cdots, \lambda_n) = \frac{1}{2}$ for almost all $(\lambda_2, \cdots, \lambda_n)$ under $P_0^{Y_2-Y_1,\cdots,Y_n-Y_1}$.

For this example we can construct the unique maximal $\mathcal{P}^{(n)}$-invariant sub-σ-field containing $\mathfrak{A}_I$. Consider the class of all $A\, \varepsilon\, \mathfrak{A}^{(n)}$ such that $P_\theta(A \mid \mathfrak{A}_I)$ is free of θ and call it $\mathfrak{B}$. It follows from Lemma 5 and the proof of Theorem 15 that $\mathfrak{B}$ is the unique maximal $\mathcal{P}^{(n)}$-invariant σ-field containing $\mathfrak{A}_1$. We conjecture that $\mathfrak{B}$ is in fact the class of all $\mathcal{P}^{(n)}$-invariant sets.

The conjecture suggests the following general question. When is the class of all $\mathcal{P}^{(n)}$-invariant sets a σ-field? For $Y_1, \cdots, Y_n$ normally distributed with mean $-\infty < \theta < \infty$ and unit variance, our corollary shows the answer is no. But as we have seen even for $n = 1$ the general problem is hard to solve.

8. Problems and speculations. We have already mentioned a few problems above. We list some more below. We confine ourselves to the real line unless otherwise stated. Also we assume $[\mu]$ is dominated.

The main problem is to produce at least one weakly incomplete dominated family of probability measures $[\mu_f]$ which is not of the kind considered in 3.3. If no such example exists most of the questions asked below would have a trivial answer.

Let E be a set with finite positive Lebesgue measure and μ the normalized restriction of Lebesgue measure to E. When is it true that $[\mu]$ is boundedly (weakly) complete? If E is an interval (parallelogram if we are on the plane) then we know from Lemma 5 that $[\mu]$ is weakly incomplete. What is the state of affairs if E is a circle or a triangle on the plane?

What are the translation invariant sub-σ-fields not covered by Theorem 6?

What are the analogues of Theorems 7 and 8 for such sub-σ-fields? If $[\mu]$ has non-trivial invariant sets of all sizes then does $\mathfrak{A}[\mu]$ include a translation invariant sub-σ-field? Does it have a factor which is a uniform distribution over some interval?

What is the class of all sets A with non-empty $M(A) \cap M_p$? What are the extreme points of $M(A) \cap M_p$ for such an A?

Acknowledgment. We would like to thank Professor M. Rajagopalan for some helpful discussions, and Mr. B. V. Rao who showed us how the proof of Theorem 6 can be simplified.

REFERENCES

[1] ADLER, R. A. (1964). Invariant and reducing subalgebras of measure preserving transformations. *Trans. Amer. Math. Soc.* **110** 350–360.

[2] BAHADUR, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25** 423–462.

[3] BLACKWELL, D. (1956). On a class of probability spaces. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **2** 1–6. Univ. of California Press.

[4] BASU, D. (1959). The family of ancillary statistics. *Sankhyā* **21** 247–256.

[5] BASU, D. (1964). Recovery of ancillary information. *Sankhyā* **26** 3–16.

[6] BASU, D. (1965). Problems related to the existence of maximal and minimal elements in some families of statistics (sub-fields). *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 41–50. Univ. of California Press.

[7] DVORETZKY, A., WALD, A. and WOLFOWITZ, J. (1951). Eliminations of randomization in certain statistical decision procedures and zero sum two person games. *Ann. Math. Statist.* **22** 1–21.

[8] FERGUSON, T. S. (1963). Location and scale parameters in exponential families of distributions. *Ann. Math. Statist.* **33** 986–1001.

[9] GHOSH, J. K. and SINGH, R. (1966). Unbiased estimation of location and scale parameters. *Ann. Math. Statist.* **37** 1671–1675.

[10] HALMOS, P. R. (1948). The range of vector-measure. *Bull. Amer. Math. Soc.* **54** 416–421.

[11] HALMOS, P. R. (1950). *Measure Theory.* Van Nostrand, Princeton.

[12] KAGAN, A. M. and LINNIK, YU. V. (1964). A class of families of distributions with similar regions (in Russian). Vestnik Lenningrad Univ., *Ser. Mat. Meh. Astronom.* **19** 16–18.

[13] LEHMANN, E. L. and SCHEFFÉ, HENRY (1950). Completeness similar regions and unbiased estimation—Part I. *Sankhyā* **10** 305–340.

[14] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[15] LIAPUNOV, A. (1940). Sur les functions-vecteurs completement additives. *Izvestiya Akad Nauk, SSSr. Ser. Mat.* **4** 465–478.

[16] MACKEY, G. M. (1957). Borel structures in groups and their duals. *Trans. Amer. Math. Soc.* **85** 134–165.

[17] PITCHER, T. S. (1957). Positivity in H. Systems and sufficient statistics. *Trans. Amer. Math. Soc.* **85** 166–175.

[18] RUDIN, H. (1962). *Fourier Analysis on Groups.* Interscience, New York.

[19] SCHEFFÉ, H. (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* **18** 434–438.

ROLE OF THE SUFFICIENCY AND LIKELIHOOD PRINCIPLES IN SAMPLE SURVEY THEORY*

By D. BASU

University of New Mexico and *Indian Statistical Institute*

SUMMARY. In this paper, the statistical model for sample surveys is first put in the conventional set-up of $(\Omega, \alpha, \mathcal{P})$, and it is shown that a maximal sufficiency reduction is always possible for a sample survey model. The corresponding minimal sufficient statistic is derived. We examine the role of the sufficiency and likelihood principles in the analysis of survey data and arrive at the revolutionary but reasonable conclusion that, once the sample has been drawn, the inference should not depend in any way on the sampling design. This poses the problem of designing a survey which will yield a good (representative) sample. The randomisation principle is examined from this view point and it is noticed that there is very little, if any, use for it in survey designs.

1. Introduction

This article was written with the object of emphasizing the following four points.

(a) The first point is only of pedagogical interest. Recently, a series of interesting papers have appeared [Pathak (1964), Godambe (1966), Hanurav (1968), Joshi (1968) to mention only a few] in which the statistical model for sample surveys has been so formulated as to confuse conventional statistical mathematicians [ordinarily incapable of speculating about anything excepting the trinity of $(X, \alpha, \mathcal{P})$!] into the belief that the analysis of survey type data falls outside the mainstream of the theory of statistical analysis. In these formulations, one sees on the surface a 'sample space' S (of possible samples s) with just one probability measure p on S. [How can there be any inference with just one measure ?!] The pair (S, p) is called the sampling design. A typical sample $s \epsilon S$ is a subset of (or a finite sequence with its members drawn from) a fixed population Π of individuals $1, 2, 3, \ldots, N$. The parameter is an unknown vector $\theta = (Y_1, Y_2, \ldots, Y_N)$. A statistic is a very special kind of a function of the sample s and the parameter θ. [How can a statistic be anything but a function defined on the sample space ?!] And so on and on it (the new formulation) goes, apparently blazing a new trail in the wilderness of statistical thought. In this article we point out that it is not really necessary to formulate the survey model in the above 'unfamiliar' manner. We need not abandon the trinity $(X, \alpha, \mathcal{P})$!

(b) The second point emphasized here is also of a purely academic nature. If we assume that the set of 'possible' values for the parameter θ is uncountable, then the family $\mathcal{P}$ in the sample survey model $(X, \alpha, \mathcal{P})$ would be typically undominated. This raises the possibility that there may not exist a maximal sufficiency reduction of the survey data (and other hair raising possibilities!). But the saving grace for the survey model is that each member of $\mathcal{P}$ is always a discrete measure. The existence

* This research was partially supported by Research Grant No. ZU-2582 of the National Science Foundation.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_22, © Springer Science+Business Media, LLC 2011

of the maximal sufficiency reduction of the data (the minimal sufficient statistic) is always assured if we take every set as measurable. Also it is very easy to characterize and use the minimal sufficient statistic.

(c) In this article we examine the role of the twin principles of sufficiency and likelihood in the analysis of survey data and arrive at the revolutionary but entirely reasonable conclusion that at the analysis stage the statistician should not pay any attention to the nature of the sampling design. Indeed, the analyst need not even know the sampling design that produced the data.

(d) It goes without saying that there is a great need for designing the survey very carefully. How else can we expect to get a good (representative) sample ? Currently, survey statisticians make extensive use of the random number tables. In this article, the author very briefly examines the randomization principle and comes to the conclusion that there is very little (if any) use for it in survey designs.

2. STATISTICAL MODELS AND SUFFICIENCY

The notion of a sampling (or statistical) experiment is idealized as a statistical model $(X, \alpha, \mathcal{P})$ where

(i) X is the sample space,

(ii) α is a fixed σ-field of subsets of X, called the measurable sets or the events, and

(iii) $\mathcal{P} = \{P_\theta \mid \theta \in \Omega\}$ is a fixed family of probability measures P_θ on α.

The family $\mathcal{P}$ is indexed by the unknown state of nature (the parameter) θ. The set of all the possible values of θ is the parameter space Ω.

By the term statistic we mean a characteristic of the sample x. A general and abstract formulation of the notion of a statistic is that of a mapping of X onto a space Y. Thus, a statistic $T = T(x)$ is an arbitrary function with X as its domain. Every statistic T defines an equivalence relation [$x \sim x'$ if $T(x) = T(x')$] on the sample space X. This leads to a partition of X into equivalent classes of sample points. As we need not distinguish between statistics that induce the same partition of X, it is convenient to think of a statistic T as a partition $\{\pi\}$ of X into a family of mutually exclusive and collectively exhaustive parts π.

The statistic (partition) $T = \{\pi\}$ is said to be wider (larger) than the statistic $T^* = \{\pi^*\}$ if every π is a subset of some π^*—in other words, if every π^* is a union of a number of π's. Given a statistic $T = \{\pi\}$, consider the class of all measurable sets (members of α) that are unions of some π's. They constitute a sub-σ-field (sub-field) of α and is denoted by α_T. We call α_T the sub-field induced by T. If T is wider than T^* then $\alpha_T \supset \alpha_{T^*}$.

An abstract and very general formulation of the notion of sufficient statistic is the following :

Definition : The statistic T is sufficient $[\alpha, \mathcal{P}]$ if, corresponding to every real-valued bounded, α-measurable function f, there exists an α_T -measurable f^* such that for all $B \in \alpha_T$ and $\theta \in \Omega$

$$\int_B f \, d P_\theta \equiv \int_B f^* \, dP_\theta.$$

The notion of sufficiency has been studied in great details in statistical literature. In the particular case where the family $\mathcal{P}$ of probability measures is dominated by a σ-finite measure λ, we have the following factorization theorem of fundamental importance.

Theorem : *Let $p_\theta = dP_\theta | d\lambda$ be a fixed version of the Radon-Nikodym derivative of P_θ wrt λ. A necessary and sufficient condition for the sufficiency of the statistic T is that there exists, for each $\theta \in \Omega$, an α_T-measurable function g_θ and a fixed α-measurable function h such that, for each $\theta \in \Omega$,*

$$p_\theta(x) = g_\theta(x) h(x) \ \ aew \, [\lambda].$$

In a dominated set-up, most of the properties of sufficient statistics flow from the above factorization theorem. For example, if T is sufficient then any statistic T^* that is wider (larger) than T is also sufficient. Again, with a separability condition on $\mathcal{P}$, it is true that there exists a sufficient statistic T which is essentially smaller (narrower) than every other sufficient statistic T^*. Such a sufficient statistic is called the minimal (or least) sufficient statistic.

That neither of the above two propositions need hold for general undominated set-ups has been exhibited by Burkholder (1961) and Pitcher (1957). Consider the following two examples.

Example 1: Let X be the real line, α the σ-field of Borel sets and $\mathcal{P}$ the class of all discrete two-point probability distributions P_θ on the line that are symmetric about the origin. [That is, the entire mass of P_θ is equally distributed over the two points $-\theta$ and θ, where $\theta > 0$.] Let E be a non-Borel set that excludes the origin but is symmetric about it. Let $T(x) = |x|$ and let

$$T^*(x) = \begin{cases} |x| & \text{if} \quad x \in E \\ x & \text{if} \quad x \notin E. \end{cases}$$

Clearly, T^* is wider than T. However, in this example, T is sufficient but T^* is not.

Example 2 : Let X, α and E be as in the previous example and let $\mathcal{P} = \{P_\theta\}$ be defined as follows. If $\theta \in E$ then the whole mass of P_θ is equally distributed over the points $-\theta$ and θ. If $\theta \notin E$, then P_θ is degenerate at θ. In this example, there does not exist a minimal sufficient statistic.

In each of the above two examples, we are dealing with a family of measures each member of which is discrete. In example 1, each measure has its entire mass concentrated at two points only; in example 2, each measure has its entire mass

distributed over at most two points. True, we are dealing, in each case, with an undominated family of measures. But that is not where the real trouble lies. In these examples, our difficulties stem from our artificially restricting ourselves to Borel sets only. If in the above two examples we take α to be the class of all subsets, then we do not have to face the above kind of anomalous situations. The natural domain of definition of discrete measures is the σ-field of all subsets. In sample survey theory, we need not consider non-discrete probability measures. By a discrete model we mean the following.

Definition : The statistical model (X, α, P_θ), $\theta \in \Omega$, is called a discrete model if

(i) each P_θ is a discrete measure,

(iii) α is the class of all subsets of X, and.

(iii) for each $x \in X$, there exists a $\theta \in \Omega$, such that, $P_\theta(\{x\}) > 0$.

[*Remark* : Condition (iii) only ensures that we do not entangle ourselves with possibilities that have zero probabilities for each possible value of the parameter θ. Condition (ii) ensures that all sets and functions are measurable.]

We, henceforth, deal with discrete models only. A discrete model is undominated if and only if X is uncountable. The Burkholder-Pitcher type pathologies cannot occur in discrete models (Basu and Ghosh, 1967).

3. SUFFICIENCY IN DISCRETE MODELS

Let (X, α, P_θ), $\theta \in \Omega$, be a discrete model. For each $x \in X$ let

$$\Omega_x = \{\theta \mid P_\theta(x) > 0\}.$$

[We, henceforth, write $P_\theta(x)$ for $P_\theta(\{x\})$.] The set Ω_x is the set of parameter points that are consistent with the observation (sample point) x. No Ω_x is vacuous.

For discrete models, the minimal sufficient statistic always exists and is uniquely defined as follows. Consider the binary relation on X : "$x \sim x'$ if $\Omega_x = \Omega_{x'}$ and $P_\theta(x) \mid P_\theta(x')$ is a constant in θ for all $\theta \in \Omega_x = \Omega_{x'}$".

The above is an equivalence relation on X. The partition (statistic) induced by the equivalence relation is the minimal (least) sufficient statistic. This is an easy consequence of the following factorization theorem (Basu and Ghosh, 1967).

Theorem : *If (X, α, P_θ), $\theta \in \Omega$, be a discrete model, then a necessary and sufficient condition for a statistic (partition) $T = \{\pi\}$ to be sufficient is that there exists a function g on X such that, for all $\theta \in \Omega$ and $x \in X$,*

$$P_\theta(x) \equiv g(x) P_\theta(\pi_x),$$

where π_x is that part of the partition $\{\pi\}$ that contains x.

The above factorization theorem is a direct and easy consequence of the definitions of sufficient statistics and discrete models (as stated in Section 2).

If $T = \{\pi\}$ be a sufficient partition and if g be defined as in the previous theorem, then it follows that $g(x) > 0$ for all $x \in X$ and that, for each π,

$$\sum_{x \in \pi} g(x) = 1.$$

Each part of a sufficient partition must be countable. What the above factorization theorem is telling us is nothing but the intuitively obvious proposition that $\{\pi\}$ is sufficient if and only if, for each part π, it is true that the conditional distribution of the sample x given π is θ-free. This is indeed the original definition of sufficiency as proposed by Fisher.

Another consequence of the above theorem is that if $T = \{\pi\}$ is a sufficient statistic then any statistic T^* that is wider than T is necessarily sufficient. It also follows that (for discrete models) there exists a one-one correspondence between sufficient statistics (partitions) and sufficient sub-fields (Basu and Ghosh, 1967).

An alternative (but equivalent) way of characterizing the minimal sufficient statistic for a discrete model is the following. For each $x \in X$ let $L_x(\theta)$ stand for the likelihood function, i.e.

$$L_x(\theta) = \begin{cases} P_\theta(x) & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{for} \quad \theta \notin \Omega_x. \end{cases}$$

Let us standardize the likelihood function as follows.

$$\bar{L}_x(\theta) = \frac{L_x(\theta)}{\sup_\theta L_x(\theta)}.$$

Consider the mapping

$$x \to \bar{L}_x(\theta),$$

a mapping of X into a class of real-valued functions on Ω. This mapping is the minimal sufficient statistic. [A little reflection would show that the partition (of X) induced by the above mapping is the same as the one induced by the equivalence relation described earlier in this section.]

4. The sample survey models

The principal features of a sample survey situation are as follows. There exists a well-defined population Π — a finite set of distinguishable objects called the (sampling) units. Typically, there exists a list of these units—the so-called sampling frame. Let us list the population as

$$\Pi = (1, 2, 3, \ldots, N).$$

The unit i has an unknown characteristic Y_i. The unknown state of nature is

$$\theta = (Y_1, Y_2, \ldots, Y_N).$$

The statistician has some prior information or knowledge K about θ. This knowledge K is largely of a qualitative and speculative nature. For example, the statistician knows that θ is a member of a well-defined set Ω (the parameter space). He also knows, for each unit i, some characteristic A_i of the unit i. Let us denote this set of known auxiliary characteristics by

$$\boldsymbol{A} = (A_1, A_2, \ldots A_N).$$

Thus, $\boldsymbol{A}$ is a principal component of K. In K is also embedded what the statistician thinks (knows) to be the true relationship between the unknown θ and the known $\boldsymbol{A}$.

It is within the powers of the statistician to find out or "observe" the characteristic Y_i for any chosen unit i. A survey problem arises when the statistician plans to gain further "information" about some function $\tau = \tau(\theta)$ of the parameter θ by observing the Y-characteristics of a set (sequence)

$$\boldsymbol{i} = (i_1, i_2, \ldots, i_n)$$

of units selected from Π.

Let us denote the observed Y-characteristics by

$$\boldsymbol{y} = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n}).$$

The problem is to make a "suitable" choice of $\boldsymbol{i}$ and then to make a "proper" use of the observations $x = (\boldsymbol{i}, \boldsymbol{y})$ in conjunction with the prior "knowledge" K to arrive at a "reasonable" "judgement" about τ.

Now, let us examine how probability theory comes into the picture. If we ignore observation errors, then there is no discernable source of randomness in the above general formulation of a survey problem (excepting some very intangible quantities like "belief", "knowledge" etc. which the Bayesians try to formalize as probability.) In any survey situation there are bound to be some observation errors (the so-called non-sampling errors). Unfortunately, in current sample survey research it is not often that we find mention of this source of randomness. It is tacitly assumed that the observation errors are negligible in comparison with the so-called "sampling error". This sampling error is the distinguishing feature of the current sample survey theory. Here is a phenomenon of randomness that is not inherent to the problem but is artificially injected into the problem by the statistician himself. The survey statistician does not lean on probability theory for the purpose of understanding and controlling the mess created by an unavoidable source of randomness or uncertainty (observation errors). He uses his knowledge of probability theory to introduce into the problem a well-understood (fully controlled) element of randomness and seems to derive all his strength (intellectual conviction) from that.

The "sampling error" is the randomness that the statistician injects into the problem by selecting the set (sequence) $\boldsymbol{i} = (i_1, i_2, \ldots, i_n)$ in a random manner. Given a sampling plan $\mathcal{S}$, for each possible $\boldsymbol{i}$ there exists a number $p(\boldsymbol{i})$ which is the probability of ending up with $\boldsymbol{i}$. Usually, this $p(\boldsymbol{i})$ does not depend on the parameter θ, although quite often it is made to depend on the auxiliary information $\boldsymbol{A}$. [However, one may consider sequential sampling plans for which $p(\boldsymbol{i})$ depends on θ. For instance, consider the sampling plan—"Choose unit 1 and observe Y_1. If Y_1 (which we suppose is real valued) is larger than b then choose unit 2, otherwise choose unit N". For this plan $\boldsymbol{i}$ is either (1,2) or (1, N) and $p(\boldsymbol{i})$ depends on θ through Y_1.] Typically, the random choice of $\boldsymbol{i}$ is made in the statistical laboratory well in advance of the time that the observation job is in progress. For such typical sampling plans, the probability $p(\boldsymbol{i})$ for any possible $\boldsymbol{i}$ does not depend on θ at all. However, even if we agree to consider sequential sampling plans of the type described within the parenthesis before, it is clear that $p(\boldsymbol{i})$ for such plans can depend on $\theta = (Y_1, Y_2, \ldots, Y_N)$ only through $y = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n})$. As we shall presently see, this remark is important. In the sequel we write $p(\boldsymbol{i}\,|\,\theta)$ for $p(\boldsymbol{i})$.

The sample is $x = (\boldsymbol{i}, \boldsymbol{y})$, the set $\boldsymbol{i}$ together with the observation $\boldsymbol{y}$. [For some sampling plans—like sampling with replacements—it is more natural to think of $\boldsymbol{i}$ as a finite sequence of units with repetitions allowed.] The sample space X is the set of all possible samples x.

Now, each x, when observed, tells us the exact Y-value of some population units, i.e., tells us about some coordinates of the vector θ. Let Ω_x be the set of parameter points θ that are consistent with a given sample x. If $P_\theta(x)$ be the probability that the sampling plan ends up with sample $x = (\boldsymbol{i}, \boldsymbol{y})$, then it is clear that

$$P_\theta(x) = \begin{cases} p(\boldsymbol{i}\,|\,\theta) & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{otherwise.} \end{cases}$$

Thus, Ω_x is also the set of all parameter points that allot non-zero probabilities to x.

As we have said before, in typical sampling plans $p(\boldsymbol{i}\,|\,\theta)$ does not depend on θ. In sequential sampling plans (where the choice of a population unit at any stage is made to depend on the observed Y-values of the previously selected units) we have noted before that $p(\boldsymbol{i}\,|\,\theta)$ depends on θ through y. Thus, we make the following important observation that, for any sampling plan,

$$P_\theta(x) = \begin{cases} \text{constant} & \text{for} \quad \theta \in \Omega_x \\ 0 & \text{for} \quad \theta \notin \Omega_x. \end{cases}$$

This leads us to the following general characterization of a sample survey model.

Definition : The model (X, α, P_θ), $\theta \in \Omega$ is called an SS-model if the model is discrete and if $P_\theta(x)$ is a constant for all $\theta \in \Omega_x$, where

$$\Omega_x = \{\theta \,|\, P_\theta(x) > 0\}.$$

From what we have said in Section 3, it then follows that

Theorem : *If* (X, α, P_θ), $\theta \in \Omega$ *be an SS-model, then the minimal (least) sufficient statistic is the mapping** $x \to \Omega_x$.

The distinguishing feature of an SS-model is that for every possible sample the likelihood function is flat. That is, for every $x \in X$ the likelihood function $L_x(\theta)$ is zero for all θ outside a set Ω_x and is a constant for $\theta \in \Omega_x$. The following is an example of a non-discrete model with the above feature.

Example 3 : Let $x = (x_1, x_2, \ldots, x_n)$ be n independent observations on a random variable that has a continuous and uniform distribution over the interval $(\theta - 1/2, \theta + 1/2)$, where θ is the parameter $(-\infty < \theta < \infty)$. Let Ω_x be the interval $(m(x), M(x))$ where $m(x) = \max x_i - 1/2$ and $M(x) = \min x_i + 1/2$. Here, $L_x(\theta) = 1$ for all $\theta \in \Omega_x$ and is zero for $\theta \notin \Omega_x$. The mapping $x \to (m(x), M(x)) = \Omega_x$ is the minimal sufficient statistic.

5. THE SUFFICIENCY AND LIKELIHOOD PRINCIPLES

The twin principles of sufficiency and likelihood both attempt to answer the same question. The likelihood principle, however, goes a great deal further in its assertion.

The question is : "What characteristic of the sample x is relevant for making an inference about the parameter θ?" In general, the sample x is a very complex entity. Must we take into account the sample x in all its detail ? Could it be that some characteritics of x are totally irrelevent for making any inference about the state of nature θ ? For instance, if in the observation x we have incorporated the outcome u from a number of tosses of a symmetric coin, then it seems very reasonable to argue that the characteristic u of x is totally irrelevent and must be ignored.

The sufficiency principle is the following. If $T = T(x)$ be a sufficient statistic, then only the characteristic $T(x)$ of x is relevant for inference making. That is, if $T(x) = T(x')$ then the inference about θ should be the same whether the sample is x or x'. The relevant information core of x is then the statistic $T_0(x)$, we T_0 is the minimal sufficient statistic.

The sufficiency principle has gained rather wide acceptance. The Neyman-Pearson school of statisticians tend to justify the principle by proving some complete class theorem that tells us that it is not necessary to consider decision rules (inference procedures) that do not depend on x through $T_0(x)$. On the other hand the Bayesians have no objection to the sufficiency principle as they point out that the posterior distribution for the parameter θ—whatever be its prior distribution—depends on x only through the minimal sufficient statistic $T_0(x)$.

As we have stated in Section 3, the mapping $x \to \bar{L}_x(\theta)$, where $\bar{L}_x(\theta)$ is the standardized (modified) likelihood function, is the minimal sufficient statistic. Thus,

*In typical survey situations, the minimal sufficient statistic (the information core of the sample) is the set of (distinct) population unit-labels that are drawn in the sample together with the corresponding Y-values.

according to the sufficiency principle, two sample points x and x' are equally informative if

$$\bar{L}_x(\theta) \equiv \bar{L}_{x'}(\theta) \text{ for all } \theta.$$

Note that the sufficiency principle does not tell us anything about the nature of the information supplied by x. The likelihood principle takes a big step forward and asserts that the information supplied by x *is* the likelihood function $\bar{L}_x(\theta)$. Whereas the sufficiency principle can compare two possible samples x and x' only when they are points in the same sample space, the likelihood principle can compare them even when they are points in different sample spaces. Consider the following example.

Example 4 : Let θ be the unknown probability of head for a given coin. The following is a list of three different experiments (among the many that one can think of) that one may perform for the purpose of eliciting information about the unknown θ.

$\mathscr{E}_1$: Toss the coin 5 times

$\mathscr{E}_2$: Toss the coin until there are 3 heads

$\mathscr{E}_3$: Toss the coin until there are 2 consecutive heads.

We give below an example x_i of a possible sample point for each experiment $\mathscr{E}_i$ $(i = 1, 2, 3)$. [H = head, T = tail]

x_1 : $H\ T\ H\ H\ T$

x_2 : $T\ T\ H\ H\ H$

x_3 : $H\ T\ T\ H\ H$

[Note that the sample spaces for the three experiments are different from one another. Also note that x_1 cannot be a sample point for either $\mathscr{E}_2$ or $\mathscr{E}_3$. Similarly, x_2 cannot be a sample point for $\mathscr{E}_3$.]

It is easy to check that the likelihood function for x_i (when it is referred to experiment $\mathscr{E}_i$) is $\theta^3(1-\theta)^2$ and this is irrespective of whether i is 1, 2, or 3. The principle of likelihood tells us that sample x_1 for experiment $\mathscr{E}_1$ gives the same information about θ as does sample x_2 from $\mathscr{E}_2$ and sample x_3 from $\mathscr{E}_3$.

From the Bayesian point of view the likelihood principle is almost a truism. The starting point for a Bayesian (in his inference making effort) is a prior probabiltiy distribution over the parameter space Ω. Let $q = q(\theta)$ be the prior probability frequency function. Having observed the sample x, the Bayesian uses the likelihood function $\bar{L}_x(\theta)$ to arrive at the posterior distribution

$$q_x^*(\theta) = \frac{q(\theta)\bar{L}_x(\theta)}{\Sigma_\theta\, q(\theta)\, \bar{L}_x(\theta)}.$$

To a Bayesian, the role of the sample x is only to change his prior scale of preference (probability distribution) $q = q(\theta)$, for various possible values of θ, to the posterior scale $q^* = q_x^*(\theta)$. And this change is effected through the likelihood function $\bar{L}_x(\theta)$. Possible sample points x and x' (whatever sampling experiments might generate them) are equivalent as long as they induce identical (modified) likelihood functions. The likelihood principle is essentially a Bayesian principle. It is hard to justify the principle under the Neyman-Pearson set-up.

6. Role and choice of the sampling plan

Let $\mathcal{S}$ be the chosen sampling plan and let $x = (\boldsymbol{i}, \boldsymbol{y})$ be the data (sample) generated by $\mathcal{S}$. In the matter of analyzing the data, how relevant is the plan $\mathcal{S}$?

If $\boldsymbol{i} = (i_1, i_2, \ldots, i_n)$ and $\boldsymbol{y} = (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_n})$, then Ω_x is the set of all $\theta \epsilon \Omega$ whose j-th co-ordinate is Y_j $(j = i_1, i_2, \ldots i_n)$—the set of θ's that are consistent with the data. Note that Ω_x depends only on x and Ω, it has nothing to do with the plan $\mathcal{S}$ The minimal sufficient statistic is the mapping $x \rightarrow \Omega_x$ and the likelihood function $\bar{L}_x(\theta)$ is

$$\bar{L}_x(\theta) = \begin{cases} 1 & \text{for} \quad \theta \, \epsilon \, \Omega_x \\ 0 & \text{otherwise.} \end{cases}$$

If $q = q(\theta)$ be the Bayesian prior distribution over Ω, then the posterior distribution is

$$q_x^*(\theta) = \frac{q(\theta)\bar{L}_x(\theta)}{\Sigma_\theta q(\theta)\, \bar{L}_x(\theta)} = \begin{cases} c(x)\, q(\theta) & \text{for} \quad \theta \, \epsilon \, \Omega_x \\ 0 & \text{otherwise.} \end{cases}.$$

The posterior distribution $q_x^*(\theta)$ is nothing but the restriction of q to the set Ω_x. And the plan $\mathcal{S}$ does not enter into the definition of Ω_x. Thus, from the Bayesian (and the likelihood principle) point of view, once the data x is before the statistician, he has nothing to do with the plan $\mathcal{S}$. He does not even need to know what the plan $\mathcal{S}$ was. [This is because, in sample survey situations, the plan $\mathcal{S}$ is an artificial source of randomness. In other statistical situations, where randomness is unavoidable and is an inherent part of the observation process, the statistician has to "understand" the process well enough to be able to arrive at his likelihood function.]

In the Neyman-Pearson type of analysis of the data, the statistician considers not only the data x in hand but also pays a great deal of attention to what other data x' he might have obtained. In other words, he needs to know the model $(X, \alpha, \mathcal{P})$ as well as the sample x. The Bayesian needs to know only the likelihood function $\bar{L}_x(\theta)$, which, in a sample survey situation, is entirely independent of the model (the sampling plan $\mathcal{S}$). The author does not think that any reconcilliation between the two approaches to data analysis is possible.

SUFFICIENCY AND LIKELIHOOD PRINCIPLES IN SAMPLE SURVEY THEORY

A majority of statisticians of the Neyman-Pearson school would readily agree to the proposition that the Bayesian analysis of the data is sensible (acceptable) when the following condition holds :

Condition B : It is reasonable to think of the parameter θ as a random variable, and the random process governing θ is at least partially discernable.

However, it is hard to understand how such statisticians reconcile themselves to the contrary positions: (a) Only the data x (the likelihood function) is relevant (for inference making) when condition B holds and (b) the whole sample space X(the model) is relevant when B does not hold. There exists a continuous spectrum of conditions between the extremes of B and not-B. But the shift of emphasis from the sample x to the sample space X is not continuous. [Fisher with his theory of ancillary statistics and choice of reference sets, made a bold but unsuccessful (see Basu, 1964) attempt to bridge the gap between the above polarities in statistical theory.]

It seems to the author that the Bayesian analysis of the data x is very appropriate in sample survey situation. Given the data $x = (\boldsymbol{i}, \boldsymbol{y})$ the sampling plan $\mathcal{S}$—the model $(X, \alpha, \mathcal{P})$—ceases to be of any relevance for inference making about the parameter θ. Given the data x the statistician arrives at his posterior preference scale $q_x^*(\theta)$ for the parameter θ. If $\tau = \tau(\theta)$ be the parameter of interest, then the statistician can compute the marginal posterior distribution $q_x^*(\tau)$ of the variable τ. The question, "Given x, how much information we have about τ ?", can then be answered by first agreeing upon a suitable definition of information. [For example, we may agree to work with the Shannon definition of information or with the posterior variance (or its reciprocal) of τ.]

Given a sample x, we can now tell how good (informative) the sample is. The object of planning a survey should be to end up with a good sample The term "representative sample" has often been used in sample survey terminology. But no one has cared to give a precise definition of the term. It is implicitly taken for granted that the statistician with his biased mind is unable to select a representative sample. So a simplistic solution is sought by turning to an unbiased die (the random numder tables). Thus, a deaf and dumb die is supposed to do the job of selecting a "representative sample" better than a trained statistician. It is, however, true that we do not really train our statisticians for the job of selecting and observing survey type data. [In contrasst, the medical practitioner is given a much more meaningful training in understanding the many variables and their interrelations in his chosen field of specialization.]

In a Bayesian plan for selecting the sample x, there is no place for the symmetric die. Very little attention has so far been paid to the problem of devising suitable sampling strategies from this point of view. In a later document the author would elaborate some of his own ideas on the problem. We end this section by describing

a Bayesian sampling strategy for the very simple case where the statistician wants to select and observe only one unit. Suppose his prior probability distribution is $q(\theta)$. If he selects unit i and observes Y_i then his posterior (marginal) distribution for τ would be, say, $q^*(\tau \mid i, Y_i)$. Once a suitable definition of "information" is agreed upon, he can use the above distribution to compute the quantity $I(i, Y_i)$—the information about τ gained from the sample (i, Y_i). At the planning stage of the experiment, the statistician does not know the value of Y_i that he is going to observe for the unit i. Let $J(i)$ be the average value of $I(i, Y_i)$ when the averaging is done over all possible values of Y_i (weighted by the prior distribution of Y_i). Thus, $J(i)$ is the "expected" information to be gained from observing unit i. Faced with the problem of deciding which unit i to select (and then observe), the statistician would not be acting unreasonably if he selects the unit i that has maximum $J(i)$. [What if the $J(i)$'s are all equal ? Such would be the case if the prior distribution of $\theta = (Y_1, Y_2, \ldots, Y_N)$ is symmetric in the coordinates. In this situation the statistician is indifferent as to which i is selected for observation. In principle, he cannot object now to a random (with equal or unequal probabilities) selection prodedure for i. However, this does not mean that he will be willing to let another person (say, a field investigator) make the choice for him. If for nothing else, a scientist ought to be always on his guard against letting an unknown element enter into the picture.]

Of course, a non-Bayesian would sneer at the arbitrariness inherent in the difinition of $J(i)$. But the procedure described above is certainly more justifiable than our current naive reliance on the symmetric die. Any reasonable Bayesian sampling strategy would have the following characteristics. (a) The sampling plan would usually be sequential. The statistician would continue sampling (one or a few units at a time) until he is satisfied with the information thus obtained or until he reaches the end of his rope (time and cost). His decision to select the units for a particular sampling stage would depend (non-randomly) on the sample obtained in the previous stages. (b) The probability that the statistician would end up observing the units $\boldsymbol{i} = (i_1, i_2, \ldots, i_n)$ in this order, would depend on $\boldsymbol{i}$ and the state of nature θ. This probability would be degenerate, i.e., zero for some values of θ and unity for the rest of the values of θ.

7. SOME CONCLUDING REMARKS

(a) Godambe (1966a) noted that the application of the likelihood principle in the sampling situation would mean that the sampling design is irrelevant for data analysis. On page 317 he writes, "One implication of this, as can be seen from (4), is that the inference about θ must not depend on the sampling design even through the probability $p(\boldsymbol{i})$ of the $\boldsymbol{i}$ that has actually been drawn. In particular, the estimator of τ should not depend on $p(\boldsymbol{i})$ or the sampling design". [In this and in the following quotation the author has taken the liberty of changing some of the notations. This was done for the sake of bringing them in line with the notations used in this article]. It is interesting to observe that Godambe immediately shies away from the revolu-

tionary implication of his remark and tries to find some excuses for not applying the likelihood principle in the sampling situation. He writes (p 317, Godambe, 1966a), "In connection with the likelihood principle, it may be further noted that here θ is the parameter and $(\boldsymbol{i}, \boldsymbol{y})$ is the sample. Thus, possibly there is some kind of relationship between the parametric space and the sample space (when the sample is observed, the parameter cannot remain completely unknown) which forbids the use of the likelihood principle. The relationships between parametric and sample spaces restricting the use of the likelihood principle are referred to by Barnard, Jenkins and Winsten (1962)". In the two 1966 papers referred to here, Godambe tries very hard to justify a particular linear estimator as the only reasonable one for the population total. Godambe's estimator depends on the sampling design. The author finds Godambe's arguments very obscure.

(b) Let us repeat once again that the posterior distribution of τ depends only on the prior distribution q (on Ω) and the sample $x = (\boldsymbol{i}, \boldsymbol{y})$. It does not depend on the sampling design $\mathcal{S}$. Thus, any fixed q on Ω would give rise to a Bayes 'estimation procedure' B_q that would tell us how to estimate τ for each possible sample x —no matter what design $\mathcal{S}$ is used to arrive at x. [Note that B_q is well-defined as a function on the union of sample spaces for all designs $\mathcal{S}$.] Now, if we consider B_q in relation to a fixed design $\mathcal{S}$, then it would be classified as an admissible estimator in the sense of Wald. The findings of Godambe (1960) and Joshi (1968) therefore appear to the author as rather obvious in nature. It is so easy to reel off any number of such universally admissible estimation procedures.

(c) The mathematical content of this article is summarized in the theorem of Section 4. This result has been known (but never explicitly proved) to the author (and among others to Godambe, Hájek, Hanurav and Pathak) for the past eleven years or so. In an yet unpublished article, entitled "Classical sufficiency and its application to sampling theory", Pathak has proved this result for a particular (non-sequential) case.

References

Basu, D. (1958): On sampling with and without replacements. *Sankhyā*, **20**, 287–294.

Basu, D. (1964): Recovery of ancillary information. *Sankhyā*, **25**, 3–16.

Basu, D. and Ghosh, J. K. (1967): Sufficient statistics in sampling from a finite universe. *Bull. Int. Stat. Inst.*, **42**, BK. 2, 850–859.

Burkholder, D. L. (1961): Sufficiency in the undominated case. *Ann. Math. Statist.*, **32**, 1191–1200.

Godambe, V. P. (1960): An admissible estimate for any sampling design. *Sankhyā*, **22**, 285-288.

——— (1966a): A new approach to sampling from finite populations, I: Sufficiency and linear estimation. *JRSS* (series B), **28**, 310-319.

——— (1966b): A new approach to sampling from finite populations, II: Distribution-free sufficiency. *JRSS* (series B), **28**, 320-328.

Hanurav, T. V. (1964): Hyper-admissibility and optimum estimators for sampling finite populations, *Ann. Math. Statist.*, **39**, 621-642.

JOSHI, V. M. (1968) : Admissibility of the sample mean as estimate of the mean of a finite population. *Ann. Math. Statist.*, **39**, 606-620.

PATHAK, P. K. (1964) : Sufficiency in sampling theory. *Ann. Math. Statist.*, **35**, 785-809.

PITCHER, T. S. (1957) : Sets of measures not admitting necessary and sufficient statistics or subfields. *Ann. Math. Statist.*, **28**, 267-268.

Paper received ; November, 1968.

Revised : May, 1969.

On Sufficiency and Invariance

D. BASU*, *University of Chicago and Indian Statistical Institute*

1. SUMMARY

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a given statistical model and let $\mathcal{G}$ be the class of all one-to-one, bimeasurable maps g of $(\mathcal{X}, \mathcal{A})$ onto itself such that g is measure-preserving for each $P \epsilon \mathcal{P}$, i.e. $Pg^{-1} = P$ for all P. Let us suppose that there exists a least (minimal) sufficient sub-field $\mathcal{L}$. Then, for each $L \epsilon \mathcal{L}$, it is true that $g^{-1}L$ is $\mathcal{P}$-equivalent to L for each $g \epsilon \mathcal{G}$, i.e., the least sufficient sub-field is almost $\mathcal{G}$-invariant. It is demonstrated that, in many familiar statistical models, the least sufficient sub-field and the sub-field of all almost $\mathcal{G}$-invariant sets are indeed $\mathcal{P}$-equivalent. The problem of data reduction in the presence of nuisance parameters has been discussed very briefly. It is shown that in many situations the principle of invariance is strong enough to lead us to the standard reductions. For instance, given n independent observations on a normal variable with unknown mean

* This research was supported by Research Grant No. NSF GP-3707 from the Division of Mathematical, Physical and Engineering Sciences of the National Science Foundation, and by the Statistics Branch, Office of Naval Research.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_23,

(the nuisance parameter) and unknown variance, it is shown how the principle of invariance alone can reduce the data to the sample variance.

2. DEFINITIONS AND PRELIMINARIES

(a) The basic probability *model* is denoted by $(\mathscr{X}, \mathscr{A}, \mathscr{P})$, where $\mathscr{X} = \{x\}$ is the sample space, $\mathscr{A} = \{A\}$ the σ-field of events and $\mathscr{P} = \{P\}$ the family of probability measures.

(b) By *set* we mean a typical member of $\mathscr{A}$. By *function* (usually denoted by f) we mean a measurable mapping of $(\mathscr{A}, \mathscr{A})$ into the real line.

(c) A set A is *$\mathscr{P}$-null* if $P(A) = 0$ for all $P \in \mathscr{P}$. Two sets A_1 and A_2 are *$\mathscr{P}$-equivalent* if their symmetric difference is $\mathscr{P}$-null. Two functions f_1 and f_2 are $\mathscr{P}$-equivalent if the set of points where they differ is $\mathscr{P}$-null. The relation symbol $\sim$ stands for $\mathscr{P}$-equivalence.

(d) By *sub-field* we mean a sub-σ-field of $\mathscr{A}$. A *statistic* is a measurable mapping of $(\mathscr{X}, \mathscr{A})$ into any measurable space. We identify a statistic with the sub-field it induces (see pp. 36–39 of [6]).

(e) By the *$\mathscr{P}$-completion* $\overline{\mathscr{A}_0}$ of a sub-field $\mathscr{A}_0$ we mean the least sub-field that contains $\mathscr{A}_0$ and all $\mathscr{P}$-null sets. Observe that $\overline{\mathscr{A}_0}$ may also be characterized as the class of all sets that are $\mathscr{P}$-equivalent to some member of $\mathscr{A}_0$. Two sub-fields are $\mathscr{P}$-equivalent if they have identical $\mathscr{P}$-completions.

(f) By *transformation* (usually denoted by g) we mean a one-to-one, bimeasurable mapping g of $(\mathscr{X}, \mathscr{A})$ onto itself such that the family

$$\mathscr{P}g^{-1} = \{Pg^{-1} \mid P \in \mathscr{P}\}$$

of induced probability measures is the same as the family $\mathscr{P}$. A transformation g is called *model-preserving* if

$$Pg^{-1} \equiv P, \text{ for all } P \in \mathscr{P}.$$

Observe that, if g is any transformation, then so also is g^n for each integral (positive or negative) n and that the identity map is always model-preserving. Also observe that any transformation carries $\mathcal{P}$-null sets into $\mathcal{P}$-null sets.

(g) Given a transformation g, the sub-field $\mathcal{A}(g)$ of *g-invariant* sets is defined as

$$\mathcal{A}(g) = \{A \mid g^{-1}A = A\}.$$

The $\mathcal{P}$-completion $\bar{\mathcal{A}}(g)$ of $\mathcal{A}(g)$ is then the class of all *essentially g-invariant* sets, i.e., sets that are $\mathcal{P}$-equivalent to some g-invariant set.

(h) The set A is *almost g-invariant* if $g^{-1}A \sim A$. It is easy to demonstrate that every almost g-invariant set is also essentially g-invariant and vice versa (see Lemma 1 for a sharper result). Thus, $\bar{\mathcal{A}}(g)$ is also the class of all almost g-invariant sets.

(i) Given a class $\mathcal{G}$ of transformations g, the three sub-fields of α) *$\mathcal{G}$-invariant*, β) *essentially $\mathcal{G}$-invariant* and γ) *almost $\mathcal{G}$-invariant* sets are defined as follows :

α) $\mathcal{A}(\mathcal{G}) = \bigcap \mathcal{A}(g)$, ($\mathcal{G}$-invariant)

β) $\bar{\mathcal{A}}(\mathcal{G}) = \mathcal{P}$-completion of $\mathcal{A}(\mathcal{G})$, (essentially $\mathcal{G}$-invariant) and

γ) $\tilde{\mathcal{A}}(\mathcal{G}) = \bigcap \bar{\mathcal{A}}(g)$, (almost $\mathcal{G}$-invariant).

Observe that $\mathcal{A}(\mathcal{G}) \subset \bar{\mathcal{A}}(\mathcal{G}) \subset \tilde{\mathcal{A}}(\mathcal{G})$. With some assumptions on $\mathcal{G}$, one can prove (see Theorem 4 on p. 225 in [6]) the equality of $\bar{\mathcal{A}}(\mathcal{G})$ and $\tilde{\mathcal{A}}(\mathcal{G})$. That $\bar{\mathcal{A}}(\mathcal{G})$ can be a very small sub-field compared to $\tilde{\mathcal{A}}(\mathcal{G})$ is shown in example 1.

(j) A function f is α) *$\mathcal{G}$-invariant*, β) *essentially $\mathcal{G}$-invariant*, or γ) *almost $\mathcal{G}$-invariant*, according as

α) $f = f(g)$, for all $g \in \mathcal{G}$,

β) $f \sim$ some $\mathcal{G}$-invariant function,

or

γ) $f \sim f(g)$, for all $g \in \mathcal{G}$.

Observe that f satisfies the definitions α), β) or γ) above if and only if f is measurable with respect to the corresponding sub-field defined in (i).

(k) When the class $\mathcal{G}$ of transformations g happens to be a group (with respect to the operation of composition of transformations), the sub-field $\mathcal{A}(\mathcal{G})$ of $\mathcal{G}$-invariant sets is easily recognized as follows. For each $x \epsilon \mathcal{X}$, define the *orbit* O_x as

$$O_x = \{x' \mid x' = gx \text{ for some } g \epsilon \mathcal{G}\}.$$

The orbits define a partition of $\mathcal{X}$, and the sub-field $\mathcal{A}(\mathcal{G})$ is the class of all (measurable) sets that are the unions of orbits. The sub-field of essentially $\mathcal{G}$-invariant sets is then the $\mathcal{P}$-completion of $\mathcal{A}(\mathcal{G})$. Our main concern, in this paper, is the sub-field $\tilde{\mathcal{A}}(\mathcal{G})$ and this is not so easily understood in terms of the orbits—unless $\mathcal{G}$ has a structure simple enough to ensure the equality of $\bar{\mathcal{A}}(\mathcal{G})$ and $\tilde{\mathcal{A}}(\mathcal{G})$ (see lemma 1 and example 1).

3. A MATHEMATICAL INTRODUCTION

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a given probability model and let g be a fixed model-preserving transformation [definition 2(f)]. We remark that the identity map is trivially model-preserving. In many instances of statistical interest there exist fairly wide classes of such transformations. However, it is not difficult to construct examples where no non-trivial transformation is model-preserving. See for instance example 5.

For each bounded function f [definition 2(b)], define the associated sequence $\{f_n\}$ of (uniformly bounded) functions as follows :

$$f_n(x) = [f(x) + f(gx) + \ldots + f(g^n x)]/(n+1), \qquad n = 1, 2, 3, \ldots$$

Since g is measure-preserving for each $P \epsilon \mathcal{P}$, the pointwise ergodic theorem tells us that the set N_f where $\{f_n(x)\}$ fails to converge is $\mathcal{P}$-null [definition 2(c)].

If we define f^* as

$$f^*(x) = \begin{cases} \lim f_n(x) & \text{when } x \notin N_f, \\ 0 & \text{otherwise} \end{cases}$$

then, it is easily seen that

i) f^* is $\mathcal{A}(g)$-measurable [definition 2(g)], i.e., it is g-invariant [definition 2(i)], and that

ii) for all $B \epsilon \mathcal{A}(g)$ and $P \epsilon \mathcal{P}$

$$\int_B f dP = \int_B f^* dP.$$

This is another way of saying that f^* is the conditional expectation of f, given the sub-field $\mathcal{A}(g)$ of g-invariant sets. Since the definition of f^* does not involve P, we have the following theorem (lemma 2 in [4]).

Theorem 1. *If the transformation g preserves the model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, then the sub-field $\mathcal{A}(g)$ is sufficient.*

[*Remark*: Note that in the proof of theorem 1 we have not used the assumptions that g is a one-to-one map and that it is bimeasurable. The proof remains valid for any measurable mapping of $(\mathcal{X}, \mathcal{A})$ into itself that is measure-preserving for each $P \epsilon \mathcal{P}$. A similar remark will hold true for a number of other results to be stated later. However, in a study of the statistical theory of invariance (see [6]) it seems appropriate to restrict our attention to one-to-one, bimeasurable maps of $(\mathcal{X}, \mathcal{A})$ onto itself that preserve the model either wholly or partially.]

Now, given a class $\mathcal{G}$ of model-preserving transformations g, what can we say about the sufficiency of the sub-field

$$\mathcal{A}(\mathcal{G}) = \bigcap \mathcal{A}(g)$$

of $\mathcal{G}$-invariant [definition 2(i)] sets ? The intersection of two sufficient sub-fields is not necessarily sufficient. However, it is known (see Theorem 4 and Corollary 2 of [3]) that the intersection of the $\mathcal{P}$-completions [definition 2(e)] of a countable number of sufficient sub-fields is sufficient. Using this result, we have the following theorem (theorem 2 in [4]).

Theorem 2. *If the class $\mathcal{G}$ of model-preserving transformations g is countable, then the sub-field*

$$\bar{\mathcal{A}}(\mathcal{G}) = \bigcap \bar{\mathcal{A}}(g)$$

of almost $\mathcal{G}$-invariant sets [definition 2(i)(γ)] *is sufficient.*

Given a countable class $\mathcal{G}$ of transformations, consider the larger class $\mathcal{G}^*$ of transformations of the type $\alpha_1\alpha_2 \ldots \alpha_n$, where each α_i is such that either α_i or α_i^{-1} belongs to $\mathcal{G}$, and n is an arbitrary positive integer. The following properties of $\mathcal{G}^*$ are easy to check :

a) $\mathcal{G}^*$ is a group (the group operation being composition of transformations),

b) $\mathcal{G}^*$ is countable,

c) $\mathcal{A}(\mathcal{G}^*) = \mathcal{A}(\mathcal{G})$ and $\tilde{\mathcal{A}}(\mathcal{G}^*) = \tilde{\mathcal{A}}(\mathcal{G})$.

Now, let A be an arbitrary almost $\mathcal{G}^*$-invariant set, i.e., $g^{-1}A \sim A$ (equivalently, $gA \sim A$) for all $g \in \mathcal{G}^*$. Consider the set

$$B = \bigcap gA$$

where the intersection is taken over all $g \in \mathcal{G}^*$. Since $\mathcal{G}^*$ is a group, the set B must be $\mathcal{G}^*$-invariant. Again, since $\mathcal{G}^*$ is countable, and each gA is $\mathcal{P}$-equivalent to A, we have $B \sim A$ We have thus established the following lemma.

Lemma 1. *For any countable class $\mathcal{G}$ of transformations* [*definition 2(f)*], *every almost $\mathcal{G}$-invariant set is essentially $\mathcal{G}$-invariant, i.e., $\tilde{\mathcal{A}}(\mathcal{G}) = \mathcal{A}(\mathcal{G})$.*

Theorem 2, together with lemma 1 and the observation that a sub-field that is $\mathcal{P}$-equivalent to a sufficient sub-field is itself sufficient, leads to the following theorem.

Theorem 3. *If $\mathcal{G}$ is a countable class of model-preserving transformations, then the sub-field $\mathcal{A}(\mathcal{G})$ of $\mathcal{G}$-invariant sets is sufficient.*

[*Remark* : Note that in Theorem 3 we have used the one-to-oneness and bimeasurability of our transformations.]

Before proceeding further, let us consider an example which shows that Theorem 3 is no longer true if we drop the condition that $\mathcal{G}$ is countable.

Example 1. Let $\mathcal{P}$ be a family of continuous distributions on the real line $\mathcal{X}$ where $\mathcal{A}$ is the σ-field of Borel-sets. Let $\mathcal{G}=\{g\}$ be the class of all one-to-one maps of $\mathcal{X}$ onto $\mathcal{X}$ such that $[x: gx \neq x]$ is finite. Clearly, $\mathcal{G}$ is a group and every member of it is model-preserving. It is easy to check that the sub-field $\mathcal{A}(\mathcal{G})$ of $\mathcal{G}$-invariant sets consists of ϕ and $\mathcal{X}$ only, and hence is not sufficient. The sub-field $\tilde{\mathcal{A}}(\mathcal{G})$ of almost $\mathcal{G}$-invariant sets is the same as $\mathcal{A}$ and is sufficient.

Consider another example.

Example 2. Let $\mathcal{X}$ be the n-dimensional Euclidean space and $\mathcal{P}$ the class of all probability measures (on the σ-field $\mathcal{A}$ of all Borel-sets) that are symmetric (in the co-ordinates). If $\mathcal{G}=\{g\}$ is the group of all permutations (of the co-ordinates) then $\mathcal{A}(\mathcal{G})$ is the sub-field of all sets that are symmetric (in the co-ordinates) and is sufficient. Since the empty set is the only $\mathcal{P}$-null set, the two sub-fields $\mathcal{A}(\mathcal{G})$ and $\tilde{\mathcal{A}}(\mathcal{G})$ are the same here.

Given a probability model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ we ask ourselves the following questions.

1. How wide is the group $\mathcal{G}$ of all model-preserving transformations ?
2. Is the sub-field $\tilde{\mathcal{A}}(\mathcal{G})$ of almost $\mathcal{G}$-invariant sets sufficient?
3. If a least (minimal) sufficient sub-field $\mathcal{L}$ exists, then is it true that $\mathcal{L} \sim \tilde{\mathcal{A}}(\mathcal{G})$?
4. What is $\tilde{\mathcal{A}}(\mathcal{G})$, when $\mathcal{G}$ is the class of all transformations that partially preserve the model in a given manner ?

4. STATISTICAL MOTIVATION

Suppose we have two different systems of measurement co-ordinates for the outcome of a statistical experiment, so that, if a typical outcome is recorded as x under the first system, the same outcome is recorded as gx under the second system. Let us suppose further that the two statistical variables x and gx

have the same domain (sample space) $\mathscr{X}$, the same family $\mathscr{A}$ of events, and the same class of probability measure $\mathscr{P}$. Furthermore, if $P \epsilon \mathscr{P}$ holds for x, then the same probability measure P also holds for gx. The second system of measurement co-ordinates may then be represented mathematically as a model-preserving transformation g of the statistical model $(\mathscr{X}, \mathscr{A}, \mathscr{P})$. The principle of invariance then stipulates that the decision rule should be invariant with respect to the transformation g—and this irrespective of the actual decision problem. If the choice of a new system of measurement co-ordinates leaves the problem (whatever it is) entirely unaffected, then so also should be every reasonable inference procedure.

Thus, if g is model-preserving, the principle of invariance leads to the invariance reduction of the model $(\mathscr{X}, \mathscr{A}, \mathscr{P})$ to the simpler model $(\mathscr{X}, \mathscr{A}(g), \mathscr{P})$, where $\mathscr{A}(g)$ is the sub-field of g-invariant sets. To put it differently, the principle of invariance requires every decision function to be g-invariant or $\mathscr{A}(g)$-measureable. Consider now the class $\mathscr{G}$ of all model-preserving transformations g. Must we, following the principle of invariance, insist that every decision rule be $\mathscr{G}$-invariant (i.e. g-invariant for every $g \epsilon \mathscr{G}$)? We have already noted in example 1 that, when $\mathscr{P}$ consists of a family of non-atomic measures, the class $\mathscr{G}$ is large enough to reduce the sub-field $\mathscr{A}(\mathscr{G})$ of $\mathscr{G}$-invariant sets to the trivial one (consisting of only the empty set and the whole space). Obviously, we cannot (must not) reduce $\mathscr{A}$ all the way down to $\mathscr{A}(\mathscr{G})$ or even to $\bar{\mathscr{A}}(\mathscr{G})$—the sub-field of essentially $\mathscr{G}$-invariant sets. A logical compromise (with the principle) would be to reduce $\mathscr{A}$ to the sub-field $\tilde{\mathscr{A}}(\mathscr{G})$ of all almost $\mathscr{G}$-invariant sets—that is to insist upon the decision function to be almost $\mathscr{G}$-invariant.

The principle of sufficiency is another reduction principle of the omnibus type. If $\mathscr{S}$ be a sufficient sub-field, then this principle tells us not to use a decision function that is not $\mathscr{S}$-measurable. Suppose there exists a least (minimal) sufficient sub-field $\mathscr{L}$. Following the principle of sufficiency, we reduce the model $(\mathscr{X}, \mathscr{A}, \mathscr{P})$ to the model $(\mathscr{X}, \mathscr{L}, \mathscr{P})$.

Which of the two reductions (invariance or sufficiency) is more extensive ? In other words, what is the relation between $\widetilde{\mathscr{A}}(\mathscr{G})$ and $\mathscr{L}$?

Theorem 1 tells us that $\mathscr{A}(g)$ is sufficient (for each g in $\mathscr{G}$). Since $\mathscr{L}$ is the least sufficient sub-field we have (from the definition of $\mathscr{L}$)

$$\overline{\mathscr{L}} \subset \overline{\mathscr{A}(g)} \text{ for all } g \in \mathscr{G}.$$

Theorem 4 follows at once.

Theorem 4. $\quad \overline{\mathscr{L}} \subset \bigcap \overline{\mathscr{A}(g)} = \overline{\widetilde{\mathscr{A}}(\mathscr{G})}.$

[*Remark*: Theorem 4 does not establish the sufficiency of $\widetilde{\mathscr{A}}(\mathscr{G})$. [See example 1 in [3].]

We thus observe that the invariance reduction (in terms of the group $\mathscr{G}$ of all model-preserving transformations) of a model can never be more extensive than its maximal sufficiency reduction (if one such reduction is available). The principal question raised in this paper is, "When is $\widetilde{\mathscr{A}}(\mathscr{G})$ equal to $\overline{\mathscr{L}}$?" We shall show later that in many familiar situations the sub-fields $\widetilde{\mathscr{A}}(\mathscr{G})$ and $\mathscr{L}$ are essentially equal. This raises the question about the nature of $\widetilde{\mathscr{A}}(\mathscr{G})$, where $\mathscr{G}$ is the class of all transformations that preserve the model partially in some well-defined manner. This question is discussed in a later section. In the next two sections we give two alternative approaches to Theorem 4.

5. WHEN A BOUNDEDLY COMPLETE SUFFICIENT SUB-FIELD EXISTS

Let us suppose that the sub-field $\mathscr{L}$ is sufficient and boundedly complete. We need the following lemma.

Lemma 2. *If z is a bounded $\mathscr{A}$-measurable function such that*

$$E(z \mid P) \equiv 0 \text{ for all } P \in \mathscr{P},$$

then, for all bounded $\mathscr{L}$-measurable functions f, it is true that

$$E(zf \mid P) \equiv 0 \quad \text{for all } P \in \mathscr{P}.$$

The proof of this well-known result is omitted. Now, let S be an arbitrary member of $\mathscr{L}$, and let g be an arbitrary model-preserving transformation. Let $S_0 = g^{-1}S$. Since g is model-preserving we have

$$P(S) \equiv Pg^{-1}S \equiv P(S_0) \quad \text{for all} \quad P \epsilon \mathscr{P}.$$

Writing I_S for the indicator of S, and noting that $I_S - I_{S_0}$ and I_S satisfy the conditions for z and f in Lemma 2, we at once have

$$E[(I_S - I_{S_0})I_S | P] \equiv 0, \text{ for all } P \epsilon \mathscr{P},$$

or

$$P(S) \equiv P(SS_0), \quad \text{for all } P \epsilon \mathscr{P}.$$

Hence,

$$\begin{aligned} P(S \Delta S_0) &\equiv P(S) + P(S_0) - 2P(SS_0) \\ &\equiv 2[P(S) - P(SS_0)] \\ &\equiv 0, \quad \text{for all } P \epsilon \mathscr{P}. \end{aligned}$$

That is, for all $S \epsilon \mathscr{L}$, the two sets S and $g^{-1}S$ are $\mathscr{P}$-equivalent. In other words,

$$\mathscr{L} \subset \bar{\mathscr{A}}(g),$$

and since g is an arbitrary element of $\mathscr{G}$ (the class of all model-preserving transformations) we have the following theorem.

Theorem 4 (a). *If $\mathscr{L}$ is a boundedly complete sufficient sub-field then $\mathscr{L} \subset \tilde{\mathscr{A}}(\mathscr{G})$.*

[*Remark*: Since bounded completeness of $\mathscr{L}$ implies that $\mathscr{L}$ is the least sufficient sub-field, Theorem 4(a) is nothing but a special case of Theorem 4. The proof of Theorem 4(a) is simple and amenable to a generalization to be discussed later.]

6. THE DOMINATED CASE

We make a slight digression to state a useful lemma. Let T be a measurable map of $(\mathscr{X}, \mathscr{A})$ into $(\mathscr{X}, \mathscr{B})$. Let P and Q be two probability measures on $\mathscr{A}$ and let PT^{-1} and QT^{-1} be the corresponding measures on $\mathscr{B}$. Suppose that Q dominates P and

let $f = dP/dQ$ be the Radon-Nikodym derivative defined on $\mathcal{X}$. It is then clear that QT^{-1} dominates PT^{-1}. Let $h = (dPT^{-1})/(dQT^{-1})$. The function hT on $\mathcal{X}$ (defined as $hT(x) = h(Tx)$) is $T^{-1}(\mathcal{B})$-measurable and satisfies the following relation.

Lemma 3. $\quad hT = E(f \mid T^{-1}(\mathcal{B}), Q),$

i.e., hT is the conditional expectation of f, given $T^{-1}(\mathcal{B})$ and Q.

The proof of this well-known lemma consists of checking the identity

$$\int_B f dQ \equiv \int_B hT dQ, \quad \text{for all } B \in T^{-1}(\mathcal{B}).$$

Corollary. *If $T^{-1}(\mathcal{B})$ is Q-equivalent to $\mathcal{A}$, then f and hT are Q-equivalent.*

Now, returning to our problem, let $\mathcal{G}$ be the class of all transformations g that preserves the model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Let us suppose that $\mathcal{P}$ is dominated by some σ-finite measure. It follows that there exists a countable collection $P_1, P_2, \ldots$, of elements in $\mathcal{P}$ such that the convex combination

$$Q = \Sigma c_i P_i, \quad c_i > 0, \quad \Sigma c_i = 1,$$

dominates the family $\mathcal{P}$. Let $f_P = dP/dQ$ be a fixed version of the Radon-Nikodym derivative of P with respect to Q. The factorization theorem for sufficient statistics asserts that a subfield $\mathcal{A}_0$ is sufficient if and only if f_P is $\overline{\mathcal{A}_0}$-measurable for every $P \in \mathcal{P}$, where $\overline{\mathcal{A}_0}$ is the $\mathcal{P}$-completion of $\mathcal{A}_0$. We now prove that f_P is $\widetilde{\mathcal{A}}(\mathcal{G})$-measurable for every P. (The $\mathcal{P}$-completion of $\widetilde{\mathcal{A}}(\mathcal{G})$ is itself.)

Since $Pg^{-1} = P$ for all $P \in \mathcal{P}$, it follows that $Qg^{-1} = Q$. From Lemma 3 we have

$$f_P g = E(f_P \mid g^{-1}(\mathcal{A}), Q).$$

The assumption that g is one-to-one and bimeasurable implies that $g^{-1}(\mathcal{A}) = \mathcal{A}$. Hence $f_P g = f_P$ a.e.w. [Q]. Since Q dominates $\mathcal{P}$, it follows that f_P is almost g-invariant, i.e., is $\overline{\mathcal{A}(g)}$ mea-

surable. Since, in the above argument, g is an arbitrary element of $\mathscr{G}$, we have the following theorem. (See problem 19 on p. 253 in [6].)

Theorem 4(b). *If $\mathscr{P}$ is dominated, then $\tilde{\mathscr{A}}(\mathscr{G})$ is sufficient.*

[*Remark*: Since, in the dominated set-up, the least sufficient sub-field $\mathscr{L}$ always exists and since, in this set-up, any sub-field containing $\mathscr{L}$ is necessarily sufficient, it is clear that Theorem 4(b) is nothing but an immediate corollary to Theorem 4. Also note that in the present proof (of Theorem 4(b)) we had to draw upon our supposition that g is one-to-one and bimeasurable. This section has been written only with the object of drawing attention to some aspects of the problem.]

7. EXAMPLES

Example 3: Let y be a real random variable having a uniform distribution over the unit interval. For each c in $[0, 1)$ define the transformation g_c as

$$g_c y = y + c \pmod 1$$

In this example, $\mathscr{P}$ consists of a single measure and each g_c is model preserving (measure-preserving). If $\mathscr{G}_0 = \{g_c \mid 0 < c < 1\}$, then the only $\mathscr{G}_0$-invariant sets are the empty set and the whole of the unit interval. Here, $\mathscr{A}(\mathscr{G}_0)$ is sufficient.

[*Remark*: In this case, there are a very large number of measure-preserving transformations that are not one-to-one maps of $[0, 1]$ onto itself. For example, let $a_n(y)$ be the nth digit in the decimal representation of y and let

$$gy = \sum_{k=1}^{\infty} \frac{a_{n_k}(y)}{10^k}$$

where $\{n_k\}$ is a fixed increasing sequence of natural numbers.]

Again, if x has a fixed continuous distribution on the real line with cumulative distribution function F, then the class $\mathscr{G}_0$ of transformations g_c defined as

$$g_c x = F^{-1}[F(x) + c \pmod 1], \quad 0 < c < 1$$

are all model-preserving for x. [In case F is not a strictly increasing function of x, we define $F^{-1}(y) = \inf\{x \mid F(x) = y\}$.]

Thus, for any fixed continuous distribution on the real line, there always exists a large class of measure-preserving transformations.

Example 4. Let $\mathcal{X} = [0, \infty)$ and let x have a uniform distribution over the interval $[0, \theta)$, where θ is an unknown positive integer. Here $\mathcal{P}$ consists of a countable infinity of probability measures. For each c in $[0, 1)$, define the transformation g_c as

$$g_c x = [x] + \{x - [x] + c \pmod 1\},$$

where $[x]$ is the integer-part of x.

Here, each g_c is model-preserving. The minimal (least) sufficient statistic $[x]$ is also the maximal-invariant with respect to the group $\mathcal{G}_0 = \{g_c\}$ of model-preserving transformations defined above. Thus, if $\mathcal{L}$ is the sub-field (least sufficient) generated by $[x]$ and $\mathcal{A}(\mathcal{G}_0)$ is the sub-field of $\mathcal{G}_0$-invariant sets, we have

$$\mathcal{L} = \mathcal{A}(\mathcal{G}_0). \tag{a}$$

Now, if $\mathcal{G}$ is the class of all model-preserving transformations, then (as we have seen in Example 1) the sub-field $\mathcal{A}(\mathcal{G})$ will reduce to the level of triviality. (It will consist of only the empty set and the whole set $\mathcal{X}$.) However, from Theorem 4 we have

$$\bar{\mathcal{L}} \subset \tilde{\mathcal{A}}(\mathcal{G}). \tag{b}$$

Since the group $\mathcal{G}_0$ has a *decent* structure, we can apply Stein's theorem (theorem 4 on p. 225 in [6]) to prove that

$$\mathcal{A}(\mathcal{G}_0) \sim \tilde{\mathcal{A}}(\mathcal{G}_0). \tag{c}$$

Since $\mathcal{G}_0 \subset \mathcal{G}$, we at once have

$$\tilde{\mathcal{A}}(\mathcal{G}) \subset \tilde{\mathcal{A}}(\mathcal{G}_0). \tag{d}$$

Putting the relations (a), (c), and (d) together we have

$$\tilde{\mathscr{A}}(\mathscr{G}) \subset \bar{\mathscr{L}} \tag{e}$$

Putting (b) and (e) together we finally have

$$\bar{\mathscr{L}} = \tilde{\mathscr{A}}(\mathscr{G}),$$

i.e., the least-sufficient sub-field (rather, the $\mathscr{P}$-completion of any version of the least sufficient sub-field) and the sub-field of almost $\mathscr{G}$-invariant sets are identical.

The chain of arguments, detailed as above, is of a general nature and will be used repeatedly in the sequel.

Example 5. Let x be a normal variable with unit variance and mean equal to either μ_1 or μ_2. Does there exist a non-trivial transformation of $\mathscr{X}$ (the real line) into itself that preserves each of the two measures ? That the answer is "no" is seen as follows. Let $\mathscr{G}$ be the class of all model-preserving transformations. In view of theorem 4, $\tilde{\mathscr{A}}(\mathscr{G})$ contains the least sufficient sub-field $\mathscr{L}$. But, in this example, the likelihood ratio (which is the least sufficient statistic) is

$$\exp\left[(\mu_1-\mu_2)x - \frac{1}{2}(\mu_1^2-\mu_2^2)\right],$$

and this is a one-to-one measurable function of x. Thus, every set is almost $\mathscr{G}$-invariant. And this implies that every g in $\mathscr{G}$ must be equivalent to the identity map.

Example 6. Let $x_1, x_2, \ldots, x_n$ be n independent observations on a normal variable with known mean and unknown standard deviation σ. Without loss of generality we may assume the mean to be zero. Let $\mathscr{G}_0$ be the group of all linear orthogonal transformations of the n-dimentional Euclidean space onto itself. Clearly, every member of $\mathscr{G}_0$ is model-preserving. That the class $\mathscr{G}$ of all model-preserving transformations is much wider than $\mathscr{G}_0$ is seen as follows. Let

$$y_i = \phi_i(x_i)\,|x_i|, \quad i = 1, 2, \ldots, n,$$

where $\phi_1, \phi_2, \ldots, \phi_n$ are arbitrary skew-symmetric (i.e., $\phi(x) = -\phi(-x)$ for all x) functions on the real line that take only the two values -1 and $+1$. It is easily checked that, whatever the value of σ, the two vectors $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are identically distributed, i.e., the above transformation (though non-linear) is model-preserving. However, the sub-group $\mathcal{G}_0$ is large enough to lead us to Σx_i^2—which is the least sufficient statistic—as the maximal invariant. In view of the decent structure of the sub-group $\mathcal{G}_0$, the arguments given for example 4 are again available to prove the equality of $\bar{\mathcal{L}}$ and $\tilde{\mathcal{A}}(\mathcal{G})$.

8. TRANSFORMATIONS OF A SET OF NORMAL VARIABLES

This section is devoted to a study of the special case (model) of n independent normal variables $x_1, x_2, \ldots, x_n$ with equal unknown means μ and equal unknown standard deviations σ. Hence Σx_i and Σx_i^2 jointly constitute the least sufficient statistic.

If $\bar{\mathcal{L}}$ is the $\mathcal{F}$-completion of the sub-field $\mathcal{L}$ induced by $(\Sigma x_i, \Sigma x_i^2)$, then we know from Theorem 4, that

$$\bar{\mathcal{L}} \subset \tilde{\mathcal{A}}(\mathcal{G})$$

where $\mathcal{G}$ is the class of all the model-preserving transformations of $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ to $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. For any model-preserving transformation from $\boldsymbol{x}$ to $\boldsymbol{y}$, we, therefore, have

$$\Sigma x_i \sim \Sigma y_i$$

and

$$\Sigma x_i^2 \sim \Sigma y_i^2$$

i.e., the statistic $(\Sigma x_i, \Sigma x_i^2)$ is almost $\mathcal{G}$-invariant. If we can demonstrate the existence of a '*decent*' sub-group $\mathcal{G}_0$ of $\mathcal{G}$ for which the statistic $(\Sigma x_i, \Sigma x_i^2)$ is the maximal invariant, then (following the method of proof indicated in examples 4 and 6) we can show that $\bar{\mathcal{L}}$ is indeed equal to $\tilde{\mathcal{A}}(\mathcal{G})$.

Let $\mathcal{G}_0$ be the sub-group of all linear model-preserving transformations. Do there exist non-trivial linear model-preserving

transformations (i.e., linear transformations that are not a permutation of the co-ordinates) ? That the answer is "yes" is seen as follows. Let $\mathcal{M} = \{M\}$ be the family of all orthogonal $n \times n$ matrices with the initial row as

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).$$

Now, if $x = (x_1, x_2, \ldots, x_n)$ are independent $N(\mu, \sigma)$'s and we define

$$y' = Mx'$$

(where x' is the corresponding column vector), then the y_i's are independent normal variables with equal standard deviation σ and with means as follows :

$$E(y_1) = \sqrt{n}\,\mu, \quad E(y_i) = 0, \qquad i = 2, 3, \ldots, n.$$

Thus, for an arbitrary pair of members M_1 and M_2 in $\mathcal{M}$, we note that M_1x' and M_2x' are identically distributed (whatever the values of μ and σ). Therefore, the two vectors x' and $M_2^{-1}M_1x$ are identically distributed for each μ and σ. In other words the linear transformation defined by the matrix

$$M_2^{-1}M_1 \qquad (= M_2'M_1,$$

since M_2 is orthogonal) is model-preserving for each pair M_1, M_2 of members from $\mathcal{M}$.

[*Remark* : Later on we shall have some use for this way of generating members of $\mathcal{G}_0$.]

For example, the 4×4 matrix

$$\begin{pmatrix} 1/2 & 1/2 & 1/2 & -1/2 \\ 1/2 & 1/2 & -1/2 & 1/2 \\ 1/2 & -1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix}$$

defines a member of $\mathcal{G}_0$ for $n = 4$.

Let H be a typical $n \times n$ matrix that defines a model-preserving linear transformation. We are going to make a brief digression about the nature of H. From the requirement that each co-ordinate of

$$y' = Hx'$$

has mean μ and standard deviation σ, we at once have that the elements in each row of H must add up to unity and that their squares also must add up to unity. From the mutual independence of the y_i's, it follows that the row vectors of H must be mutually orthogonal. Thus, H must be an orthogonal matrix with unit row sums. Now, for each model-preserving H, its inverse

$$H^{-1} \ (= H', \text{ since } H \text{ is orthogonal})$$

is necessarily also model-preserving and, thus, the columns of H must also add up to unity, etc.

It is easily checked that the sub-group $\mathcal{G}_0 = \{H\}$ of linear model-preserving transformations of the n-space onto itself may also be characterized as the class of all linear transformations that preserve both the sum and the sum of squares of the co-ordinates. The author came to learn that G. W. Haggastrom of the University of Chicago had come upon these matrices from this point of view and had a brief discussion on them in an unpublished work of his. We call such matrices by the name Haggstrom-matrix.]

Going back to our problem, we have to demonstrate that the statistic $T = (\Sigma x_i, \Sigma x_i^2)$ is a maximal invariant with respect to the sub-group $\mathcal{G}_0 = \{H\}$ of model-preserving linear transformations. For this we have only to prove that if $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$ are any two points in n-space such that $T(\mathrm{a}) = T(\mathrm{b})$ then there exists a Haggstrom-matrix H such that

$$\boldsymbol{b}' = H\boldsymbol{a}'.$$

Let M_1 and M_2 be two arbitrary members of $\mathcal{M}$—the class of all orthogonal $n \times n$ matrices with the leading row as $(1/\sqrt{n}, \ldots,$

$1/\sqrt{n}$). If $\boldsymbol{\alpha} = M_1\boldsymbol{a}'$ and $\boldsymbol{\beta} = M_2\boldsymbol{b}'$, then we have, from $T(\boldsymbol{a}) = T(\boldsymbol{b})$, that $\alpha_1 = \beta_1$ that and that

$$\sum_2^n \alpha_i^2 = \sum_2^n \beta_i^2.$$

It follows that there exists an $(n-1)\times(n-1)$ orthogonal matrix that will transform $(\alpha_2, \alpha_3, \ldots, \alpha_n)$ to $(\beta_2, \beta_2, \ldots, \beta_n)$. And this, in turn, implies that there exists an $n\times n$ orthogonal matrix K, with the first row as $(1, 0, 0, \ldots, 0)$, such that

$$\boldsymbol{\beta}' = K\boldsymbol{\alpha}'.$$

Thus, the transformation $M_2^{-1}KM_1$ takes $\boldsymbol{a}$ into $\boldsymbol{b}$. Now, note that, since K is an orthogonal matrix with the initial row as $(1, 0, 0, \ldots, 0)$, the matrix KM_1 is orthogonal with its initial row as $(1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n})$, i.e. $KM_1 \in \mathcal{M}$. In other words, there exists a matrix of the form $M_2^{-1}M$ (with M_2 and M belonging to $\mathcal{M}$) which transforms $\boldsymbol{a}$ into $\boldsymbol{b}$. We have already noted that all such matrices belong to $\mathcal{G}_0$, and this completes our proof that $(\Sigma x_i, \Sigma x_i^2)$ is a maximal invariant with respect to $\mathcal{G}_0$.

In example 6 of the previous section, we considered the particular case of the foregoing problem where μ is known. Consider now the other particular case where σ is known and μ is the only unknown parameter. In this case Σx_i is the least sufficient statistic. Now, it is no longer possible to produce a sub-group $\mathcal{G}_0$ of linear model-preserving transformations such that Σx_i is the maximal invariant with respect to $\mathcal{G}_0$. This is because every linear model-preserving transformation must necessarily be orthogonal and would thus preserve Σx_i^2 also. We proceed as follows.

Suppose (without loss of generality) that $\sigma = 1$. Let $\boldsymbol{y}' = M\boldsymbol{x}'$ where M is a fixed member of the class $\mathcal{M}$ of orthogonal $n\times n$ matrices with the initial row as $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$. Observe that the y_i's are mutually independent and that $y_2, y_3, \ldots, y_n$ are standard normal variables. Let F stand for the cumulative distribution function of a standard normal variable. Let c_2,

$c_2, \ldots, c_n$ be arbitrary constants in $[0, 1)$. If we define $z_1 = y_1$ and

$$z_i = F^{-1}[F(y_i)+c_i \pmod 1], \qquad i = 2, 3, \ldots, n,$$

then $z = (z_1, z_2, \ldots, z_n)$ has the same distribution as that of y and so it follows that x' and $M^{-1}z'$ are identically distributed. Observe that we have described above a model-preserving transformation corresponding to each $(n-1)$-vector $(c_2, c_3, \ldots, c_n)$ with the c_i's in $[0, 1)$. It is easily checked that Σx_i is a maximal invariant with respect to the above class of model-preserving transformations.

9. PARAMETER-PRESERVING TRANSFORMATIONS

Let $\gamma = \gamma(P)$ be the parameter of interest. That is, the experimenter is interested only in the characteristic $\gamma(P)$ of the measure P (that actually holds) and considers all other details about P to be irrelevent (nuisance parameters). We define a γ-preserving transformation as follows.

Definition. The transformation [see Definition 2(f)] g is γ-preserving if $\gamma(Pg^{-1}) = \gamma(P)$, for all $P \in \mathcal{P}$.

If g is γ-preserving then so also is g^{-1}. The composition of any two γ-preserving transformations is also γ-preserving. Let $\mathcal{G}_\gamma$ be the group of all γ-preserving transformations.

In the particular case where $\gamma(P) = P$, the γ-preserving transformations are what we have so far been calling model-preserving. If $\mathcal{G}$ is the class of all model-preserving transformations, then note that $\mathcal{G} \subset \mathcal{G}_\gamma$ for every γ.

If γ is the parameter of interest, then the principle of invariance leads to the reduction of $\mathcal{A}$ to the sub-field $\tilde{\mathcal{A}}(\mathcal{G}_\gamma)$. This section is devoted to a study of the sub-field $\tilde{\mathcal{A}}(\mathcal{G}_\gamma)$.

Since $\mathcal{G} \subset \mathcal{G}_\gamma$ we have

$$\tilde{\mathcal{A}}(\mathcal{G}_\gamma) \subset \tilde{\mathcal{A}}(\mathcal{G}).$$

Let us suppose that the least sufficient sub-field $\mathcal{L}$ exists. We have discussed a number of examples where $\tilde{\mathcal{A}}(\mathcal{G})$ and $\mathcal{L}$ are

identical. [The author believes that the above identification is true under very general conditions.] In all such situations we therefore have

$$\tilde{\mathcal{A}}(\mathcal{G}_\gamma) \subset \bar{\mathcal{L}}.$$

That is, when the interest of the experimenter is concentrated on some particular characteristic $\gamma(P)$ or P, the principle of invariance will usually reduce the data more extensively than the principle of sufficiency.

Theorems 4, 4(a) and 4(b) tell us that

$$\bar{\mathcal{L}} \subset \tilde{\mathcal{A}}(\mathcal{G}).$$

The following theorem gives us a similar lower bound for $\tilde{\mathcal{A}}(\mathcal{G}_\gamma)$.

A set A is called γ-oriented if, for every pair $P_1, P_2 \epsilon \mathcal{P}$ such that $\gamma(P_1) = \gamma(P_2)$, it is true that $P_1(A) = P_2(A)$. In other words, a γ-oriented set is one whose probability depends on P through $\gamma(P)$. A sub-field is γ-oriented if every member of the sub-field is. The following theorem* is a direct generalization of Theorem 4(a).

Theorem 5. *Let $\mathcal{G}_\gamma$ be the class of all γ-preserving transformations and let $\mathcal{B}$ be a sub-field that is*

i) *γ-oriented and*

ii) *contained in the boundedly complete sufficient sub-field $\mathcal{L}$ (which exists).*

Then,

$$\mathcal{B} \subset \tilde{\mathcal{A}}(\mathcal{G}_\gamma).$$

Proof : Let $g \epsilon \mathcal{G}_\gamma$ and $B \epsilon \mathcal{B}$ and let $B_0 = g^{-1} B$

Then

$$P(B_0) = P(g^{-1}B) \qquad (\because \quad B_0 = g^{-1}B)$$

$$= Pg^{-1}(B).$$

* The author wishes to thank Professor W. J. Hall of the University of North Carolina for certain comments that eventually led to this theorem.

Since g is γ-preserving, we have $\gamma(Pg^{-1}) = \gamma(P)$. And since B is γ-oriented we have

$$Pg^{-1}(B) = P(B).$$

Therefore,

$$P(B_0) = P(B) \quad \text{or} \quad R(I_B - I_{B_0} | P) \equiv 0.$$

The rest of the proof is the same as in Theorem 4(a).

In the next section we repeatedly use the above theorem.

10. SOME TYPICAL INVARIANCE REDUCTIONS

Let $x_1, x_2, \ldots, x_n$ be n independent and identical normal variables with unknown μ and σ. Suppose the parameter of interest is σ. Let $\mathcal{G}_\sigma$ be the class of all σ-preserving transformations. If $\bar{x}$ is the mean and s, the standard deviation of the n observations, then the pair $(\bar{x}, s)$ is a complete sufficient statistic. Also s is σ-oriented. From Theorem 5, we then have that s is almost $\mathcal{G}_\sigma$-invariant. Thus, the principle of invariance cannot reduce the data beyond s. That s is indeed the exact (upto $\mathcal{P}$-equivalence) attainable limit of invariance reduction is shown as follows.

Let $\{H\}$ be the class of all $n \times n$ Haggstrom matrices (see section 8), i.e., each H is an orthogonal matrix with unit row and column sums. Let c be an arbitrary real number and let $\boldsymbol{i}$ stand for the n-vector $(1, 1, \ldots, 1)$. Consider all linear transformations of the type

$$\boldsymbol{y}' = H\boldsymbol{x}' + (c\boldsymbol{i})'.$$

If $\mathcal{G}_\sigma^*$ is this class of transformations, then it is easily verified that $\mathcal{G}_\sigma^*$ is a sub-group of $\mathcal{G}_\sigma$. That $\Sigma(x_i - \bar{x})^2$ is a maximal invariant with respect to the sub-group $\mathcal{G}_\sigma^*$ of σ-preserving transformations is seen as follows.

Let $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$ be two arbitrary n-vectors such that

$$\Sigma(a_i - \bar{a})^2 = \Sigma(b_i - \bar{b})^2.$$

Let $c = \bar{b} - \bar{a}$. Then the two vectors $\boldsymbol{a} + c\boldsymbol{i}$ and $\boldsymbol{b}$ have equal sums and sums of squares (of co-ordinates). Hence there exists a

Haggstrom matrix H that transforms $\boldsymbol{a}+c\boldsymbol{i}$ into $\boldsymbol{b}$. In other words, the transformation

$$\boldsymbol{y}' = H\boldsymbol{x}' + (c\boldsymbol{i})'$$

maps $\boldsymbol{a}$ into $\boldsymbol{b}$.

If $\mathcal{B}(s)$ is the sub-field generated by s, then we have just proved that $\mathcal{B}(s) = \mathcal{A}(\mathcal{G}_\sigma^*)$. The proof of the $\mathcal{P}$-equivalence of

$$\mathcal{B}(s) \quad \text{and} \quad \tilde{\mathcal{A}}(\mathcal{G}_\sigma)$$

will follow the familiar pattern set up earlier for example 4 in section 7.

Now, suppose that our parameter of interest is $\gamma = \mu/\sigma$. The statistic $\bar{x}/s$ is γ-oriented and hence (Theorem 5) is almost $\mathcal{G}_\gamma$-invariant, where $\mathcal{G}_\gamma$ is the group of all γ-preserving transformations.

Consider now the sub-group $\mathcal{G}_\gamma^*$ of all linear transformations of the type

$$\boldsymbol{y}' = c H \boldsymbol{x}'$$

where $c > 0$ and H is a Haggstrom matrix.

It is easily checked that the maximal invariant with respect to the sub-group $\mathcal{G}_\gamma^*$ is the statistic $\bar{x}/s$ and hence, the sub-field $\mathcal{B}(\bar{x}/s)$ generated by $\bar{x}/s$ is $\mathcal{P}$-equivalent to the sub-field $\tilde{\mathcal{A}}(\mathcal{G}_\gamma)$ of all almost $\mathcal{G}_\gamma$-invariant sets.

The case where our parameter of interest is $|\mu|/\sigma$ is very similar. Once again we observe that the statistic $|\bar{x}|/s$ is oriented towards $|\mu|/\sigma$. Hence, making use of Theorem 5 and observing that $|\bar{x}|/s$ is the maximal invariant with respect to the sub-group of linear parameter-preserving transformations of the form

$$\boldsymbol{y}' = cH\boldsymbol{x}' \quad (c \neq 0,\ H \text{ a Haggstrom matrix}),$$

we are able to show that the invariance reduction of the data is to the statistic $|\bar{x}|/s$.

11. SOME FINAL REMARKS

a) In statistical literature, the principle of invariance has been used in a rather half-hearted manner (see, for example, [5] and [6]). We do not find any consideration given to the present project of reducing the data with the help of the whole class $\mathcal{G}_\gamma$ of γ-preserving transformations. In fact, the question of how extensive the class $\mathcal{G}_\gamma$ can be has escaped general attention. One usually works in the framework of a relatively small and simple sub-group of $\mathcal{G}_\gamma$ and invokes simultaneously the two different principles of sufficiency and invariance for the purpose of arriving at a *satisfactory* data reduction. The main object of the present article is to investigate how far the principle of invariance by itself can take us.

b) The main limitation of the principle of sufficiency is that it does not recognize nuisance parameters. Several attempts have been made to generalize the idea of sufficiency so that one gets an effective data reduction in the presence of nuisance parameters. Not much success has, however, been achieved in this direction.

c) On the other hand, the invariance principle usually falls to pieces when faced with a discrete model. The Bernoulli experimental set-up is one of the rare discrete models that the invariance principle can tackle. If $x_1, x_2, \ldots, x_n$ are n independent zero-one variables with probabilities θ and $1-\theta$, then all permutations (of co-ordinates) are model-preserving. And they reduce the data directly to the least sufficient statistic $r = x_1 + \ldots + x_n$. Take however, the following simple example. Let x and y be independent zero-one variables, where

$$P(x = 0) = \theta, \quad 0 < \theta < 1,$$

and

$$P(y = 0) = 1/3.$$

Now the identity-map is the only available model-preserving transformation. The principle of sufficiency reduces the data immediately to x.

d) The object of this article was not to make a critical evaluation of the twin principles of data reduction. Yet, the author finds it hard to refrain from observing that both the principles of sufficiency and invariance are extremely sensitive to changes in the model. For example, the spectacular data reductions we have achieved in the many examples considered here become totally unavailable if the basic normality assumption is changed ever so slightly.

References

[1] Bahadur, R. R. "Sufficiency and Statistical Decision Functions," *Ann. Math. Statist.*, 25 (1954), 423–462.

[2] Basu, D. "Sufficiency and Model-Preserving Transformations," Inst. of Statistics, Mimeo Series No. 420, Univ. of North Carolina, Chapel Hill, 1965.

[3] Burkholder, D. L. "Sufficiency in the Undominated Case," *Ann. Math. Statist.*, 32 (1961), 1191–1200.

[4] Farrell, R. H. "Representation of Invariant Measures," *Illinois J. of Math.*, 6 (1962), 447–467.

[5] Hall, W. J., R. A. Wijsman, and J. K. Gosh. "The Relationship between Sufficiency and Invariance," *Ann. Math. Statist.*, 36 (1965), 575–614.

[6] Lehmann, E. L. *Testing Statistical Hypotheses*, John Wiley and Sons, New York, 1959.

(*Received Sept. 17, 1965.*)

AN ESSAY ON THE LOGICAL FOUNDATIONS OF SURVEY SAMPLING, PART ONE*

D. Basu
The University of New Mexico and
Indian Statistical Institute

1
An Idealization of the Survey Set-up

It is a mathematical necessity that we idealize the real state of affairs and come up with a set of concepts that are simple enough to be incorporated in a mathematical theory. We have only to be careful that the process of idealization does not distort beyond recognition the basic features of a survey set-up, which we list as follows:

(a) There exists a population—a finite collection $\mathcal{P}$ of distinguishable objects. The members of $\mathcal{P}$ are called the (sampling) units. [Outside of survey theory the term population is often used in a rather loose sense. For instance, we often talk of the infinite population of all the heads and tails that may be obtained by repeatedly tossing a particular coin. Again, in performing a Latin-square agricultural experiment the actual yield from a particular plot is conceived of as a sample from a conceptual population of yields from that plot. It is needless to mention that such populations are not real. The existence of a down-to-Earth finite population is a principal characteristic of the survey set-up.]

(b) There exists a sampling frame of reference. By this we mean that the units in $\mathcal{P}$ are not only distinguishable pairwise, but are also *observable* individually; that is, there exists a list (frame of reference) of the units in $\mathcal{P}$ and it is within the powers of the surveyor to pre-select any particular unit from the list and then observe its characteristics. Let us assume that the units in $\mathcal{P}$ are listed as

$$1, 2, 3, \ldots, N,$$

*Research supported in part by NSF Grant GP-9001.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_24, © Springer Science+Business Media, LLC 2011

where N is finite and is known to the surveyor. [We are thus excluding from our survey theory such populations as, for example, the insects of a particular species in a particular area or the set of all color-blind adult males in a particular country. Such populations as above can, of course, be the subject matter of a valid statistical inquiry but the absence of a sampling frame makes it impossible for such populations to be *surveyed* in the sense we understand the term survey here.]

(c) Corresponding to each unit $j \in \mathscr{P}$ there exists an unknown quantity Y_j in which the surveyor is interested. The unknown Y_j can be made known by observing (surveying) the unit j. The unknown state of nature is the vector quantity

$$\theta = (Y_1, Y_2, \ldots, Y_N).$$

However, the surveyor's primary interest is in some characteristic (parameter)

$$\tau = \tau(\theta)$$

of the state of nature θ. [Typically, the Y_j's are vector quantities themselves and the surveyor is seeking information about a multiplicity of τ's. However, for the sake of pinpointing our attention to the basic questions that are raised here, we restrict ourselves to the simple case where the Y_j's are scalar quantities (real numbers) and $\tau = \Sigma Y_j$.]

(d) The surveyor has prior knowledge K about the state of nature θ. This knowledge K is a multi-dimensional complex entity and is largely of a qualitative and speculative nature. We consider here the situation where K has at least the following two well-defined components. The surveyor *knows* the set Ω of all the *possible* values of the state of nature θ and, for each unit j, ($j = 1,2,\ldots,N$), he has access to a record of some known auxiliary characteristic A_j of j. [Typically, each A_j is a vector quantity. However, in our examples we shall take the A_j's to be real numbers.] The set Ω and the vector

$$\alpha = (A_1, A_2, \ldots, A_N)$$

are the principal measurable components of the surveyor's prior knowledge K. Let us denote the residual past of the knowledge by R and write

$$K = (\Omega, \alpha, R).$$

(e) The purpose of a survey is to gain further knowledge (beyond what we have described as K) about the state of nature θ and, therefore, about the parameter of interest $\tau = \tau(\theta)$. Since the surveyor is supposed to know the set Ω of all the possible values of θ, he knows the set $\mathscr{T}$ of all the possible values of τ. Initially, the surveyor's *ignorance* about τ is, therefore, *spread* over the set $\mathscr{T}$. [Later on, we shall quantify this initial *spread of ignorance* as a prior probability distribution.] In theory, the surveyor can dispel this ignorance and gain complete knowledge by making a total survey (complete enumeration) of $\mathscr{P}$. If he observes the Y-characteristic of every

unit j ($j = 1,2,\ldots,N$), then he knows the actual value of $\theta = (Y_1, Y_2, \ldots, Y_N)$ and, therefore, that of $\tau(\theta)$. We are, however, considering the case where a total survey is impracticable. By a *survey* of the population $\mathcal{P}$ we mean the selection of a (usually small) subset

$$u = (u_1, u_2, \ldots, u_n)$$

of units from $\mathcal{P}$ and then observing the corresponding Y-values

$$y = (Y_{u_1}, Y_{u_2}, \ldots, Y_{u_n})$$

of units in the subset u.

(f) We make the simplifying assumption that there are no *non-response* and *observation* errors; that is, the surveyor is able to observe every unit that is in the subset u, and when he observes a particular unit j, he finds the true value of the hitherto unknown Y_j without any error.

(g) The surveyor's blueprint for the survey is usually a very complicated affair. The survey plan must take care of myriads of details. However, in this article we idealize away most of these details and consider only two facets of the survey project, namely, the *sampling plan* and the *fieldwork*. The sampling plan is the part of the project that yields the subset u of $\mathcal{P}$ and fieldwork generates the observations y on members of u. The data (sample) generated by the survey is

$$x = (u, y) .$$

For reasons that will be made clear later, it is important to distinguish between the two parts u and y of the data x.

(h) Let $\mathcal{S}$ stand for the sampling plan of the surveyor. The plan (when set in motion) produces a subset u of the population $\mathcal{P} = (1,2,\ldots,N)$. We write $u = (u_1, u_2, \ldots, u_n)$, where $u_1 < u_2 < \ldots < u_n$ are members of $\mathcal{P}$. The fieldwork generates the vector $y = (Y_{u_1}, Y_{u_2}, \ldots, Y_{u_n})$ which we often write as $(y_1, y_2, \ldots, y_n)$. [Occasionally, we shall consider sampling plans that introduce a natural selection order among the units that are selected. For such plans it is more appropriate to think of u, not as a subset of $\mathcal{P}$, but as a finite sequence of elements $u_1, u_2, \ldots, u_n$ drawn from $\mathcal{P}$ in that selection order. In rare instances, the sampling plan may allow the possibility of a particular unit appearing repeatedly in the sequence $(u_1, u_2, \ldots, u_n)$. From the description of the sampling plan $\mathcal{S}$ it will usually be clear if we intend to treat u as a set or a sequence. In either case, we can think of u as a vector $(u_1, u_2, \ldots, u_n)$ and y as the corresponding observation vector $(y_1, y_2, \ldots, y_n)$ where $y_i = Y_{u_i}$.]

(i) *Summary*: Our idealized survey set-up consists of the following:

(i) A finite population $\mathcal{P}$ whose members are listed in a sampling frame as $1,2,\ldots,N$. Availability of each $j \in \mathcal{P}$ for observation.

(ii) The unknown state of nature $\theta = (Y_1, Y_2, \ldots, Y_N)$ and the parameter of interest $\tau = \tau(\theta)$.

(iii) The prior knowledge $K=(\Omega, \alpha, R)$.
(iv) Absence of non-response and observation errors.
(v) Choice of a sampling plan $\mathcal{S}$ as part of the survey design.
(vi) Putting the sampling plan $\mathcal{S}$ and the fieldwork into operation, thus arriving at the data (sample)

$$x=\{u=(u_1,u_2,...,u_n), \quad y=(y_1,y_2,...,y_n)\}$$

where $u_i \in \mathcal{P}$ and $y_i = Y_{u_i} (i=1,2,...,n)$.
(vii) Making a *proper* use of the data x in conjunction with the prior knowledge K to arrive at a *reasonable judgment* (or decision) related to the parameter τ.

The operational parts of the survey are its design (v), the actual survey (vi) and the data analysis (vii). In this article we are concerned only with the design and the analysis of a survey.

2
Probability in Survey Theory

We posed the survey set-up as a classic problem of inductive inference—a problem of inferring about the whole from observations on only a part. The basic questions are: Which part does one observe? Does the part (actually observed) tell us anything about the whole? and, then the main question, Exactly what does it tell? Let us now examine how probability enters into the picture.

There are three different ways in which probability theory finds its way into the mathematical theory of survey sampling. First, there is the time honored way through a probabilistic model for observation errors. Indeed, this is how probability theory first infiltrated the sacred domain of science. When we observe the Y-value Y_j of unit j, there is bound to be some observation error. In current survey theory we classify this kind of error as *non-sampling* error. In this article we have idealized away this kind of probability by assuming that there exists no observation error. We have deliberately taken this simplistic view of the survey set-up. The idea is to concentrate our attention on the other two sources of probability.

In current survey theory, the main source of probability is randomization, which is an artificial introduction (through the use of random number tables) of randomness in the sampling plan $\mathcal{S}$. Randomization makes it possible for the surveyor to consider the set (or sequence) u, and therefore the data $x=(u,y)$, as random elements. With an element of randomization incorporated in the sampling plan $\mathcal{S}$, the surveyor can consider the space U of all the possible values of the random element u and then the probability distribution p_θ of u over U. [For sampling plans usually discussed in survey textbooks, the probability distribution p_u of u is uniquely determined (by the plan) and is, therefore, independent of the state of nature θ. In part two of this article we shall take a broader view of the subject and also consider plans for which the probability distribution of u involves θ.] Now, let X be the space of all the possible values of the data (sample) $x=(u, y)$ of which

we have already recognized (thanks to randomization) the part u to be a random element. The space X is our sample space. Let P_u be the probability distribution of x over the sample space X. If $T = T(x)$ is an estimate of τ, then (prior to sampling and fieldwork) we can consider T to be a random variable and speculate about its sampling distribution and its average performance characteristics (as an estimator of τ) in an hypothetical sequence of repeated experimentations. This decision-theoretic approach is not possible unless we regard randomization as the source of probability in survey theory. From the point of view of a frequency-probabilist, there cannot be a statistical theory of surveys without some kind of randomization in the plan $\mathcal{S}$.

Apart from observation errors and randomization, the only other way that probability can sneak into the argument is through a mathematical formalization of what we have described before as the residual part R of the prior knowledge $K = (\Omega, \alpha, R)$. This is the way of a subjective (Bayesian) probabilist. The formalization of R as a prior probability distribution of θ over Ω makes sense only to those who interpret the probability of an event, not as the long range relative frequency of occurrence of the event (in an hypothetical sequence of repetitions of an experiment), but as a formal quantification of the illusive (but nevertheless very real) phenomenon of *personal belief* in the truth of the event. According to a Bayesian, probability is a mathematical theory of belief and it is with this kind of a probability theory that one should seek to develop the guidelines for inductive behavior in the presence of uncertainty. The purpose of this essay is not to examine the logical basis of Bayesian probability nor to describe how one may arrive at the actual qualification of R into a prior probability distribution of θ over Ω. [Of late, a great deal has been written on the subject. See, for instance, I. R. Savage's delightfully written new book, *Statistics: Uncertainty and Behavior.*]

Can the two kinds of probability co-exist in our survey theory? This is what we propose to find out.

3
Non-Sequential Sampling Plans and Unbiased Estimation

By a non-sequential sampling plan we mean a plan that involves no fieldwork. If the sampling plan $\mathcal{S}$ is non-sequential, then the surveyor can (in theory) make the selection of the set (or sequence) u of population units right in his office and then send his field investigators to the units selected in u and thus obtain the observation part y of the data $x = (u, y)$. A great majority of survey theoreticians have so far restricted themselves to non-sequential plans that involve an element of randomization in it. In this section we consider such plans only. The essence of non-sequentialness of a plan $\mathcal{S}$ is that the probability distribution of u does not involve the state of nature θ. Thus, the sampling plan where we continue to draw a unit at a time with equal probabilities and with replacements until we get ν distinct units is a non-sequential plan.

Given $u = (u_1, u_2, ..., u_n)$, the observation part $y = (y_1, y_2, ..., y_n)$, where $y_i = Y_{u_i} (i = 1, 2, \ldots, n)$, is obtained through the fieldwork and is uniquely determined by the state of nature $\theta = (Y_1, Y_2, \ldots, Y_N)$. The conditional

probability distribution of y given u is degenerate, the point of degeneration depending on θ. That is, for all y, u and θ

$$\text{Prob}(y \mid u, \theta) = 0 \text{ or } 1. \tag{3.1}$$

[We are taking the liberty of using the symbols u, y, x and θ both as variables and as particular values of the variables.]

For each sampling plan $\mathcal{S}$ we have the space U of all the possible values of u. The probability distribution p of u over U is θ-free, that is, is uniquely defined by the plan $\mathcal{S}$. The probability distribution p is clearly discrete. There is no loss of generality in assuming that $p(u) > 0$ for all $u \in U$. If the non-sequential plan is *purposive* (that is, the plan involves no randomization) then U is a single-point set and the distribution of u is degenerate at that point.

Let X be the sample space, the set of all possible samples (data) $x = (u, y)$ where u is generated by the plan $\mathcal{S}$ and y by the fieldwork. For each $\theta \in \Omega$, we have a probability distribution P_θ over X. Whatever the plan $\mathcal{S}$, the probability distribution P_θ is necessarily discrete. We write $P_\theta(x)$ or $P_\theta(u, y)$ for the probability of arriving at the data $x = (u, y)$ when θ is the true value of the state of nature. Clearly,

$$P_\theta(x) = P_\theta(u, y) = p(u) \text{ Prob } (y \mid u, \theta). \tag{3.2}$$

The surveyor takes a peep at the unknown $\theta = (Y_1, Y_2, \ldots, Y_N)$ through the sample $x = (u, y)$. Prior to the survey, the surveyor's ignorance about θ was spread over the space Ω. Once the data x is at hand, the surveyor has exact information about some coordinates of the vector θ. These are the coordinates that correspond to the distinct units that are in u. The data x rules out some points in Ω as clearly inadmissible. Let Ω_x be the subset of values of θ that are consistent with the data x. In other words, $\theta \in \Omega_x$ if $P_\theta(x) > 0$; that is, it is possible to arrive at the data x when θ is the true value of the state of nature. The subset Ω_x of Ω is well-defined for every $x \in X$. Without any loss of generality we may assume that no Ω_x is vacuous. From (3.1) and (3.2) it follows that the likelihood function $L(\theta)$ is given by the formula

$$L(\theta) = P_\theta(x) = \begin{cases} p(u) & \text{for all } \theta \in \Omega_x \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

In other words, whatever the data x, the likelihood function $L(\theta)$ is flat (a positive constant) over the set Ω_x and is zero outside Ω_x. This remark holds true for sequential plans also (Basu, 1969). The importance of the remark will be made clear later on.

A major part of survey theory is concerned with unbiased estimation. A statistic is a characteristic of the sample x. An estimator $T = T(x)$ is a statistic that is well-defined for all $x \in X$ and is used for estimating a parameter $\tau = \tau(\theta)$. By an unbiased estimator of $\tau(\theta)$ we mean an estimator T that satisfies the identity

$$E(T \mid \theta) = \sum_x T(x) P_\theta(x) \equiv \tau(\theta), \text{ for all } \theta \in \Omega. \tag{3.4}$$

Let $w(t, \theta)$ be the loss function. That is, $w(t, \theta)$ stands for the surveyor's assessment of the magnitude of error that he commits when he estimates the parameter $\tau = \tau(\theta)$ by the number t. We assume that

$$w(t, \theta) \geq 0 \text{ for all } t \text{ and } \theta,$$

the sign of equality holding only when $t = \tau(\theta)$. The risk function $r_T(\theta)$ associated with the loss function w and the estimator T is then defined as the expected loss

$$r_T(\theta) = E[w(T, \theta) \mid \theta] = \sum_x w(T(x), \theta) P_\theta(x). \tag{3.5}$$

[If the reader is not familiar with the decision-theoretic jargons of loss and risk, he may restrict himself to the particular case where $w(t, \theta)$ is the squared error $(t - \tau(\theta))^2$ and the risk function $r_T(\theta)$ is the variance $V(T \mid \theta)$ of the unbiased estimator T.] The following theorem proves the non-existence of a uniformly minimum risk (variance) unbiased estimator of τ.

Theorem. Given an unbiased estimator T of τ and an arbitrary (but fixed) point $\theta_0 \in \Omega$, we can always find an unbiased estimator T_0 (of τ) such that $r_{T_0}(\theta_0) = 0$, that is, T_0 has zero risk at θ_0.

Proof: We find it convenient to write $T(u, y)$ and $P_\theta(u, y)$ for $T(x)$ and $P_\theta(x)$ respectively. It has been noted earlier that the conditional distribution of y given u is degenerate at a point that depends on θ. Let $y_0 = y_0(u)$ be the point of degeneration of y, for given u, when $\theta = \theta_0$. Consider the statistic

$$T_0 = T(u, y) - T(u, y_0) + \tau(\theta_0). \tag{3.6}$$

The statistic $T(u, y_0)$ is a function of u alone and so its probability distribution, and therefore its expectation are θ-free. Indeed,

$$\begin{aligned} E[T(u, y_0)] &= \sum_u T(u, y_0) p(u) \\ &= \sum_{u,y} T(u, y) P_{\theta_0}(u, y) \\ &= E[T(u, y) \mid \theta_0] \\ &= \tau(\theta_0). \end{aligned}$$

Thus, the statistic T_0 as defined in (3.6) is an unbiased estimator of τ. Now, when $\theta = \theta_0$, the statistics $T(u, y)$ and $T(u, y_0)$ are equal with probability one, and so $T_0 = \tau(\theta_0)$ with probability one. This proves the assertion that $r_{T_0}(\theta_0) = 0$.

The impossibility of the existence of a uniformly minimum risk unbiased estimator follows at once. For, if such an estimator T exists then $r_T(\theta)$ must be zero for all $\theta \in \Omega$. That is, whatever the value of the state of nature θ it should be possible to estimate $\tau(\theta)$ without any error (loss) at all. Unless the sampling plan $\mathcal{S}$ is equivalent to a total survey of the population, such a T clearly cannot exist for a parameter τ that depends on all the coordinates of θ. The following two examples will clarify the theorem further.

Example 1. Consider the case of a simple random sample of size one from the population $\mathcal{P} = (1, 2, ..., N)$. The sample is (u, y) where u has a uniform probability distribution over the N integers 1, 2, ..., N and $y = Y_u$. Let the population mean

$$\overline{Y} = \frac{1}{N}(Y_1 + ... + Y_N)$$

be the parameter to be estimated. Clearly, y is an unbiased estimator of $\overline{Y}$. Let $\theta_0 = (a_1, a_2, \ldots, a_N)$ and $\bar{a} = (\Sigma a_j)/N$. The statistic

$$T_0 = y - a_u + \bar{a} \tag{3.7}$$

is an unbiased estimator of $\overline{Y}$ with zero risk (variance) when $\theta = \theta_0$. The variance of y is $\Sigma(Y_j - \overline{Y})^2/N$ and that of T_0 is $\Sigma(Z_j - \overline{Z})^2/N$ where $Z_j = Y_j - a_j$ $(j = 1, 2, \ldots, N)$.

Example 2. Let $\mathcal{S}$ be an arbitrary non-sequential sampling plan that allots a positive selection probability to each population unit. That is, the probability Π_j that the unit j appears in the set (or sequence) u is positive for each $j (j = 1, 2, \ldots, N)$. Since $\mathcal{S}$ is non-sequential, the vector

$$\Pi = (\Pi_1, \Pi_2, \ldots, \Pi_N)$$

is θ-free. Let Y be the population total ΣY_j. A particular unbiased estimator of Y that has lately attracted a great deal of attention is the so-called Horvitz-Thompson (HT) estimator (relative to the plan $\mathcal{S}$). The HT-estimator is defined as follows. Let $u_1 < u_2 < \ldots < u_\nu$ be the distinct population units that appear in u and let $\hat{y} = (y_1, y_2, \ldots, y_\nu)$ be the corresponding observation vector. Let $p_i = \Pi u_i$ $(i = 1, 2, \ldots, \nu)$. The HT-estimator H is then defined as

$$H = \frac{y_1}{p_1} + \ldots + \frac{y_\nu}{p_\nu}. \tag{3.8}$$

That H is an unbiased estimator of Y will be clear when we rewrite (3.8) in a different form. Let $E_j (j = 1, 2, \ldots, N)$ stand for the event that the plan $\mathcal{S}$ selects unit j, and let I_j be the indicator of the event E_j. That is, $I_j = 1$ or 0 according as unit j appears in u or not. It is now easy to check that

$$H = \sum_{j=1}^{N} \Pi_j^{-1} I_j Y_j. \tag{3.9}$$

That H is an unbiased estimator of Y follows at once from the fact that

$$\begin{aligned} E(I_j) &= \text{Prob}\,(E_j) \\ &= \Pi_j \quad (j = 1, 2, \ldots, N). \end{aligned}$$

Now, let $\theta_0=(a_1, a_2, \ldots, a_N)$ be a point in Ω that is selected by the surveyor (prior to the survey) and let H_0 be defined as

$$H_0=H-\Sigma\,\Pi_j^{-1} I_j a_j+\Sigma a_j. \tag{3.10}$$

It is now clear that H_0 is an unbiased estimator of Y and that $V(H_0 \mid \theta)=0$ when $\theta=\theta_0$. Since the variances of H and H_0 are continuous functions of θ, it follows that

$$V(H_0 \mid \theta) < V(H \mid \theta) \tag{3.11}$$

for all θ in a certain neighborhood Ω_0 of the point θ_0. If the surveyor has the prior knowledge that the true value of θ lies in Ω_0, then the modified Horvitz-Thompson estimator H_0 is uniformly better than H. The estimator H_0 will look a little more reasonable if we rewrite it as

$$H_0=[\Sigma\,\Pi_j^{-1} I_j(Y_j-a_j)]+\Sigma a_j. \tag{3.12}$$

If (3.9) is a reasonable estimator of $Y=\Sigma Y_j$, then the variable part of the right hand side of (3.12) is an equally reasonable estimator of

$$\Sigma(Y_j-a_j)=Y-\Sigma a_j.$$

The strategy of a surveyor who advocates the use of (3.12) [in preference to that of (3.9)] as an estimator of $Y=\Sigma Y_j$ is quite clear. Instead of defining the state of nature as

$$\theta=(Y_1, Y_2, \ldots, Y_N)$$

he is defining it as

$$\theta'=(Y_1-a_1, Y_2-a_2, \ldots, Y_N-a_N).$$

Suppose the surveyor has enough prior information about the state of nature, so that by a proper choice of the vector $(a_1, a_2, \ldots, a_N)$ he can make the coordinates of θ' much less variable than that of θ. He is then in a better position to estimate the total of the coordinates of θ' than one who is working with θ. Consider the situation where the surveyor knows in advance that the j^{th} coordinate Y_j of θ lies in a small interval around the number a_j ($j=1, 2, \ldots, N$). In such a situation the surveyor ought to shift the origin of measurement (for θ) to the point $(a_1, a_2, \ldots, a_N)$ and represent the state of nature as

$$\theta'=(Y_1-a_1, Y_2-a_2, \ldots, Y_N-a_N).$$

If the numbers $a_1, a_2, \ldots, a_N$ has a large dispersion, then shifting the origin of measurement to $(a_1, a_2, \ldots, a_N)$ will cut down the variability in the coordinates of the state of nature to a large extent. The effect will be similar to what is usually achieved by stratification.

At this point one may very well raise the questions: Why must the surveyor choose his knowledge vector $(a_1, a_2, \ldots, a_N)$ before the survey?

Is it not more reasonable for him to wait until he has the survey data at hand and then take advantage of the additional knowledge gained thereby? Once the data is at hand, the surveyor knows the exact values of the surveyed coordinates of θ. The natural post-survey choice of a_j for any surveyed j is, therefore, Y_j. For a non-surveyed j, the surveyor's best estimate a_j of the unknown Y_j would still be of a speculative nature. If in formula (3.12) we allow the surveyor to insert a post-survey specification of the vector $(a_1, a_2, \ldots, a_N)$, then the first part of the right hand side of (3.12) will vanish and the estimator will look like

$$\begin{aligned} H_* &= \Sigma\, a_j \\ &= (Y_{u_1} + Y_{u_2} + \ldots + Y_{u_\nu}) + \sum_{j \notin u} a_j \qquad (3.13) \\ &= S + S^*, \end{aligned}$$

where S is the sum total of the Y-values of the distinct surveyed units and S^* is the surveyor's post-survey estimate of the total Y-values of the non-surveyed units.

A decision-theorist will surely object to our derivation of formula (3.13) as naive and incompetent. He will point out that we have violated a sacred cannon of inductive behavior, namely *never select the decision rule after looking at the data*. He will also point out that S^* in (3.13) is, as yet, not well-defined (as a function on the sample space X), and he will reject H_* (as an estimator of Y) with the final remark that the whole thing stinks of Bayesianism!

Nevertheless, the fact remains that formula (3.13) points to the very heart of the matter of estimating the population total Y. A survey leads to a complete specification of a part of the population total. This part is the sample total S as defined in (3.13). At the end of the survey the remainder part Y-S is still unknown to the surveyor. If the surveyor insists on putting down T as an estimate of Y, then he is in effect saying that he has reason to believe that T-S is close to Y-S. And then he should give a reasonable justification for his belief. Of course we can write any estimate T in the form

$$T = S + S^*$$

where S is the sample total and $S^* = T - S$. But then, for some T, the part S^* (of T) would appear quite preposterous as an estimate of the unknown part $Y - S$ of Y. The following two examples will make clear the point that we are driving at.

Example 3. The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the

elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take 50 y (where y is the present weight of Sambo) as an estimate of the total weight $Y=Y_1+\ldots+Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive samplings plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why? The estimate ought to be $50y$ of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is $100y/99$ and not $50y$." "And, how would you have estimated Y," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!)

Example 4. Sampling with unequal probabilities has been recommended in situations that are less frivolous than the one considered in the previous example but the recommended unbiased estimators for such plans sometimes look hardly less ridiculous than the one just considered. Let us consider the so-called pps (probability proportional to size) plans about which so many research papers have been written in the past 20 years. A pps sampling plan is usually recommended in the following kind of situation. Suppose for each population unit j we have a record of an auxiliary characteristic A_j (the size of j). Also suppose that each A_j is a positive number and that the surveyor has good reason to believe that the ratios

$$\Lambda_j = Y_j/A_j \quad (j=1,2,\ldots,N) \tag{3.14}$$

are nearly equal to each other. In this situation it is often recommended that the surveyor adopts the following without replacement pps.

Sampling plan. Let $A=\Sigma A_j$ and $P_j=A_j/A$ $(j=1,2,\ldots,N)$. Choose a unit (say, u_1) from the population $\mathcal{P}=(1,2,\ldots,N)$ following a plan that allots a selection probability P_j to unit j $(j=1,2,\ldots,N)$. The selected unit u_1 is then removed from the sampling frame and a second unit (say, u_2) is selected

from the remaining $N-1$ units with probabilities proportional to their sizes (the auxiliary characters A_j). This process is repeated n times so that the surveyor ends up with n distinct units

$$u_1, u_2, \ldots, u_n$$

listed in their natural selection order. After the fieldwork the surveyor has the sample

$$x = \{(u_1, y_1), \ldots, (u_n, y_n)\}$$

where $y_i = Y_{u_i} (i = 1, 2, \ldots, n)$. Let us write p_i for $P_{u_i} (i = 1, 2, \ldots, n)$ and

$$x^* = \{(u_1, y_1), \ldots, (u_{n-1}, y_{n-1})\}$$

for the vector defined by the first $n-1$ coordinates of x. It is then easy to see that (see Desraj [6] Theorem 3.13)

$$E\left(\frac{y_n}{p_n} \middle| x^*\right) = \Sigma' \frac{Y_j}{P_j} \cdot \frac{P_j}{1-p_1-p_2-\ldots-p_{n-1}}$$

$$= (\Sigma' Y_j)/(1-p_1-\ldots-p_{n-1}),$$

where the summation is carried over all j that are different from $u_1, u_2, \ldots, u_{n-1}$. Since $\Sigma' Y_j = Y-(y_1+\ldots+y_{n-1})$ it follows at once that

$$E\left(y_1+\ldots+y_{n-1}+\frac{y_n}{p_n}(1-p_1-\ldots-p_{n-1}) \mid x^*\right) = Y. \tag{3.15}$$

Therefore, the unconditional expectation of the lefthand side of (3.15) is also Y. And so we have the so-called Desraj estimator

$$D = y_1+\ldots+y_{n-1}+\frac{y_n}{p_n}(1-p_1-\ldots-p_{n-1}), \tag{3.16}$$

which is an unbiased estimator of Y. Writing S for the sample total $y_1+\ldots+y_n$ we can rewrite (3.16) as

$$D = S + S^* \tag{3.17}$$

where

$$S^* = \frac{y_n}{p_n}(1-p_1-\ldots-p_n).$$

Let us examine the face-validity of S^* as an estimate of Y^*, the total Y-values of the unobserved population units.
Writing $A = \Sigma A_j$, $a_i = A_{u_i}$ $(i = 1, 2, \ldots, n)$ and $A^* = A-a_1-\ldots-a_n$ (the total A-value of the unobserved units), we have

$$S^* = \frac{y_n}{p_n}(1-p_1-\ldots-p_n) \tag{3.18}$$

$$= \frac{y_n}{a_n} A^* \quad (\text{since } p_i = \frac{a_i}{A}).$$

Clearly, S^* would be an exact estimate of Y^* if and only if

$$\frac{y_n}{a_n}=\frac{Y^*}{A^*}=\frac{\Sigma' Y_j}{\Sigma' A_j} \tag{3.19}$$

(the summation is over the unobserved j's).

Now, if the surveyor claims that according to his belief (3.17) is a good estimate of Y, then that claim is equivalent to an assertion of belief in the near equality of the two ratios

$$\frac{y_n}{a_n} \text{ and } \frac{Y^*}{A^*}.$$

What can be the logical basis for such a belief? We started with the assumption that the surveyor has prior knowledge of near equality in the N ratios in (3.14). At the end of the survey, the surveyor has observed exactly n of these ratios and they are

$$\frac{y_1}{a_1}, \frac{y_2}{a_2}, \ldots, \frac{y_n}{a_n}. \tag{3.20}$$

The surveyor is now in a position to check on his initial supposition that the ratios in (3.14) are nearly equal. Suppose he finds that the observed ratios in (3.20) are indeed nearly equal to each other. This will certainly add to the surveyor's conviction that the unobserved ratios Λ_j (where j is different from $u_1, u_2, \ldots, u_n$) are nearly equal to each other and that they lie within the range of variations of the observed ratios in (3.20). Now, Y^*/A^* is nothing but a weighted average of the unobserved ratios (the weights being the sizes of the corresponding units). It is then natural for the surveyor to estimate Y^*/A^* by some sort of an average of the observed ratios. For instance, he may choose to estimate Y^*/A^* by $(y_1+\ldots+y_n)/(a_1+\ldots+a_n)$. This would lead to the following modification of the Desraj estimate (3.17):

$$D_1=S+\frac{y_1+\ldots.+y_n}{a_1+\ldots.+a_n}A^* \tag{3.21}$$

$$=\frac{y_1+\ldots.+y_n}{a_1+\ldots.+a_n}A$$

(and this we recognize at once as the familiar ratio estimate). Alternatively, the surveyor may choose to estimate the ratio Y^*/A^* by the simple average

$$\frac{1}{n}\left(\frac{y_1}{a_1}+\ldots.+\frac{y_n}{a_n}\right)$$

of the observed ratios. This will lead to another variation of the Desraj estimate, namely

$$D_2=(y_1+\ldots.+y_n)+\frac{1}{n}\left(\frac{y_1}{a_1}+\ldots.+\frac{y_n}{a_n}\right)A^*. \tag{3.22}$$

What we are trying to say here is the simple fact that both (3.21) and (3.22) have much greater face validity as estimates of Y than the Desraj estimate (3.17). In the Desraj estimate we are trying to evaluate Y^*/A^* by the n^{th} observed ratio $y_n \mid a_n$ and are taking no account of the other $n-1$ ratios. This is almost as preposterous as the estimate suggested by the circus statistician in the previous example. Suppose the surveyor finds that the n observed ratios $y_i \mid a_i$ $(i = 1, 2, \ldots, n)$ are nearly equal alright, but $y_n \mid a_n$ is the largest of them all. In this situation how can he have any faith in the Desraj estimate

$$D = S + \frac{y_n}{a_n} A^*$$

being nearly equal to Y? [Remember, the factor A^* will usually be a very large number.] Again, what does the surveyor do when he discovers that his initial supposition that the ratios Y_j/A_j $(j = 1, 2, \ldots, N)$ are nearly equal, was way off the mark? Will it not be ridiculous to use the Desraj estimate in this case? Here we are concerned not with the mathematical property of unbiasedness of an estimator but with the hard-to-define property of face validity of an estimate. An estimate T of the population total Y has little face validity if after we have written T in the form

$$T = S + S^*$$

we are hard put to find a reason why the part S^* should be a good estimate of Y^*.

4
The Label-Set and The Sample Core

We have noted elsewhere that, for a non-sequential sampling plan $\mathcal{P}$, the label part u of the data $x = (u, y)$ is an ancillary statistic; that is, the sampling distribution of the statistic u does not involve the state of nature θ. The sampling distribution of u is uniquely determined by the plan. It is therefore obvious that the label part of the data cannot, by itself, provide any information about θ. Knowing u, we only know the names (labels) of the population units that are selected for observation. [When u is a sequence, we also know the order and the frequency of appearance of each selected unit in u.] With a non-sequential plan $\mathcal{P}$, the knowledge of u alone cannot make the surveyor any wiser about θ. The surveyor may, and often does, incorporate his prior knowledge of the auxiliary characters $\alpha = (A_1, A_2, \ldots, A_N)$ in the plan $\mathcal{P}$. But this does not alter the situation a bit. The label part u of the data x will still be an ancillary statistic.

If the label part u is informationless, then can it be true that the observation part y of the data $x = (u, y)$ contains all the available information about θ? A little reflection will make it abundantly clear that the answer must be an emphatic, no. A great deal of information will be lost if the label part of the data is suppressed. Without the knowledge of u, the surveyor cannot relate the components of the observation vector y to the population units and so

he cannot make any use of the auxiliary characters $\alpha = (A_1, A_2, \ldots, A_N)$ and whatever other prior knowledge he may have about the relationship between θ and α.

Let us call a statistic $T = T(u, y)$ *label-free* if T is a function of y alone. So far, the only label-free estimator that we have come across is the estimator y of $\bar{Y}$ in Example 1. If in this case the surveyor has prior knowledge that the true value of θ lies in the vicinity of the point $\theta_0 = (a_1, a_2, \ldots, a_N)$, then he would naturally prefer the estimator (3.7) as an unbiased estimator of $\bar{Y}$. The surveyor can arrive at an estimate like (3.7) only if he has access to the information contained in u. In survey literature, we find several attempts at justifying label-free estimates. But a reasonable case for a label-free estimate can be made only under the assumption of a near complete ignorance in the mind of the surveyor. But, in these days of extreme specialization, who is going to entrust an expensive survey operation in the hands of a very ignorant surveyor?! To remain in survey business, the surveyor has to carefully orient himself to each particular survey situation, gather a lot of auxiliary data $A_1, A_2, \ldots, A_N$ about the population units, and then make intelligent use of such data in the planning of the survey and in the analysis of the survey data. Considerations of label-free estimates are, therefore, of only an academic interest in survey theory.

Let us denote by $\hat{u}$ the set of distinct population units that are selected (for survey) by the sampling plan. The set $\hat{u}$ is a statistic—a characteristic of the sample $x = (u, y)$. We call $\hat{u}$ the label-set and find it convenient to think of $\hat{u}$ as a vector

$$\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_\nu),$$

where $\hat{u}_1 < \hat{u}_2 < \ldots < \hat{u}_\nu$ are the ν distinct unit-labels that appear in u, arranged in an increasing order of their label values. The *observation-vector* $\hat{y}$ is then defined as

$$\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_\nu),$$

where $\hat{y}_i = Y_{\hat{u}_i}$ $(i = 1, 2, \ldots, \nu)$.

We denote the pair $(\hat{u}, \hat{y})$ by $\hat{x}$ and call it the *sample-core*. For each sample $x = (u, y)$ we have a well-defined sample-core $\hat{x} = (\hat{u}, \hat{y})$. In the literature the sample-core has been called by other fancy names like *order statistic* or *sampley*, etc. [It should be noted that, though we can think of $\hat{u}$ as a subset of $\mathcal{P}$, we cannot think of $\hat{y}$ as a set, because the values in $\hat{y}$ need not be all different. Even if it were possible to think of $\hat{y}$ as a set, it would not be fruitful to do so. For, if $\hat{u}$ and $\hat{y}$ are both conceived as sets, then we have no way to relate a member of $\hat{y}$ to the corresponding label in $\hat{u}$. This is the reason why we prefer to think of the label-set $\hat{u}$ as a vector and of $\hat{y}$ as the corresponding observation-vector.]

The sample core $\hat{x}$ is a statistic. The mapping $x \to \hat{x}$ is usually many-one. For instance, in the pps plan of Example 4, the number ν (of distinct units selected) is the same as n, but, for each value of the label-set $\hat{u}$, there are exactly $n!$ values of u (corresponding to the $n!$ different selection-orders in which the n units might have been selected). Here the mapping $x \to \hat{x}$ is $n!$ to 1.

In part two of this essay we shall establish the fact that the sample-core $\hat{x}$ is a sufficient statistic. This means that, given $\hat{x}$, the conditional distribution of the sample x is uniquely determined (does not involve the unobserved part of the state of nature θ). The widely accepted principle of sufficiency tells us that if T be a sufficient statistic then every reasonable estimator (of every parameter τ) ought to be a function of T. Following Fisher we may call an estimator H *insufficient* if H is not a function of the sufficient statistic T. The Desraj estimator (3.17) of the population total Y is then an insufficient estimate. If we rewrite the Desraj estimate as

$$D = S + \frac{y_n}{a_n} A^*$$

where S is the sample Y- total and A^* is the total A-values of the unobserved units, then it is clear that both S and A^* are functions of the sample-core $\hat{x} = (\hat{u}, \hat{y})$. [Indeed, S is a function of $\hat{y}$ and A^* is a function of $\hat{u}$.] However, $y_n|a_n$ (the ratio corresponding to the last unit drawn in the without replacement pps plan) is not a function of $\hat{x}$. Knowing $\hat{x}$, we only know that the ratio $y_n|a_n$ may have been any one of the n ratios

$$\hat{y}_i|\hat{a}_i \quad (i = 1, 2, \ldots, n)$$

where $\hat{y}_i = Y_{\hat{u}_i}$ and $a_i = A_{\hat{u}_i}$.

If we define λ_i $(i = 1, 2, \ldots, n)$ as the conditional probability of $u_n|a_n$ being equal to $\hat{y}_i|\hat{a}_i$, then the λ_i's are well-defined (θ-free) constants, $\Sigma\lambda_i = 1$ and

$$\bar{D} = E(D \mid \hat{x}) = S + \left[\sum_1^n \lambda_i \frac{\hat{y}_i}{\hat{a}_i}\right] A^*. \tag{4.1}$$

Since D is an unbiased estimator of $Y = \Sigma Y_j$, so also is the estimator $\bar{D}$. From the Rao-Blackwell theorem it follows that (if $n > 1$) the variance of $\bar{D}$ is uniformly smaller than that of D. The estimator $\bar{D}$ has been variously called in the literature, the *symmetrized* or the *un-ordered* Desraj estimator. In view of what we explained in the previous section, the symmetrized Desraj estimator (4.1) looks much better than the original Desraj estimator on the score of face-validity. However, the coefficients $\lambda_1, \lambda_2, \ldots, \lambda_n$ in (4.1) are much too complicated to make $\bar{D}$ an acceptable estimator of Y. The estimates (3.21) and (3.22) have about the same face-validity as that of (4.1) and are much simpler to compute. However, (4.1) scores over the other two estimates on the dubious criterion of unbiasedness!

The estimator (4.1) cannot be the only unbiased estimator of Y that is a function of the sample core $\hat{x}$. Consider the estimator

$$\frac{y_1}{p_1} = \frac{y_1}{a_1} A\,, \tag{4.2}$$

where y_1 is the Y-value of the first unit that was drawn (by the pps plan of Example 4) and a_1 is the corresponding A-value. Clearly, (4.2) is an *insufficient* unbiased estimator of Y. The symmetrized version of (4.2) will be

$$\left(\Sigma\mu_i \frac{\hat{y}_i}{\hat{a}_i}\right) A, \tag{4.3}$$

where

$$\mu_i = P\left(\frac{y_1}{a_1} = \frac{\hat{y}_i}{\hat{a}_i} \,\middle|\, \hat{x}\right) \quad (i = 1, 2, \ldots, n).$$

The estimator (4.3) is unbiased and is a function of $\hat{x}$.

Of late, quite a few papers have been written in which the main idea is the above described method of *un-ordering an ordered estimate*, that is, making use of the Rao-Blackwell theorem and the sufficiency of the sample core $\hat{x}$. Whatever the sampling plan $\mathcal{J}$ is, the sample-core is always sufficient. Indeed, the sample-core is (in general) the minimum (minimal) sufficient statistic. However, for a non-sequential sampling plan $\mathcal{J}$, the sufficient statistic $\hat{x}$ is never *complete*. By the *incompleteness* of $\hat{x}$ we mean the existence of non-trivial functions of $\hat{x}$ whose expectations are identically zero for all possible values of the state of nature θ. This is because (when the plan is non-sequential) the label-set $\hat{u}$ (which is a component of $\hat{x}$) is an ancillary statistic. For every parameter of interest $\tau(\theta)$, there will exist an infinity of unbiased estimators each of which is sufficient in the sense of Fisher (that is, is a function of the minimal sufficient statistic $\hat{x}$.)

5
Linear Estimation in Survey Sampling

During the past several years a great many research papers have been written dealing exclusively with the topic of linear estimation of the population mean $\overline{Y}$ or, equivalently, the population total Y. Some confusion has, however, been created by the term *linear*. An estimator is a function on the sample space X. Unless X is a linear space we cannot, therefore, talk of a linear estimator. In our formulation, X is the space of all samples $x = (u, y)$ and so X is not a linear space. How then are we to reconcile ourselves to the classical statement that, in the case of a simple random sampling plan, the sample mean is the best unbiased linear estimate of the population mean? We have the often quoted contrary assertion from Godambe that in no realistic sampling situation (whatever the plan $\mathcal{J}$) can there exist a best estimator in the class of linear unbiased estimators of the population mean. This section is devoted entirely to the notions of the so-called linear estimates.

Consider first the case of a simple random sampling plan in which a number n (the sample size) is chosen in advance and then a subset of n units is selected from the population $\mathcal{P}$ in such a manner that all the $\binom{N}{n}$ subsets of $\mathcal{P}$ with n elements are alloted equal selection probabilities. Let us suppose that the plan calls for a selection of the n sample units one by one without replacements and with equal probabilities, so that we can list the selected units in their natural selection order as $u_1, u_2, \ldots, u_n$. The label part of the data is then the sequence $u = (u_1, u_2, \ldots, u_n)$ and the observation part is the corresponding observation vector $y = (y_1, y_2, \ldots, y_n)$. Clearly, the y_i's are identically distributed (though not mutually independent) random variables with

$$E(y_i) = (\Sigma Y_j)/N = \overline{Y} \quad (i = 1, 2, \ldots, n).$$

Now, if the surveyor chooses to ignore the label part u of the data, then he can define a linear estimator of $\bar{Y}$ as a linear function

$$T = b_0 + b_1 y_1 + b_2 y_2 + \ldots + b_n y_n \tag{5.1}$$

of the observation vector y, where the coefficients $b_0, b_1, \ldots, b_n$ are preselected constants. All estimators of the above kind are label-free estimators. Let L be the class of all unbiased estimators of $\bar{Y}$ that are of the type (5.1). In other words, L is the class of all estimators of the type (5.1) with

$$b_0 = 0 \text{ and } b_1 + \ldots + b_n = 1 . \tag{5.2}$$

The sample mean $\bar{y} = (\Sigma y_i)/n$ is a member of L. The classical assertion that we referred to before is to the effect that, in the class L, there exists a uniformly minimum variance estimator and that is the sample mean $\bar{y}$. This result is well-known and a fairly straightforward proof may be given for the particular case where we define variance as the mean square deviation from the mean. We, however, consider it appropriate to sketch a proof that ties in well with the general spirit of this article.

Consider the sample core $\hat{x} = (\hat{u}, \hat{y})$ where we write the label-set $\hat{u}$ as a sequence $(\hat{u}_1, \ldots, \hat{u}_n)$ with $\hat{u}_1 < \hat{u}_2 < \ldots < \hat{u}_n$ and look upon $\hat{y}$ as the corresponding observation vector $(\hat{y}_1, \ldots, \hat{y}_n)$. Note that the mapping $x \to \hat{x}$ is $n!$ to 1 and that the vector $\hat{y}$ is obtained from the vector y by rearranging its coordinates in an increasing order of their corresponding unit labels. Now, given $\hat{x}$, the conditional distribution of x is equally distributed over the $n!$ possible values of x and so it follows that

$$E(y_i | \hat{x}) = (\Sigma \hat{y}_i)/n = \bar{y} \quad (i = 1, 2, \ldots, n) . \tag{5.3}$$

Thus, if $T = \Sigma a_i y_i$, with $\Sigma a_i = 1$, is any member of L then from (5.3) it follows that

$$E(T | \hat{x}) = \Sigma (a_i \bar{y}) = \bar{y}. \tag{5.4}$$

And so from the Rao-Blackwell theorem it follows that $\bar{y}$ is better than T [and this is irrespective of the loss function $w(t, \theta)$ (see §3) as long as $w(t, \theta)$ is convex (from below) in t for each fixed value of θ]. Observe that, in the class L, the sample mean

$$\bar{y} = (\Sigma y_i)/n = (\Sigma \hat{y}_i)/n$$

is the only one that is a function of the sample core $\hat{x}$. Every other member of L is insufficient in the sense explained in the earlier section. And so it is no wonder that $\bar{y}$ beats every other member of L in its performance characteristics. The class $L - \{\bar{y}\}$ is certainly not worth any consideration at all.

At this stage one may ask: Why not consider the class of all linear functions of the vector $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$? The snag is that the variables $\hat{y}_1, \ldots, \hat{y}_n$ have very complicated distributions and their expectations are not easy to obtain. For instance, the variable $\hat{y}_1$ can take only the values $Y_1, Y_2, \ldots, Y_{N-n+1}$ and its expectation is a complicated linear function of these $N - n + 1$

values. It is, therefore, not easy to characterize the class of unbiased estimators of $\overline{Y}$ that are linear functions of the vector $\hat{y}$. In any case, our representation of $\hat{y}$ as a vector is a rather artificial one and it is difficult to see why we should consider linear functions of the vector $\hat{y}$.

Let us look at the problem from another angle. True, the sample space X is not linear, but the parameter space Ω of all the possible values of the state of nature $\theta=(Y_1, \ldots, Y_N)$ is a part of the N-dimensional linear space R_N. A linear function on Ω is a function of the type

$$B_0+B_1 Y_1+\ldots+B_N Y_N \tag{5.5}$$

where $B_0, B_1, \ldots, B_N$ are constants. But (5.5) is a linear function of the parameter θ and cannot be conceived of as a statistic. Consider, however, a modification of (5.5) where we replace the coefficient B_j by the variable $B_j I_j$ where I_j is the indicator of the event E_j that the unit j is selected by the sampling plan $\mathcal{S}(j=1,2,\ldots,N)$. For each set of coefficients $B_0, B_1, \ldots, B_N$ we then have a sort of a linear function [see formula (3.9)]

$$T=B_0+\sum_1^N B_j I_j Y_j \tag{5.6}$$

on Ω, where the coefficients $B_j i_j (j=1,2,\ldots,N)$ are random variables. The indicator I_j is a function of the label-set $\hat{u}$ [$I_j(\hat{u})=1$ or 0 according as j is a member of $\hat{u}$ or not]. It is easy to recognize T as a statistic—indeed as a function of the sample core $\hat{x}=(\hat{u}, \hat{y})$. Only observe that we may rewrite (5.6) as

$$T=B_0+\sum_1^n b_i \hat{y}_i \tag{5.7}$$

where $b_i=B_{\hat{u}_i}$ $(i=1, 2, \ldots, n)$. Let us repeat once again that T is not a linear function on the sample space X, but that we may stretch our imagination a little bit to conceive of T as a random linear function on Ω with coefficients that are determined by the label-set $\hat{u}$. If T is defined as in (5.6) then

$$E(T)=B_0+\Sigma B_j \Pi_j Y_j \tag{5.8}$$

where $\Pi_j=E(I_j)=P(E_j)$. And so T is an unbiased estimator of $\overline{Y}$ if and only if [we are assuming that each $\Pi_j>0$ and that Ω does not lie in a subspace (of R_N) of dimension lower than N]

$$B_0=0 \text{ and } B_j=(N\Pi_j)^{-1}(j=1,2,\ldots,N). \tag{5.9}$$

If we define a linear estimator as in (5.6), then it follows that the Horvitz-Thompson estimator [see (3.8) and (3.9)] is the only unbiased linear estimator of $\overline{Y}$. Following Godambe we, therefore, take one step further and define the class of linear estimators in the following manner:

Definition. Let $\beta_0, \beta_1, \ldots, \beta_N$ be well-defined functions of the label-set $\hat{u}$. By a generalized linear estimator T we mean a statistic that may be represented as

$$T=\beta_0+\sum_1^N \beta_j I_j Y_j. \tag{5.10}$$

Note that the β_j's and I_j's are functions of $\hat{u}$ and that it is only the observed Y_j's that really enter into the definition of T. We may rewrite T in the alternative form

$$T=\beta_0+\sum_1^n \beta_{\hat{u}_i}\hat{y}_i \tag{5.11}$$

and thus recognize it as a function of the sample core $\hat{x}$.

The generalized linear estimator T [as defined in (5.10)] is an unbiased estimator of $\bar{Y}$ if and only if

$$E(\beta_0)=0 \text{ and } E(\beta_j I_j)=N^{-1} \text{ for all } j. \tag{5.12}$$

Let us denote by $\mathcal{L}'$ the class of generalized linear unbiased estimators of $\bar{Y}$. If each $\Pi_j>0$, then $\mathcal{L}'$ is never vacuous, for we have already recognized the Horvitz-Thompson estimator

$$H=\Sigma(N\Pi_j)^{-1}I_jY_j \tag{5.13}$$

as a member of $\mathcal{L}'$. If $\theta_0=(a_1, a_2, \ldots, a_N)$ be a fixed point in Ω and we define H_0 as

$$H_0=\Sigma(N\Pi_j)^{-1}I_j(Y_j-a_j)+(\Sigma a_j)/N \tag{5.14}$$

then H_0 is a member of $\mathcal{L}'$ and has zero risk (variance) when $\theta=\theta_0$. It follows that in the class $\mathcal{L}'$ of generalized linear unbiased estimators of $\bar{Y}$ there cannot exist a best (uniformly minimum risk) estimator. This then is the celebrated Godambe assertion that we referred to in the opening paragraph of this section.

If we go back to the case of simple random sampling and compare the two classes L and $\mathcal{L}'$ [defined in (5.2) and (5.12) respectively] then we shall observe that the two classes have precisely one member in common, namely the sample mean

$$\bar{y}=\left(\sum_1^n y_i\right)/n=\sum_1^N (n^{-1}I_jY_j).$$

The Godambe class $\mathcal{L}'$ of generalized linear estimators is not an extension of the class L. The two classes L and $\mathcal{L}'$ are essentially different in character and scope. Thus, the classical assertion that the sample mean is the best linear unbiased estimate of the population mean and Godambe's denial that no such best linear unbiased estimate can ever exist are both true (each rather trivially) in their separate contexts.

Following Hanurav, we may extend the Godambe class of linear estimators by defining a linear estimate as

$$T^*=\beta_0^*+\Sigma\beta_j^* I_jY_j \tag{5.15}$$

where $\beta_0^*, \beta_1^*, \ldots, \beta_N^*$ are well-defined functions of u—the label part of the data $x=(u, y)$. The only difference between (5.10) and (5.15) is that in the former the β's are functions of $\hat{u}$, whereas in the latter, the β^*'s are functions of u. Once we remember that the I_j's are functions of the label-set $\hat{u}$, it

follows at once that

$$E(T^* \mid \hat{x}) = \beta_0 + \Sigma \beta_j I_j Y_j \tag{5.16}$$

where

$$\beta_j = E(\beta_j^* \mid \hat{x}) = E(\beta_j^* \mid \hat{u})$$

is a function of the label-set $\hat{u}$ $(j = 0, 1, 2, \ldots, N)$. Thus, the conditional expectation of each T^*, given the sufficient statistic (sample-core) $\hat{x}$, is a T as defined in (5.10). From the Rao-Blackwell theorem it then follows that for each estimator of type (5.15) we can find an estimator of type (5.10) with a performance characteristic that is at least as good as (uniformly) that of the former. From the decision theoretic point of view the extension of the class (5.10) by the class (5.15) is, therefore, sort of vacuous.

6
Homogeneity, Necessary Bestness and Hyper-Admissibility

During the past few years, altogether much too much has been written on the subject of linear estimates of the population total Y. The original sin was that of Horvitz and Thompson who in 1952 sought to give a classification of linear estimates of Y. The tremendous paper-writing pressure of the past decade has taken care of the rest. For a plan $\mathcal{S}$ that requires that the n sample units be drawn one at a time, without replacements, and with equal or unequal probabilities, Horvitz and Thompson called an estimator T to be of T_1-type if T be of the form (5.1) with $b_0 = 0$, where $b_1, b_2, \ldots, b_n$ are pre-fixed constants and $y_1, y_2, \ldots, y_n$ are the n observed Y-values in their natural selection order. An estimator of the type (5.6) with $B_0 = 0$ (definable for an arbitrary plan $\mathcal{S}$) was classified as a T_2-type estimator. By a T_3-type estimator, Horvitz and Thompson meant an estimator $T = \beta S$, where S is the sample total and $\beta = \beta(\hat{u})$ is an arbitrary function of the label-set $\hat{u}$. That is, a T_3-type estimator is of the form (5.10) with $\beta_0 = 0$ and $\beta_1 = \beta_2 = \ldots = \beta_N$. Prabhu Ajgaonkar (1965) combined the features of the T_2 and T_3 type estimators to define his T_5-type (someone else must have defined the T_4-type!) estimators as estimators of the type (5.10) with

$$\beta_0 = 0 \text{ and } \beta_j = \beta B_j \quad (j = 1, 2, \ldots, N) \tag{6.1}$$

where β is a function of $\hat{u}$ and $B_1, B_2, \ldots, B_N$ are pre-fixed constants. With the exception of the T_1-type estimators, all the other types are subclasses of the Godambe class of linear homogeneous estimators, that is, estimators of the type (5.10) with $\beta_0 = \beta_0(\hat{u}) \equiv 0$ for all values of $\hat{u}$. Let us denote the Godambe class of linear homogeneous unbiased estimators of Y by $\mathcal{L}_0$. The rest of this section is devoted to a study of the class $\mathcal{L}_0$.

The class $\mathcal{L}_0$ is the class of all estimators of the form

$$T = \Sigma \beta_j I_j Y_j, \tag{6.2}$$

where β_j is a function of the label-set $\hat{u}$, I_j is the indicator of the event $j \in \hat{u}$ and

$$E(\beta_j I_j) = 1 \quad (j = 1, 2, \ldots, N). \tag{6.3}$$

Let us count the degrees of freedom that we have in setting up an estimator in $\mathscr{L}_0$. Let U be the set of all the possible values (given a plan $\mathscr{S}$) of the label-set $\hat{u}$ and let U_j be the subset of those $\hat{u}$'s that include the unit j. [The event E_j that $j \in \hat{u}$ is then the same as $\hat{u} \in U_j$.] Let m_j be the number of members in the set U_j. We are assuming that no U_j is vacuous; that is, no E_j is an impossible event; that is, $\Pi_j = P(E_j) > 0$ for all j. Thus,

$$m = \Sigma m_j \geq N. \tag{6.4}$$

For defining a T in $\mathscr{L}_0$ we need to define the N functions $\beta_1, \beta_2, \ldots, \beta_N$ on U. Since $I_j = I_j(\hat{u}) = 0$ for all $\hat{u} \notin U_j$, it is clear that we really need to define β_j on the set U_j only ($j = 1, 2, \ldots, N$). [The values of β_j outside the set U_j have no bearing on the statistic T as defined in (6.2).] Thus, we can think of each β_j as an m_j-dimensional vector. Now (6.3) is, in reality, a linear restriction on the m_j-dimensional vector β_j. We, therefore, have $m_j - 1$ degrees of freedom in our choice of the function (vector) β_j and so we have in all

$$\Sigma(m_j - 1) = m - N \tag{6.5}$$

degrees of freedom in our selection of a T in $\mathscr{L}_0$. We may visualize $\mathscr{L}_0$ as an $m - N$ dimensional surface (plane) in the m-dimensional Euclidean space R_m.

Let us stop for a moment to consider the extreme (and rather trivial) situation where $m = N$, that is, $m_j = 1$ for all j. This is the case of a unicluster (the terminology is Hanurav's) sampling plan, that is, a plan $\mathscr{S}$ that partitions the population $\mathscr{P}$ into a number of mutually exclusive and collectively exhaustive parts and then selects just one of these parts as the label-set $\hat{u}$. In this case we have no degree of freedom in the selection of a T; that is, the class $\mathscr{L}_0$ is a one point set consisting only of the Horvitz-Thompson estimator

$$T_0 = \Sigma \Pi_j^{-1} I_j Y_j. \tag{6.6}$$

Let us return to the non-trivial case where $m > N$. As we remarked before, a member T in $\mathscr{L}_0$ is then determined by our choice of $(\beta_1, \beta_2, \ldots, \beta_N)$ which we may look upon as an m-dimensional vector lying in an $m - N$ dimensional plane. The problem is to choose a T in $\mathscr{L}_0$ that has minimum variance. Now, if T be as in (6.2) then

$$V(T) = \sum_j V(\beta_j I_j) Y_j^2 + 2 \sum_{j<k} \mathrm{Cov}(\beta_j I_j, \beta_k I_k) Y_j Y_k \tag{6.7}$$

which depends on the state of nature $\theta = (Y_1, Y_2, \ldots, Y_N)$. For each $\theta \in \Omega$, it is then clear that $V(T)$ is a (positive semi-definite) quadratic form in the m-dimensional vector $(\beta_1, \beta_2, \ldots, \beta_N)$. For each θ in Ω, there clearly exists a choice of the vector $(\beta_1, \beta_2, \ldots, \beta_N)$ that minimizes (6.7). Except in some very special situations (with Ω a very small set), there cannot exist a choice of $(\beta_1, \beta_2, \ldots, \beta_N)$ that will minimize (6.7) uniformly for all $\theta \in \Omega$. In the class $\mathscr{L}_0$ of all linear homogeneous unbiased estimators of Y there does not exist a uniformly minimum variance unbiased estimator (Godambe, 1955).

So the search was on for some other performance criterion that would uphold some estimator as the best in the class $\mathscr{L}_0$ (or in some other smaller or

larger class). Of late two rather curious such criteria have been proposed for consideration. They are (a) Ajgaonkar's criterion of *necessary bestness* and (b) Hanurav's criterion of *hyper-admissibility*. Let us first consider necessary bestness, the curiouser of the two criteria.

"In order to choose a serviceable estimator from the practical point," writes Ajgaonkar (1965, p.638), "we propose the following criterion of the necessary best estimator."

Definition (Ajgaonkar). Between two unbiased estimators T and T' (of the population total Y) with variances

$$V(T)=\Sigma a_j Y_j^2+2\sum_{j<k} a_{jk}Y_jY_k$$

and

$$V(T')=\Sigma b_j Y_j^2+2\sum_{j<k} b_{jk}Y_jY_k$$

the estimator T is *necessary better* than T' if $a_j \leq b_j$ for all j. The estimator T (in the class C) is *necessary best in* C if it is necessary better than every other estimator in C.

From (6.7) and the above definition it then follows that the estimator $T=\Sigma\beta_j I_j Y_j$ is necessary best in the class $\mathcal{L}_0$ if and only if $V(\beta_j I_j)$ is uniformly minimum for all j. From the Schwarz inequality we have

$$V(\beta_j I_j)\,V(I_j) \geq [\mathrm{Cov}(\beta_j I_j, I_j)]^2=[E(\beta_j I_j^2)-E(\beta_j I_j)\,E(I_j)]^2. \qquad (6.8)$$

Since $I_j^2 = I_j$, $E(I_j)=\Pi_j$, $V(I_j)=\Pi_j(1-\Pi_j)$ and $E(\beta_j I_j)=1$ for all j, we at once have

$$V(\beta_j I_j) \geq (1-\Pi_j)/\Pi_j \quad (j=1,2,\ldots,N). \qquad (6.9)$$

The sign of equality holds for all j in (6.9) if we select

$$\beta_j=\Pi_j^{-1} \quad (j=1,2,\ldots,N),$$

that is, if T is the Horvitz-Thompson estimator. Thus, in $\mathcal{L}_0$ there exists a unique necessary best estimator and that is the Horvitz-Thompson estimator (5.22). [Ajgaonkar (1965) gave a very complicated looking proof of the necessary bestness of (6.6) in the subclass of T_5-type estimators as defined in (6.1), and for a particular class of sampling plans. The present proof is a simplification of a proof suggested by Hege (1967).]

But why necessary bestness? It is hard to figure out how Ajgaonkar stumbled across this curious name and definition. Let us hazard a guess. We begin with a most unrealistic assumption that the space Ω contains points of the type

$$(0,\ldots,0,Y_j,0,\ldots,0),$$

that is, vectors with only one non-zero coordinate Y_j ($j=1,2,\ldots,N$), and let

Ω_0 be the subset of all points of the above kind. For a typical $\theta \in \Omega_0$ the variance of $T = \Sigma \beta_j I_j Y_j$ is equal to

$$V(\beta_j I_j) Y_j^2 \text{ (for some } j \text{ and } Y_j).$$

Hence, if we restrict our attention to the subset Ω_0 of Ω, the necessary best estimator in $\mathcal{L}_0$ is also the uniformly minimum variance estimator. The Horvitz-Thompson estimator has uniformly minimum variance (in $\mathcal{L}_0$) over the subset Ω_0.

Let us now consider the hyper-admissibility thesis of Hanurav (1968). Hyper-admissibility as the name suggests, is a strengthening of the decision-theoretic notion of admissibility. In order not to draw the attention of the reader away from the present context, let us define admissibility in the narrow framework of unbiased point estimation (of the population total Y) with variance as the risk function. Let T_0 and T_1 be unbiased estimators of Y.

Definition. T_0 is *uniformly better* than T_1 if

$$V(T_0) \leq V(T_1) \quad \text{for all } \theta \in \Omega$$

with the strict sign of inequality holding for at least one $\theta \in \Omega$.

Let C be a class of unbiased estimators of Y. We tacitly assume that C is a convex class, that is, when T_0 and T_1 are both members of C then so also is $(T_0 + T_1)/2$. For instance, the class $\mathcal{L}_0$ is convex.

Definition. $T_0 \in C$ is *admissible* in C if there does not exist a $T_1 \in C$ that is uniformly better than T_0.

If T_0 is admissible in C, then for any alternative $T_1 \in C$ it must be true that T_1 is not uniformly better than T_0; that is, either

(a) $V(T_0) \equiv V(T_1)$ for all $\theta \in \Omega$, or

(b) $V(T_0) < V(T_1)$ for at least one $\theta \in \Omega$.

Now, in view of the admissibility of T_0 and the convexity of C, the alternative (a) is impossible. Suppose (a) holds. Consider the estimator

$$T_* = (T_0 + T_1)/2$$

and observe that

$$\begin{aligned} V(T_*) &= \tfrac{1}{4}\{V(T_0) + V(T_1)\} + \tfrac{1}{2}\rho \sqrt{V(T_0)V(T_1)} \\ &= V(T_0)(1+\rho)/2 \\ &\leq V(T_0), \end{aligned}$$

where ρ is the correlation coefficient between T_0 and T_1. Since T_0 is admissible, it follows that $V(T_*) \equiv V(T_0)$ for all θ; that is, $\rho \equiv 1$ for all θ. Therefore, $T_0 = a + bT_1$. Since, T_0 and T_1 are both unbiased estimators of Y, it follows that $a = 0$ and $b = 1$. This contradicts the initial supposition that T_0 and T_1 are different estimators.

Thus, in our present context, we may redefine admissibility as

Definition (Hanurav). $T_0 \in C$ is admissible in C if, for any other $T_1 \in C$, it is true that

$$V(T_0) < V(T_1)$$

for at least one value of θ, say θ_{01}, in Ω. [The point θ_{01} will usually depend on T_0 and T_1.}

It is clear that the admissibility of an estimator T_0 depends on two things, namely, (a) the extent of the class C that T_0 is referred to and (b) the extent of the space Ω in which θ is supposed to lie. The smaller the class C and the larger the space Ω, the easier it is to establish the admissibility of a T_0 in C. A little while ago we noted that, in the class $\mathcal{L}_0$, the Horvitz-Thompson estimator (6.6) is the only one that has uniformly minimum variance over the set Ω_0 of all points θ with only one non-zero coordinate. If we are allowed to make the unrealistic assumption that $\Omega \supset \Omega_0$, then the admissibility of (6.6) in $\mathcal{L}_0$ follows at once. Godambe and Joshi (1965) proved the admissibility of (6.6) in the wider class of all unbiased estimators of Y, under the very unrealistic assumption that $\Omega = R_N$. As we have noted earlier [see (3.10) and (3.11)], the Horvitz-Thompson estimator is no longer admissible (even in the small class $\mathcal{L}$ of all linear unbiased estimators of Y) if it is known that Ω is a small neighborhood of a point $\theta_0 = (a_1, a_2, \ldots, a_N)$.

Hanurav sought to strengthen the notion of admissibility as follows. Following Godambe, he made the unrealistic assumption that $\Omega = R_N$ and then defined a *principal hyper-surface (phs)* of Ω as a linear subspace of all points $\theta = (Y_1, Y_2, \ldots, Y_N)$ with

$$Y_{j_1} = Y_{j_2} = \ldots = Y_{j_k} = 0$$

where $0 \leq k < N$ and $(j_1, \ldots, j_k)$ is a subset of $(1, 2, \ldots, N)$. [The whole space Ω corresponds to the case $k = 0$. There are $2^N - 1$ phs's of Ω.] Let Ω^* be a typical phs in Ω. Let C be a class of unbiased estimators of Y.

Definition (Hanurav). $T_0 \in C$ is hyper-admissible in C if, for every phs $\Omega^* \in \Omega$, it is true that T_0 is admissible in C when we restrict θ to Ω^*.

It follows at once that the H-T estimator $T_0 = \Sigma \Pi_j^{-1} I_i Y_j$ is the unique hyper-admissible estimator in $\mathcal{L}_0$. Suppose $T = \Sigma \beta_j I_j Y_j$ is hyper-admissible in $\mathcal{L}_0$. Consider the phs Ω_j^* of all points θ with $Y_i = 0$ for all $i \neq j$. For a typical $\theta \in \Omega_j^*$

$$V(T) = V(\beta_j I_j) Y_j^2$$

and this [as we have noted in (6.9)] is greater than

$$V(T_0) = \Pi_j^{-1}(1-\Pi_j)\, Y_j^2$$

unless $\beta_j = \beta_j(\hat{u}) = \Pi_j^{-1}$. Thus, the admissibility of T in each phs Ω_j^*implies that $T = T_0$. That T_0 is hyper-admissible, that is, is admissible on each phs, is equally trivial. Let Ω^* be a typical phs and let $T^* = \Sigma\beta_j^* I_j Y_j$ be a member of $\mathscr{L}_0$ such that

$$V(T^*) \leq V(T_0) \quad \text{for all } \theta \in \Omega^*. \tag{6.10}$$

For each one-dimensional phs $\Omega_j^* \in \Omega^*$, we must have the sign of equality in (6.10) for all $\theta \in \Omega_j^*$, and so it follows that $\beta_j^* = \Pi_j^{-1}$ for each j such that $\Omega_j^* \in \Omega^*$. Therefore, the sign of equality holds in (6.10) for all $\theta \in \Omega^*$. In other words, it is impossible to find an estimator T^* in $\mathscr{L}_0$ that is uniformly better than T_0 in the phs Ω^*; that is, T_0 is admissible (in the class $\mathscr{L}_0$) when we restrict θ to Ω^*.

In the context of the class $\mathscr{L}_0$, the twin criteria of *necessary bestness* and *hyper-admissibility* are mathematically equivalent. Before we proceed to examine the logical basis of the criterion of hyper-admissibility, let us point out a curious error committed by Hanurav (1968, p. 626). In his relation (3.2) Hanurav mistakenly asserts that T_0 is hyper-admissible (in C) if and only if, for every alternative $T_1 \in C$ and every phs $\Omega^* \subset \Omega$, we can find a point $\theta_{01} \in \Omega^*$ such that

$$V(T_0|\theta = \theta_{01}) < V(T_1|\theta = \theta_{01}). \tag{6.11}$$

We give an example to contradict the above assertion. Consider T_0 and T_1 where T_0 is as in (6.6) and

$$T_1 = \beta_1 I_1 Y_1 + \sum_{j=2}^{N} \Pi_j^{-1} I_j Y_j \text{ with } E(\beta_1 I_1) = 1.$$

In the phs Ω^* of all θ's with $Y_1 = 0$, it is clear that

$$V(T_0) \equiv V(T_1).$$

So in Ω^* we cannot find a point θ_{01} satisfying (6.11) and this in spite of T_0 being hyper-admissible in $\mathscr{L}_0$ and T_1 being an alternative member of $\mathscr{L}_0$.

The main result of Hanurav is to the effect that, for any nonunicluster sampling plan $\mathscr{L}$, the H-T estimator (6.6) is the unique hyper-admissible estimator in the class $\mathscr{U}^*$ of all polynomial unbiased estimators of Y. A quadratic estimator of Y is a statistic T of the form

$$T = \beta_0 + \Sigma\beta_j I_j Y_j + \Sigma\beta_{jk} I_{jk} Y_j Y_k \tag{6.12}$$

where $I_{jk} = I_j I_k$ is the indicator of the event that both j and k are in the label set $\hat{u}$, and the β's are functions of $\hat{u}$ with the (unbiasedness) conditions

$$E(\beta_0) = 0, \quad E(\beta_j I_j) \equiv 1, \quad E(\beta_{jk} I_j I_k) \equiv 0 \quad \text{for all } j \text{ and } k.$$

A polynomial estimator is similarly defined.

Now, let us examine the logical content of the hyper-admissibility criterion. Let $\mathscr{P}^*$ be an arbitrary but fixed subset (subpopulation) of the population $\mathscr{P}$ and let Y^* be the total Y-value of the units in $\mathscr{P}^*$; that is,

$$Y^* = \sum_{j \in \mathscr{P}^*} Y_j. \tag{6.12}$$

Suppose, along with an estimate of Y, the surveyor also needs to estimate the parameter Y^*. Once the surveyor has decided upon an estimator $T = T(\hat{u}, \hat{y})$ for Y, he may choose to derive an estimate T^* for Y^* in the following manner. Recall that $\hat{y}$ is the vector $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_\eta)$ of the (observed) Y-values of the μ distinct units $\hat{u}_1 < \hat{u}_2 < \ldots < \hat{u}_\eta$ in the label set $\hat{u}$. Define y^* as the vector

$$y^* = (y_1^*, y_2^*, \ldots, y_\eta^*),$$

where y_i^* is y_i or zero according as $\hat{u}_i$ is or is not a member of $\mathscr{P}^*$. In other words, we derive y^* by substituting by zeros those coordinates of the observation vector $\hat{y}$ that corresponds to units that are outside the subpopulation $\mathscr{P}^*$. Now define

$$T^* = T(\hat{u}, y^*). \tag{6.13}$$

If T is an unbiased estimator of Y, then it is almost a truism that T^* is an unbiased estimator of Y^*. If T is the linear homogenous estimator $\Sigma \beta_j I_j Y_j$, then T^* is the estimator $\Sigma^* \beta_j I_j Y_j$, where the summation Σ^* extends over all j that belong to $\mathscr{P}^*$. In particular, if $\mathscr{P}^*$ is the single member subpopulation consisting of the unit j alone, then the H-T estimator $T_0 = \Sigma \Pi_j^{-1} I_j Y_j$ gives rise to the estimator

$$T_0^* = \Pi_j^{-1} I_j Y_j = \begin{cases} \Pi_j^{-1} Y_j & \text{if } j \text{ is surveyed} \\ 0 & \text{otherwise} \end{cases} \tag{6.14}$$

for the parameter $Y^* = Y_j$.

The estimate (6.14) is similar to the one considered in example 3 of Section 3 and is, of course, utterly ridiculous. But in the makebelieve world of mathematicians, we are allowed to make any supposition. Let us pretend that when a surveyor estimates Y by T, he naturally commits himself to estimating each of the $2^N - 1$ subtotals Y^* by the corresponding derived estimate T^*. Given a class $C = \{T\}$ of estimators of Y, let us consider, for each subtotal Y^*, the class $C^* = \{T^*\}$ of derived estimators of Y^*. The estimator $T_0 \in C$ is hyper-admissible in C if, for each subtotal Y^*, the derived estimator T_0^* is admissible in C^*. According to Hanurav, if the sampling plan $\mathscr{S}$ is nonunicluster, then given any linear (or polynomial) unbiased estimator T that is different from the H-T estimator, he can always find another unbiased estimator T_1 and a subtotal Y^* such that the derived estimator T_1^* (for Y^*) is uniformly better than the derived estimator T^*.

7
Linear Invariance

We have idealized away many of the mathematically intractable features of the survey operation. But even with our oversimplified mathematical framework, the dimension N of the state of nature θ will usually run into several hundred thousands. It is clear that we are dealing with a most complex inference situation. A typical survey operation is an essentially non-repeatable, once in a lifetime affair. The surveyor, who is a specialist in the particular survey area, plans the survey, collects and analyzes the huge survey data and then arrives at his estimates of the various parameters of interest. Why does he need to consult a mathematician? How can the deductive processes of mathematics be of any use to the surveyor in his purely inductive inference making efforts? The author suspects that the answer lies in the general concensus among the scientific community that the mathematicians are the true watchdogs of rationalism. It may well be argued that this great reverence for mathematicians, this identification of rationalism with deduction, this over-eagerness to put every argument (be it in the realm of economics, psychology, survey theory, even philosophy) in the mold of pure deduction have done more harm than good to the general growth of knowledge. True, a good mathematician, having sharpened his mind with constant exercises in deductive reasoning, will often be able to comb out many a tangle created by unclear thinking on the part of the scientist. But new tangles are created by our over-eagerness to force a mathematical model for a situation that is essentially non-mathematical in nature. We close Part One of our essay with one more example of such a tangle in survey theory.

When the surveyor calls upon a decision-theorist (let us abbreviate the name to DT) to audit his survey work, the DT does not attempt to evaluate the thought process by which the surveyor arrived at his estimate T (for, say, the population total Y) from the data x. Indeed, the DT denies the very existence of a rational thought process that may lead us from the particular data x to the estimate T. [So far we have been freely using the two terms *estimate* and *estimator* and did not care to distinguish between them. But the whole controversy that is now raging in survey theory may be summarized as the difference between the estimate and the estimator. To the surveyor, the parameter Y is an unknown variable and the estimate T is a constant suggested by the data x at hand. The DT thinks of Y as an unknown constant and looks upon T as a random variable—a function on the sample space X.] As the DT cannot evaluate the estimate T, he proceeds to force an estimator out of the surveyor. For this he needs the sampling plan f to be randomized and, preferably, non-sequential. Once the DT has figured out the space X of all the possible data x (that the surveyor might have obtained from the survey), he would ask the surveyor to answer the impossible question of how he would have estimated Y for each x in X. If the function $T(x)$ is very complicated (as it would usually be) then that would be the end of the DT's audit. The estimator $T(x)$ better be simple enough so that the DT can evaluate the risk function—the average performance characteristics of the estimator. But before the risk function is evaluated, the DT would like to know the surveyor's loss function which again better be a simple one. As the DT

cannot answer the question: How good (rational) is the estimate T?, he evades the issue and proceeds to answer what he thinks to be a nearly equivalent question: How good is the average performance characteristic of the estimator T?

Instead of looking at the average performance characteristics of T, the DT may try to evaluate the estimator T by examining it directly as a function on X. A criterion that is frequently used for such direct evaluation of the estimator T is the criterion of linear invariance. The DT tries to find out if the surveyor's estimate T of the population total depends in some way on the scale in which the population values (the state of nature θ) are measured. With a linear shift (change of origin and scale) in the measurement scale for the population values, the state of nature $\theta=(Y_1, Y_2, \ldots, Y_N)$ will be shifted to $\theta' = (a+bY_1, a+bY_2, \ldots, a+bY_n)$ and the parameter Y will be shifted to $Y' = Na + bY$. With the same shift in the measurement scale the data

$$x = \{(u_1, u_2, \ldots, u_n), (y_1, y_2, \ldots, y_n)\}$$

will appear as

$$x' = \{(u_1, u_2, \ldots, u_n), (a+by_1, \ldots, a+by_n)\}. \tag{7.1}$$

Since x and x' represent the same data (in two different scales) it is natural to require that they lead to the same estimate (in the two scales) of the population total. This leads us to the following

Definition. The estimator $T = T(x)$ is origin and scale invariant if

$$T(x') \equiv Na + bT(x) \tag{7.2}$$

for all x, a, and $b > 0$, where x' is defined as in (7.1). We call T scale invariant if the above identity holds with $a = 0$.

One reason why there is so much interest (see Section 6) in linear homogeneous estimators is that they are supposed to be scale invariant. [As we shall presently point out, the above supposition is true only under some qualifications.] The Horvitz-Thompson estimator $T_0 = \Sigma\Pi_j^{-1} I_j Y_j$ is clearly scale invariant. It will be origin invariant only if

$$\Sigma\Pi_j^{-1} I_j \equiv N \tag{7.3}$$

for all samples. Since the expected value of the left hand side is clearly equal to N, we may restate the identity (7.3) as $V[\Sigma\Pi_j^{-1} I_j] = 0$ or equivalently

$$\begin{aligned} N^2 &= E[\Sigma\Pi_j^{-1} I_j]^2 \\ &= \Sigma\Pi_j^{-1} + \sum_{j \neq k} [\Pi_{jk}/(\Pi_j \Pi_k)] \end{aligned} \tag{7.4}$$

where Π_{jk} is the probability that both j and k are in the sample. In the case of simple random sampling with sample size n, it is clear that $\Pi_j = n/N$ for all j and so the H-T estimator reduces to the simple origin and scale invariant estimator

$$G = (N S)/n \tag{7.5}$$

where S is the sample total.

So far mathematicians have generally avoided the non-homogeneous linear estimators of the type

$$\beta_0 + \Sigma \beta_j I_j Y_j \tag{7.6}$$

in the mistaken belief that such estimators cannot possibly be scale invariant. It is tacitly assumed that any function $\beta = \beta(\hat{u})$ of the label-set $\hat{u}$ is necessarily *scale-free*; that is, the value of $\beta(\hat{u})$ depends only on $\hat{u}$ and not on the scale in which the population values are measured. That this need not be so is seen as follows. Suppose the surveyor defines β as

$$\beta(\hat{u}) = \Sigma I_j a_j \tag{7.7}$$

where $\theta_0 = (a_1, a_2, \ldots, a_N)$ is a pre-selected fixed point in the space Ω. The function β is clearly scale invariant. That is, if the surveyor is told that, in the new measurement scale, each of the population values is to be multiplied by the scaling factor b, then he (the surveyor) will automatically represent the point θ_0 as $(ba_1, ba_2, \ldots, ba_n)$ and re-compute $\beta(\hat{u})$ as $b\beta(\hat{u})$. Let us look back on the modified H-T estimator

$$H_0 = \Sigma \Pi_j^{-1} (Y_j - a_j) + \Sigma a_j \tag{7.8}$$

that we had considered earlier in (3.12), where $\theta_0 = (a_1, a_2, \ldots, a_N)$ is a pre-selected fixed point in Ω. A surveyor using (7.8) as his estimating formula for Y can never be accused of violating the canon of linear invariance. [We are not saying that H_0 is a respectable or a reasonable estimator of Y. We are only saying that, apart from being an unbiased estimator of Y with zero variance when $\theta = \theta_0$, the estimator H_0 is origin and scale invariant.] It has been repeatedly asserted by Godambe (see either of his 1968 papers) that, in the class of all estimators that are functions of the label-set $\hat{u}$ and the sample total S, the estimator $G = (NS)/n$ (where n is the number of units in $\hat{u}$) is the unique origin and scale invariant one. However, observe that if $\beta = \beta(\hat{u})$ is any scale-free function of $\hat{u}$ and

$$\beta_0(\hat{u}) = \Sigma a_j - \beta(\hat{u}) \Sigma I_j a_j ,$$

where $\theta_0 = (a_1, a_2, \ldots, a_N)$ is a fixed point in Ω, then the estimator

$$G_0 = \beta_0 + \beta S = \Sigma a_j + \beta \Sigma I_j (Y_j - a_j) \tag{7.9}$$

is an origin and scale invariant function of $\hat{u}$ and S.

References

1. Ajgaonkar, S.G. Prabhu, "On a Class of Linear Estimates in Sampling with Varying Probabilities without Replacements," *Journal of the American Statistical Association*, 60, 637-642, 1965.
2. Basu, D., "On Sampling With and Without Replacements," *Sankhyā*, 20, 287-294, 1958.

3. Basu, D., "Recovery of Ancillary Information," *Sankhyā*, 26, 3-16, 1964.
4. Basu, D. and Ghosh, J.K., "Sufficient Statistics in Sampling from a Finite Universe," *Proceedings of the 36th Session of Int. Stat. Inst.*, 850-859, 1967.
5. Basu, D., "Role of the Sufficiency and Likelihood Principles in Sample Survey Theory," *Sankhyā*, 31, 441-454, 1969.
6. Desraj, *Sampling Theory*, McGraw-Hill, 1968.
7. Godambe, V.P., "A Unified Theory of Sampling from Finite Populations," *Journal of the Royal Statistical Society, B*, 17, 269-278, 1955.
8. Godambe, V.P., "An Admissible Estimate for Any Sampling Design," *Sankhyā*, 22, 285-288, 1960.
9. Godambe, V.P. and Joshi, V.M., "Admissibility and Bayes Estimation in Sampling Finite Populations – Part I, II, and III," *Annals of Mathematical Statistics*, 36, 1707-1742, 1965.
10. Godambe, V.P., "Contributions to the United Theory of Sampling," *Rev. Int. Stat. Inst.*, 33, 242-258, 1965.
11. Godambe, V.P., "A New Approach to Sampling From a Finite Universe, Part I and II," *J. Roy. Statist. Soc.*, 28, 310-328, 1966.
12. Godambe, V.P., "Bayesian Sufficiency in Survey-Sampling," *Ann. Inst. Stat. Math.* (Japan), 20, 363-373, 1968.
13. Godambe, V.P., "Some Aspects of the Theoretical Developments in Survey Sampling," *New Developments in Survey Sampling*, Wiley-Interscience, 27-53, 1968-69.
14. Hájek, J., "Optimum Strategy and Other Problems in Probability Sampling," *Casopis Pest. Math.*, 84, 387-423, 1959.
15. Hanurav, T.V., "On Horvitz and Thompson Estimator," *Sankhyā, A*, 24, 429-436, 1962.
16. Hanurav, T.V., "Hyper-Admissibility and Optimum Estimators for Sampling Finite Populations," *Ann. Math. Statist.*, 39, 621-642, 1968.
17. Hege, V.S., "An Optimum Property of the Horvitz-Thompson Estimate," *Journal of the American Statistical Association*, 62, 1013-1017, 1967.
18. Horvitz, D.G. and Thompson, D.J., "A Generalization of Sampling Without Replacements from a Finite Universe," *J. Am. Stat. Ass.*, 47, 663-685, 1952.
19. Joshi, V.M., "Admissibility of the Sample Mean as Estimate of the Mean of a Finite Population," *Ann. Math. Statist.*, 39, 606-620, 1968.
20. Midzuno, H., "On the Sampling System with Probability Proportionate to Sum of Sizes," *Ann. Inst. Stat. Math.* (Japan), 3, 99-107, 1952.
21. Murthy, M.N., "On Ordered and Unordered Estimators," *Sankhyā, A*, 20, 254-262, 1958.
22. Pathak, P.N., "Sufficiency in Sampling Theory," *Ann. Math. Statist.*, 35, 795-808, 1964.
23. Raiffa, H. and Schlaifer, R.O., *Applied Statistical Decision Theory*, Boston Division of Research, Graduate School of Business Administration, Harvard University, 1961.
24. Roy, J. and Chakravarti, I.M., "Estimating the Mean of a Finite Population," *Ann. Math. Statist.*, 31, 392-398, 1960.
25. Zacks, S., "Bayes Sequential Designs for Sampling Finite Populations," *J. Am. Stat. Ass.*, 64, 1969.

[D. Basu presented a very brief summary of his paper and then went on to make some additional remarks. The following is an abstract of these additional remarks.

The sample core $\hat{x} = (\hat{u}, \hat{y})$ is always a sufficient statistic. In general, it is minimal sufficient. The sufficiency principle tells us to ignore as irrelevant all details of the sample $x = (u, y)$ that are not contained in the sample core $\hat{x}$. The likelihood principle tells us much more. The (normalized) likelihood function is the indicator of the set Ω_x of all parameter points θ that are consistent with the sample x. The set Ω_x depends on the sample x only through the sample-core $\hat{x}$. Given $\hat{x}$, the set Ω_x has nothing to do with the sampling plan $\mathcal{J}$. And this is true even for sequential sampling plans. For one who believes in the likelihood principle (as all Bayesians do) the sampling plan is no longer relevant at the data analysis stage.

A major part of the current survey sampling theory was dismissed by D. Basu as totally irrelevant. He stressed the need for science oriented, down to earth data analysis. The survey problem was posed as a problem of extrapolation from the observed part of the population to the unobserved part. An analogy was drawn between the problem of estimating the population total ΣY_j and the classical problem of numerical integration. In the latter the problem is to 'estimate' the value of the integral $\int_a^b Y(u)du$ by 'surveying' the function $Y(u)$ at a number of 'selected points' $u_1, u_2, \ldots, u_n$. Which points to select and how many of them, are problems of 'design'. Which integration formula to use and how to assess the 'error' of estimation, are problems of 'analysis'. True, it is possible to set up a statistical theory of numerical integration by forcing an element of randomness in the choice of the points. But, how many numerical analysts will be willing to go along with such a theory?

The mere artifact of randomization cannot generate any information that is not there already. However, in survey practice, situations will occasionally arise where it will be necessary to insist upon a random sample. But this will be only to safeguard against some unknown biases. In no situation, is it possible to make any sense of unequal probability sampling.

The inner consistency of the Bayesian point of view is granted. However, the analysis of the survey data need not be Bayesian. Indeed, who can be a true Bayesian and live with thousands of parameters? According to the author, survey statistics is more an art than a science.]

COMMENTS

G. A. Barnard:

First, a point of detail. Dr. Basu suggested, in the unwritten part II of his paper, a method of estimation using an assistant and dividing the data into two parts, D_1 and D_2. He said that no estimate of error would be available. I simply want to point out that an error estimate could be obtained, in an obvious way, if Dr. Basu has a twin, Basu[1], and his assistant also has a twin, assistant[1]. Then the data are divided into four sets, D_1, D_1', D_2, and D_2'.

Second, a general point. Dr. Basu and others here are concerned particularly with the problems which arise when it is necessary to make use of the additional information or prior knowledge α. In many sample survey situations

this prior knowledge of individuals is negligible (at least for a large part of the population under discussion) and in this case the classical procedures, in particular the Horvitz-Thompson estimators, apply in a sensible manner. It is important that we should not appear in this conference to be casting doubt on procedures which experience has shown to be highly effective in many practical situations.

The problem of combining external information with that from the sample is in general difficult to solve. For instance, in ordinary (distribution-free) least squares theory, the additional information that one or more of the unknown parameters has a bounded range makes the usual *justifications* inapplicable and no general theory appears to be possible, though it is easy to see what we should do in some particular cases.

With Dr. Basu's elephants, a realistic procedure (on the data he has given) would seem to be to think of the measurement to be made on one elephant as providing some estimate of how the animals have put on weight, or lost it, during the past three years. The circus owner should be able to give good advice on how elephants grow, but in the absence of this it would seem plausible to assume that the heaviest elephant, being fully mature, will have gained nothing, and that the percentage growth will be a linear function of weight three years ago. It would then be wise to select an elephant somewhat lighter than Sambo for weighing. The estimation procedure is clear.

Evidently, as Dr. Basu suggests, no purely mathematical theory is ever likely to be able to account for an estimation procedure such as that suggested. But I do not think this implies that all mathematical theories are time wasting in this context. A judicious balance is necessary. In particular, as I have said, we should not throw overboard the classical theory, or the work of Godambe, Horvitz, Thompson and others, just because we can envisage situations where these results would clearly not be applicable.

V. P. Godambe:

Professor Basu has given a very interesting presentation of some ideas in survey sampling theory which many of us have been contemplating for some time. I find it difficult however to agree with him in one respect. The likelihood principle, which does not permit the use of the sampling distribution generated by randomization for inference purposes, is unacceptable to me in relation to survey sampling. It seems as though the likelihood principle has different implications for two intrinsically similar situations: for the coin tossing experiment the likelihood principle allows the use of binomial distributions while infering about the binomial parameter but if the experiment is replaced by one of the drawing balls from a bag containing black and white balls, the likelihood principle does not allow the use of corresponding binomial (or hypergeometric if sampling is without replacement) distribution to infer the unknown proportion of white balls in the bag.

Professor Basu's comments on HT-estimator (example of weighing elephant and so on) are humourous and I wonder if he wants us to take them at all seriously. The comments fail to take into account the fact that the inclusion probabilities involved in HT-estimator are inseparably tied to the prior knowledge represented or approximated by a *class* of (indeed a very very wide one)

prior distributions (Godambe, *J. Roy. Statist. Soc.*, 1955) on the parametric space. I believe the only way of making sense of sampling practice and theory is through studying the frequency properties implied by the distributions generated by different modes of randomization (that is, different sampling designs) of the estimators obtained on the basis of the considerations of prior knowledge; of course one should also study the implications of reversing the role of *frequency properties* and *prior knowledge* (reference: section 7, Godambe's and Thompson's Symposium paper).

At the end of his paper Professor Basu comments on "origin and scale invariant estimator" in my paper, "Bayesian Sufficiency in Survey-Sampling", *Ann. Inst. Stat. Math.*, 1968. My assertion about the uniqueness in the paper is certainly true. Basu's comments suggest a different type of invariance which is already discussed in our (Godambe and Thompson) symposium paper (section 3).

J. Hájek:

Professor Basu and myself both like the likelihood function connected with sampling from finite populations, but for opposite reasons. He likes it to support the likelihood principle in sample surveys, and I like it to discredit this principle by showing its consequences in the same area. We both are wrong, because the probabilities of selection of samples are in a vague sense dependent on the unknown parameter, because they depend on the same prior facts (prior means and expectations, etc.) that have influenced the values under issue. Consequently, we do not have exactly the situation assumed in applications of the likelihood and conditionality principles. Of course this dependence of parameter and sample strategy is hard to formalize mathematically. My recognition of this dependence is due to a discussion I had recently with Professor Rubin on the conditionality principle.

As to the Horvitz-Thompson estimate, its usefulness is increased in connection with ratio estimation. For example, if the probabilities of inclusion are π_i and we expect the Y_i's to be proportionate to A_i, then we should use the estimate

$$\left(\sum_{i=1}^{N} A_i\right) \frac{\sum_{i \in s} Y_i/\pi_i}{\sum_{i \in s} A_i/\pi_i},$$

which would save the statistician's circus job. This estimate is not unbiased but the bias is small, and the idea of unbiasedness is useful only to the extent that greatly biased estimates are poor no matter what other properties they have.

J.C. Koop:

Professor Basu's essay is very stimulating and sometimes also provocative.

Regarding Sambo, I find the choice of selection probability for him (equal to 99/100) rather unwise in the face of the existence of a list of elephants' weights taken three years ago in the owner's possession. Sambo, we are told, was a middle-sized elephant, and knowing the existence of Jumbo in the herd, it might have been wiser to choose the selection probabilities directly propor-

tional to the respective weights of the elephants according to the available records. The reason being that if the elephants grew such that their present weights are directly proportional to their weights three years ago, then the variance of the estimate (equal to the selected elephant's weight divided by its selection probability), is zero. The circus statistician ought to have known better, and one should not be surprised that he was fired!

I am in complete agreement with him that the label of each unit in a sample (or in my terminology, the identity of a unit) cannot be discarded on the ground that it does not provide information. His discussion on this important point is very clear and can be read with profit.

However, I am somewhat surprised at his lack of appreciation for the basic ideas contained in Horvitz and Thompson's path-breaking paper of 1952 as evidenced by the following statement in section 6 of his paper: "During the past few years, altogether too much has been written on the subject of linear estimators of the population total Y. The original sin was that of Horvitz and Thompson who in 1952 sought to give a classification of linear estimates of Y. The tremendous paper-writing pressure of the past decade has taken care of the rest." These two writers constructed three linear estimators, each depending on one of the following three basic features of what I subsequently termed as the axioms sample formation in selecting units one at a time, namely, (i) the order of appearance of a unit in a sample, (ii) the presence or absence of a unit in the sample and (iii) the identity of the sample itself. Sample survey theorists have since benefited from their work. I for one felt in 1956 that the various types of estimators in the literature of that time needed classification and starting with these three features of sample formation, showed that $2^3-1=7$ types or classes of linear estimators, $T_1, T_2, \ldots, T_7$ were possible for one-stage sampling, three of which were those of Horvitz and Thompson. Godambe in his fundamental paper of 1955 found what I subsequently classified as the T_5-type of estimator, which should certainly not be attributed to Ajgaonkar, whose work began much later. In the process of this classification, it was found that an estimator given in the early pages of Sukhatme's text book of 1954 is of the T_4-type, that is, an estimator where the coefficients attached to the variate-values (observations in Basu's terminology) depended on the identity of the unit (label) and the order of appearance of the unit. Among other things, all this work was described in a thesis accepted by the North Carolina State University in 1957 and published in its *Institute of Statistics Mimeo Series* as No. 296, in 1961. Subsequently in 1963 I revised some of this work and amplified some of its ramifications in a paper in *Metrika*, Vol. 7(2) and (3).

One may ask what is the use of recognizing the three features of sample formation? In the context of the real world of sample surveys it must be said that they have physical meaning, which has some bearing on how an estimator may be constructed. Equally important, they point to the information supplied by the sample even before (field) observations on its members are made. In discussing Dr. C.R. Rao's excellent paper, I constructed a class of estimators where two of the features of sample formation were used, viz., (ii) which is equivalent to recognizing the identity of the distinct units (labels) and (iii) the identity of the sample itself, to show that an estimator of this class can have smaller M.S.E. than the U.M.V. estimator, derived through an appeal to the

principles of maximum likelihood, sufficiency and completeness, carried over almost bodily from classical estimation theory, thus bringing into question the extent of relevance of these principles in estimation theory for sample surveys of a finite universe. (It must be stressed that this does not detract from Dr. Rao's valuable paper which I interpret as a probe to uncover the difficulties of the subject.)

R. Royall:

Although I agree with much of what is said in this paper, I must take exception to one fundamental point. In section 2 Professor Basu states that: "From the point of view of a frequency probabilist, there cannot be a statistical theory of surveys without some kind of randomization in the plan S."

"Apart from observation errors and randomization, the only other way that probability can sneak into the argument is through a mathematical formalization of what we have described before as the residual part R of the prior knowledge, $K = (\Omega, \alpha, R)$. This is the way of a subjective (Bayesian) probabilist. The formalization of R as a prior probability distribution of θ over Ω makes sense only to those who interpret the probability of an event, not as the long range relative frequency of occurrence of the event (in a hypothetical sequence of repetitions of an experiment), but as a formal quantification of the ... phenomenon of *personal belief* in the truth of the event."

It seems frequently to be true that at some time before the values $y_1, y_2, \ldots, y_N$ are fixed it is natural and generally acceptable to consider these numbers as values, to be realized, of random variables $Y_1, Y_2, \ldots, Y_N$. For instance, these might be the numbers of babies born in each of the N hospitals in the state during the next month. What particular values will appear is uncertain, and this uncertainty can be described probabilistically. Although subjectivists would presumably accept these statements, in many finite populations such models are precisely as *objective* as those used everyday by frequentists. If such a model is appropriate before the y's are realized, it seems to be equally appropriate after they are fixed but unobserved. If a fair coin is flipped, the probability that it will fall heads is one half; if the coin was flipped five minutes ago, but the outcome has not yet been observed, my statement that the probability of heads is one half is no less objective now than it was six minutes ago. The state of uncertainty is not transformed from objective to subjective by the single fact that the event which determines the outcome has already occurred.

It can be argued that since the event has already occurred, the outcome should be treated as a fixed but unknown constant (so that now the probability of heads is one if the fixed but unknown outcome *is* heads and otherwise is zero). Such an argument leads back to the conventional model but rests on an unduly restrictive notion of the scope of objective probability theory.

The probability of one half for heads arises from my failure to notice that the coin is slightly warped. It can be argued that all probability models for real phenomena are likewise conditioned on personal knowledge and should therefore be called subjective. Be that as it may, (i) many statisticians do not consider themselves to be subjectivists and (ii) *super-population* models are frequently as objective as any other probability models used in applied sta-

tistics. Since such models, in conjunction with non-Bayesian statistical tools, can be extremely useful in practice as well as in theory, it seems to me to be a mistake to insist that they are available only to subjective Bayesians without pointing out that in this context the term applies to essentially all practicing statisticians.

REPLY

Professor Koop and Professor Godambe seem to think that the real difficulty in the elephant problem lies in the 'unrealistic' sampling plan—a plan that is 'not related' to the background knowledge. I always thought that the real purpose of a sampling plan is to get a good representative sample. If the owner knows how to relate the present weight of the representative elephant Sambo to the total weight of his fifty elephants, then he ought to go ahead and select Sambo. Why does he need a randomized sampling plan? Professor Koop wants to allot larger selection probability to Jumbo, the large elephant. Does he really prefer to have Jumbo rather than Sambo in his sample? I think Professor Koop is actually indifferent as to which elephant he selects for weighing. He *knows* more about the circus elephants than the circus owner. He 'knows' that the 50 ratios of the present and past weights of the elephants are nearly equal. Therefore, he has made up his mind that the ratio estimate is a good one irrespective of which elephant is selected. But he is not prepared to go all the way with me and assert the goodness of the ratio estimate irrespective of the selection plan. Professor Koop needs to allot unequal selection probabilities (proportional to their known past weights) to the 50 elephants so that he can mystify his non-statistical customers with the assertion that his estimate is then an unbiased one. As a scientist he has been trained to make a show of objectivity. May I ask what Professor Koop would do if the elephant trainer informs him that Jumbo (the big elephant) is on hunger strike for the past 10 days? Should he not try to avoid selecting Jumbo? He should, because now he does not know how to relate the present weight of Jumbo to the total weight of the 50 elephants.

In survey literature, we often come across the term *representative sample.* But to my knowledge the term has never been properly defined. At one time it used to be generally believed that the simple random sampling plan yields a representative sample. However, the difficulty with this naive sampling plan was soon recognized and so surveyors turned to stratification and other devices (like ratio and regression estimation) to exploit their background information about a specific survey problem. It is not easy to understand how surveyors got messed up with the idea of unequal probability sampling. I think it started with the idea of making the ratio estimate look unbiased. Thus Lahiri devised his method of using the random number tables in such a manner that the probability of selecting a particular sample set of units is proportional to the total 'size' of the units. This plan made the ratio estimate look 'good'. The flood-gate of unequal probability sampling was then opened and a surprisingly large number of learned papers has been published on the subject. What is even more surprising is that no one seems to worry about the fact that the surveyor can allot only one set of selection probabilities $\pi_1, \pi_2, \ldots, \pi_N$, but

that he has usually to estimate a vast number of different population totals. For each particular population total the surveyor may be able to find an appropriate ratio (or regression) estimate. But how can he possibly make all these different ratio estimates look 'good'?

Of late, a great deal has been written about the Horvitz-Thompson estimate. A little while ago Professor Rao proved an optimum property of the method. But to me the H-T estimate looks particularly curious. Here is a method of estimation that sort of contradicts itself by alloting weights to the selected units that are inversely proportional to their selection probabilities. The smaller the selection probability of a unit, that is, the greater the desire to avoid selecting the unit, the larger the weight that it carries when selected.

The question that Professor Hájek raised in the first part of his comments is exceedingly important and is one that, at one time, had given me a great deal of trouble. As Professor Hájek admitted, the question is hard to formulate and is even harder to answer. In the second part of my essay, I shall discuss the problem in greater detail. To-day, let us try to understand the difficulty in the context of the circus elephants. Suppose the surveyor (the owner) selects three elephants u_1, u_2, and u_3 with probabilities proportional to their past weights (and, say, with replacements) so that the data is $x = [(u_1,y_1), (u_2,y_2), (u_3,y_3)]$. In this case, the selection probability of the labels $u = (u_1,u_2,u_3)$ depends on the past weights of the 50 elephants and, therefore, also *depends* on their present weights—the state of nature $\theta = (Y_1, Y_2,..., Y_{50})$. If the selection probability of u depends on θ, then the very fact of its selection gives the surveyor some information about θ. Should the surveyor ignore this fact and act as if he always wanted to select this set of labels u and analyze the data x on that basis? This is precisely what I am advising the surveyor to do and this is what Professor Hájek thinks to be an error. But let us stop and think for a moment. Does the information that u is selected tell the surveyor anything (about θ) that the surveyor did not know already? When the question is phrased this way, one will be forced to admit that there is no real difference between the above plan and a simple random sampling plan. Indeed, the important point that I am trying to make is this, that even when the sampling plan is sequential, the relevant thing is the data generated by the plan and the likelihood function (which depends only on the data and has nothing whatsoever to do with the plan).

However, contrast the above sampling plan with a plan where the owner asks the elephant trainer to give him the names of three elephants that come first to his mind. If (u_1,u_2,u_3) are the three elephants that are selected by the above plan, then the surveyor does not really know how he got the labels (u_1,u_2,u_3) and so he cannot analyze the data x. Could it be that the three elephants were refusing to eat for some time and that is why they were on the trainer's mind at the time? If the owner must depend on the trainer for the names and present weights of three sample elephants, and if he does not have the sampling frame (so that he cannot select the labels himself), then he may be well advised to instruct the trainer to select the three sample labels at random. Randomness is a devil no doubt, but this is a devil that we understand and have learnt to live with. It is easier to trust a known devil than an unknown saint!

The second point raised by Professor Hájek is easier to deal with. If the

surveyor *knows* that the ratios of the present and past weights of the elephants are nearly equal, then why does he not use the ratio estimate itself? I do not see any particular merit in the estimate suggested by Professor Hájek.

Now, let us turn to Professor Godambe's objection to the likelihood principle in the context of survey sampling. It will be easier for us to understand Godambe's point if we examine the following example. In a class there are 100 students. An unknown number τ of these students have visited the musical show *Hair*. Suppose we draw a simple random sample of 20 students and record for each student, not his (or her) name, but only whether he has seen *Hair*. The likelihood is then a neat (hypergeometric) function involving only the parameter of interest τ. Godambe likes this likelihood function. However, if we had also recorded the name of each of the selected students, then the likelihood function would have been a lot messier. It would no longer have been a direct function of τ, but would have been a function of the state of nature $\theta = (Y_1, Y_2, ..., Y_N)$, where Y_j is 1 or 0 according as the student j has or has not seen *Hair*. Godambe does not know how to make any sense of this likelihood function. My advice to Professor Godambe will be this: "If the names (labels) are 'not informative', if there is no way that you can relate the labels to the state of nature θ, then do not make trouble for yourself by incorporating the labels in your data". After all, isn't this what we are doing all the time? When we toss a coin several times to determine the extent of its bias, do we record for each toss the exact time of the day or the face that was up when the coin was stationary on the thumb? We throw out such details from our data in the belief that they are not relevant (informative). Statistics is both a science and an art. It is impossible to rationalize everything that we do in statistics. These days we are hearing a lot of a new expression—rationality of type II. It is this second kind of rationality that will guide a surveyor in the matter of selection of his sample and the recording of his data.

The final remark of Professor Godambe seems to suggest that he has not quite understood what I said in the last paragraph of my essay. It is simply this that the constants in the estimating formula of the surveyor need not be (indeed, they should not be) pure numbers like π and e. The estimating formula (estimator) that the surveyor chooses surely depends on the particular inference situation. If the mathematician wishes to find out how the estimator behaves in the altered situation where the population values are measured in a different scale, he should first ascertain from the surveyor whether he (the surveyor) would like to adjust the constants in his formula to fit the new scale. When the surveyor is given this freedom, then it is no longer true that $G = NS/n$ is the only linearly invariant estimator in the class of all estimators that depend only on the label-set and the sample total. The Godambe assertion holds true only in the context of a severely restricted choice.

If I have understood Professor Royall correctly, then he claims that his super-population models for the parameter $\theta = (Y_1, Y_2, ..., Y_N)$ are non-Bayesian in the sense that such models do have objective frequency interpretations. His contention about the tossed coin in the closed palm is somewhat misleading. Let us examine a typical super-population model in which the Y_j's are assumed to be independent random variables with means αA_j and variances βA_j^{γ}, where A_j is a known auxiliary character of unit j and α, β, γ are known (or unknown) constants ($j = 1,2,...,N$). To me, such a model looks

exactly like a Bayesian formalization of the surveyor's background knowledge or information. Certainly, there is nothing objective about the above model. Indeed, is any probability model objective? When a scientist makes a probability assumption about the observable X, he is supposed to be very objective about it. But as soon as he makes a similar statement about the state of nature θ he is charged with the unmentionable crime of subjectivity. Mr. Chairman, you have always been telling us that the ultimate decision is an 'act of will' on the part of the decision (inference) maker. Isn't it equally true that the choice of the probability model for the observable X is also an act of will on the part of the statistician? Equally subjective is the choice of the 'performance characteristics'. A true scientist has to be subjective. Indeed, he is expected to draw on all his accumulated wisdom in the field of his specialization. My own subjective assessment of the present day controversy on objectivity in science and statistics is this that the whole thing is only a matter of semantics.

If we define mathematics as the art and science of deductive reasoning—an effort at deducing theorems from a set of basic postulates, using only the three laws of logic—then statistics (the art and science of induction) is essentially anti-mathematics. A mathematical theory of statistics is, therefore, a logical impossibility!

Sankhyā : The Indian Journal of Statistics
1975, Volume **37**, Series A, Pt. 1, pp. 1-71.

STATISTICAL INFORMATION AND LIKELIHOOD*

By D. BASU

University of Manchester and *Indian Statistical Institute*

PART 1 : PRINCIPLES

SUMMARY. In part one of this essay the notion of 'statistical information generated by a data' is formulated in terms of some intuitively appealing principles of data analysis. The author comes out very strongly in favour of the unrestricted likelihood principle after demonstrating (to his own satisfaction) the reasonableness of the Bayes-Fisher postulate that, within the framework of a particular statistical model, the 'whole of the relevant information in the data' must be supposed to be summarised in the likelihood function generated by the data.

Part two begins with a brief discussion on some non-Bayesian likelihood methods of data analysis that originated in the writings of R. A. Fisher. The central Fisher-thesis on likelihood that it is only a point function is challenged. The principle of maximum likelihood is questioned and the limitations of the method exposed.

Part three of the essay is woven around some paradoxical counter examples. The author demonstrates (again to his own satisfaction) how such examples discredit the fiducial argument, underline the impropriety of improper Bayesianism, expose the naivety of standard statistical practices like (pin-point) null-hypothesis testing, 3σ-likelihood interval estimates, etc. and how at the same time they illuminate and strengthen the likelihood principle by putting it into its true Bayesian perspective.

1. Statistical information

The key word in Statistics is information. After all, this is what the subject is all about. A problem in statistics begins with a state of nature, a parameter of interest ω about which we do not have enough information. In order to generate further information about ω, we plan and then perform a statistical experiment $\mathcal{E}$. This generates the sample x. By the term 'statistical data' we mean such a pair $(\mathcal{E}, x)$ where $\mathcal{E}$ is a well-defined statistical experiment and x the sample generated by a performance of the experiment. The problem of data analysis is to extract 'the whole of the relevant information'—an expression made famous by R. A. Fisher—contained in the data $(\mathcal{E}, x)$ about the parameter ω. But, what is information ? No other concept in statistics is more elusive in its meaning and less amenable to a generally agreed definition.

---

*This essay is dedicated to the memory of the Late Professor Prasanta Chandra Mahalanobis.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_25,

To begin with, let us agree to the use of the notation

$$\text{Inf } (\mathscr{E}, x)$$

only as a pseudo-mathematical short hand for the ungainly expression : 'the whole of the relevant information about ω contained in the data $(\mathscr{E}, x)$'. At this point an objection may well be raised to the following effect : The concept of information in the data $(\mathscr{E}, x)$ makes sense only in the context of (i) the 'prior-information' q (about ω and other related entities) that we must have had to begin with and (ii) the particular 'inferential problem' II (about ω) that made us look for further information.

While agreeing with the criticism that it is more realistic to look upon 'information in the data' as a function with four arguments II, q, $\mathscr{E}$ and x, let us hasten to point out that at the moment we are concerned with variations in $\mathscr{E}$ and x only and so we are holding fixed the other two elements of II and q. That $\text{Inf}(\mathscr{E}, x)$ may depend very critically on x, is well-illustrated by the following simple example.

Example 1 : Suppose an urn contains 100 tickets that are numbered consecutively as $\omega+1$, $\omega+2$, ... $\omega+100$ where ω is an unknown number. Let $\mathscr{E}_n$ stand for the statistical experiment of drawing a simple random sample of n tickets from the urn and then recording the sample as a set of n numbers $x_1 < x_2 < \ldots < x_n$. If at the planning stage of the experiment, we are asked to choose between the two experiments $\mathscr{E}_2$ and $\mathscr{E}_{25}$ then, other things being equal, we shall no doubt prefer $\mathscr{E}_{25}$ to $\mathscr{E}_2$. Consider now the hypothetical situation where $\mathscr{E}_2$ has been performed resulting in the sample $x = (17, 115)$. How good is $\text{Inf}(E_2, x)$? A quick analysis of the data will reveal that ω has to be an integer and must satisfy both the inequalities

$$\omega+1 \leqslant 17 \leqslant \omega+100 \quad \text{and} \quad \omega+1 \leqslant 115 \leqslant \omega+100.$$

In other words, $\text{Inf}(\mathscr{E}_2, x)$ tells us categorically that $\omega = 15$ or 16. Now, contrast the above with another hypothetical situation where $\mathscr{E}_{25}$ has been performed and has yielded the sample $x' = (17, 20, \ldots, 52)$, where 17 and 52 are respectively the smallest and the largest number drawn. With $\text{Inf}(\mathscr{E}_{25}, x')$ we can now only assert that ω is an integer that lies somewhere in the interval $[-48, 16]$. While it is clear that, in some average sense, the experiment $\mathscr{E}_{25}$ is 'more informative' than $\mathscr{E}_2$, it is equally incontrovertable that the particular sample (17, 115) from experiment $\mathscr{E}_2$ will tell us a great deal more about the parameter than will the sample (17, 20, ..., 52) from $\mathscr{E}_{25}$. To be more specific, with

$$\text{Inf}\{\mathscr{E}_n, (x_1, x_2, \ldots, x_n)\}$$

we know without any shadow of doubt that the true value of ω must belong to the set

$$A = \{x_1-1, x_1-2, \ldots, x_1-m\}$$

where $m = 100-(x_n-x_1)$. In the present case the likelihood function (for the parameter ω) is 'flat' over the set A and is zero outside (a situation that is typical of all survey sampling set-ups) and this means that the sample $(x_1, x_2, \ldots, x_n)$ from experiment $\mathcal{E}_n$ 'supports' each of the points in the set A with equal intensity. Therefore, it seems reasonable to say that we may identify the information supplied by the data $\{\mathcal{E}_n, (x_1, x_2, \ldots, x_n)\}$ with the set A and quantify the magnitude of the information by the statistic $m = 100-(x_n-x_1)$—the samller the number m is, the more precise is our specification of the unknown ω. Once the experiment $\mathcal{E}_n$ is performed and the sample $(x_1, x_2, \ldots, x_n)$ recorded, the magnitude of the information obtained depends on the integer m (which varies from sample to sample) rather than on the constant n.

Among contemporary statisticians there seems to be a complete lack of consensus about the meaning of the term 'statistical information' and the manner in which such an important notion may be meaningfully formalized. As a first step towards finding the greatest common factor among the various opinions held on the subject, let us make a beginning with the following loosely phrased operational definition of equivalence of two bits of statistical information.

Definition : By the equality or equivalence of $\mathrm{Inf}(\mathcal{E}_1, x_1)$ and $\mathrm{Inf}(\mathcal{E}_2, x_2)$ we mean the following :

(a) the experiment $\mathcal{E}_1$ and $\mathcal{E}_2$ are 'related' to the same parameter of interest ω, and

(b) 'everything else being equal', the outcome x_1 from $\mathcal{E}_1$ 'warrants the same inference' about ω as does the outcome x_2 from $\mathcal{E}_2$.

We plan to make an evaluation of several guidelines that have been suggested from time to time for deciding when two different bits of information ought to be regarded as equivalent. But before we proceed with that project, let us agree on a few definitions.

2. Basic definitons and relations

In contrast to the situation regarding the notion of statistical information, there exists a general consensus of opinion among present-day statisticians regarding a mathematical framework for the notion of a statistical experiment. We formalize a statistical experiment $\mathcal{E}$ as a triple $(\mathcal{X}, \Omega, p)$ where

(i) $\mathscr{X}$, the *sample space*, is the set of all the possible *samples* (outcomes) x that a particular performance of $\mathscr{E}$ may give rise to,

(ii) Ω, the *parameter space*, is the set of all the possible values of an entity ω that we call the *universal parameter* or the *state of nature*, and

(iii) $p = p(x \mid \omega)$, the probability function, is a map $p : \mathscr{X} \times \Omega \to [0, 1]$ that satisfies the identity

$$\sum_{x \in \mathscr{X}} p(x \mid \omega) \equiv 1 \qquad \text{for all} \quad \omega \in \Omega.$$

To avoid being distracted by measurability conditions, we stipulate from the beginning that both $\mathscr{X}$ and Ω are finite* sets. There is no loss of generality in the further assumption that

$$\sum_{\omega \in \Omega} p(x \mid \omega) > 0 \qquad \text{for all} \quad x \in \mathscr{X}.$$

It will frequently happen that we are not really interested in ω itself, but rather in some characteristic $\theta = \theta(\omega)$ of the universal parameter. In such cases we call θ the *parameter of interest* and denote its range of values by Θ. If there exists a set Φ of points ϕ such that we can write

$$\Omega = \Theta \times \Phi \quad \text{and} \quad \omega = (\theta, \phi),$$

we then call $\phi = \phi(\omega)$ the *nuisance parameter*.

With reference to an experiment $\mathscr{E} = (\mathscr{X}, \Omega, p)$, we define a *statistic* T as a map $T : \mathscr{X} \to \mathscr{T}$ of $\mathscr{X}$ into a space $\mathscr{T}$ of points t. Every point $t \in \mathscr{T}$ defines a subset $\mathscr{X}_t = \{x \mid T(x) = t\}$ of $\mathscr{X}$ and the family $\{\mathscr{X}_t \mid t \in \mathscr{T}\}$ of all these subsets defines a *partition* of $\mathscr{X}$. Conversely, every partition of $\mathscr{X}$ is induced by some suitably defined statistic. It is convenient to visualize a statistic T as a partition of the sample space $\mathscr{X}$.

Given $\mathscr{E}$ and a statistic T, we define the *marginal experiment* $\mathscr{E}_T$ as

$$\mathscr{E}_T = (\mathscr{T}, \Omega, p_T)$$

where the map $p_T : \mathscr{T} \times \Omega \to [0, 1]$ is given by

$$p_T(t \mid \omega) = \sum_{x \in \mathscr{X}_t} p(x \mid \omega).$$

Operationally, we may define $\mathscr{E}_T$ as 'perform $\mathscr{E}$ and then observe only $T = T(x)$.'

*The author holds firmly to the view that this contingent and cognitive universe of ours is in reality only finite and, therefore, discrete. In this essay we steer clear of the logical quick sands of 'infinity' and the 'infinitesimal'. Infinite and continuous models will be used in the sequel, but they are to be looked upon as mere approximations to the finite realities.

Still taking T as above, we may define, for each $t \epsilon \mathcal{T}$, a (conceptual) experiment

$$\mathcal{E}_t^T = (\mathcal{X}_t, \Omega, p_t^T)$$

where the map $p_t^T : \mathcal{X}_t \times \Omega \to [0, 1]$ is given by the formula

$$p_t^T(x \mid \omega) = p(x \mid \omega) / \sum_{x' \epsilon \mathcal{X}_t} p(x' \mid \omega)$$

for all $x \epsilon \mathcal{X}_t$ and $\omega \epsilon \Omega$. [The usual care needs to be taken about a possible zero denominator here.] We call $\mathcal{E}_t^T$ the *conditional experiment* given that $T(x) = t$. The experiment $\mathcal{E}_t^T$ may be loosely characterized as : 'Reconstruct the sample x from the information that $T(x) = t$'. [In a later section we examine the question whether such a reconstruction is operationally meaningful.] With each statistic T we may then associate a conceptual decomposition of the experiment $\mathcal{E}$ into a two-stage experiment : 'First perform $\mathcal{E}_T$ and then perform $\mathcal{E}_t^T$ where t is the outcome of $\mathcal{E}_T$.'

We now briefly list a set of well-known definitions and theorems.

Definition 1 (*A partial order*) : The statistic $T : \mathcal{X} \to \mathcal{T}$ is *wider* or *larger* than the statistic $T' : \mathcal{X} \to \mathcal{T}'$, if for each $t \epsilon \mathcal{T}$ there exists a $t' \epsilon \mathcal{T}'$ such that $\mathcal{X}_t \subset \mathcal{X}_{t'}$, that is, if the partition of $\mathcal{X}$ induced by T is a sub-partition of the one induced by T'.

Definition 2 (*Non-informative experiments*) : An experiement $\mathcal{E} = (\mathcal{X}, \Omega, p)$ is statistically trivial or non-informative (about the universal parameter ω) if, for each $x \epsilon \mathcal{X}$, the function $\omega \to p(x \mid \omega)$ is a constant.

Definition 3 (*Ancillary statistic*) : The statistic $T : \mathcal{X} \to \mathcal{T}$ is called an *ancillary* statistic (w.r.t. ω) if the marginal experiment $\mathcal{E}_T$ is non-informative (about ω).

Definition 4 (*Sufficient statistic*) : The statistic T is called a sufficient statistic (for ω) if, for all $t \epsilon \mathcal{T}$, the conditional experiment $\mathcal{E}_t^T$ is non-informative (about ω).

Definition 5 (*Likelihood function*) : When an experiment $\mathcal{E} = (\mathcal{X}, \Omega, p)$ is performed resulting in the outcome $x \epsilon \mathcal{X}$, the function $\omega \to p(x \mid \omega)$ is called the *likelihood function* generated by the data $(\mathcal{E}, x)$ and is variously denoted in the sequel as L, $L(\omega)$, $L(\omega \mid x)$ or $L(\omega \mid \mathcal{E}, x)$.

Definition 6 (*Equivalent likelihoods*) : Two likelihood functions L_1 and L_2 defined on the same parameter space Ω [but possibly corresponding to two different pairs $(\mathcal{E}_1, x_1)$ and $(\mathcal{E}_2, x_2)$ respectively] are said to be equivalent if

there exists a constant $c > 0$ such that $L_1(\omega) = cL_2(\omega)$ for all $\omega\epsilon\Omega$. [The constant c may, of course, depend on $\mathscr{E}_1$, $\mathscr{E}_2$, x_1 and x_2]. We write $L_1 \sim L_2$ to indicate the equivalence of the likelihood functions.

Definition 7 (*Standardized likelihood*) : Each likelihood function L on Ω gives rise to an equivalent *standardized* likelihood function $\bar{L}$ on Ω defined as

$$\bar{L}(\omega) = L(\omega) / \sum_{\omega' \epsilon \Omega} L(\omega').$$

Note that our earlier assumptions about Ω and p preclude the possibilities of the denominator being zero or infinite.

Theorem 1 : *A statistic T is sufficient if and only if, for $x_1, x_2 \epsilon \mathscr{X}$, $T(x_1) = T(x_2)$ implies $L(\omega \mid x_1) \sim L(\omega \mid x_2)$.*

In other words, a statistic $T : \mathscr{X} \to \mathscr{T}$ is sufficient if and only if, for every $t\epsilon\mathscr{T}$, it is true that all points x on the T-surface $\mathscr{X}_t$ generate equivalent likelihood functions. The following result is then an immediate consequence of the above.

Theorem 2 : *For any experiment $\mathscr{E} = (\mathscr{X}, \Omega, p)$ the map (statistic) $x \to \bar{L}(\omega \mid x)$, from x to a (standardized) likelihood function $\bar{L}$ on Ω, is the minimal sufficient statistic, that is, the above statistic is sufficient and every other sufficient statistic is wider than it.*

Definition 8 (*Mixture of experiments*) : Suppose we have a number of experiments $\mathscr{E}_i = (\mathscr{X}_i, \Omega, p_i)$, $i = 1, 2, \ldots$, with the same parameter space Ω, to choose from. And let $\pi_1, \pi_2, \ldots$ be a pre-assigned set of non-negative numbers summing to unity. The *mixture* $\mathscr{E}$ of the experiments $\mathscr{E}_1, \mathscr{E}_2, \ldots$ according to mixture (selection) probabilities $\pi_1, \pi_2, \ldots$ is defined as a two-stage experiment that begins with (i) a random selection of one of the experiments $\mathscr{E}_1, \mathscr{E}_2, \ldots$ with selection probabilities $\pi_1, \pi_2, \ldots$, followed by (ii) the performing of the experiment selected in stage (i). Clearly, the sample space $\mathscr{X}$ of the mixture experiment $\mathscr{E} = (\mathscr{X}, \Omega, p)$ is the set of all pairs (i, x_i) with $i = 1, 2, \ldots$ and $x_i \epsilon \mathscr{X}_i$ (that is, $\mathscr{X}$ is the disjoint union of the sets $\mathscr{X}_1, \mathscr{X}_2, \ldots$). And the probability function $p : \mathscr{X} \times \Omega \to [0, 1]$ is given by

$$p(x \mid \omega) = \pi_i p_i(x_i \mid \omega)$$

when $x = (i, x_i)$.

It is important to note our stipulation that the mixture probabilities $\pi_1, \pi_2, \ldots$ are pre-assigned numbers and, therefore, unrelated to the unknown

parameter ω. Given an experiment $\mathcal{E} = (\mathcal{X}, \Omega, p)$ and an ancillary statistic $T : \mathcal{X} \to \mathcal{T}$, we may view $\mathcal{E}$ as a mixture of the family

$$\{\mathcal{E}_t^T : t \epsilon \mathcal{T}\}$$

of conditional experiments, with mixture probabilities

$$\pi_t = p_T(t \mid \omega), \quad t \epsilon \mathcal{T}$$

which do not depend on ω since T is ancillary.

Definition 9 (*Similar experiments*) : The experiments $\mathcal{E}_1 = (\mathcal{X}_1, \Omega, p_1)$ and $\mathcal{E}_2 = (\mathcal{X}_2, \Omega, p_2)$ with the same parameter space Ω are said to be *similar* or *statistically isomorphic* if there exists a one to one and onto map $g : \mathcal{X}_1 \to \mathcal{X}_2$ such that

$$p_1(x_1 \mid \omega) = p_2(gx_1 \mid \omega)$$

for all $x_1 \epsilon \mathcal{X}_1$ and $\omega \epsilon \Omega$. The function g is then called a *similarity map.*

We end this section with a definition, due to D. Blackwell (1950), of the sufficiency of an experiment for another experiment and a few related remarks.

Definition 10 (*Blackwell sufficiency*) : The experiment $\mathcal{E}_1 = (\mathcal{X}_1, \Omega, p_1)$ is *sufficient* for the experiment $\mathcal{E}_2 = (\mathcal{X}_2, \Omega, p_2)$ if there exists a *transtion function* $\pi : \mathcal{X}_1 \times \mathcal{X}_2 \to [0, 1]$ (with the usual condition that $\sum_{x_2} \pi(x_1, x_2) = 1$ for all $x_1 \epsilon \mathcal{X}_1$) which satisfies the additional requirement that

$$p_2(x_2 \mid \omega) = \sum_{x_1} p_1(x_1 \mid \omega)\pi(x_1, x_2)$$

for all $\omega \epsilon \Omega$ and $x_2 \epsilon \mathcal{X}_2$.

The sufficiency of $\mathcal{E}_1$ for $\mathcal{E}_2$ means exactly this : that the experiment $\mathcal{E}_2$ may be simulated by first performing $\mathcal{E}_1$ and noting its outcome x_1, and then obtaining a point x_2 in $\mathcal{X}_2$ via a secondary randomization process that is defined in terms of the transition function $\pi(x_1, \cdot)$. Note that, for each $x_1 \epsilon \mathcal{X}_1$, the function $\pi(x_1, \cdot)$ defines a probability distribution on $\mathcal{X}_2$ that is free of the unknown ω. We refer to Blackwell (1950) for an alternative but equivalent formulation of Definition 10 in terms of the average performance characteristics of statistical decision functions.

If for experiment $\mathcal{E} = (\mathcal{X}, \Omega, p)$ the statistic $T : \mathcal{X} \to \mathcal{T}$ is sufficient (Definition 4), then the marginal experiment $\mathcal{E}_T = (\mathcal{T}, \Omega, p_T)$ is sufficient (Definition 10) for $\mathcal{E}$. The converse proposition is also true. If $\mathcal{E}_1$ and $\mathcal{E}_2$ are similar (Definition 9) experiments with $g : \mathcal{X}_1 \to \mathcal{X}_2$ as a similarity map, then the Kronecker delta function $\delta(gx_1, x_2)$ may be taken as the transition function

$\pi(x_1, x_2)$ to prove the sufficiency of $\mathcal{E}_1$ for $\mathcal{E}_2$. In a like manner the similarity map $g^{-1} : \mathcal{X}_2 \to \mathcal{X}_1$ proves the sufficiency of $\mathcal{X}_2$ for $\mathcal{X}_1$. Furthermore, any decision function δ_2 for $\mathcal{E}_2$ can be completely matched (in terms of its average performance characteristics) by the decision function δ_1 for $\mathcal{E}_1$ defined as

$$\delta_1(x_1) = \delta_2(gx_1) \quad \text{for all} \quad x_1 \epsilon \mathcal{X}_1.$$

3. Some principles of inference

Instead of plunging headlong into a controversial definition of Inf $(\mathcal{E}, x)$, let us follow a path of less resistance and formulate, on the model of A. Birnbaum (1962) some guidelines for the recognition of equivalence of two different bits of statistical information. Each such guideline is stated here as a Principle (of statistical inference).

Looking back on definition 9 of the previous section, it is clear that two similar experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ are identical in all respects excepting in the manner of labelling their sample points. Since the manner of labelling the sample points of an experiment should not have any effect on the actual information obtained in a particular trial, the following principle is almost self-evident.

Principle $\mathcal{I}$ (*The invariance or similarity principle*) : If $\mathcal{E}_1 = (\mathcal{X}_1, \Omega, p_1)$ and $\mathcal{E}_2 = (\mathcal{X}_2, \Omega, p_2)$ are similar experiments with $g : \mathcal{X}_1 \to \mathcal{X}_2$ as a similarity map of $\mathcal{E}_1$ onto $\mathcal{E}_2$, then

$$\text{Inf}(\mathcal{E}_1, x_1) = \text{Inf}(\mathcal{E}_2, x_2)$$

if $gx_1 = x_2$.

Now, suppose the two points x' and x'', in the sample space of an experiment $\mathcal{E} = (\mathcal{X}, \Omega, p)$, give rise to *identical* likelihood functions, that is, $p(x' \mid \omega) = p(x'' \mid \omega)$ for all $\omega \epsilon \Omega$. We can then define a similarity map $g : \mathcal{X} \to \mathcal{X}$ of $\mathcal{E}$ onto itself in the following manner :

$$gx = \begin{cases} x \text{ if } x \notin \{x', x''\} \\ x' \text{ or } x'' \text{ acc. as } x = x'' \text{ or } x'. \end{cases}$$

The following is then a specialization of principle $\mathcal{I}$ to the case of a single experiment $\mathcal{E}$.

Principle $\mathcal{I}'$ (*A weak version of* $\mathcal{I}$) : If $p(x' \mid \omega) = p(x'' \mid \omega)$ for all $\omega \epsilon \Omega$, then

$$\text{Inf}(\mathcal{E}, x') = \text{Inf}(\mathcal{E}, x'').$$

Principle $\mathcal{I}'$ induces the following equivalence relation on the sample space of an experiment : The two points x' and x'' in the sample space $\mathcal{X}$ of an

experiment $\mathscr{E}$ are equivalent or equally informative if they generate *identical* likelihood functions.

Let us look back on Definition 2 in Section 2 and re-assert the almost self-evident proposition : 'No additional information can be generated about a partially known parameter ω by performing a statistically trivial experiment $\mathscr{E}$.' It follows then that once an experiment $\mathscr{E}_1$ has been carried out resulting in the outcome y, it is not possible to add to the information $\mathrm{Inf}(\mathscr{E}_1, y)$ so obtained by carrying out a further 'post-randomization' exercise—that is, by performing a secondary experiment $\mathscr{E}_{(y)}$ whose randomness structure may depend on the outcome y of $\mathscr{E}_1$ but is completely known to the experimenter. Let us formally rewrite the above in the form

$$\mathrm{Inf}(\mathscr{E}_1, y) = \mathrm{Inf}\{(\mathscr{E}_1 \rightarrow \mathscr{E}_{(y)}),\ (y, z)\}$$

where $(\mathscr{E}_1 \rightarrow \mathscr{E}_{(y)})$ stands for the composite experiment '$\mathscr{E}_1$ followed by $\mathscr{E}_{(y)}$' and y, z are the outcomes of $\mathscr{E}_1$ and $\mathscr{E}_{(y)}$ respectively.

Now let $T : \mathscr{X} \rightarrow \mathscr{T}$ be a sufficient statistic for $\mathscr{E} = (\mathscr{X}, \Omega, p)$ and let $\mathscr{E}_T$ and $\{\mathscr{E}_t^T : t \epsilon \mathscr{T}\}$ be respectively the marginal experiment and the family of conditional experiments as defined in Section 2. Now, we may look upon a performance of $\mathscr{E}$ and the observation of the outcome x as 'a performance of the marginal experiment $\mathscr{E}_T$, observation of its outcome $t = T(x)$, followed by a post-randomization exercise $\mathscr{E}_t^T$ of identifying the exact location of x on the surface $\mathscr{X}_t = \{x' \mid T(x') = t\}$'. Since T is sufficient, the conditional experiment $\mathscr{E}_t^T$ is statistically trivial for every $t \epsilon \mathscr{T}$. Looking back on the argument of the previous paragraph, one may now claim that the following principle has been sort of 'proved by analogy'.

Principle $\mathscr{S}$ (*The sufficiency principle*) : If, in the context of an experiment $\mathscr{E}$, the statistic T is sufficient then, for all $x \epsilon \mathscr{X}$ and $t = T(x)$,

$$\mathrm{Inf}(\mathscr{E}, x) = \mathrm{Inf}(\mathscr{E}_T, t).$$

If T is sufficient and $\mathscr{X}_t$ a particular T-surface, then from $\mathscr{S}$ it follows that $\mathrm{Inf}(\mathscr{E}, x)$ is the same for all $x \epsilon \mathscr{X}_t$. In the literature we often find the sufficiency principle stated in the following alternative (and perhaps a trifle less severe) form :

Principle $\mathscr{S}'$ (*Alternative version of* $\mathscr{S}$) : $\mathrm{Inf}(\mathscr{E}, x') = \mathrm{Inf}(\mathscr{E}, x'')$ if for some sufficient statistic T it is true that $T(x') = T(x'')$.

From Theorems 1 and 2 of Section 2 it follows at once that the following is an equivalent version of $\mathscr{S}'$:

Principle $\mathscr{L}'$: (*The weak likelihood principle*) : $\mathrm{Inf}(\mathscr{E}, x') = \mathrm{Inf}(\mathscr{E}, x'')$ if the two sample points x' and x'' generate equivalent likelihood functions, that is, if $L(\omega \mid x') \sim L(\omega \mid x'')$.

Clearly, $\mathscr{L}'$ implies $\mathscr{I}'$. Before we turn our attention to some other guiding principles of statistical inference, let us summarize our findings about the logical relationships among the principles $\mathscr{I}$, $\mathscr{I}'$, $\mathscr{S}$, $\mathscr{S}'$ and $\mathscr{L}'$ in the following :

Theorem 1 : $\mathscr{I} \Longrightarrow \mathscr{I}'$, $\mathscr{S} \Longrightarrow \mathscr{S}' \Longleftrightarrow \mathscr{L}' \Longrightarrow \mathscr{I}'$.

Whereas the sufficiency principle warns us to be vigilant against any 'post-randomization' in the statistical experiment and advises us to throw away the outcome of any such exercise as irrelevant to the making of inference, the conditionality principle concerns itself in a like manner with any 'pre-randomization' that may have been built into the structure of an experiment. Consider an experiment $\mathscr{E} = (\mathscr{X}, \Omega, p)$ which is a mixture (Definition 8, Section 2) of the two experiments $\mathscr{E}_i = (\mathscr{X}_i, \Omega, p_i)$, $i = 1, 2$, where the mixture probabilities π and $1-\pi$ are known. A typical outcome of $\mathscr{E}$ may then be represented as $x = (i, x_i)$, where $i = 1, 2$ and $x_i \epsilon \mathscr{X}_i$. Now, having performed the mixture experiment $\mathscr{E}$ and recognizing the sample as $x = (i, x_i)$, the question that naturally arises is whether we should present the data (for analysis) as $(\mathscr{E}, x)$ or in the simpler form of $(\mathscr{E}_i, x_i)$. To the author it seems almost axiomatic that the second form of data presentation should not entail any loss of information and this is precisely the content of the following.

Principle $\mathscr{C}'$ (*The weak conditionality principle*) : If $\mathscr{E}$ is a mixture of $\mathscr{E}_1$, $\mathscr{E}_2$ as described above, then for any $i \epsilon \{1, 2\}$ and $x_i \epsilon \mathscr{X}_i$

$$\mathrm{Inf}(\mathscr{E}, (i, x_i)) = \mathrm{Inf}(\mathscr{E}_i, x_i).$$

In the literature we frequently meet a much stronger version of $\mathscr{C}'$ which may be stated as follows :

Principle $\mathscr{C}$. (*The conditionality principle*) : If $T : \mathscr{X} \rightarrow \mathscr{T}$ is an ancilary statistic (Definition 3, Section 2) associated with the experiment $\mathscr{E} = (\mathscr{X}, \Omega, p)$, then, for all $x \epsilon \mathscr{X}$ and $t = T(x)$,

$$\mathrm{Inf}(\mathscr{E}, x) = \mathrm{Inf}\,(\mathscr{E}_t^T, x).$$

[For a discussion of $\mathscr{C}$ in a somewhat related context see Basu (1964).] We are now ready to state the centre-piece of our discussion in this essay—the likelihood principle. Let $\mathscr{E}_1$, $\mathscr{E}_2$ be any two experiments with the same parameter space and let x_i be a typical outcome of $\mathscr{E}_i$ $(i = 1, 2)$.

Principle $\mathscr{L}$ (*The likelihood principle*) : If the data $(\mathscr{E}_1, x_1)$ and $(\mathscr{E}_2, x_2)$ generate equivalent likelihood functions on Ω, then $\text{Inf}(\mathscr{E}_1, x_1) = \text{Inf}(\mathscr{E}_2, x_2)$.

Before going into the far-reaching implications of $\mathscr{L}$, let us briefly examine the logical relationships in which $\mathscr{L}$ stands vis a vis the principles stated earlier. That $\mathscr{L} \Longrightarrow \mathscr{I}$ follows at once from the definition of similar experiments. From the definition of a sufficient statistic it follows that the likelihood functions $L(\omega \mid \mathscr{E}, x)$ and $L(\omega \mid \mathscr{E}_T, t)$ are equivalent, whenever T is sufficient and $t = T(x)$. So $\mathscr{L} \Longrightarrow \mathscr{S}$. Likewise, when T is ancillary, the likelihood functions generated by the data $(\mathscr{E}, x)$ and $(\mathscr{E}_t^T, x)$ are equivalent, whenever $t = T(x)$. Therefore, $\mathscr{L} \Longrightarrow \mathscr{C}$. The following theorem asserts that the two weak principles $\mathscr{I}'$ and $\mathscr{C}'$ are together equivalent to $\mathscr{L}$.

Theorem 2 : $(\mathscr{I}'$ *and* $\mathscr{C}') \Longrightarrow \mathscr{L}$.

Proof : Suppose the data $(\mathscr{E}_1, x_1)$ and $(\mathscr{E}_2, x_2)$ generate equivalent likelihood functions, that is, there exists $c > 0$ such that

$$L(\omega \mid \mathscr{E}_1, x_1) = cL(\omega \mid \mathscr{E}_2, x_2) \qquad \ldots \quad (*)$$

for all $\omega \epsilon \Omega$. Using $\mathscr{I}'$ and $\mathscr{C}'$ we have to prove the equality $\text{Inf}(\mathscr{E}_1, x_1) = \text{Inf}(\mathscr{E}_2, x_2)$. To this end let us contemplate the mixture experiment $\mathscr{E}$ of $\mathscr{E}_1$ and $\mathscr{E}_2$ with mixture probabilities $c/(1+c)$ and $1/(1+c)$ respectively. Now, $(1, x_1)$ and $(2, x_2)$ are points in the sample space of the mixture experiment $\mathscr{E}$. In view of $(*)$ and our choice of the mixture probabilities, it is clear that the data $(\mathscr{E}, (1, x_1))$ and $(\mathscr{E}, (2, x_2))$ generate *identical* likelihood functions, and so from $\mathscr{I}'$ it follows that

$$\text{Inf}(\mathscr{E}, (1, x_1)) = \text{Inf}(\mathscr{E}, (2, x_2)).$$

Now, applying $\mathscr{C}'$ to each side of the above equality we arrive at the desired equality.

Since $\mathscr{S}' \Longrightarrow \mathscr{I}'$, we immediately arrive at the following corollary which was proved earlier by A. Birnbaum (1962).

Corollary : $(\mathscr{S}'$ *and* $\mathscr{C}') \Longrightarrow \mathscr{L}$.

4. Information as a function

From our exposition so far it should be amply clear that we are looking upon 'statistical information'—in the context of a particular problem of inference about a partially known state of nature ω—as some sort of a function that maps the space D of all conceivably attainable data $d = (\mathscr{E}, x)$ related to ω into an yet undefined range space Λ. For the logical development of any

concept it is important to agree in advance upon a 'universe of discourse'. In our case it is the space of all attainable data $d = (\mathcal{E}, x)$, where $\mathcal{E} = (\mathcal{X}, \Omega, p)$ is a typical statistical experiment concerning ω and x a typical outcome that may arise when $\mathcal{E}$ is performed. But what data are attainable, in other words, what triples $(\mathcal{X}, \Omega, p)$ correspond to performable statistical experiments ? The question is a tricky one and has escaped the general attention of statisticians.

Given a state of nature ω, not all conceivable triples $(\mathcal{X}, \Omega, p)$ can be models of performable statistical experiments. The situation is quite different in Probability Theory where we idealize the notion of a *random experiment* in terms of a single probability measure P on a measurable space $(\mathcal{X}, \mathcal{A})$. These days, with the help of powerful computers, we can simulate any reasonable random experiment upto almost any desired degree of approximation. That the situation is not quite the same with statistical experiments should be clear from the following.

Example : Let ω be the unknown probability of heads for a particular unsymmetric looking coin. One may argue that no informative (see Definition 2 in Section 2) statistical experiment concerning ω can be performed by anyone who is not in possession of the coin in queston. With the coin in possession we can plan an experiment $\mathcal{E}$ for which $\mathcal{X} = \{1, 2, 3, \ldots\}$ and $p(x\,\omega) = \omega(1-\omega)^{x-1}$, $x \in \mathcal{X}$. It is not difficult to see how we can plan a (marginal) experiment $\mathcal{E}_1$ for which $\mathcal{X}_1 = \{0, 1\}$ and $p_1(0 \mid \omega) = 1/(2-\omega)$. But can we plan an experiment $\mathcal{E}_2$ for which $\mathcal{X}_2 = \{0, 1\}$ and $p_2(0 \mid \omega) = \sqrt{\omega}$ or $\sin(\frac{1}{2}\pi\omega)$? Intuitively, we feel that such strange looking functions of ω are unlikely to appear as probabilities in 'performable' experiments. They might, and an interesting mathematical problem associated with our coin is to determine the class of functions L that can arise as likelihoods, that is

$$L(\omega) = \text{Prob}\,(A \mid \omega)$$

where A is an event defined in terms of a 'performable' experiment with the coin. But it is not easy to see how we can give a satisfactory mathematical definition of 'performability'.

If we insist on our universe of discourse to be the class of all conceivable triples $(\mathcal{X}, \Omega, p)$, then it is plausible that we shall end up with paradoxes such as those that have arisen in set theory in the past. Without labouring the point any further let us then agree that we are concerned with a rather small class $\mathfrak{E}$ of 'performable' experiments. Let this $\mathfrak{E}$ be our tongue-in-the-cheek definition of performability ! If $\mathcal{E}_1$ and $\mathcal{E}_2$ are performable experiments then

it stands to reason to claim that any mixture of $\mathcal{E}_1$, $\mathcal{E}_2$ with known mixture probabilities is also performable. In other words, we may assume that the class $\mathfrak{E}$ is convex, i.e., closed under known mixtures. It also seems reasonable to claim that our class $\mathfrak{E}$ is closed under 'marginalization', that is, if $\mathcal{E} = (\mathcal{X}, \Omega, p)$ is performable then for any statistic $T : \mathcal{X} \to \mathcal{T}$ the marginal experiment $\mathcal{E}_T = (\mathcal{T}, \Omega\, p_T)$ as defined in Section 2 is also performable. But how secure is the case for the conditional experiment $\mathcal{E}_t^T = (\mathcal{X}_t, \Omega, p_t^T)$? If T is sufficient then, for every $t \epsilon \mathcal{T}$, the conditional experiment $\mathcal{E}_t^T$ is non-informative and so is performable in a sense—the experiment can be simulated with the help of a random number table. Now, note that for a description of the general conditionality principle $\mathcal{C}$ we need to assume that for any ancillary statistic T (and every t in the range space of T) the conditional experiment $\mathcal{E}_t^T \epsilon \mathfrak{E}$. [Refer to Basu (1964) for some discussions on this assumption. In that article the author rejected the reasonableness of such an assumption and thereby sought to explain away certain anomalies that he had discovered in an unrestricted use of principle $\mathcal{C}$ in the manner advocated by R. A. Fisher. Those anomalies arose only because the author was then trying to reconcile $\mathcal{C}$ with the traditional 'sample space' analysis of data—in terms of the average performance characteristics of some inference procedures.] However, note that our description of the weaker conditionality principle $\mathcal{C}'$ and our derivation of $\mathcal{L}$ from $\mathcal{S}'$ and $\mathcal{C}'$ cannot be faulted on the ground of non-performability of any experiment. In this connection it is interesting to look back on a derivation of the above implication theorem by Hajék (1967). Not only is Hajék's proof longer and somewhat obscure, but it appears to pre-suppose (in a quite unacceptable manner) that $\mathfrak{E}$ consists of all triples $(\mathcal{X}, \Omega, p)$.

Having recognized 'information' as a function Inf with its domain as the space D of all data $d = (\mathcal{E}, x)$ with $\mathcal{E} \epsilon \mathfrak{E}$, let us finally turn our attention to the range of Inf. If we accept the likelihood principle, i.e., if we agree that

$$\mathrm{Inf}(\mathcal{E}_1, x_1) = \mathrm{Inf}(\mathcal{E}_2, x_2)$$

whenever $L(\omega \mid \mathcal{E}_1, x_1) \sim L(\omega \mid \mathcal{E}_2, x_2)$, then we may as well take a short step further and agree to view Inf as a mapping of the space D of all attainable data $d = (\mathcal{E}, x)$ onto the set Λ of all realizable likelihood functions $L = L(\omega \mid \mathcal{E}, x)$. Once again we repeat that our definition of equality on Λ is that of proportionality : $L_1 \sim L_2$ if there exists $c > 0$ such that $L_1(\omega) \equiv cL_2(\omega)$.

5. Fisher information

R. A. Fisher's controversial thesis regarding the logic of statistical inference rests on an unequivocal and complete rejection of the Bayesian point of view.

He drew the attention of the statistical community away from the Bayesian 'prior' and 'posterior' and focussed it on the likelihood function. Although we do not find the likelihood principle explicitly stated in the writings of Fisher, yet it is clear that he recognized the truth that statistical inference should be based on the 'whole of the relevant information' supplied by the data and that this information is contained in the likelihood function. However, quite a few of the many ideas formulated by Fisher are not in full accord with the above principal theme of his writings. One such idea is that of 'Fisher Information' which we discuss briefly in this section.

In the situation where the parameter of interest is a number θ belonging to an interval subset of the real line, and some regularity conditions are satisfied by $p(x \mid \theta)$ as a function of θ, the Fisher Information is defined as

$$I(\theta) = E_\theta \left\{ \frac{\partial}{\partial \theta} \log p(X \mid \theta) \right\}^2$$

$$= -E_\theta \left\{ \frac{\partial^2}{\partial \theta^2} \log p(X \mid \theta) \right\}$$

where X is regarded as a random variable ranging over $\mathscr{X}$. How did Fisher arrive at such a notion of inference that does not depend on the sample $X = x$? Has $I(\theta)$ got anything to do with the kind of information that we are talking about ? We speculate here on what might have led Fisher to the above mathematically interesting but statistically rather fruitless notion.

If $\hat{\theta} = \hat{\theta}(x)$ is the maximum likelihood estimate of θ, then the true value of θ ought to lie in some small neighbourhood of $\hat{\theta}$—at least in the large sample situation. Writing $\Lambda(\theta) = \log L(\theta)$—dealing with log-likelihood was a matter of mathematical convenience with Fisher—we can then say that

$$\Lambda(\theta) \doteq \Lambda(\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^2 \Lambda''(\hat{\theta})$$

for all θ in a small neighbourhood of $\hat{\theta}$ (where the true θ ought to be). Writing $J(\theta)$ for $-\Lambda''(\theta)$, the log-likelihood may be approximately characterized as

$$\Lambda(\theta) \doteq \Lambda(\hat{\theta}) - \frac{1}{2} (\theta - \hat{\theta})^2 J(\hat{\theta})$$

where $J(\hat{\theta})$ is (normally) a positive quantity. Now, the magnitude of the statistic $J(\hat{\theta}) = -\Lambda''(\hat{\theta})$ tells us how rapidly the likelihood function drops away from its maximum value as θ moves away from the maximum likelihood estimate. (Note that $J(\hat{\theta}) = -L''(\hat{\theta})/L(\hat{\theta})$ and this is the reciprocal of the radius of curvature of the likelihood function at its mode.) It seems clear

that Fisher recognized in $J(\hat{\theta})$ a convenient and reasonable numerical measure for the quantum of information contained in a particular likelihood function. For example, if $x = (x_1, x_2, \ldots, x_n)$ is an n-tuple of i.i.d. random variables with x_i distributed as $N(0, \sigma^2)$, then

$$\hat{\sigma}^2 = \Sigma x_i^2/n, \qquad J(\hat{\sigma}^2) = 2n/\hat{\sigma}^2$$

and the latter varies from sample to sample (as information usually should).

At some stage of the game Fisher became interested in the notion of average information available from an experiment, that is, in

$$E_\theta(J(\hat{\theta})). \qquad \ldots \quad (*)$$

It is not easy to get a neat general expression for the above, and so it seems plausible that Fisher had the inconvenient $\hat{\theta}$ in (*) substituted by θ (the true value, which ought to be near $\hat{\theta}$ anyway) and thus arriving at

$$E_\theta J(\theta) = E_\theta \left\{ -\frac{\partial^2}{\partial \theta^2} \log L(\theta \mid X) \right\}$$

$$= \sum_x \left\{ \frac{\partial}{\partial \theta} \log p(x \mid \theta) \right\}^2 p(x \mid \theta) \qquad \ldots \quad (**)$$

which is the Fisher information $I(\theta)$. At this stage one may well wonder as to whether Fisher ever thought of first re-writing $J(\hat{\theta})$ as $-L''(\hat{\theta})/L(\hat{\theta})$ and then substituting $\hat{\theta}$ by θ before computing its average value as in (**). For, in this case he would have arrived at the number zero as his average information !

6. The likelihood principle

If we adopt the Bayesian point of view, then the likelihood principle becomes almost a truism. A Bayesian looks upon the data, or rather its information content $\mathrm{Inf}(\mathcal{E}, x)$, as some sort of an operator that transforms the pattern q of his prior beliefs (about the parameter ω) into a new (posterior) pattern q^*. He formalizes the notion of a 'pattern of beliefs' about ω as a probability distribution on Ω, and postulates that probability as a 'measure of (coherent) belief' obeys the same laws as 'frequency probability' is supposed to obey. The transformation $q \to q^*$ is then effected through a formal use of the Bayes theorem (of conditional probability) as

$$q^*(\omega \mid \mathcal{E}, x) = L(\omega \mid \mathcal{E}, x) q(\omega) / \Sigma L(\omega \mid \mathcal{E}, x) q(\omega)$$
$$\sim L(\omega \mid \mathcal{E}, x) q(\omega).$$

In view of the above, a Bayesian should not have any qualms about identifying $\mathrm{Inf}(\mathcal{E}, x)$ with the likelihood function $L(\omega \mid \mathcal{E}, x)$.

Fisher was not the first statistician to look upon the sample x as a variable point in a sample space $\mathcal{X}$, but it was certainly he who made this approach

popular. He put forward the notion of 'average performance characteristics' of estimators and sought to justify his method of maximum likelihood on this basis. In the early thirties Neyman and Pearson, and then Wald (in the forties) pushed the idea of 'performance characteristics' to its natural limit. Principle $\mathscr{L}$ is in direct conflict with this neo-classical approach to statistical inference. With $\mathscr{L}$ as the guiding principle of data analysis, it no longer makes any sense to investigate (at the data analysis stage) the 'bias' and 'standard error' of point estimates, the probabilities of the 'two kinds of errors' for a test, the 'confidence-coefficients' associated with interval estimates, or the 'risk functions' associated with rules of decision making.

Principle $\mathscr{L}$ rules out all kinds of post-randomization. If, after obtaining the data d, an artificial randomization scheme (using a random number table or a modern computer) generates further data d_1, then the likelihood functions generated by d and (d, d_1) coincide (are equivalent). Since the generation of d_1 does not change the information (i.e., the likelihood function), it should not have any bearing on the inference about ω, or on any assessment of the quality of the inference actually made. Being only a principle of data analysis, $\mathscr{L}$ does not rule out the reasonableness of any pre-randomization being incorporated into the planning of experiments. However, it does follow from $\mathscr{L}$ that the exact nature of any such pre-randomization scheme is irrelevant at the data analysis stage—what is relevant is the actual outcome of the pre-randomization scheme, not its probability. [The latter appears only as a constant factor in the likelihood function eventually obtained.] This last point has a far-reaching consequence in the analysis of data produced by survey sampling. If we are not to take into account the sampling plan (the pre-randomization scheme choosing the units to be surveyed) at the data analysis stage, then we have to throw overboard a major part of the current theories regarding the analysis of survey data. [See Basu (1969) for more details regarding this.] Recently a great deal has been written on the 'randomization analysis' of experimental data. [Curiously, it was again Fisher who initiated this kind of analysis and we sometimes hear it said that this was his most important contribution to statistical theory!] Principle $\mathscr{L}$ rejects this kind of analysis of data.

No wonder then that there is so much resistance to $\mathscr{L}$ among contemporary statisticians. But it is truly remarkable how universal is the acceptance of the sufficiency principle ($\mathscr{S}$ and its variant $\mathscr{S}'$) even though, in the context of a particular experiment, the two principles $\mathscr{L}$ and $\mathscr{S}'$ are indistinguishable. The general acceptance of $\mathscr{S}$ appears to be based on a

widespread belief that the reasonableness of the principle has been mathematically justified by the Complete Class Theorems of the Rao-Blackwell vintage. Let us examine the question briefly.

In the context of some point-estimation problems, the Rao-Blackwell theorem indeed succeeds in providing a sort of decision-theoretic justification for $\mathcal{S}$. But this success is due to (i) the atypical fact that, in a point-estimation problem with a continuous parameter of interest, the action space $\mathcal{A}$ may be regarded as a convex set, and also to (ii) the somewhat arbitrary assumption that the loss function $W = W(\omega, a)$ is convex in a (the action) for each fixed $\omega \epsilon \Omega$. Now, let us formalize the notions of (i) a statistical decision problem as a quintuple

$$\mathcal{D} = (\mathcal{X}, \Omega, p, \mathcal{A}, W),$$

(ii) a non-randomized decision function as a point map of $\mathcal{X}$ into $\mathcal{A}$ and (iii) a randomized decision function as a transition function mapping points in $\mathcal{X}$ into probability measures on $\mathcal{A}$. Let $T : \mathcal{X} \to \mathcal{T}$ be a sufficient statistic for the experiment $(\mathcal{X}, \Omega, p)$. Then, for each decision function δ, we can find an equivalent (in the sense that they generate identical risk functions) decision function δ^* which depends on the sample x only through its T-value $T(x)$. But the snag in this kind of Rao-Blackwellization is that δ^* will typically be a randomized decision function and so its use for decision making will entail a direct violation of $\mathcal{S}$ (which is nothing but a rejection of all post-randomizations). How can a principle be justified by an argument that invokes its violation ? !

It is difficult to understand why among contemporary statisticians the support for $\mathcal{S}$ is so overwhelming and unequivocal, and yet that for $\mathcal{L}$ is so lukewarm. In a joint paper with Jenkins and Winsten, it was argued by Barnard (1962) that $\mathcal{S}'$ implies $\mathcal{L}$. Although this attempted deduction of $\mathcal{L}$ from $\mathcal{S}'$ turned out to be fallacious, the fact remains that even as late as 1962 Barnard found it hard to distinguish between the twin principles of sufficiency and likelihood. [In the writings of Fisher also it is very hard to find an instance where he has stated $\mathcal{L}$ separately from $\mathcal{S}$. It seems to the author that Fisher always meant by a sufficient statistic T the minimal sufficient statistic and invariably visualized it as that characteristic of the sample knowing which the likelihood function can be determined upto an equivalence.] In view of the many 'unpleasant' consequences of $\mathcal{L}$, Barnard seems to have lost a great deal of his early enthusiasm for $\mathcal{L}$ though his conviction in $\mathcal{S}$ remains unshaken. Birnbaum (1962) deduced $\mathcal{L}$ from $\mathcal{S}'$ and $\mathcal{C}'$ and stated that $\mathcal{S}'$ can be deduced from $\mathcal{C}'$, implying thereby that $\mathcal{C}'$ implies $\mathcal{L}$. In 1962 Birnbaum found in $\mathcal{C}'$ a statistical principle that is almost axiomatic in its import and was,

therefore, duly impressed by $\mathscr{L}$ which he (mistakenly) thought to be a logical equivalent of $\mathscr{C}'$. At present Birnbaum too seems to have lost his earlier enthusiasm for $\mathscr{L}$, though it is not clear to the author whether his conviction in $\mathscr{C}'$ has suffered in the process or not.

Let us look back on the simplest (and perhaps the least controversial) of the eight principles stated in Section 3, namely, the invariance principle. To the author, principle $\mathscr{I}$ seems axiomatic in nature. Yet one may argue that $\mathscr{I}$ is far from convincing under the following circumstances. Let $\mathscr{E}_1 = (\mathscr{X}_1, \Omega, p_1)$ and $\mathscr{E}_2 = (\mathscr{X}_2, \Omega, p_2)$ be two statistically isomorphic or similar experiments with $g : \mathscr{X}_1 \to \mathscr{X}_2$ as the similarity map. Principle $\mathscr{I}$ then asserts the equality

$$\text{Inf}(\mathscr{E}_1, x_1) = \text{Inf}(\mathscr{E}_2, x_2)$$

for each $x_1 \epsilon \mathscr{X}_1$ and $x_2 \epsilon \mathscr{X}_2$ such that $x_2 = gx_1$. Now, suppose the sample space $\mathscr{X}_1$ is endowed with an order structure that is in some way related to some natural order structure in the parameter space Ω, whereas the sample space $\mathscr{X}_2$ has no such discernable order structure. For example, suppose $\mathscr{X}_1$ consists of the six numbers 1, 2, 3, 4, 5 and 6 whereas $\mathscr{X}_2$ consists of the six qualities R (red), W (white), B (black), G (green), Y (yellow) and V (violet). If the statistician feels that he knows how to 'relate' the points in $\mathscr{X}_1$ with the unknown ω in Ω, and if he also feels that he does not know how to 'relate' the points in $\mathscr{X}_2$ with points in Ω (excepting through what he knows about the similarity map $g : \mathscr{X}_1 \to \mathscr{X}_2$), then he may 'feel more informed' about ω when $\mathscr{E}_1$ is performed resulting in x_1 than when $\mathscr{E}_2$ is performed resulting in $x_2 = gx_1$. When it comes to a matter of feeling, not much can be done about it. It is however difficult to see how one can build up a coherent theory of 'information in the data' that will allow one to discriminate between the data $(\mathscr{E}_1, x_1)$ and its g-image $(\mathscr{E}_2, gx_1)$, where g is a similarity map.

Perhaps the point can be emphasized more forcefully in terms of the weak invariance principle $\mathscr{I}'$. [In Section 3 we recognized $\mathscr{I}'$ as a corollary to both $\mathscr{I}$ and the sufficiency principle.] If the two points x and x' in the sample space of the experiment $\mathscr{E} = (\mathscr{X}, \Omega, p)$ generate identical likelihood functions, i.e., if $p(x \mid \omega) = p(x' \mid \omega)$ for all $\omega \epsilon \Omega$, then $\mathscr{I}'$ asserts the equality of $\text{Inf}(\mathscr{E}, x)$ and $\text{Inf}(\mathscr{E}, x')$. Now, a statistician with a strong intuitive feeling for the relevance of 'related order structures' in $\mathscr{X}$ and Ω will perhaps rebel against principle $\mathscr{I}'$ if he is confronted with the following kind of a situation. Suppose the statistical problem is the traditional one of testing a simple null-hypothesis H_0 about the probability distribution of a one-dimensional random variable X on the basis of the experiment $\mathscr{E}$ that consists of taking a single observation on X. Let $(\mathscr{X}, \Omega, p)$ be a suitable statistical model (for the

experiment $\mathcal{E}$) that subsumes H_0 as the hypothesis $\omega = \omega_0$. Consider now the case where we recognize two points x and x' in $\mathcal{X}$ such that they both generate identical likelihood functions and yet x is near the centre (say, the mean) of the distribution of X under H_0 whereas x' is out at the right tail-end (say, the 1% point) of the same distribution. Notwithstanding $\mathcal{I}'$, which asserts that x and x' are equally informative, our statistician (with the strong intuition) may well assert that x (being near the centre of X under H_0) sort of confirms H_0, whereas x' (being out in the tail area) sort of disproves the null-hypothesis !

One may take an uncharitable view about the above kind of discriminatory feeling and lightly dismiss the whole matter as a prejudice that has been nurtured in the classical practice of null-hypothesis testing (formulated without any explicit mention of the plausible alternatives). It will, however, be charitable to concede that in great many situations it is true that points in the tail-end of the distribution of X under H_0 differ greatly in their information aspects from points in the centre part of the same distribution. We should also concede that our formulation of the equality of statistical information in the data $(\mathcal{E}, x)$ and $(\mathcal{E}, x')$ was made relative to a particular model $(\mathcal{X}, \Omega, p)$ for the experimental part of the data. It is now plausible to suggest that our statistician (with the strong intuition) is not really rejecting principle $\mathcal{I}'$ in the present instance, but is only doubting the adequacy or appropriateness of the particular statistical model $(\mathcal{X}, \Omega, p)$.

This points to the very heart of the difficulty. All statistical arguments are made relative to some statistical model and there is nothing very sacred and irrevocable about any particular model. When an inference is made about the unknown ω, the fact should never be lost sight of that, with a different statistical model for $\mathcal{E}$, the same data $(\mathcal{E}, x)$ might have warranted a different inference. No particular statistical model is likely to incorporate in itself all the knowledge that the experimenter may have about the 'related order structure' or any other kind of relationship that may exist between the sample space and the parameter space. But if we agree to the proposition that our search for the 'whole of the relevant information in the data' must be limited to within the framework of a particular statistical model, then the author is hard put to find any cogent reason, for not identifying the 'information in the data' with the likelihood function generated by it. If in a particular instance the experimenter feels very upset by the look of the likelihood function generated by the data, then he may (and indeed should) re-examine the validity and adequacy of the model itself. A strange-looking likelihood function does not necessarily destroy the likelihood principle. [Later on, we shall take up several such cases of apparent likelihood principle paradox.]

On p. 334 of Barnard, Jenkins and Winsten (1962) we find the following astonishing assertion which is in the nature of a blank cheque for all violations against $\mathscr{L}$. [In this and in the following two quotations from Barnard, we have taken the liberty of slightly altering the notations so as to bring them in line with those in this article.]

"In general, it is only when the triplet $(\mathscr{X}, \Omega, p)$ can by itself be regarded as specifying all the inferential features of an experimental situation that the likelihood principle applies. If $\mathscr{X}$ and Ω are provided with related ordering structures, or group structures, or perhaps other features, it may be reasonable to apply a form of argument which would not apply if these special features were not present. The onus will, of course, be on anyone violating the likelihood principle to point to the special feature of this experiment and to show that it justifies his special argument."

Does it ever happen that a triple $(\mathscr{X}, \Omega, p)$ specifies 'all the inferential features' of an experimental situation ? Can any experimenter be ever so dumb as not to be able to recognize some 'related order structure or group structures or perhaps other features' connecting $\mathscr{X}$ and Ω ? If we are to take the above assertion at its face value, then we must conclude that under hardly any circumstances is Barnard willing to place his immense authority unequivocally behind the likelihood principle ! As a discussant of Birnbaum (1962, p. 308), Barnard made the point once again as follows :

"The qualification concerns the domain of applicability of the principle of likelihood. To my mind, this applies to those situations, and essentially to only those situations, which are describable in terms which Birnbaum uses —that is, in terms of the sample space $\mathscr{X}$, and the parameter space Ω and a probability function p of x and ω defined for x in $\mathscr{X}$ and ω in Ω. If these elements constitute the whole of the data of a problem, then it seems to me the likelihood principle is valid. But there are many problems of statistical inference in which we have less than this specified, and there are many other problems in which we have more than this specified. In particular, the simple tests of significance arise, it seems to me, in situations where we do not have a parameter space of hypotheses; we have a single hypothesis essentially, and the sample space then is the only space of variables present in the problem. The fact that the likelihood principle is inconsistent with significance test procedures in no way, to my mind, implies that significance tests should be thrown overboard; only that the domain of applicability of these two ideas should be carefully distinguished. We also, on the other hand, have situations where more is given than simply the sample space and the parameter space. We may have properties of invariance, and such things, which enable us to

make far wider, firmer assertions of a different type; for example, assertions that produce a probability when these extra elements are present. And then, of course, there are the decision situations where we have .loss functions and other elements given in the problem which may change the character of the answers we give".

If, following Barnard, we set up the test of significance problem in the classical manner of Karl Pearson and R. A. Fisher—with a single probability distribution on the sample space and without any tangible parameter space——then the sample will not produce any likelihood function. Without a likelihood function how can we possibly violate principle $\mathscr{L}$? In the other kind of situations, where we have 'invariance and such other things', Barnard says that we can make assertions that are 'far wider and firmer'. But, wider and firmer than what ? What does $\mathscr{L}$ assert that is not sufficiently firm or wide ? We must recognize this basic fact that $\mathscr{L}$ does not assert anything that can be measured in terms of its operating characteristics. It appears that in this instance Barnard is confusing principle $\mathscr{L}$ with a set of his favourite likelihood methods of inference (see Section 7) and it is this set of likelihood methods that he is now finding to be generally lacking in width and firmness. Before returning to the question of the true implication of $\mathscr{L}$, let us quote once again from Barnard and Sprott (1971, p. 176) :

"$\mathscr{L}$ applies to problems for which the model consists of a sample space $\mathscr{X}$, a parameter space Ω and a family of probability functions $p : \mathscr{X}\times\Omega\rightarrow R^{+}\ldots$. For two such problems $(\mathscr{X}, \Omega, p)$ and $(\mathscr{X}', \Omega, p')$, principle $\mathscr{L}$ asserts that if $x\epsilon\mathscr{X}$ and $x'\epsilon\mathscr{X}'$ and $p(x\,|\,\omega)/p'(x'\,|\,\omega)$ is independent of ω, then the inference from x must be the same as the inference from x'. We may distinguish three forms of $\mathscr{L}$:

1. *Strongly restricted* $\mathscr{L}$: Principle $\mathscr{L}$ applicable only if $(\mathscr{X}, \Omega, p) = (\mathscr{X}', \Omega, p')$. This is equivalent to the sufficiency principle.

2. *Weakly restricted* $\mathscr{L}$: Principle $\mathscr{L}$ applicable (a) whenever $(\mathscr{X}, \Omega, p) = (\mathscr{X}', \Omega, p')$ and (b) when $(\mathscr{X}, \Omega, p) \neq (\mathscr{X}', \Omega, p')$ but there are no structural features of $(\mathscr{X}, \Omega, p)$ (such as group structures) which have inferential relevance and which are not present in $(\mathscr{X}', \Omega, p')$.

3. *Unrestricted* $\mathscr{L}$: Principle $\mathscr{L}$ applicable to all situations which can be modelled as above.

4. *Totally unrestricted* $\mathscr{L}$: As in 3, but, further, all inferential problems are describable in terms of the model given.

As we understand the situation, almost everyone would accept 1, while full Bayesians would accept 4. George Barnard's own position is now, and has been since 1957, 2."

The distinction that Barnard is trying to make above between the two forms (3 and 4) of unrestricted $\mathscr{L}$ is not clear and is perhaps not relevant to our present discussion. In 1 Barnard recognizes the equivalence of $\mathscr{S}$ and $\mathscr{L}$ in the context of a single experiment and appears to have no reservations about $\mathscr{S}$. But in 2 we once again come across the same astonishing blank cheque phrased this time in terms of the all-embracing double negatives : 'there are no structural features...which are not present'.

Later on, we shall discuss in some detail the two principal sources of Barnard's discomfiture with the unrestricted likelihood principle—the Stein Paradox and the Stopping Rule Paradox. For the moment, let us briefly discuss what we consider to be the real implication of $\mathscr{L}$.

Apart from identifying the information content of the data $(\mathscr{E}, x)$ with the likelihood function $L(\omega \mid \mathscr{E}, x)$ generated by it, principle $\mathscr{L}$ tells us hardly anything else. It certainly does not tell us how to make an inference (based on the likelihood function) in any particular situation. It is best to look upon $\mathscr{L}$ as a sort of code of conduct that ought to guide us in our inference making behaviour. In this respect it is analogous to the unwritten medical code that requires a Doctor to make his diagnosis and treatment of a patient dependent wholly on (i) the case history of and the outcomes of some diagnostic tests carried out on that particular patient, and (ii) all the background information that the Doctor (and his consultants) may have on the particular problem at hand. It is this same unwritten code that disallows a Doctor to include a symmetric die or a table of random numbers as a part of his diagnostic gadgets. It also forbids him to allow his judgement about a particular patient to be coloured by any speculations on the types and number of patients that he may have later in the week. [Of course, like any other rule the above must also have its exceptions. For instance, if our Doctor in a far away Pacific island is running short of a drug that is particularly effective against a prevalent disease, he may then be forgiven for treating a less severely affected patent in an unorthodox manner.]

In the colourful language of J. Neyman, the making of inference is nothing but an 'act of will'. And this act is no more (and no less) objective than that of a medical practitioner making his routine diagnoses. We are all too familiar with the beautiful mathematical theory of Neyman-Pearson-Wald about what is generally recognized as correct inductive behaviour. In principle $\mathscr{L}$ we recognize only a preamble to an anti-thesis to the currently popular N.P.W. thesis. [For a well-stated version of $\mathscr{L}$ from the Bayesian point of view, refer to Lindley (1965, p. 59) or Savage (1961).]

PART 2 : METHODS

7. NON-BAYESIAN LIKELIHOOD METHODS

In Part I of this article our main concern was with the notion of statistical information in the data, and with some general principles of data analysis. Now we turn our attention from principles to a few methods of data analysis. By a *non-Bayesian likelihood method* we mean any method of data analysis that neither violates $\mathscr{L}$—the likelihood principle—nor explicitly incorporates into its inference-making process any prior information (that the experimenter may have about the parameter ω) in the form of a prior probability distribution over the parameter space Ω. The origin of most of such methods may be traced back to the writings of R. A. Fisher. In this section we list several such methods. To fix our, ideas let us suppose that Ω is either a discrete or an interval subset of the real line. In the latter case, we shall also suppose that the lkelihood function $L(\omega)$ is a smooth function and has a single mode (whenever such an assumption is implicit in the method) and so on.

(a) *Method of maximum likelihood* : Estimate the unknown ω by that point $\hat{\omega} = \hat{\omega}(x)$ where the likelihood function $L(\omega)$, generated by the data $(\mathscr{E}, x)$, attains its maximum value. Fisher tried very hard to elevate this method of point estimation to the level of a statistical principle. Though it has since fallen from that high pedestal, it is still widely recognized as the principal method of point estimation. Note that this method is in conformity with $\mathscr{L}$ as long as we do not try to understand and evaluate the precision of the maximum likelihood estimate $\hat{\omega} = \hat{\omega}(x)$ in terms of the sampling distribution of the 'estimator' $\hat{\omega}$. However, most users of this method quite happily violate $\mathscr{L}$ in order to do just that.

(b) *Likelihood interval estimates* : Choose and fix a fairly large number λ (20 or 100 are usually recommended values) and consider the set

$$I_\lambda = \{\omega : L(\hat{\omega})/L(\omega) \leqslant \lambda\}$$

where $\hat{\omega}$ is the maximum likelihood estimate of ω. If the likelihood function is unimodal then the set I_λ is a sub-interval of Ω and is intended to be used as a sort of 'likelihood confidence interval' for the parameter ω.

(c) *Likelihood test of a null-hypothesis* : If the null-hypothesis to be tested is defined as H_0 = Hypothesis that $\omega = \omega_0$, then the method is : Reject H_0 if and only if ω_0 does not belong to the likelihood interval I_λ defined in (b) above. As before, 20 or 100 are recommended values. [The numbers 20 and 100 correspond roughly to the mystical 5% and 1% of the classical tests of significance.]

(d) *Likelihood ratio method*: If Ω consists of exactly two points ω_0 and ω_1 then $\mathscr{L}$ implies that the likelihood ratio $\rho = L(\omega_1)/L(\omega_0)$ generated by the data $(\mathscr{E}, x)$ should provide the sole basis for making judgements about whether the true ω is ω_0 or ω_1. The method is : Choose and fix λ (20 or 100 say) and then reject the hypothesis $\omega = \omega_0$ if $\rho \geqslant \lambda$, and accept the hypothesis $\omega = \omega_0$ if $\rho \leqslant \lambda^{-1}$, but do not make any judgement if $\lambda^{-1} < \rho < \lambda$. Wald's method of sequential probability ratio test is really an outgrowth of the above. However, in a later section we shall discuss how principle $\mathscr{L}$ is frequently violated in Wald's analysis of sequentially observed data.

(e) *General likelihood ratio method*: In a general testing situation with two composite hypotheses

$$H_0 = \text{Hypothesis that } \omega \epsilon \Omega_0 \subset \Omega$$

and

$$H_1 = \text{Hypothesis that } \omega \epsilon \Omega_1 = \Omega - \Omega_0,$$

the method requires computation of a ratio statistic $\rho = \rho(x)$, defined as the ratio $L(\Omega_1)/L(\Omega_0)$

$$L(\Omega_i) = \sup_{\omega \epsilon \Omega_i} L(\omega) \qquad (i = 0, 1)$$

and then rejecting the null-hypothesis H_0 if and only if the ratio ρ is considered to be too large—greater than a pre-fixed critical value λ. [This method, along with the methods (b), (c) and (d) given above, draws its inspiration from the maximum likelihood method of point situation.] The method has great practical (computational) advantages when the basic statistical model is that of a multivariate normal distribution (with some unknown parameters). Indeed, a major part of the classical theory of multivariate analysis is nothing but a systematic exploitation of the method in a variety of situations. We should not however lose sight of the fact that in these applications of the method the critical value λ for the ratio ρ is determined (almost universally) with reference to the sampling distribution (under H_0) of the ratio statistic ρ and that this constitutes (almost invariably) a violation of $\mathscr{L}$.

(f) *Nuisance parameter elimination method*: Consider the situation where $\omega = (\theta, \phi)$, θ is the parameter of interest and, therefore, ϕ is the nuisance parameter. From the data $(\mathscr{E}, x)$ we have a likelihood function $L(\theta, \phi)$ that involves the nuisance parameter. The following is a very popular method of eliminating ϕ from L. Maximise $L(\theta, \phi)$ w.r.t. ϕ thus arriving at the eliminated likelihood function

$$L_e(\theta) = \sup_{\phi} L(\theta, \phi)$$

where e denotes the fact of elimination. Having eliminated ϕ from the likelihood function, the method then requires that all inferences about θ should be carried out with the eliminated likelihood function $L_e(\theta)$ along the lines suggested earlier. Method (f) may be looked upon as a natural generalization of method (e).

Let us end this section with a few comments on some common features of these methods.

(i) For going through the motions of any of these methods, it is not necessary to know any details of the sample x other than the likelihood function generated by it. In their pure (that is, uncontaminated by the Neyman-Pearson type arguments) forms, the methods are in conformity with principle $\mathscr{L}$. However, it should be borne in mind that none of the above methods can be logically deduced from $\mathscr{L}$ by itself.

(ii) In none of the methods we find any mention of the two elements q and Π that we briefly talked about in Section 1. Let us recall that in q we have incorporated all the background (prior) information that the experimenter has about ω and other related entities. In Π is incorporated all other particular features (such as, the relative hazards of making wrong inferences of various kinds etc.) of the inferential problem at hand. The likelihood methods of this section differ from standard Bayesian methods mainly in their failure (rather, refusal) to recognize the relevance of q and Π.

(iii) In their pure forms, these methods do not require the evaluation of the average performance characteristics of anything. This, however, does not mean that we should not speculate about long term characteristics of such methods. Advocates of likelihood methods are surely not averse to the idea of comparing their methods with any other well-defined method on the basis of their average performance characteristics in a hypothetical sequence of repeated applications of the methods. [Even Bayesians, who do not usually care for the frequency interpretation of probability, do care very much about one kind (perhaps, the only kind that is relevant) of frequency, namely, the long term success ratio of their methods. After all, the real proof of the pudding lies in the eating.] From our description of the non-Bayesian likelihood methods, it is not clear with what kind of average performance characteristics in mind these methods were initially proposed. Indeed, in some later sections we shall give examples of situations where simple-minded applications of these methods will have disastrous long-term performance characteristics. Such examples will not, however, disprove $\mathscr{L}$ because the methods do not follow from $\mathscr{L}$ by itself.

(iv) The differences between the Bayesian and the (non-Bayesian) Likelihood schools of data-analysis may be summarised as follows : Whereas, the Bayesian looks upon the likelihood function $L(\omega)$ as an intermediate step —a link between the prior and the posterior—the Likelihoodwallah* looks upon $L(\omega)$ as a sort of an end in itself. Furthermore, the latter looks upon $L(\omega)$ as a point function —$L(\omega)$ is the relative magnitude (or intensity) with which the data supports the point ω—that should never (well, almost never) be looked upon as something that can generate a measure of support (for subsets of Ω that are not single-point sets). In the next section we discuss this point in some detail.

8. Likelihood —a point-function or a measure?

It was R. A. Fisher who first thought of likelihood as an alternative measure of rational belief. The following quotation clearly spells out Fisher's own ideas on the subject. [These remarks of Fisher appear to have greatly influenced the thinking processes of many of our contemporary statisticians.] Discussing the likelihood function, Fisher (1930, p. 532) wrote :

"The function of the θ's maximised is not however a probability and does not obey the laws of probability; it involves no differential element $d\theta_1\, d\theta_2\, d\theta_3 \ldots$; it does none the less afford a rational basis for preferring some values of θ, or combination of values of the θ's, to others. It is, just as much as a probability, a numerical measure of rational belief, and for that reason called the likelihood of $\theta_1, \theta_2, \theta_3, \ldots$ having given values, to distinguish it from the probability that $\theta_1, \theta_2, \theta_3, \ldots$ lie within assigned limits, since in common speech both terms are loosely used to cover both types of logical situation.

If A and B are mutually exclusive possibilities the probability of "A or B" is the sum of the probabilities of A and of B, but the likelihood of A or B means no more than "the stature of Jackson or Johnson", you do not know what it is until you know which is meant. I stress this because inspite of all the emphasis that I have always laid upon the difference between probability and likelihood there is still a tendency to treat likelihood as though it were a sort of probability.

The first result is that there are two different measures of rational belief appropriate to different cases. Knowing the population we can express our incomplete knowledge of, or expectation of, the sample in terms of probability;

* 'Wallah' in Hindi means a peddler and is a non-derogatory term. The name, Likelihoodwallah, then denotes a peddler of an assortment of non-Bayesian likelihood methods.

knowing the sample we can express our incomplete knowledge of the population in terms of likelihood. We can state the relative likelihood that an unknown correlation is +0.6, but not the probability that it lies in the range .595-.605".

From the above it is clear that Fisher intended his notion of likelihood to be used as some sort of a measure of (the degree of) rational belief. But all the same he was very emphatic in his denial that likelihood is not a measure like probability—it is not a set function but only a point function. It is not however clear why this data-induced likelihood measure of rational belief (about various simple hypotheses related to the populaton) must differ from the other measure of rational belief (namely, probability) in being non-additive. Why can't we talk of the likelihood of a composite hypothesis in the same way we talk about the probability of a composite event ?

In our quotation we find Fisher lightly dismissing the question with the curious analogy of "the stature of Jackson or Johnson, you do not know what it is until you know which is meant". Twentysix years later we find Fisher (1956, p. 69) still persisting with the same analogy—only this time it was "the income of Peter or Paul". These analogies are particularly inept and misleading. Both stature and income are some kind of measure—the former of size and the latter of earning power. Why can't we talk of the total stature or the total income of a group of people ? It should be noted that when Fisher is talking of 'Jackson or Johnson' he is using the conjunction 'or' in its everyday disjunctive sense of 'either-or'. On the other hand, when we talk about the degree of rational belief (probability or likelihood) in 'A or B' the 'or' is the logical (set-theoretic) connective 'and/or' (union).

Ian Hacking (1965) in his very interesting and informative book, *Logic of Statistical Inference*, has given a detailed and eminently readable account of how this Fisher-project of building an alternative likelihood framework for a measure of 'rational belief' may be carried out. The expression 'rational belief' sounds a little awkward in the present context as the whole exercise is about a mathematical theory of what 'the data has to tell' rather than about what 'the experimenter ought to believe'. Hacking therefore suggests an alternative expression, 'support-by-data'. About this theory of 'support' Hacking (1965, p. 32) writes :

"The logic of support has been studied under various names by a number of writers. Koopman called it the logic of intuitive probability; Carnap of confirmation. Support seems to be the most general title. ... I shall use only the logic of comparative support, concerned with assertions that one proposition

is better or worse supported by one piece of evidence, than another proposition is by other or the same evidence. ... . The principles of comparative support have been set out by Koopman; the system of logic which he favours will be called Koopman's logic of support".

The Fisher-project of building an alternative likelihood framework for 'support-by-data' is then carried out by Hacking as follows. Hacking begins with Koopman's postulates of intuitive probability—the logic of support—and enriches it with an additional postulate, which he calls the Law of Likelihood. A rough statement of the law may be given as follows :

Law of likelihood : Of two hypotheses that are consistent with given data, the better supported (by the data) is the one that has greater likelihood.

In terms of our notations, the Law tells us the following : If $L(\omega_1) > L(\omega_2)$ then the data $(\mathcal{E}, x)$ supports the hypothesis $\omega = \omega_1$ better than the hypothesis $\omega = \omega_2$. The Law sets up a linear order on the parameter space Ω. Any two simple hypotheses $\omega = \omega_1$ and $\omega = \omega_2$ may be compared on the basis of the intensity of their support by the data. But how about composite hypotheses like $\omega = \omega_1$ or ω_2 ? Suppose $A = \{\omega_1, \omega_2\}$ and $B = \{\omega_1', \omega_2'\}$ and suppose further that $L(\omega_i) > L(\omega_i')$, $i = 1, 2$. Would the statistical intuition of Sir Ronald have been outraged by the suggestion that, under the above circumstances, it is right to say that the data supports the hypothesis $\omega \epsilon A$ better than the hypothesis $\omega \epsilon B$?. The author thinks not.

At the risk of scandalizing some staunch admirers of Sir Ronald, the author now suggests a stronger version of Hacking's law of likelihood.

The strong law of likelihood : For any two subsets A and B of Ω, the data supports the hypothesis $\omega \epsilon A$ better than the hypothesis $\omega \epsilon B$ if

$$\sum_{\omega \epsilon A} L(\omega) > \sum_{\omega \epsilon B} L(\omega).$$

[Let us recall the assumption (rather, assertion) in Section 2 that all our sets (the sample space, the parameter space etc.) are finite. Because of this we run into no definition trouble.] Before looking into the possibility of any inconsistencies that may arise out of this Strong Law of Likelihood, let us consider some of its consequences.

With the Strong Law of Likelihood incorporated into Koopman's logic of support, we can now identify the notion of 'support-by-data' for the hypothesis $\omega \epsilon A$ with its likelihood $L(A)$ defined as

$$L(A) = \sum_{\omega \epsilon A} L(\omega).$$

Given a data d, its support for various hypotheses about the population is then a true measure—the likelihood measure of Fisher. Since a scaling factor in the likelihood function does not alter its character, we may as well work with the standardized likelihood function

$$\bar{L}(\omega) = L(\omega)/L(\Omega),$$

and then the corresponding set function $A \to \bar{L}(A)$ gets endowed with all the characteristics of a probability measure.

No Likelihoodwallah can possibly object to our scaling of the likelihood to a total of unity. They can however challenge the Strong Law of Likelihood. But observe that the Strong Law is nothing but the Law of Likelihood (which all Likelihoodwallahs accept) together with an additivity postulate for the logic of support-by-data. [It should be noted that the additivity postulate is not in the set that Hacking (1965, p. 33) borrowed from Koopman's logic of intuitive probability. However, in a later part (Chapter IX) of his book, Hacking introduced this postulate in his logic of support with a view to developing the idea as a sort of "consistent explication of Fisher's hitherto inconsistent theory of fiducial probability". The author had difficulties in following this part of Hacking's arguments.] One may ask : "How can you assume that data support hypotheses in an additive fashion ?" But then the same question may be asked about the other postulates also.

The author is willing to postulate additivity because (i) it is not in conflict with his own intuition on the subject, (ii) it makes the logic of support neat and useful, but mainly because (iii) he does not know how to 'prove' it ! The author is not a logician. The long-winded 'proofs' that some subjective probabilists give about the additivity of their measure of 'rational belief' leave the author bewildered and bemused. He finds it a lot easier to accept additivity as a primary postulate for probability. When it comes to likelihood (a measure of support-by-data) he finds it equally easy to accept it as additive. If we can accept that the mind of a rational *homo sapien* ought to work in an additive fashion when it comes to his pattern of belief in various events, why can't we also accept that the inanimate data should lend its support to various hypotheses in a similarly additive manner ? Let us not forget that Fisher used the term 'rational belief' and not 'support-by-data'. The 'belief' of what rational mind was he contemplating ? Certainly, not that of the statistician (experimenter). Because he is a rational being, the experimenter cannot (and must not) forget all the other (prior) information that he has on the subject. It seems Fisher was contemplating an extremely intelligent being—a Martian perhaps—who at the same time is totally devoid of any background

information about ω other than what is contained in the description of the statistical model $(\mathcal{X}, \Omega, p)$ for the experiment $\mathcal{E}$ and the data $(\mathcal{E}, x)$. Our intelligent Martian objectively weighs all the evidence given by the data and then makes up his own mind about the various possibilities related to ω. Fisher wanted to distinguish this posterior pattern of the Martian's 'rational belief' with the ordinary kind of 'rational belief', which we call probability, by calling the former likelihood. But why did he insist so vehemently that likelihood is not additive ?

The answer lies in Fisher's preoccupation with the illusory notions of the infinite and the infinitesimal. Suppose we have formulated in our mind an infinite set of hypotheses $H_1, H_2, H_3 \ldots$ and suppose our experiment is the trivial one of tossing a symmetric coin once, resulting in the sample H (= head). Now, the data equally support each member of our infinite set of hypotheses. There is no difficulty in visualising the likelihood as a nice, flat point-function. But how can we convert this into an ordinary kind of a probability measure ? Even Hacking, the logician, seems to have been taken in by the force of this argument. On p. 52 of his book Hacking writes : "Likelihood does not obey Kolmogoroff's axioms. There might be continuously many possible hypotheses; say, that $P(H)$ lies anywhere on the continuum between 0 and 1. On the data of two consecutive heads, each of this continuum of hypotheses (except $P(H) = 0$) has likelihood greater than zero. Hence the sum of the likelihoods of mutually exclusive hypotheses is not 1, as Kolmogoroff's axioms demand; it is not finite at all".

The author finds the above remark all the more surprising because in the very next paragraph Hacking writes : "... , in any real experimental situation, there are only a finite number of possible outcomes of a measurement of any quantity, and hence a finite number of distinguishable results from a chance set-up. Continuous distributions are idealizations." If Hacking is willing to concede that all sample spaces are in reality only finite, why does he not agree to the proposition that the parameter space also is in reality only finite ?

A finite and, therefore, realistic version of the Hacking-idealization of the parameter $\theta = P(H)$ lying "anywhere on the continuum between 0 and 1" may be set up as follows : Stipulate that θ varies over some finite and evenly spread out set like $J = \{.00, .01, .02, \ldots .99, 1.00\}$. On the basis of the data (of two consecutive heads in two throws) our Martian then works out his likelihood measure over the set J in terms of the standardized likelihood function $\bar{L} : J \to [0, 1]$ defined as

$$(*) \qquad \bar{L}(\theta) = \theta^2 / \sum_{\theta \in J} \theta^2,$$

Now, the above discrete likelihood measure can be reasonably (and rather usefully) approximated by a continuous (likelihood) distribution over the unit interval [0, 1] that is defined by the density function

$$(**) \qquad \bar{l}(\theta)d\theta = 3\theta^2 d\theta$$

Note that the (true) likelihood function $\bar{L}(\theta)$ in (*) has no differential element attached to it, whereas its idealized counterpart in (**) has. In order to avoid the logical hazards of the infinitesimal, it is better to look upon the density function $\bar{l}(\theta)$ only as a convenient tool and nothing else.

Now, let us examine how our clever but very ignorant Martian reacts to a re-statement of the statistical model in terms of a transformation of the parameter θ. Suppose we write $\phi = \theta^2$ and describe the model in terms of the parameter ϕ. In order to be consistent with our earlier stipulation that $\theta \epsilon J$, we have to inform the Martian that $\phi \epsilon J_1$ where $J_1 = \{(.00)^2, (.01)^2, \ldots (.99)^2, (1.00)^2\}$. Looking at the data of two consecutive heads, the Martian will now arrive at his likelihood measure on J_1 on the basis of the standardized likelihood function $\bar{L}_1$ defined as

$$\bar{L}_1(\phi) = \phi / \sum_{\Phi \epsilon J_1} \phi.$$

And this measure on J_1 is entirely consistent with the measure on J obtained earlier in (*). In view of the fact that the set J_1 is not evenly spread out over the interval [0, 1], the idealized limiting version of the above discrete distribution on J_1 is not given by the density $2\phi\, d\phi$ but by the natural progeny of (**) obtained in the usual manner as

$$\bar{l}_1(\phi)d\phi = \bar{l}_1(\theta) \left| \frac{d\theta}{d\phi} \right| d\phi$$

$$= \frac{3}{2} \sqrt{\phi}\, d\phi, \qquad 0 \leqslant \phi \leqslant 1.$$

It should be noted that the function $\bar{l}_1(\phi) = \frac{3}{2}\sqrt{\phi}$ has no likelihood interpretation as a point function. However, for reasonable sets A, the integral $\int_A \bar{l}_1(\phi)d\phi$ may be interpreted as the likelihood of the hypothesis $\phi \epsilon A$ but then only as an approximation.

At this point one may ask the question : "Why is it that the Martian is reacting differently to the two parametrizations of the model in terms of θ

and ϕ ?" In the first case we find that the likelihood function $\bar{L}(\theta)$ is proportional to the likelihood density $\bar{l}(\theta)$. But in the second case the two functions $\bar{L}_1(\phi)$ and $\bar{l}_1(\phi)$ are not proportional. The answer lies of course in the fact that the parameter spaces J and J_1 are differently oriented. Suppose, instead of telling the Martian that $\phi \epsilon J_1$, we leave him to his own devices with the vague assertion that ϕ lies somewhere in the continuous interval [0, 1]. Now the computer-like mind of the Martian will immediately translate our vague (infinitesimal) statement about ϕ into a finite (realistic) statement like $\phi \epsilon J = \{.00, .01, \ldots .99, 1.00\}$ and proceed to evaluate the evidence of the data in precisely the same way as he did for θ. His likelihood function $\bar{L}_2 : J \rightarrow (0, 1]$ will now be defined as

$$(*) \qquad \bar{L}_2(\phi) = \phi / \sum_{\Phi \epsilon J} \phi$$

and its idealized continuous version will be described in terms of the density function

$$\bar{l}_2(\phi)d\phi = 2\phi d\phi, \quad 0 \leqslant \phi \leqslant 1.$$

The fact that the density function $\bar{l}_2(\phi)d\phi$ is not consistent with the density function $\bar{l}(\theta)d\theta$ was the principal reason why Fisher rejected the idea of likelihood as an additive measure. His mind probably worked in the following fashion : The map $\theta \rightarrow \theta^2 = \phi$ sets up a one-one correspondence between the intervals [0, 1] and [0, 1]. The statements $\theta \epsilon [0, 1]$ and $\phi \epsilon [0, 1]$ are therefore *equivalent in every way.* If on the basis of equivalent background information the Martian is liable to arrive at different (inconsistent) measures of rational belief, then it is clear that we cannot trust his methods for converting the likelihood function into an additive measure. It is therefore safer to regard likelihood only as a point function. This way we cannot possibly land ourselves into paradoxes of the above kind.

Let us analyse the flaw in the above argument. The assertion that $\theta \rightarrow \phi$ is a one-one map is strictly true only in the idealized continuous case. To recognize this we have only to look at a finite (non-infinitesimal) version, say, J of [0, 1]. For each θ (in J) there is a ϕ (in J), which is well-defined as $\phi = \theta^2$ correct to its second decimal place. But now the correspondence is many-one and not onto. For example, the statement $\phi = 0$ is the union of the eight statements $\theta = .00$, $\theta = .01$, ... $\theta = .07$, and the statement $\phi = .99$ corresponds to no elementary statement about θ. The assertions $\theta \epsilon J$ and $\phi \epsilon J$ are therefore quite different (both logically and statistically) in nature and our

Martian cannot be faulted for reacting differently to two different bits of information. Even in the idealized continuous case, the two statements $\theta\epsilon[0, 1]$ and $\phi\epsilon[0, 1]$ are equivalent only in a logical sense. It is certainly not true that the two statements are equally informative in a statistical sense.

Let us look back on the passage that we quoted in the beginning of this section from Fisher (1930). Curiously enough, it was in this 1930 paper that Fisher first introduced us to his fiducial probability methods for constructing an additive measure of support-by-data, which according to him must be recognized as ordinary frequency probability. It now appears that Fisher was only protesting too much when he so severely deplored the "tendency to treat likelihood as though it were a sort of probability".

The author can find no logical justification for the often repeated assertion that likelihood is only a point function and not a measure. He does not see what inconsistencies can arise from the postulation of the Strong Law of Likelihood in the Koopman-Hacking logic of support-by-data. On the other hand, we shall show later on how some of the non-Bayesian likelihood methods get into serious trouble because of their non-recognition of the additivity of the likelihood measure.

9. Maximum likelihood

Volumes have been written seeking to justify in one way or another the maximum likelihood (ML) method of point estimation (and its sister method —the likelihood ratio method for test of hypotheses), and yet the author cannot find any logical justification for upholding the method as anything but a simplistic tool that may (with some reservations) be used for routine data analysis in situations where the sample size is not too small and the statistical model not too shaky (unrobust). By definition, the ML estimate $\hat{\omega}$ (of the value of ω that obtains) is the point in the parameter space that is best supported by the data. But what logical compulsions guide us to the *maximum likelihood principle* :"The best (or most reasonable) estimate of a parameter is that value (of the parameter) which is best supported by the data" ? If we contemplate for a moment our very ignorant Martian, who is trying to make sense of data related to a parameter about which he has absolutely no pre-conceived notions, then we ought to be more prepared in our mind to accept the reverse proposition : "The most reasonable estimate of a parameter will rarely coincide with the one that has the greatest support from the data".

If Fisher ever thought in terms of the idealization of a Martian, then he must have visualized him (the Martian) as a rational being who not only is

very ignorant (about the parameter of interest) but is also endowed with very limited capabilities. Fisher's Martian does not know how to add likelihoods, he can only compare them. His recognition of points in the parameter space is only microscopic (pointwise). He compares parameter points pairwise—he can only tell how much more likely a particular point is compared to another. Given two composite hypotheses $\omega \epsilon A$ and $\omega \epsilon B$, the only thing that he can do, in the way of comparing the likelihoods (of the composite-hypotheses being true), is to compare the likelihoods of the best supported points $\hat{\omega}_1$ and $\hat{\omega}_2$ in A and B respectively. This is the Martian's Likeliood Ratio method for testing a composite hypothesis against a composite alternative and is analogous to a child's method for picking the winning team in a tug-of-war contest by concentrating his whole attention on the anchors of the two teams ! He has no understanding of any natural topology on the parameter space that may exist. And finally, he does not know anything about the relative hazards of incorrect inferences. The six likelihood methods that we have described in Section 7 are geared to the needs and limitations of such a Martian. It is easy to construct examples where uncritical uses of such methods will lead to disastrously inaccurate inferences. Here is one such.

Example : 1 An urn contains 1000 tickets, 20 of which are marked θ and the remaining 980 are marked 10θ, where θ is the parameter of interest. A ticket is drawn at random and the number x on the ticket is observed. The ML estimate of θ is then $x/10$. In this case, the ML estimation procedure leads to an exact estimate with a probability of .98. So everything seems to be as it should be. But consider a slight variant of the urn-model, where we still have 20 tickets marked θ, but the remaining 980 tickets are now marked $\theta a_1, \theta a_2, \ldots \theta a_{980}$ respectively, and where the 980 constants $a_1, a_2, \ldots, a_{980}$ are all known, distinct from each other, and all of them lie in the short neighbourhood (9.9,10.1) of the number 10. The situation is not very different from the one considered just before, but now look what happens to our Martian. Noting that the likelihood function is

$$L(\theta \mid x) = \begin{cases} .02 & \text{for } \theta = x \\ .001 & \text{for } \theta = xa_i^{-1}, \quad i = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

the Martian now recognizes x as the ML estimate of θ. He also declares (see method (b) of Section 7) that x is at least 20 times more likely than any other point in the parameter space and, therefore, identifies the single-point set $\{x\}$ as the likelihood interval I_λ with $\lambda = 20$. Irrespective of what the true value of θ is, the ML method now over-estimates it with a factor of nearly 10

and with a pobability of .98. As a confidence interval the likelihood interval I_λ (with $\lambda = 20$) has a confidence coefficient of .02.

The source of the Martian's trouble with this example is easy to fathom. If he knew how to add his likelihood measure, then he would have recognized that the likelihood of the true θ lying in the interval $J = (x/(10.1), x/(9.9))$ is .98. Furthermore, if he could recognize that (for medium sized x) the interval J is a narrow one and that small errors in estimation are much less hazardous than an over-estimate with a factor of 10, then he would surely have recognized the reasonableness of estimating the true θ by a point like $x/10$ rather than by the ML estimated x.

We all know that under certain circumstances the ML method works rather satisfactorily in an asymptotic sense. But the community of practising statisticians are not always informed of the fact that under the same circumstances the Bayesian method : "Begin with a reasonable prior measure q of your belief in the various possible values of θ, match it with the likelihood function generated by the data, and then estimate θ by the mode of the posterior distribution so obtained", will work as well as the ML method, because the two methods are asymptotically equivalent.

And once we take the final Bayesian step of 'matching the likelihood function with some reasonably formulated prior measure of our personal belief', we can then orient the task of inference making to all the realities —ω, Ω, q, Π, $\mathfrak{E}$, x, etc.—of the particular situation. If we look back on the six likelihood methods described in Section 7, it will then appear that, excepting for method (d)—the likelihood ratio method of testing a simple hypothesis against a simple alternative—all the other methods are too simplistic and rather disoriented towards the complex realities of the respective inference making situations.

We end this section with another example to demonstrate how disastrously disoriented the Martian can get (in his efforts to evaluate the likelihood evidence given by the data) because of his inability to add likelihoods. Let us look back on methods (e) and (f) described in Section 7 and then consider the following.

Example 2 : The universal parameter ω is (θ, ϕ), where θ (the parameter of interest) lies in the two-point set $I = \{-1, 1\}$ and the nuisance parameter ϕ lies somewhere in the set $J = \{1, 2, \ldots 980\}$. Our task is to draw a ticket at random from an urn containing 1000 tickets and then to guess the true value of θ on the basis of the observed characteristics of the sample ticket.

About the 1000 tickets in the urn we have the information that (i) the number θ is written in large print on exactly 980 tickets and the number $-\theta$ appears in large print in the remaining 20 tickets, and (ii) the 980 tickets marked θ carry the distinguishing marks 1, 2, ... 980 respectively in microscopic print, whereas, the remaining 20 tickets carry the mark ϕ in microscopic print (where the unknown $\phi \epsilon J$). Let x and y be the numbers in large and small print respectively on our sample ticket. Our sample space then is $I \times J$, which is also our parameter space.

Let us suppose for a moment that either we do not have a magnifying glass to read the small print y or for some reason we consider it right to suppress this part of the data from our Martian. The Martian will then be very pleased to discover that his likelihood function (based on x alone) does not depend on the nuisance parameter and is

$$(*) \qquad L(\theta) = L(\theta, \phi \mid x) = \begin{cases} .98 & \text{when} \quad \theta = x \\ .02 & \text{when} \quad \theta = -x \end{cases}$$

and so he will come out strongly in support of the guess : 'the true θ is x'. No doubt we should feel proud of our clever Martian because, irrespective of what θ is, the probability of his guessing right in the above circumstances is .98.

But see what happens when we can read y and cannot find any good reason for suppressing this part of the data. With the full sample (x, y) in his possession, the Martian will routinely analyse the data by first setting up the likelihood function as

$$(**) \qquad L(\theta, \phi \mid x, y) = \begin{cases} .001 & \text{when} \quad \theta = x, \quad \phi \epsilon J \\ .02 & \text{when} \quad \theta = -x, \ \phi = y \\ 0 & \text{otherwise} \end{cases}$$

and then eliminating ϕ from (**) as per method (f) of Section 7. The eliminated likelihood function is

$$(***) \qquad L_e(\theta) = \sup_{\phi} L(\theta, \phi \mid x, y) = \begin{cases} .001 & \theta = x \\ .02 & \theta = -x \end{cases}$$

and so this time the Martian comes out strongly in support of the guess $\theta = -x$. With the full data, the performance characteristic of the Martian's method is now 'only 2% probability of success' !

It should be observed that the real source of the Martian's debacle lies in his inability to add likelihoods. Before the data was available, the Martian's ignorance about the parameter $\omega = (\theta, \phi)$ extended over the 2×980 points of the set $I \times J$. With the sample reading (x, y), the Martian correctly recognized in (**) that his ignorance about (θ, ϕ) is cut down to the smaller set $A \bigcup B$ where

$$A = \{(x, 1), (x, 2), \ldots (x, 980)\}$$

and

$$B = \{(-x, y)\}$$

and that the likelihood of each of the 980 points in A is .001 and that of the single point in B is .02. From the Strong Law of Likelihood (see Section 8) it follows that the likelihood support (by the data) for the composite-hypothesis $\omega \epsilon A$ (that is, $\theta = x$) should have been worked out as

$$L(A) = \sum_{(\theta, \Phi) \epsilon A} L(\theta, \phi \mid x, y) = .98$$

and this compares very favourably with the likelihood support of .02 for the hypothesis $\omega \epsilon B$ (that is, $\theta = -x$).

The elimination of the nuisance parameter ϕ by the above method of addition (of the likelihood function over the range of ϕ for fixed θ) certainly smacks of Bayesianism, but it appears to be a much more natural thing to do than the Fisher-inspired elimination method by maximization (w.r.t. ϕ for fixed θ). [In the present example, it so happens that the 'addition method' of elimination (of ϕ) leads to the same eliminated likelihood function as was achieved earlier in (*) by the 'marginalization method' of suppressing the y-part of the data. However, the author cannot see how a good case can be made for such a marginalization procedure, even though the distribution of x (as a random variable) depends only on the parameter of interest θ, and that of y depends on the nuisance parameter ϕ alone. Note that, for fixed (θ, ϕ), the statistics x and y are not stochastically independent. It follows that, even when the parameters θ and ϕ are entirely unrelated (independent a-priori), suppression of y may lead to valuable loss of information. In order to see this, suppose that we knew for sure that $\phi = 1$ or 2. Now, the statistic y will give us extra information about θ —if $y > 2$ then we know for sure that $\theta = x$ etc.]

Let us close this section with the remark that, however well-suited the 'addition method' (of elimination) may be to the needs and capabilities of our ignorant Martian, the method is not being recommended here as a routine statistical procedure to be adopted by any knowledgeable scientist.

PART 3 : PARARDOXES

10. A FALLACY OF FIVE TERMS

The author vividly recalls an occasion in late 1955 when Sir Ronald (then visiting the Indian Statistical Institute, Calcutta and giving a series of seminars based on the manuscript of his forthcoming book) got carried away by his own enthusiasm for fiducial probability and tried to put the fiducial argument in the classical form of the Aristotelian syllogism known as Barbara: 'A is B, C is A, therefore C is B'. The context (data) was : A random variable X which is known to be normally distributed with unit variance and unknown mean θ about which the only information that we have is, $-\infty < \theta < \infty$. The variable X is observed and the observation is 5. Sir Ronald declared that the following constitutes a 'proof' :

Major premise : Probability that the variable X exceeds θ is 1/2.

Minor premise : The variable X is observed and the observation is 5.

Conclusion : Probability that θ is less than 5 is 1/2.

We know that in Aristotelian logic an argument of the kind : 'Caesar rules Rome, Cleopatra rules Caesar, therefore, Cleopatra rules Rome', is classified as a 'fallacy of four terms'—the four terms being (i) Caesar, (ii) one who rules Rome, (iii) Cleopatra, and (iv) one who rules Caesar. Sir Ronald is perhaps the only person (in the history of scientific thought) who ever dared (even in a moment of euphoria) to suggest a three-line proof involving five different terms—the terms being (i) $\Pr(X > \theta)$, (ii) 1/2, (iii) the observed value of X, (iv) 5, and (v) $\Pr(\theta < 5)$!

About Fisher's fiducial argument Hacking (p. 133) writes : "No branch of statistical writing is more mystifying than that which bears on what he calls the fiducial probabilities reached by the fiducial argument. Apparently the fiducial probability of an hypothesis, given some data, is the degree of trust you can place in the hypothesis if you possess only the given data." The confusion has been further compounded by Fisher's repeated assertions that in those circumstances where he considers it right to talk about fiducial probabilities, the notion should be understood in exactly the same way as a gambler understands his (frequency) probability. Neyman's theory of confidence intervals arose from his efforts to understand the fiducial argument and to re-interpret the concept in terms of frequency probability. Recently, Fraser, with his structural probability methods, is trying to build a mathematical framework for Fisher's ideas on fiducial probabilities. Whereas Neyman never had had any illusions about his 'confidence coefficients' being the same

as ordinary probabilities, it appears that Fraser (like Fisher) deos not make any logical distinction between ordinary and structural (fiducial) probabilities.

On the surface the fiducial method may appear to be of the true likelihood vintage—an exercise in analysing the mind of the Martian (the particular data at hand). A little reflection (see Anscombe [1957] in this connection) however will prove otherwise. Consider the context where the variable X is known to have a $N(\theta, 1)$ distribution, the only background information about θ is that $-\infty < \theta < \infty$, and the observed value of X is x. The fiducial argument leads to the fiducial distribution $N(x, 1)$ for θ. The argument has hardly anything to do with the fact that the data generates the likelihood function $\exp\{-(\theta-x)^2/2\}$, but is based on (i) the fortuitous discovery of the pivotal quantity $X-\theta$ with a standard normal distribution, (ii) a re-interpretation of our lack of prior information about θ, and of course (iii) that X is observed as x. The fiducial argument clearly does not respect the likelihood principle.

In the present context we have two unobservable entities—the parameter θ and the (pivotal) quantity $Y = X-\theta$. About θ the statistician (rather, the Martian) is supposed to know nothing other than that the parameter lies in (varies over) the infinite interval $(-\infty, \infty)$. About Y, on the other hand, he has the very precise information that $Y \bigcap N(0, 1)$ irrespective of what value θ takes. In a sense we may then say that the (unobservable) random quantity Y is stochastically independent of the parameter θ. Now, the sum $\theta+Y = X$ is observable and has actually been observed as x. The fiducial argument then somehow justifies the assertion that the observation $\theta+Y = x$ altered the logical status of the parameter θ from that of an unknown quantity lying somewhere in the interval $(-\infty, \infty)$ to that of a random variable with the probability distribution $N(x, 1)$. In particular, the argument seeks to prove $\Pr(\theta < x) = 1/2$. Following Neyman, we may interpret the above only to mean that if, under the above kind of situation, we always assert $\theta < x$ then, in a long sequence of (independent) such situations—with the unobservable θ's varying in an arbitrary manner and with varying observations x—we shall be right in approximately 50% of cases. But Fisher (also Fraser) seems to be saying something more than this. In effect he is saying that the observation $X = x$ does not have any effect on the probability distribution of the quantity $Y = X-\theta$—that is, given $X = x$ the quantity $Y = X-\theta \bigcap N(0, 1)$ In other words, Fisher is saying that Y is independent of $X(= \theta+Y)$. Note the inherent contradiction between this assertion of independence and our earlier stipulation that Y is independent of θ. If θ has the character of a random variable and is independent of Y, then Y and $Y+\theta$ can never be

independent of each other unless Y is a constant (which it is not). If not, then it is not clear what we are talking about.

Let us try to understand in another way what Fisher really had in mind when he said (in the context of our present X and θ) to the effect : When X is observed as x, we can regard θ as a random variable with $\Pr(\theta < x) = 1/2$, and this irrespective of what x is. Furthermore, the statement $\Pr(\theta < x) = 1/2$ can be interpreted in the same way as we interpret the statement : "For a fair coin Pr(Head) = 1/2."

In order to do so, let us see if we can distinguish between the following two guessing situations :

Situation I : Every morning Peter confronts Paul with an integral number x that he (Peter) has freshly selected that very morning, and then challenges Paul to hazard a guess (on the basis of the number x) about the outcome Y of a single toss of a fair coin (to be carried out immediately afterwards). Clearly, the number x gives Paul no information whatsoever about Y. And if we are to believe in the fairness of the coin (as the frequency probabilists understand it), then there exists no guessing strategy for Paul that, in the long run, will make him guess correctly in more (or less) than 50% of the mornings on which he chooses to hazard a guess. In the language of Fisher, Paul cannot 'recognize' any subsequence of mornings on which the long run relative frequency of occurance of heads will be different from 1/2.

Now consider

Situation II : Every morning Peter confronts Paul with a bag containing two tickets numbered respectively as $\theta-1$ and $\theta+1$, where the number θ is an integer that has been selected by Peter that very morning. Each morning Paul's task is to draw a ticket at random from the bag, observe the number x on the ticket drawn, and then hazard a guess on whether the number θ (the mean of the two numbers in the bag) is $x-1$ or $x+1$.

Clearly, situation II is a simplified (integral) version of the Fisher-problem we started this section with. Let us suppose that Paul has no idea whatsoever about how θ gets selected on any particular morning. He only knows that the unobservable θ can take any value in the infinite set $\{0, \pm 1, \pm 2, \ldots\}$. He also knows that for given θ, the observable X takes only the two values $\theta-1$ and $\theta+1$ with equal probabilities. As before we have the unobservable (pivotal) quantity $Y = X-\theta$ with a well-defined probability distribution. In accordance with the Fisher logic, the only thing that the data $X = x$ tells on any morning about the particular θ that obtains, is simply this : θ is either $x-1$ or $x+1$ with equal probabilities. It seems to the author that Fisher

would not have recognized any qualitative difference between the two situations. If Paul cannot read the mind of Peter then there is no way he can guess right in more (or less) than 50% of the mornings that he chooses to guess on.

Now, let us look at the following interesting argument given by Buehler (1971, p. 337). That Paul can do better than being right in only 50% of the guesses that he is going to make, is shown by Buehler as follows. Suppose Paul refuses to guess whenever $x < 0$, but always guesses θ as $x-1$ whenever $x \geqslant 0$. Now, let us classify all future mornings of Paul on the basis of the values of θ (that Peter is going to select) as follows :

$$M_1(\theta \leqslant -2), \quad M_2(\theta = -1 \text{ or } 0), \quad M_3(\theta \geqslant 1).$$

On M_1-mornings, Paul never guesses and, therefore, is never wrong. Paul makes a guess on 50% of the M_2-mornings and is always right on such occasions. On M_3-mornings Paul always makes a guess and is right in only 50% of such guesses.

No doubt the Buehler argument will be endlessly debated by the advocates of the fiducial and structural probability methods. But let us point out that the argument is in the nature of a broadside against the improper Bayesians also. An improper Bayesian is one who systematically exploits the mathematical advantages of neat improper 'priors' and generally ignores the first requirement of Bayesian data analysis, namely, that the 'prior' ought to be an honest representation of the Bayesian's prior pattern of belief. Observe that in situation II above, an improper Bayesian will note with great relish the fact that the data allows him to assume that the parameter space is the unrestricted set I of all integers and that the likelihood function generated by the observation $X = x$ has the simple form

$$L(\theta \mid x) = \begin{cases} \frac{1}{2} & \text{when} \quad \theta \epsilon \{x-1, x+1\} \\ 0 & \text{for all other } \theta \text{ in I.} \end{cases}$$

He will now simplify everything by starting with the uniform prior over the infinite set I (an impropriety of the highest order according to the author). thus arriving at a posterior distribution which is the same as the uniform fiducial distribution over the two point set $\{x-1, x+1\}$.

11. The stopping rule paradox

The controversy about the relevance of the stopping rule at the data analysis stage is best illustrated by the following simple example :

Example : Suppose 10 tosses of a coin, with an unknown probability θ for landing heads, resulted in the outcome

$$x = THTTHHTHHH.$$

Now, for each of the following four experimental procedures :

E_1 : Toss the coin exactly 10 times;

E_2 : Continue tossing until 6 heads appear;

E_3 : Continue tossing until 3 consecutive heads appear;

E_4 : Continue tossing until the accumulated number of heads exceeds that of tail by exactly 2;

and indeed for any sequential sampling procedure (of the usual kind, with prescience denied) that could have given rise to the above sequence of heads and tails, the likelihood function (under the usual assumption of independence and identity of tosses) is the same, namely,

$$L(\theta \mid x) = \theta^6(1-\theta)^4.$$

From the likelihood principle ($\mathscr{L}$) it then follows that at the time of analysing the information contained in the data $(\mathscr{E}, x)$, we need not concern ourselves about the exact nature of the experiment $\mathscr{E}$—our whole attention should be rivetted on the likelihood function $\theta^6(1-\theta)^4$, which does not depend on the stopping rule. In general terms, we may state the following principle due to George Barnard :

Stopping rule principle (for a sequential sampling plan) : Ignore the sampling plan at the data analysis stage.

This suggestion will no doubt shock and outrage anyone whose statistical intuition has been developed within the Neyman-Pearson-Wald framework. Even some enthusiastic advocates of $\mathscr{L}$ find the stopping rule principle embarrasingly hard to swallow. It will be quite interesting to make a survey of contemporary practising statisticians with a suitably framed questionnaire based on the above example. However the matter cannot be settled democratically ! Dennis Lindley, having seen an earlier draft of this article, wrote to say the following : "You may like to know that in my third-year course I have, for many years now, given the class the results of an experiment like you give, and ask them if they need any more information before making an inference. I have *never* had a student ask what the sample space was. I then point out to them that they could not construct a confidence interval, do a significance test, etc., etc. Although they are not practising statisticians, they have had two years of statistics. They just don't feel the sample space is relevant. I have tried this out with more experienced audiences and only occasionally had an enquiry about whether it was direct or inverse sampling".

The rest of this section is devoted to a detailed discussion of the famous Stopping Rule Paradox*, which is generally believed to have knocked out the logical basis of principle $\mathscr{L}$. In order to isolate the various issues involved, it will help if we denote by $\mathscr{F}$ the following set of three classical (Fisherian) methods of statistical inference.

The $\mathscr{F}$ methods : The data consists of the pre-fixed number n of independent observations on a random variable X that is known to be normally distributed with unknown mean $\theta(-\infty < \theta < \infty)$ and known variance 1. The data then generates the information (likelihood function)

$$\text{(i)} \qquad L(\theta) \sim \exp\{-n(\theta-\bar{x}_{(n)})^2/2\}$$

where $\bar{x}_{(n)} = (x_1+x_2+\ldots+x_n)/n$. Under the above circumstances, let $\mathscr{F}$ consist of the trilogy of statistical methods :

$\mathscr{F}(a)$: If $|\bar{x}_{(n)}-\theta_0| > 3/\sqrt{n}$, then reject the null-hypothesis H_0 : $\theta = \theta_0$ and declare that the data is highly significant.

$\mathscr{F}(b)$: The statement $\theta \epsilon (\bar{x}_{(n)}-3/\sqrt{n},\ \bar{x}_{(n)}+3/\sqrt{n})$ may be made with a great deal (well over 99%) of 'self-assurance' or 'confidence'.

$\mathscr{F}(c)$: The sample mean $\bar{x}_{(n)}$ is the most 'appropriate' point estimate of θ and the estimate is associated with a 'standard error' of $1/\sqrt{n}$.

Now consider the sequential sampling procedure based on the stopping rule :

$\mathscr{R}$: Continue observing X until the sample mean $\bar{x}_{(n)}$ satisfies the inequality $|\bar{x}_{(n)}| > 3/\sqrt{n}$.

If N is the (random) sample size associated with our rule $\mathscr{R}$, then it is easy to prove that N is finite with probability one if $\theta \neq 0$, and when $\theta = 0$ this conclusion still holds. [The latter may be deduced from the Law of the Iterated Logarithms, but can be proved much more easily directly. It should be noted, however, that $E(N \mid \theta)$ is finite only when $\theta \neq 0$.] Thus our rule $\mathscr{R}$ is mathematically well-defined in the sense that N is finite with probability one for all possible values of θ. Suppose, following the rule $\mathscr{R}$, we generate the sample $x_1, x_2, \ldots x_N$. Our N is now random (not pre-fixed) but somehow the likelihood function fails to recognize this fact, for it is in the familiar form (see (i) above)

$$\text{(ii)} \qquad L(\theta) = (\sqrt{2\pi})^{-N} \exp\{-\Sigma(x_i-\theta)^2/2\}$$

$$\sim \exp\{-N(\theta-\bar{x}_{(N)})^2/2\}.$$

*The author is unaware of who first formulated this clever paradox.

Now, if we combine $\mathscr{L}$ with $\mathscr{F}$, then looking back on (i) and (ii), we shall be forced to admit that, even when the sample $x_1, x_2, \ldots x_N$ is generated by the sequential sampling rule $\mathscr{R}$, the following two inferences are also appropriate :

(a′) The null-hypothesis $H_0 : \theta = 0$ should be rejected, at a very high level of significance (assurance), since $|\bar{x}_{(N)}| > 3/\sqrt{N}$ holds by definition.

(b) We ought to place more than 99% confidence or assurance in the truth of the assertion that the true value of θ lies in the interval $(\bar{x}_{(N)} - 3/\sqrt{N}, \bar{x}_{(N)} + 3/\sqrt{N})$.

The paradox : The stopping rule paradox lies in the observation that method (a′) leads to a sure rejection of hypothesis H_0 (at a high level of significance) even when H_0 is true. Also observe that the confidence interval $\bar{x}_{(N)} \pm 3/\sqrt{N}$ constructed for the unknown θ surely excludes the point $\theta = 0$ even when H_0 is true. Clearly, there must be something very wrong with principle $\mathscr{L}$!

For the moment let us only reverse the charge and claim that the stopping rule paradox, instead of discrediting $\mathscr{L}$, ought to strengthen our faith in the principle by exposing the naivete of certain standard statistical methods that are not truly in accord with the spirit of $\mathscr{L}$. To prove our claim, let us first of all concentrate our attention on the $\mathscr{F}(a)$ method of testing the null hypo thesis $H_0 : \theta = 0$.

Intuitively, it seems that the sequential sampling rule $\mathscr{R}$ used above is especially well-suited to the problem of obtaining information on whether the hypothesis $H_0 : \theta = 0$ is true or not. When θ is appreciably different from zero we do not need too many observations on X before we lose faith in H_0, whereas when θ is nearly zero, we need quite a large sample before we could be reasonably sure that H_0 is false.

Why then should a 'reasonable' sampling plan $\mathscr{R}$, when coupled with $\mathscr{L}$ and the standard method $\mathscr{F}(a)$, lead us to a testing procedure (a') with a power function

$$\pi(\theta) = \Pr(\text{Test ends with rejection of } H_0 \mid \theta)$$

that is uniformly equal to one ? Is there any paradox at all ?

Could the trouble lie in the fact that our rule $\mathscr{R}$ is not bounded above and, therefore, is perhaps a non-performable experiment ? To see if this might be so, let us define a bounded version $\mathscr{R}_M$ *of* $\mathscr{R}$ as follows :

$\mathscr{R}_M$: Continue observing X until the sample mean $\bar{x}_{(n)}$ satisfies the inequality $|x_{(n)}| > 3/\sqrt{n}$ or $n = M$, whichever happens first. Our M is a fixed

but possibly very large integer. With such a 'performable' rule $\mathcal{R}_M$ replacing $\mathcal{R}$, our power function $\pi_M(\theta)$ will now have the familiar U-shape that many of us like so much. Now, one might argue that it is only in the idealized limiting situation ($M \to \infty$) that our test becomes endowed with the (very desirable) property of having maximum power* of discernment against H_0, when the hypothesis is false, coupled with the (rather undesirable !) property of non-recognition of H_0 when it is true. Let us look at the problem from another angle.

Is it not illogical to talk of a null-hypothesis H_0 that is specified by a particular value of a continuous parameter θ ? Are we not insisting from the beginning that all our realities are finite and therefore discrete ? How can a pinpointed hopothesis like $H_0: \theta = 0$ be classified as anything but an illusory idealization ? Surely, such an 'infinitesimal' hypothesis (as H_0) is 'certainly false' to begin with, and ought to be rejected out of hand however large the sample is. How can a testing procedure be faulted for suggesting just that !

In the same spirit that we replaced the unbounded stopping rule $\mathcal{R}$ by a bounded version $\mathcal{R}_M$, let us replace the infinitesimal hypothesis H_0 by a non-infinitesimal version.

$$H_\delta : \text{Hypothesis that } \theta \epsilon (-\delta, \delta),$$

where δ is some suitable positive number.

Let us see what happens to our paradox when we work with the finite (bounded) stopping rule $\mathcal{R}_M$ and finite (non-infinitesimal) hypothesis H_δ to be tested. If $x = (x_1, x_2, \ldots x_N)$ be the sample observations on X that we obtain following rule $\mathcal{R}_M$, then what is the quality and strength of our information $\text{Inf}(\mathcal{R}_M, x)$ regarding the hypothesis H_δ ? Principle $\mathcal{L}$ tells us not to take into account any details of the statistical structure of the experiment performed or of the sample obtained other than the nature of the likelihood function $L(\theta \mid x)$ generated by the data. Fortunately, $\mathcal{L}$ does not stop us from using any background (prior) information about the parameter θ that we might have had to begin with. However, only a Bayesian knows how to match his 'prior information' with the 'likelihood information' supplied by the data. [Many valiant and rather desperate attempts have been made by believers in $\mathcal{L}$—like Fisher, Barnard and others—to avoid taking this final Bayesian step, but according to the author such efforts have not met with much success.] So let us examine how the Bayesian method works in the present case.

*Indeed it was the stopping rule paradox that awakened the author (about five years ago) to the possibility of the Darling-Robbins type tests with power one for the hypothesis $\theta \leqslant 0$ against the alternative $\theta > 0$.

Suppose, for the sake of this argument, that our Bayesian decides upon a uniform distribution over the interval $(-20, 20)$ as a reasonable approximation to the information (or the general lack of it) that he has about the unknown θ. Looking back on (ii), it is clearly very unlikely that we shall end up with a likelihood function L that does not lie well within (in the obvious sense) the interval $(-20, 20)$. With L lying well within the interval $(-20, 20)$ the 'posterior density' of θ will be worked out by our Bayesian as roughly proportional to L and so he will evaluate the posterior probability of H_δ as

$$\text{(iii)} \qquad \Pr(H_\delta \mid x) = \int_{-\delta}^{\delta} \frac{\sqrt{N}}{\sqrt{2\pi}} \exp\{-N(\theta - \bar{x}_{(N)})^2/2\} d\theta$$

$$= \Pr\{-\delta\sqrt{N} - \sqrt{N}\bar{x}_{(N)} < Z < \delta\sqrt{N} - \sqrt{N}\bar{x}_{(N)}\}$$

where Z is a $N(0, 1)$ variable.

The stopping rule $\mathscr{R}_M$ is such that with a fair sized N the sample mean $\bar{x}_{(N)}$ is either roughly equal to $\pm 3/\sqrt{N}$ or is some number in between. Let us consider the situation when $\bar{x}_{(N)}$ is just above $3/\sqrt{N}$ and ignore the overshoot. Formula (iii) now becomes

$$\text{(iv)} \qquad \Pr(H_\delta \mid x) = \Pr(-\delta\sqrt{N} - 3 < Z < \delta\sqrt{N} - 3)$$

and so the 'Bayesian significance' of the data depends entirely on the size of the statistic N. In order to see this let us suppose that $\delta = 1/10$. When $N = 100$, the right hand side in (iv) becomes $\Pr(-4 < Z < -2)$ which is less than 0.025. Whereas, when $N = 10{,}000$, the expression in (iv) becomes $\Pr(-13 < Z < 7)$ which is far in excess of 0.999 !!

The point is clear : It is naive to propose $\mathscr{F}(a)$ as a realistic statistical method. It simply does not make good statistical sense to set up a pin-point (infinitesimal) null-hypothesis like $H_0 : \theta = 0$ and then to recommend its rejection whenever $|\bar{x}_{(n)}| > 3/\sqrt{n}$, where $\bar{x}_{(n)}$ is the observed mean of n (pre-fixed) independent observations on an X distributed as $N(\theta, 1)$ with $-\infty < \theta < \infty$. It should be recognized that the level of significance of the data vis a vis the hypothesis H_0 does not depend on the magnitude of $|\sqrt{n}\,\bar{x}_{(n)}|$ alone. It also depends, in a very crucial manner, on the magnitude of the sample size n. A Fisherian will perhaps feel quite satisfied with the information that $\sqrt{n}\,\bar{x}_{(n)} = 3$, and will, in any case, confidently reject the hypothesis H_0. But a Bayesian will surely enquire about the size of n (even though he may be quite uninterested at the data analysis stage to know whether n was prefixed or not). And, as we have just seen, the Bayesian's reactions to the two situations, $n = 100$ and $n = 10{,}000$, will be entirely different.

In the first case he will consider it very unlikely that the true θ lies in the interval $(-0.1, 0.1)$, whereas in the second case he will have an enormous amount of confidence in the same hypothesis.

The stopping rule paradox should really be recognized as just another paradox of the infinitesimal. To emphasize this once again, let us briefly return to that part of the paradox that refers to (b') that is, to the fact that, with $\mathcal{R}$ as the stopping rule, the 3σ likelihood interval $\bar{x}_{(N)} \pm 3/\sqrt{N}$ will always exclude the point 0 even when $\theta = 0$. This should not worry the planner of the experiment $\mathcal{R}$ if he bears in mind the fact that, in an hypothetically infinite sequence of repeated trials with θ fixed at 0, the variable N will usually take extremely large values, since $E(N \mid \theta = 0) = \infty$. For then he will recognize that the 3σ-interval $\bar{x}_{(N)} \pm 3/\sqrt{N}$ will in general be extremely short and will have its centre exceedingly near the point 0. In other words, the 3σ likelihood interval will, with a great deal of probability, overlap very largely with the experimenter's indifference zone $(-\delta, \delta)$ around the point $\theta = 0$. Let us repeat once again that the pin-point hypothesis $\theta = 0$ is only a convenient idealization and should never be mistaken for a reality.

12 The stein paradox

In 1961 L. J. Savage wrote : "The likelihood principle, with its at first surprising conclusions, has been subject to much oral discussion in many quarters. If the principle were untenable, clear-cut counter-examples would by now have come forward. But such examples seem, rather, to illuminate, strengthen, and confirm the principle". In the following year, Charles Stein (1962) took up the challenge and came up with his famous paradoxical counter-example. It is popularly believed that the Stein paradox demolishes principle $\mathcal{L}$. We propose to show here why the paradox should really be regarded as something that illuminates, strengthens and confirms the likelihood principle.

The counterexample is based on the function

$$f(y) = y^{-1} \exp\{-50(1-y^{-1})^2\}, \quad 0 < y < \infty$$

defined over the positive half-line. Note that $\lim f(y) = 0$ both when $y \to 0$ and $y \to \infty$, and that in the latter case the rate of convergence (to zero) is slow enough to make the integral $\int_0^\infty f(y)dy$ diverge. We can therefore choose a and b such that

$$\text{(i)} \qquad \int_0^b af(y)dy = 1 \quad \text{and} \quad \int_{10}^b af(y)dy = 0.99$$

In point of fact the number b is exceedingly large—larger than 10^{1000}.

Now suppose that the probability distribution of the observable Y involves the unknown θ as a scale parameter in the following manner. The probability density function of Y is given by

$$\text{(ii)} \qquad p(y\,|\,\theta) = \begin{cases} a\theta^{-1}f(y\theta^{-1}), & 0 < y < b\theta \\ 0 & y \geqslant b\theta \end{cases}$$

Let us also suppose that our only prior knowledge about θ is $0 < \theta < \infty$. With a single observation y on Y we end up with the likelihood function

$$\text{(iii)} \qquad L(\theta\,|\,y) \sim \begin{cases} \exp\{-100(\theta-y)^2/2y^2\} & yb^{-1} < \theta < \infty \\ 0 & 0 < \theta \leqslant yb^{-1} \end{cases}$$

Note that the maximum likelihood (ML) estimate of θ is y itself. But from (i) and (ii) we have

$$\text{(iv)} \qquad \Pr(Y > 10\theta\,|\,\theta) = \int_{10\theta}^{b\theta} p(y\,|\,\theta)dy = \int_{10}^{b} f(y)dy = 0.99$$

In other words, we have a situation where the ML estimator over-estimates the true θ by a factor in excess of 10 and with a degree of certainty that is 99% ! The force of this criticism is, however, not directed against principle $\mathscr{L}$. We have seen earlier in Section 9 that simple-minded, unquestioning applications of the ML method can lead us into serious trouble. The Stein example is another such sign-post warning us against uncritical use of the ML method. In this respect it is analogous to the following variant of an urn-model that we considered earlier in Section 9.

Example : Suppose $0 < \theta < \infty$ and that an urn contains 1000 tickets out of which 10 are numbered θ and the remaining 990 are marked respectively as $\theta a_1, \theta a_2, \ldots \theta a_{990}$, where the a_i's are known numbers all greater than 10. The random variable Y is the number on a ticket that is to be drawn at random from the urn. Here $\Pr(Y > 10\theta\,|\,\theta) = 0.99$; and when Y is observed as y, the unknown θ becomes 10 times more 'likely' to be equal to y than any one of the other 990 possible values, namely, ya_i^{-1} $(i = 1, 2, \ldots 990)$.

Stein's ingenious arguments against principle $\mathscr{L}$ run along the following lines : If Y were distributed as $N(\theta, \sigma)$, with $-\infty < \theta < \infty$ and σ known, then an observation y on Y would have generated the 'normal' likelihood function

$$\text{(v)} \qquad \exp\{-(\theta-y)^2/2\sigma^2\} \qquad -\infty < \theta < \infty$$

and in such a case it would have been clearly correct (method $\mathscr{F}(b)$ of Section 11) to make an assertion like

(vi) $$y-3\sigma < \theta < y+3\sigma$$

with an associated level of assurance (confidence) that is at least 99%. Now, if we look back on L in (iii) and remember that $b > 10^{1000}$, then we have to admit that, for all practical purposes and irrespective of what y is, the likelihood function L in (iii) is indistinguishable from the one in (v) above with $\sigma = y/10$. Invoking principle $\mathscr{L}$ together with the 3σ-interval method $\mathscr{F}(b)$, Stein concludes that it must then be appropriate to associate at least 99% confidence in the truth of the proposition

(vii) $$(0.7)y < \theta < (1.3)y$$

where y is the observed value of a random variable Y distributed as in (ii) and θ is the value of the unknown parameter that obtains. But from (iv) it follows that, having observed $Y = y$, we are also entitled to make the assertion

(viii) $$\theta < (0.1)y$$

with a 99% degree of confidence.

The Stein paradox then lies in the observation that the two statements (vii) and (viii) are mutually exclusive and, therefore, in no meaningful sense can they both be associated with degrees of confidence that are as high as 99%. According to Stein, this paradox clearly proves the untenability of principle $\mathscr{L}$, and a great many contemporary statisticians seem to be in wholehearted agreement with him.

A re-examination of the Stein argument will make it clear how the anomaly was forged out of the union of $\mathscr{L}$ with method $\mathscr{F}(b)$—the 3σ interval-estimation method based on an observation y on $Y \frown N(\theta, \sigma)$, with $-\infty < \theta < \infty$ and σ known. But what is the logical status of method $\mathscr{F}(b)$? And then, how compatible is $\mathscr{F}(b)$ with principle $\mathscr{L}$? We know all too well how the 3σ-interval is justified in the Neyman-Pearson theory in terms of the 'coverage probability' of the corresponding (random) interval-estimator $(Y-3\sigma,\ Y+3\sigma)$. We are also aware of the Fisher/Fraser efforts of justifying the same interval in terms of fiducial/structural probability. But such 'sample space' arguments are not compatible with $\mathscr{L}$, nor are they applicable to the present case.

There are two well-known likelihood routes following which one may seek to arrive at method $\mathscr{F}(b)$ from principle $\mathscr{L}$. The first route is briefly charted out in our description of method (b) in Section 7 —the LR (likelihood

ratio) method of interval estimation. Following this route, one first recognizes the 3σ-interval in (vi) and (vii) as the LR interval

$$I_\lambda = \{\theta : L(\hat{\theta})/L(\theta) < \lambda\}$$

where $\hat{\theta}$ $(= y)$ is the ML estimate of θ and $\lambda = e^{4.5}$, and then the argument is allowed to rest on the largeness of the number λ $(= e^{4.5})$. However, observe that the Stein paradox does not relent a bit even when one increases the λ to the staggering level of $e^{40.5}$—that is, replaces the 3σ-interval by the 9σ-interval. In Sections 8 and 9 we have argued at length against likelihood methods that are based solely on pointwise comparisons of likelihood ratios. The Stein paradox ought to be recognized as just another sign-post of warning against uncritical uses of the ML and the LR methods of Section 7.

The other slippery route that will generate the 3σ-intervals (vi) and (vii) from $\mathscr{L}$ is of course the way of the improper Bayesians. Looking at the likelihood function (v), an improper Bayesian will immediately recognize the enormous mathematical advantages of beginning his Bayesian data-anlysis rituals with the uniform prior over the infinite parameter space. This will allow him to claim that, given $Y = y$, the posterior distribution of θ is $N(y, \sigma)$. And then he will arrive at the 3σ-interval $(y-3\sigma, y+3\sigma)$ in the approved manner and associate the interval with more than 99% posterior probability. In a moment of euphoria an improper Bayesian may even put down the following as a fundamental statistical principle :

Priniple $\mathscr{IB}$: If the likelihood function L generated by the data is indistinguishable from the normal likelihood (v) above, and if our prior knowledge about the parameter θ is very diffuse, then it is right to associate over 99% confidence (probability) in the truth of the proposition that the true θ lies in the 3σ-interval (vi).

Stein's denounciation of the likelihood principle is apparently based on the supposition that $\mathscr{IB}$ is a corollary to $\mathscr{L}$. In his example, the L in (iii) is truly indistinguishable from (v) and this is so irrespective of the magnitude of the observed y. It is $\mathscr{IB}$ (and not $\mathscr{L}$) then that justifies a posterior probability measure in excess of 99% for the interval in (vii), and this for all possible observed values y for Y. Written formally as a conditional probability statement, the above will look like : If θ is uniformly distributed over the parameter space $(0, \infty)$ and if Y, given θ, is distributed as in (ii), then

(a) $$\Pr(A \mid Y = y) > 0.99 \quad \text{for all } y \epsilon (0, \infty),$$

where the event A is defined by the inequality $(0.7)Y < \theta < (1.3)Y$. But from (iv) we know that

(b) $$\Pr(A \mid \theta) < 0.01 \quad \text{for all } \theta\epsilon(0, \infty).$$

Of course, all our probabilistic intuitions will rebel against the suggestion that there can exist a random event A whose conditional probability is either uniformly greater than 0.99 or uniformly smaller than 0.01 depending on whether we choose the conditioning variable as Y or θ ! But it should be realized that the improper Bayesian has lifted the subject matter to the rarefied, metaphysical plane of infinite (improper) probabilities and so no mathematical contradictions are involved, since both θ and Y are (marginally) improper random variables and the unconditional probability of A is infinite.

To a proper Bayesian, the Stein paradox is merely another paradox of the infinite. In order to see this, let us see what happens if we couple a proper prior density function q to the likelihood function in (iii) and then obtain the shortest 99% confidence interval (in the approved Bayesian manner) as the interval $I_q(y) = (m(y), M(y))$. We now have

$$\Pr(\theta\epsilon I_q(y) \mid Y = y, q) = 0.99$$

And if we consider θ as fixed and speculate about the 'coverage probability' of the (random) interval-estimator $I_q(Y)$, then we arrive at the performance characteristic

$$\pi(\theta) = \Pr\{\theta\epsilon I_q(Y) \mid \theta\} = \Pr\{m(Y) < \theta < M(Y) \mid \theta\}.$$

Since q is a proper prior, we now recognize (thanks to Fubini) that

$$\int_0^\infty \pi(\theta)q(\theta)d\theta = 0.99$$

and we are saved from an embarrassment of the kind that the improper Bayesian suffered in (b) above—his $\pi(\theta)$ was uniformly smaller than 0.01 !

All of us have our favourite paradoxes of the infinite and the infinitesimal. The author cannot resist the temptation of setting down here his favourite paradox of the infinite.

Example : Peter and Paul are playing a sequence of even money games of chance in which the odds are heavily stacked against Paul—the games are identical and independent, and in each game Paul's chance of winning is only 0.01. Paul, however, has the choice of stakes and can decide when to stop playing. Paul considers the situation to be highly favourable to himself, but bemoans the fact that his chance of winning in a single game is not low enough

—he would have much preferred it to be, say, one in a million. Simple ! Paul trebles the stakes after each loss, and continues to play until his first (or the n-th) win. Observe that we have opened our windows to three infinities : Paul's capital, Peter's capital and the playing time—all are supposed to be unbounded.

What then is the real status of the 3σ-interval in (vii) ? Principle $\mathcal{JB}$ notwithstanding, it is certainly wrong to say : "No matter how large or small y is, the interval $J(y) = (0.7y, 1.3y)$ should be associated with a high degree of confidence/likelihood/probability for containing the true θ" Only a Bayesian, working with a honest (and, therefore, proper) measure of prior belief, is able to give a reasonable answer to the question : "Under what circumstances is it plausible to associate a 3σ-likelihood interval like (vii) with a posterior measure of belief that is in excess of 99%" His answer will be something like : "When the prior distribution is found to be nearly uniform (with a positive density) over the 3σ-interval" Suppose, for the sake of the argument, that the Bayesian regards a uniform probability distribution over the interval $(0, C)$ as a fair representation of the state of knowledge that he started with about the parameter θ. This means, in particular, that he has about 90% prior belief in the proposition $\theta > (0.1)C$. So when he plans to take an observation on the Stein variable Y he is already very confident that the observation y will fall well outside the interval $(0, C)$. He will not be at all surprised to find the 3σ-likelihood interval $J(y)$ to be disjoint with his parameter space $(0, C)$ and will naturally allot a zero measure of (posterior) belief to the 3σ-interval then.

Mathematics is a game of idealizations. We must however recognize that some idealizations can be relatively more monstrous than others. The idea of a uniform prior over a finite interval $(0, C)$ as a measure of belief is a monstrous one indeed. But the super-idealization of a uniform prior over the infinite half-line $(0, \infty)$ is really terrifying in its monstrosity. Can anyone be ever so ignorant to begin with about a positive parameter θ that he is (infinitely) more certain that θ lies in the interval (C, ∞) than in the interval $(0, C)$ —and this for all finite C however large ?! Naturally, everything goes completely haywire when such a person, with his mystical all-consuming belief in $\theta > C$ for any finite C, is asked to make an inference about θ by observing a variable Y which is almost sure to be at least 10 times larger than θ itself !

According to the author's monstrosity scale for mathematical idealizations, the uniform prior over the half-line $(0, \infty)$ is rated as only half as monstrous as the prior distribution defined in terms of the improper density function

$d\theta/\theta$. Stein cleverly exploited the logical vulnerability of the former at the infinite end. The latter is vulnerable at the zero end also. Anyone endowed with this latter kind of prior knowledge about θ must regard each of the two statements $0 < \theta < \varepsilon$ and $C < \theta < \infty$ as infinitely more probable than any statement of the kind $\varepsilon < \theta < C$ —and this for all $\varepsilon > 0$ and $C < \infty$!

However, one point in 'favour' of the measure Q on $(0, \infty)$ defined by the density $d\theta/\theta$ is that it is a (multiplicative) Haar measure on the (multiplicative) group of positive numbers—the measure is invariant for all changes of scale (transformations like $\theta \to a\theta$, with $a > 0$, of $(0, \infty)$ onto itself). This, together with the fact that θ enters into the model (for Y) as a scale parameter, make Q almost irresistible to many improper Bayesians who will somehow convince themselves of the necessity of taking Q as a prior measure of rational belief. The rest of their arguments will then follow the standard Bayesian line ending in the 99% posterior probability interval $J_Q(y)$ for θ.

With Q as the Bayesian prior, the posterior distribution of the scale parameter θ is defined in terms of the density function

$$q(\theta \mid y) = \begin{cases} a\theta^{-1} \exp\left\{-50\left(\frac{\theta}{y} - 1\right)^2\right\}, & b^{-1}y \leqslant \theta < \infty \\ 0 & 0 < \theta < b^{-1}y \end{cases}$$

and is the same as the fiducial/structural probability distribution of θ that is obtained in the usual manner from the pivotal quantity y/θ. In view of the fact that the above density function is bimodal (with modes at $b^{-1}y$ and at a point roughly equal to $99y/100$), the usual 99% posterior probability set $J_Q(y)$ will in fact be the union of two intervals and, therefore, different from the 99% confidence interval $J_S(y) = (b^{-1}y, 10^{-1}y)$ suggested by Stein. It should however be noted that the improper Bayesian will evaluate the posterior probability of the interval $J_S(y)$ as 99% and hence the two intervals J_Q and J_S must have an overlap with at least 98% posterior probability.

At this point let us take note of the fact that any recommendation for the use of the prior Q (for weighting the likelihood function) on the score of θ being a scale parameter is contrary to the spirit of principle $\mathcal{L}$. This is because the information that θ is a scale parameter cannot be deciphered from a description of the likelihood function alone. Curiously enough, of all persons George Barnard also has a lot to do with the logical monstrosity of Q. In Barnard (1962) we have a discription of how he proposes to use the posterior (fiducial) distribution q above in conjunction with the likelihood function L to arrive at a confidence interval $J_B(y)$. The interval $J_B(y)$ looks startlingly

different from $J_S(y)$ but has* the same 99% 'coverage probability' as that of the latter.

Let us close this section by asserting once again that the Stein paradox illuminates the likelihood principle by focussing our attention on the true Bayesian profile of the principle. It also strengthens principle $\mathcal{L}$ by demonstrating the logical inadequacies of some so-called likelihood methods/principles like ML, LR $\mathcal{IB}$, etc.

ACKNOWLEDGEMENT

In October/November 1972 the author gave a series of lectures on Likelihood at the University of Sheffield. This three part essay is the re-written version of part of the lecture notes circulated at that time. The author wishes to thank Terry Speed and other participants in this seminar series whose unflagging interest in the subject persuaded him to do this re-writing. In future another three parts should be added to this essay.

The attention of the author has been drawn by A. Birnbaum to a short note of his in the December 1972 issue of *JASA*. There is a certain amount of overlap between Birnbaum's note and part one of this essay.

REFERENCES

ANSCOMBE, F. J. (1957): Dependence of the fiducial argument on the sampling rule. *Biometrika*, **33**, 464-69.

BARNARD, G. A. (1962): Comments on Stein's "A remark on the likelihood principle". *J. R. Statist. Soc.*, A, **125**, 569-73.

——— (1967): The use of the likelihood function in Statistical practice. *Proc. Fifth. Berkeley Symp., U. of Calif. Press*, **1**, 27-40.

BARNARD, G. A., JENKINS, G. M. and WINSTEN, C. B. (1962): Likelihood inference and time series. *J. R. Statist. Soc.*, A, **125**, 321-372.

BARNARD, G. A., SPROTT, D. A. (1971): A note on Basu's examples of anomalous ancillary statistics. *Foundations of Statistical Inference*, Ed. Godambe and Sprott, Holt, Rinehart and Winston, 163-76.

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

——— (1969): Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā*, A, **31**, 441-54.

BIRNBAUM, A. (1962): On the foundations of Statistical inference. *J. Amer. Statist. Assoc.*, **57**, 269-326.

——— (1972): More on concepts of statistical evidence. *J. Amer. Statist. Assoc.*, **67**, 858-861.

BLACKWELL, D. (1951): Comparison of experiments. *Proc. Second Berkeley Symp.*, U. of Calif. Press, 93-102.

BUEHLER, R. J. (1971): Measuring information and uncertainty. *Foundations of Statistical Inference*; Eds: Godambe and Sprott; Holt, Rinehart and Winston.

*In an earlier version of the essay the author had mistakenly asserted that the interval J_B fails the Stein test on its coverage probability. The author is grateful to Professor Barnard for his pointing out this error.

DARLING, D. A. and ROBBINS, HERBERT (1967): series of notes published in the proc. of the *U.S. Nat. Aca. of Sciences*, beginning will *vol.* **57**, 1188-92.

EDWARDS, A. W. F. (1972): *Likelihood*, Camb. Univ. Press.

FISHER, R. A. (1930): Inverse probability. *Proc. Camb. Phil. Soc.*, **26**, 528-35. (Reprinted in Fisher 1950).

——— (1950): *Contributions to Mathematical Statistics*. John Wiley and Sons.

——— (1956): *Statistical Methods and Scientific Inference*, Oliver and Boyd.

FRASER, D. A. S. (1968): *The Structure of Inference*, John Wiley and Sons.

HACKING, I. (1965): *Logic of Statistical Inference*, Camb. Univ. Press.

HAJEK, J. (1967): On basic concepts of statistics. *Proc. Fifth Berkeley Symp., U. of Calif. Press*, **1**, 139-162.

KOOPMAN, B. O. (1940): The axioms and algebra of intuitive probability. *Ann. of Maths.*, **41**, 269-92.

LINDLEY, D. V. (1965): *Introduction to Probability and Statistics*, Part 2, Camb. U. Press.

SAVAGE, L. J. (1961): The foundations of statistics reconsidered, *Proc. Fourth Berkeley Symp., U. of Calif. Press*, **1**, 575-86.

STEIN, CHARLES (1962): A remark on the likelihood principle. *J. R. Statist. Soc.*, A, 565-68.

DISCUSSION

This three part essay was presented to the Conference on Foundational Questions in Statistical Inference held at the Institute of Mathematics of Aarhus University, Denmark between 7th and 12th May, 1973. The essay was read in two instalments on May 9 and May 12 and was followed by discussions on each occasion. The following is a consolidated account of the discussions that took place. The discussants were A. W. F. Edwards, G. A. Barnard, A. P. Dempster, G. Rasch, D. R. Cox, S. L. Lauritzen, O. Barndorff-Nielsen, P. Martin-Lof and J. D. Kalbfleisch

Edwards : Professor Basu raised the question of why Fisher felt he had to justify the method of maximum likelihood in repeated-sampling terms. I believe he did so in response to an invitation by Karl Pearson : 'If you will write me a defence of the Gaussian method [as Pearson termed maximum likelihood], I will certainly consider its publication'. Thus, ten years after he had originally proposed the method, Fisher examined its repeated-sampling properties (1922). But by 1938 he was writing 'A worker with more intuitive insight than I might perhaps have recognized that likelihood must play in inductive reasoning a part analogous to that of probability in deductive problems' (see Jeffreys (1938)).

Barnard : Concerning Fisher's 1912 paper, the justification given for maximum likelihood was to some extent its "absolute" character, in being, unlike χ^2, independent of any arbitrary grouping of the observations, or of any arbitrary choice of variables for fitting moments.

The Bayesian position cannot be reckoned as having been fully stated until they specify how the prior factor q, in the posterior Lq, is to be determined. The last posthumous paper by Jimmie Savage was a serious attempt to do this; but its very length and complexity (and that of a related paper by, I think Winckler, in *JASA*) show how much has yet to be done here. Sometimes the non-Bayesian position is attacked as leading sometimes to arbitrary conclusions; but any limited degree of arbitrariness there may be is negligible compared with the much greater arbitrariness represented by q.

It is important to realize that the L factor is capable of verification, by repeated experiments; but the q factor is not. This does not mean that the L factor must necessarily be given an oversimplified "frequency" interpretation.

Dempster : Professor Barnard appears to set up a ridiculously strict double standard by requiring that the Bayesian shall say exactly where his prior distribution comes from while assuming that the likelihood is known beyond question. In fact, it is often unclear which of the two sources of uncertainty in the model is the more dangerous.

Rasch : While, of course, admitting the benefit of prior knowledge, if available, I am disinclined to transforming "pure belief"—whether superstitious or not—into a "measure", whether "probabilistic" in some sense or not. Instead I shall ask two questions : In what does the prior information consist ? and : Just where does it come from ?

There seems to be two sources.

One is the insight—direct or indirect—in the field of inquiry of the data, such as it may have accumulated until the actual investigation.

As regards such "insight" I may be a bit more explicit : As "direct" I take, for one thing, knowledge about the conditions under which the data were in fact collected (planned experiment, survey, responses to questionnaires, routine records on the part of the Central Statistical Bureau, regular astronomical observations, or what not). For another thing it includes available theory about the subject matter in question. By "indirect" I am partly thinking of inspired analogies from related fields—more or less distant—partly of general views, e.g. philosophical and technical, both of which may influence the mathematical formalization.

As a case in point I may refer to my realizing the common structure of data on misreadings by schoolchildren exposed to two or more reading tests, and accidents occurring to the population of drivers, when they are riding on different road categories at different days. This gave rise to using the same model in the two cases (the Multiplicative Poisson Model).

However, both direct and indirect insight should, I think, enter *into the construction of the model* that is going to form the basis for the analysis.

The other source is experience with same or related sorts of data, whether it be from previous studies—whoever made them— or from parallel studies in different places (such as serological analyses of the same substances carried out at different laboratories, as organized by WHO).

But in such cases the available data, or the results of analyzing them, might simply be handled parallel to the actual data, on the basis of *models expressed in ordinary probabilistic terms*—elaborated, of course, with due respect to differences in conditions.

In principle, this point of view removes the difference between data collected in the past and in future, in one place or another. It aims at giving a model, once (tentatively) established, as broad a background as at all feasible for checking it.

As a case in point I may mention an investigation of the death rates in Denmark through 50 years which disclosed a certain structure in their dependence on age, in spite of relatively strong changes in living conditions. Afterwards the same structure was found in Sweden, and again, some years later, in United Nations data from numerous countries all over the world.

Barnard : Some notion of repeatability is involved in any form of scientific inference. We would not be interested in the behaviour of Nile floods if we knew

that the Nile would disappear tomorrow, and, along with it, the area of Abyssinia and other parts of Africa whose weather conditions largely determine the Nile floods.

A repetition need not be an exact replication. Thus a measurement of length to 1 mm may be "repeated" by a test whether the length is $>$or $<$ 100 cms. And a measurement of rainfall around the Blue Nile may indirectly "repeat" a measurement of the height of a Nile flood. The essential feature is the accumulation of *independent* pieces of evidence bearing on a given topic. And the meaning of "independence" here is not mere statistical independence (cf. my 1949 paper, pp. 119-120).

Cox : Dr. Basu has talked of analysis not involving a sample space. Yet the start of his treatment is that a parameter ω is given. Quite apart from the issue that the formulation of an appropriate ω is often a key point, how can ω be given a physical meaning without some notion of repetition, even if hypothetical, and hence how can consideration of some sample space be avoided ?

Lauritzen : It seems difficult to me to give any meaning to the parameter ω without referring to outcomes of other experiments

Rasch : Although agreeing with the view, expressed by Steffen Lauritzen, that assigning a probability distribution to a parameter in general would seem artificial, I may add that there *are* cases, albeit few in my own experience, where such a superstructure in warranted.

By way of an example I may mention measuring the diameters of 500 red blood corpuscles in each of a number of blood samples, taken in quick succession from the same normal person. Each sample shows a most beautiful normal distribution and the estimated standard deviations lie quite close to each other, but the average diameters varied much more than allowed for by the standard error. The reason for this discrepancy was, however, quite clear : During the technical preparation of a blood sample it is exposed to a certain pressure, exerted by hand—therefore sometimes a bit harder than at other times, thus influencing the sizes of all of the blood cells, but not noticeably the differences between them.

This, of course, does not turn the problem into a proper Bayesian one. In the instances of repeated sampling the model applied was : the distribution $N(\xi_i, \sigma^2)$ for diameters within sample no. i and $N(\xi, \tau^2)$ for the variation of mean values ξ_i between samples, which leaves us with an ordinary estimation problem.

Barndorff-Nielsen : In relation to Professor Barnard's remark concerning repeatability of experiments, may I make the following comment. It seems to me that there exists experiments—in the broad sense of the word—which are not repeatable in any real sense, but which do properly belong to the province of science. I am, inter alia, thinking of data pertaining to the geological history of the earth or to the theory of evolution.

Barnard : The current revival of interest in geology is due in large measure to the fact that 1) we have at last another body the moon—which is in some sense a "repetition" of the Earth, and we are beginning to obtain "geological" information

about Mars; 2) we have theories of geological processes (continental drift, etc.) which are still going on and which seem likely to enable us eventually to predict earthquakes, etc ; 3) experimental work on the behaviour of materials under ultra-light pressures, though difficult, is approaching relevance to geological processes. Thus, although the specific history of the earth is not replicated, the processes involved can be, at least to some extent.

Martin-Löf : In response to Barnard, I would like to stress that even when an experiment cannot be repeated (except in our thought as done by Gibbs and von Mises with their ensembles and Kollektivs, respectively) it may be amenable to a statistical analysis. A typical example is Lauritzen's (1973) treatment of the gravitational field of the earth as one observation of a certain Gaussian random field. It is quite enough that we can draw verifiable conclusions from the probabilistic assumptions by means of the interpretation clause which allows us to neglect events of small probability.

Barnard : Professor Basu's claim that the Bayesian will more often be right assumes that the Bayesian's prior will correspond with the actual frequencies arising in the sequence of problems dealt with. But there seems no reason to suppose this will be so. Thus the Bayesian may well be *less* often right.

Edwards : A measure of the unsatisfactory nature of the confidence estimate is its sensitivity to variation in b, a somewhat hypothetical quantity. I suspect that the likelihood interval is not so sensitive.

Dempster : I wish only to record that the Stein and Stopping Rule paradoxes no longer seem to me to deserve the name paradox. There is no mathematical reason to expect Bayesian and confidence probability levels to agree, and their predictive and post-dictive interpretations are, in any case, incommensurable. The Bayesian approach is right in principle, but may be difficult in practice. If the required prior knowledge is too weak for any reasonably objective Bayesian inference to be allowed, I would back off and use a sampling-rule dependent confidence method, carefully pointing out the tricky and weak associated meaning.

Barnard : I may be wrong, but I believe Fisher did not assert any frequency-covering properties for likelihood intervals. He simply asserted that any specific θ_1 outside the interval

$$\{\theta : L(\hat{\theta})/L(\theta) \leqslant 100\}$$

would have plausibility, relative to the maximum likelihood value $\hat{\theta}$, less than 1/100. Whenever one wishes to make frequency statements concerning a *single* parameter value θ_1, considered by itself, one must consider sampling distributions in some way (unless, of course, one is prepared to assume a distribution of θ ("prior" distribution) as true of the set of cases with reference to which the frequency is asserted.)

Dempster : I feel that the non-Bayesians in this discussion have not yet been sufficiently nudged to face the difficulties in their position. I propose therefore that

we consider a game which can actually be played, and which I believe goes to the heart of the issue. Imagine N pairs of statisticians (A_i, B_i) for $i = 1, 2, \ldots, N$ where A_i is non-Bayesian and B_i is Bayesian. Each pair engages an agent C_i to determine a parameter value θ_i where A_i and B_i have some common understanding of how the determination is to be made (e.g., asking a random man in the street for a random number) but neither A_i nor B_i are given the value θ_i. Instead, an experiment is performed, say a sequential experiment, which allows θ_i to be estimated. Both A_i and B_i have a common access to the results of the experiment. A_i then creates a 95% confidence interval I_i for θ_i, which necessarily depends on the sampling rule as well as the likelihood. B_i is then offered the choice of sides in a wager over $\theta \epsilon I_i$ and $\theta \notin I_i$ at odds of 19 to 1. A referee totals the net gain or loss of the A team from or to the B team over the N wagers, and declares the winning team accordingly.

There is of course no guarantee that either team will win, even for very large N. The defining property of the confidence intervals undeniably holds when the experimental model specification holds, but this property is inadequate to render the above game fair unless each B_i chooses his side of the wager according to a rule free from both prior knowledge and experimental data. In the real world, every scrap of available information will be used, hence the confidence interval property is inadequate for much of statistical practice. A simplistic Bayesian property also holds, namely, that the Bayesian can quite generally expect positive long run gain under his assumed probability models. But this property is also inadequate since no realistic Bayesian would expect all his model specifications to hold up in a long-run practice.

Where do we stand ! My own view is to distrust non-Bayesian decision theory since it fails to model the free choice aspect of decision-making. While there is no *carte blanche* in favour of Bayes, I do believe that the B-team will very often win in the real world precisely because it can reflect real prior knowledge, at least sufficiently well to stay in the black. This is a matter of judgement, not proof.

Kalbfleisch : Professor Dempster has raised the question as to why the many adherents to the frequentist theories of inference have raised no specific objections to this paper. For my part, I find that the paradoxes outlined in this paper are forceful and do lead me to the conclusion that $\mathcal{F}(a)$ and $\mathcal{F}(b)$ cannot be viewed as solutions to all problems. But, the arguments leading to this conclusion are themselves frequentist in nature and there is the feeling that this strengthens rather than weakens the frequentist position. The justifications for accepting the likelihood principle that Professor Basu gives are not essentially different from those given by Birnbaum, and as I have pointed out there are objections which can be raised to these arguments.

The fact that the likelihood function alone is not enough, as Basu's exposition suggests, leads us to try to supplement it—either with the prior information q or with various frequentist arguments—for the solution of certain problems. I think

much is to be said for a weaker sequence of principles (like those I have suggested) which allow for many different approaches such as tests of significance, confidence procedures, procedures of the type $\mathscr{F}(a)$ and $\mathscr{F}(b)$ and Bayesian methods, each applying to certain problems and not to others.

Edwards : Extreme paradoxes such as Stein's are intended to provide us with results so conflicting that we are bound to vote one way or the other. In practice they leave us bemused, and it may be better to focus on less extreme but more realistic examples which similarly contrast likelihood and confidence principles by making use of distributions with unusually long tails.

Consider the case in which a theoretical physicist predicts the value of a fundamental parameter to be $\mu = 0$. After many years' work practical physicists have made just two measurements, 11.5 and 13.5, and then their apparatus blew up. It is agreed that these measurements may be regarded as a random sample from a normal distribution with unknown variance. Forming the statistic t on one degree of freedom, it is 12.5, not significant at the 5% two-tailed point. But on a support test (see Table 6 of *Likelihood*) the increase in support available is $\ln(1+(12.5)^2)$, a likelihood ratio of 157.25, an impressive amount.

Barnard : Concerning Professor Basu's example about adding likelihoods, I said that the Bayesian consider it is *always* possible to add them, i.e. to find λ such that "α or β" $= \lambda\alpha+(1-\lambda)\beta$.

Dempster : Only "always-Bayesians" think it *always* possible !

Barnard : I agree.

I said it was only *sometimes* possible to add likelihoods. So long as we are considering only small sample sizes, Basu's nearly identical hypotheses give the same likelihood orderings and so they clearly can be combined. But larger samples could show up differences between the hypotheses, which could become important, and then one could not add them. Thus, in my view, one cannot always add.

Dempster (note added in written version) : What I had in mind is that some Bayesians may feel comfortable switching over to a significance testing mode to provide checks on their assumed models. Such Bayesians, including myself, are "sometimes Bayesians" (so B's in Barnard's abbreviation) rather than "always Bayesians".

Barnard : I believe the stoping rule paradox was first brought up by Bartlett in a letter to me in the middle 50's. Armitage independently raised it in the discussion initiated by Savage. Although my views on it have not always been the same, I now think it simply serves to show that likelihoods are relevant to comparisons of *pairs* of (simple) hypotheses; they cannot apply to statements involving a single hypothesis, considered on its own. For the case stated, with n fixed, and x being the variable

$$\frac{|\bar{x}|}{\sqrt{n}} \geqslant$$

rejects the hypothesis $\mu = 0$. But if $|\bar{x}|/\sqrt{n}$ is fixed, and n is variable, the test criterion becomes n; low values of n will tend to reject the hypothesis.

Author's reply : We are talking about statistical data—data equipped with statistical models. We are debating about the basic statistical question of how a given data $d = (\mathcal{E}, x)$, where $\mathcal{E} = (\mathcal{X}, \Omega, p)$ is the model and x is the sample, ought to be analysed. My submission to you is that the likelihood principle of data analysis is unexceptionable. The principle simply asserts that if our intention is not to question the validity of the model $\mathcal{E}$ but to make relative (to the model) judgements about some parameters in the model, then we should not pay attention to any characteristics of the data other than the likelihood function generated by it. From the discussions it would appear that very few amongst us is in full agreement with the above proposition. The Neyman-Pearson-Wald anti-thesis to the likelihood principle is what we may call the principle of performance characteristics which requires us to evaluate the data in full perspective of the sample space. Few, if any, amongst us seem to have any conviction in this unconditional 'sample space' approach to data analysis.

What I am saying is that, for one who truly believes in the likelihood principle, there is hardly any choice left but to act as a Bayesian. If L is the 'whole of the relevant information contained in the data' then we ought to match L with 'all other information' q on the subject. In point of fact we usually have a lot of other information. How can we ignore q ? It seems to me that only an honest Bayesian can give a sensible answer (however clumsy and incompetent it may appear to non-Bayesians) to the basic question : How to analyse a given data ?

Professor Barnard likes the likelihood factor L but does not care for the Bayesian's prior q. He is arguing that the former is verifiable but the latter is not. Our concern here is not with the verification of assumed models but with the question of data analysis relative to such models. In any event, the kind of experiments that we come across in scientific inference can hardly be called repeatable in any meaningful sense of the term. Who has ever heard of a scientific experiment being repeated a number of times with the purpose of checking on the authenticity of an assumed likelihood function ? The likelihood L is no less subjective and hardly any more verifiable than the prior q.

Irrespective of whether we believe in repeatability of experiments and frequency interpretation of probability or not, we are all immensely concerned with one kind of frequency, namely, the long run relative frequency of success in our inference making efforts. Whether the Bayesian method of data analysis is superior to any other well defined method cannot be proved mathematically. The long run success of an individual Bayesian will surely depend on his ability to come up with realistic q's and L's. Professor Barnard remarked that a Bayesian can well be *less* often right if his specification of the prior q is off the mark. He is apparently visualizing a sequence of identical experiments in which the model and, therefore, the L factor is

always right but the same off key q is being used again and again. If the Bayesian is allowed to update his prior q for each experiment in the light of his past accumulated experience, then there is no reason to believe that he will fare badly in the long run even in such an unrealistic hypothetical sequence.

In real life, a practising statistician faces a sequence of different inferential problems about different parameters. If in each case he really applies his mind to the task of constructing a realistic likelihood scale L and carefully goes about the task of quantifying the prior information q then it seems entirely believable to me that our Bayesian will fare much better than a traditional 'sample space' data analyst. For one thing, the 'sample space' analyst has to work with a plethora of likelihood functions—one for each point in his sample space. Naturally he can work with only rather simplistic (and, therefore, unrealistic) statistical models. The Bayesian is never inhibited by such constraints. Since he has to work with only one likelihood function—the one that corresponds to the observed sample—he can boldly reach for more sophisticated (and, therefore, more meaningful) statistical models.

I am certainly not averse to the idea of sample space. As Professor Cox pointed out, in some cases even the parameter (say, the true weight of the chalk stick that I am holding in my hand) cannot be defined without the idea of repeated measurements. At the time of planning a statistical experiment we of course need to speculate about its sample space. But with an experiment already planned and performed, with the sample x already before us, I do not see any point in speculating about all the other samples that might have been.

The Bayesian and the Neyman-Pearson-Wald theories of data analysis are the two poles in current statistical thought. To day, I find assembled before me a number of eminent statisticians who are looking for a via media between the two poles. I can only wish you success in an endeavour in which the redoubtable R. A. Fisher failed.

References in the discussion

Barnard, G. A. (1949): Statistical inference. *J. Roy. Statist. Soc.* Ser. B, **11**, 119-120.

Edwards, A. W. F. (1972): *Likelihood,* Cambridge University Press.

Fisher, R. A. (1912): On an absolute criterion for fitting frequency curves. *Mess. Math.*, **41**, 155-160.

Fisher, R. A. (1922): On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London*, Ser. A, **222**, 309-368.

Jeffreys, H. (1938): Maximum likelihood, inverse probability, and the method of moments. *Ann. Eugen*, **8**, 146-51.

Lauritzen, S. L. (1973): The probabilistic background of some statistical methods in physical geodesy. Meddelelse nr. 48, Geodætisk Institut, Copenhagen.

BARNARD-BASU CORRESPONDENCE

After the conference, Professor Barnard and Professor Basu corresponded on some points in the essay presented by Basu. On the proposal of Basu, and with the consent of Barnard, we reproduce the correspondence in the following. (Reference to pages in the essay are in accordance with the present numbering).

Brightlingsea, 18th May, 1973

Dear Dev,

1. It was good to see you in Aarhus, and I hope we meet again soon. I liked your paper, especially the first part, which was a very clear account of issues around the Likelihood Principle. But, as I said, I think in Part II you are not wholly fair to Fisher—and having checked with my own papers, which I could not do in Aarhus, I think you are not wholly fair to me.

2. First, on p. 23 you say "Fisher tried very hard to elevate maximum likelihood to the level of a statistical principle. Though it has since fallen...". I don't think this is true. The matter is not easy to discuss in a precise way without specifying precisely what we understand by the problem of point estimation. Nowadays there are many people who seem to identify this with the decision problem, to find a function of the observations which will minimize the mean square deviation from the true value. This is certainly not the sense in which Fisher understood the problem. But my understanding of Fisher is that he pointed to the advantages of the maximum likelihood method, in regular situations, but never claimed it as a matter of principle. For instance, the passage beginning "A realistic consideration of the problem of estimation..." on p.157 (1st Ed.) or p.160 (2nd Ed.) in *Statistical Methods and Scientific Inference* shows what I mean.

3. At the same time, I venture the following assertion about ML. : Let us call a method of estimation "algorithmic" if, given the specification of the density function of the observations (i.e. given the model), the estimate derived can be obtained by a standard mathematical process such as solution of an equation, maximisation of a given function, etc. I assert that no algorithmic method of estimation is known which is superior to ML.

4. Can you produce a counter-example ? In case you should refer to Bayes, I will accept integration of a given function as an algorithmic process; but you must also give an *algorithm* for determining a (reasonable ?) "prior".

5. Next, on your pp. 23-24 you refer to "likelihood intervals" as "likelihood confidence intervals". This would suggest that covering frequency properties are claimed for them, when in fact this is not so, except in specific cases when additional conditions are satisfied. It was, so far as I can remember, always clear both to Fisher and to me that an interval defined as your I_λ would not necessarily cover the true value with

any particular frequency. Your subsequent examples which bring this out in a very strong way should therefore, I think, make clear only that these intervals do not possess a property which was never claimed for them.

6. On p. 28 I think the statistical intuition of Sir Ronald *would* have been outraged by the suggestion you make, since the data specified are not inconsistent with the following numerical values:

$L(\omega_1) = 0.011$, $L(\omega_1') = 0.01$, $L(\omega_2) = 0.101$, $L(\omega_2') = 0.10$ and the prior probabilities 0.25, 0.005, 0.25, 0.495, respectively. A priori, the hypotheses $\omega \epsilon A$ and $\omega \epsilon B$ are equally probable, but given the data, their probabilities are 0.028 and 0.04955. Thus the data support B better than A in this case. We certainly could not say, in general, the opposite.

7. More generally, your supposition about adding likelihoods amounts to an assumption that all hypotheses are equally probably a priori. This can be made self-consistent; but I do not accept it as true, any more than Fisher did.

8. I find Fisher's analogy of "the height of Peter or Paul" a good analogy. If we were told that this was to mean "Choose Peter or Paul with equal probability, and then measure the chosen one's height", the phrase would acquire a definite meaning, as a random variable.

9. Your example on pp. 34-35 I find unconvincing, because if your Martian were prepared to regard the range (9.9, 10.1) as of negligible width, he would do this in the first place, and so reduce your second case to the first. But if (as might be), he was interested in being *exactly* right, with "a miss" being "as good as a mile" (as the saying goes), then in the second case his best bet really would be $\theta = x$.

10. In your discussion of the fiducial argument on p. 38, I think you should say, to begin with, that $X-\theta$ is $N(0, 1)$, and then proceed to discuss θ and X on a symmetrical footing. There is no particular reason to suppose that either is unobservable.

11. With Buehler's argument, on p. 41, I think you should point out that, unless the M_2 mornings have *positive* density in the long run—and there is nothing to guarantee this—then Paul will, in the long run, be right no more often than 50% of the time.

12. A small point, on your p. 45. I enclose an offprint which indicates that I was considering the stopping rule paradox before 1964, and the associated idea of tests of power 1. The priority over Darling-Robbins is unimportant, but since you have been referring to me, it should perhaps be made clear that, presumably, I have some way of dealing with the problem.

13. Finally, on p. 54, you say my likelihood interval fails the Stein test "miserably". I think you will find it meets the frequency test exactly.

George

9

Manchester, 29th May, 1973

Dear George,

1. Many thanks for your letter of May 18 which I find very interesting and informative. My views on the various issues raised by you are recorded below. Please note that the paragraphs of this letter correspond to those of yours.

2. I am reassured to learn that you regard ML only as a method of point-estimation. I am however not so sure about Sir Ronald's own views on the subject. In any case, hardly anything can be said about Sir Ronald's views on Statistical Inference that cannot be denied. In paragraph 3 of p. 49 in Hacking's book you will find a reference to the Fisher Principle of ML.

3. I am somewhat bewildered by your challenge about producing an "algorithmic" method of point-estimation that is "superior" to ML. Superior in what sense ? If you are asking for a method B that is universally (i.e., for all models) and uniformly (i.e., for all parameter values in each particular model) superior to ML in the usual sense of some average performance characteristics then I am afraid I have nothing tangible to offer. But then I can as easily counter your challenge by producing a method B and then asking you to produce something "superior" to that. In Section 9 of my essay I have elaborated at length on my objections to ML as a method. My objections stem mainly from the fact that the method has nothing to do with the two essential ingredients of inference making that are always present in some measure in every realistic situation and which I have denoted in my essay by the symbols q and Π.

4. Regarding your remark in paragraph 4, I do not know how a (honest) Bayesian's prior can be characterized in terms of the mathematical description of the model. I have made it amply clear why any such attempted characterization will violate the likelihood principle and, therefore, the very essence of Bayesianism.

5. Without disagreeing with your comments in paragraph 5, I have only to say that when I use the word "confidence" I tend to associate it with the elusive notion of a "measure of belief" rather than with that of "frequency probability". With my examples I have been trying to establish this simple fact that there exists no logical (coherent) basis for supposing that a likelihood interval I_λ with a sufficiently large λ has a claim to a large measure of assurance about the true θ lying in that interval. My examples underline the crucial (and to me self-evident) fact that the "information" contained in the likelihood function can be analysed only in the context of the background knowledge q and the inferential problem Π. I consider it utterly self-defeating to try to build a theory of inference on likelihood alone.

6. Your remarks in paragraph 6 made me happy in the knowledge that you are not averse to prior probabilities. It seems to me that we are talking on slightly different wave lengths but essentially about the same thing. We are agreed then that there are two sources of information—the prior knowledge q and the likelihood measure L of support-by-data. You have produced an example where the L-support

for the composite hypothesis A is greater than that for B, the q-support for A is the same as that for B, but the $(L+q)$-support for A is less than that for B. Where is the contradiction ?

7. Regarding your remark in paragraph 7, I shall readily concede that the supposition that the likelihood support-by-data is an additive measure is tantamount to the supposition that *to the data* (or the ignorant Martian) all simple hypotheses are equally probable a priori. Of course, I do not believe in the Martian's "equally distributed ignorance" any more than you do or Fisher did. That is why this insistance about the meaninglessness of L by itself and about the necessity of matching it with an honest prior of the seientist.

8. As regards "the height of Peter or Paul", I still fail to see why it is a better analogy to "the likelihood of A and / or B" than the natural analogy of "the probability of A and/or B". With this (false) analogy Fisher dismissed the Bayesian insight about the likelihood being something that is meant to be weighted and then accumulated. How does your random vhoice between Peter or Paul make the analogy a better one ?

9. My example on pp. 34-35 was constructed to demonstrate the fact that methods like ML, LR, etc. are disoriented to the task of inference making. Apart from the fact that such methods do not make any use of q and Π, they are also based on the popular misconception that likelihood is a point-function and as such can be interpreted only by maximization and by ratio-comparisons.

10. I must admit that I can never cease to be mystified by the fiducial/structural probability arguments of Fisher/Fraser. How can I "proceed to discuss X and θ on a symmetrical footing" when they are not? I have observed $X = x$ and am trying to make an inference about θ. I also have some pre-conceived ideas about θ. Where is the symmetry ?

11. Paul ought to be able to recognize some event like M_2 ($\theta = -1$ or 0) that has "positive density in the long run". Otherwise, his ignorance about Peter's θ is of such a monstrously all-consuming kind (a uniform prior over all integers ?) that I refuse to speculate about it.

12. It was very interesting to read through the off-print you sent me. It shows that the stopping rule paradox led you to the idea of tests with power one. I mean to find out from Professor Robbins as to how he was led to the same idea.

13. I am sorry for the error on my p. 34 where I said that the interval $J_B(y)$ fails the Stein test. Please accept my apologies. I mean to re-write p. 34 with my debt to you acknowledged in a foot-note.

With all the best,
yours sincerely,
D. Basu

Brithtlingsea, 5th June, 1973

Dear Dev,

1. I had better begin by confessing I write this under the selfimposed handicap that I lost my copy of my letter; so please forgive any resulting deviations from logical order. My comments are numbered to yours.

2. I wish I could remember what I said that can have led you to the first sentence. ML is *primarily* a method of point estimation, and as I read Fisher, this is how he understood it. On referring to Hacking, I find I have marked the passage you mention as being in error. And as far as Fisher's views on inference are concerned, I would have thought we can take his "Contributions to Mathematical Statistics", and "Statistical Methods and Scientific Inference" as representing his views, and it is reasonable to ask, if someone says Fisher took a certain view, that he should be asked to support the statement by some reference to Fisher's works—not to Hacking's, or anyone else's.

The nearest, I think, you could come to a quotation to justify your statement about Fisher is to be found on p. 100 of Anthony Edwards' book, last paragraph. But I think this clearly, in fact, shows your statement to be unjustified.

3. My challenge about producing a better algorithm than ML stands—and you can determine the sense of "superior" in any reasonable way you like, so long as you say what it is. Of course I am not asking for something that is universally and uniformly superior.

4. I agree. But since the mathematics of Bayes Theorem are very simple, within the scope of any mathematician who can integrate, acceptance of the Bayesian position means that statistics texts will need to concentrate on the very difficult task of enabling people to assess for themselves their prior distributions and their loss functions. I say very difficult because for many of us, we are unaware often of the existence of these things (and, indeed, unpersuaded).

5. I have now checked what Fisher said about likelihood intervals, and it is clear that he, no more than I, did not think that a likelihood interval I_λ would have (except in regular asymptotic cases) any particular probability of containing the true value. Thus, in arguing as you do, I think you are flogging a dead horse. But of course, the fact that I_λ has no particular probability of containing the true value do not justify your "crucial (and to you self-evident) fact". I agree with your last sentence, but nonetheless think it worthwhile to see how far we can go with a theory based on likelihood alone: and clearly I think one can go further than you suggest.

6. Considering that I advocated the use of prior probabilities in 1946 when such a point of view was far from popular I think it clear that I am not averse to them, when they exist, in the sense that they can be subjected at least in principle to some sort of objective verification. And of course I agree that there are two

sources of information. But the question is, can the prior knowledge *always* be expressed in terms of prior distribution ?

As to the example of course there is no *contradiction*; but there is a paradox. If one piece of information is neutral as between A and B and the other piece favours A, it surely is odd that the two together should favour B. Such a thing cannot happen with simple likelihoods.

7. I think we agree here.

8. You think Fisher's analogy false because you, unlike him, take a Bayesian view.

With regard to your "and/or" what you say on p. 27 of your Part 2 is, I think, false. Because A will denote a different parameter value from B, and this will imply that A and B are incompatible. Thus the "or" really is the disjunctive "or". If it were "and/or" one could say that the likelihood of "A or A or A" was 3 times the likelihood of A, whcich is absurd.

9. I agree with what you say. But I do not find your demonstration convincing.

10. The symmetry is, that I *might* have observed θ and be trying to make an inference about x. As to preconceived ideas, I may also have such ideas about x. It is part of the argument that I have no *knowledge* about θ (or, respectively x), other than that specified.

11. A uniform prior for θ over all integers is not required. I do not follow the sense of your "all-consuming". The information given by the observations is not "consumed" by the prior ignorance.

12. I would prefer the term "tests with power one" to the terms "Darling-Robbins type tests", seeing that Barnard published and used such a test in practice three years before Darling or Robbins.

The term "Darling-Robbins type tests" should, I think, be used for tests whose power function is discontinuous.

13. Many thanks.

Best regards
George Barnard

Manchester, 12th June, 1973

Dear George,

Many thanks for your letter of June 5. Excepting for two points, I must concede you the last word on all the other issues.

I find your remarks on "and/or" in paragraph 8 very confusing. May be the difficulty is only a matter of semantics. In every introductory course on probability theory, don't we always carefully explain why the expression Pr(A or B) must not

be understood to mean "probability of either A or that of B" ? We then explain that the "or" is not to be used in its usual disjunctive sense of "either-or" but in the "accumulative" sense of the set-theoretic/logical connective union/and-or. After that we have a hard time (especially if we take the subjectivist point of view) explaining why $\Pr(A \text{ or } B) = \Pr(A)+\Pr(B)$ when A and B are exclusive events. In p. 27 of my essay I only suggested that the trouble with the Fisherian analogy of "the height of Peter or Paul" for "the likelihood of A or B" lies in the fact that the "or" in the former is the disjunctive "either-or", whereas the "or" in the latter ought to be understood in the same accumulative sense as we understand it in $\Pr(A \text{ or } B)$. Why not ? After all Fisher wanted us to look upon likelihood as an "alternative measure of rational belief".

Regarding your comments in paragraph 11, the "uniform prior over the infinite set of all integers" was cited by me only as an example of a "monstrously all-consuming" (if you do not like the word "all-consuming", please read it as "all-pervading") state of Paul's prior ignorance about the integral parameter θ that makes him allot zero (relatively, that is) prior probability to every finite set of integers. That sensible looking posterior distributions (or knowledge about θ) can often be (mathematically) derived from such a monstrous lack of prior information, is nothing but a piece of mathematical curiosity to me.

With all the best,
yours sincerely,
D. Basu

Brightlingsea, 18th June, 1973

Dear Dev,

Thanks for your letter and for the copies of mine. I now have them all clipped together with your paper, so if I lose one I lose the lot.

About the "or" and "and/or", I guess I should try a different approach, along lines I gave in my second talk in Aarhus. Let us agree that a *simple statistical* hypothesis H is one which specifies *completely* a probability distribution $P(x : H)$ on a sample space S (finite, for simplicity; x is a point in S). Since x is a point, it specifies *completely* a possible result of the experiment to which H relates; it can therefore be called a *simple* event.

Now it is a property of experiments that we can *always* imagine them modified in such a way that the sample space S becomes S', where the *points* of S' correspond to the sets of a partition of S. Thus, for example, in throwing a dice, $S = \{1, 2, 3, 4, 5, 6\}$; we can imagine ourselves incapable of counting the spots, but only capable of seeing whether there is an even or an odd number of them, in which case $S' = \{E, O\}$ corresponds to the partition $S = \{2, 4, 6\} \bigcup \{1, 3, 5\}$ of S. It is reasonable

to require that the hypothesis H should specify the probability distribution on S' as well as that on S. Evidently this can be done if we use the addition rule, so that, e.g. $P(E:H) = P(2:H)+P(4:H)+P(6:H)$. This is, essentially, what leads us to add probabilities.

You will find distinctions such as those I have indicated in any careful treatment of the foundations of probability. Thus, for example, Renyi (in Foundations of Probability) distinguishes between the *outcome* of an experiment (my *simple* event) and an *event*. An outcome is a *point* in the sample space, an event is a set of points. For Renyi, an *experiment* ξ is a non-empty set $\mathscr{X}$ of elements x called outcomes of the experiment and a σ-algebra $\mathscr{A}$ of subsets of $\mathscr{X}$ called observable events. He writes $\xi = (\mathscr{X}, \mathscr{A})$.

In Renyi's terminology, what I am saying is that given any experiment $\xi = (\mathscr{X}, \mathscr{A})$, and any sub-$\sigma$-algebra $\mathscr{A}'$ of $\mathscr{A}$, there exists an experiment $\xi' = (\mathscr{X}, \mathscr{A}')$. It is this fact that gives importance to the addition rule for probabilities, in applications to experiments.

Now given a family Φ of simple hypotheses, with $H \epsilon \Phi$, what general logical process is there that corresponds to going from $\mathscr{A}$ to $\mathscr{A}'$? I assert that in general there is no such process, although in special cases there may be.

Specifically, given the experiment $\xi = (\mathscr{X}, \mathscr{A})$, and a family Φ of (simple) hypotheses (completely) specifying probability distributions on $\mathscr{X}$, I say that a subset of Φ is a *disjunctive* subset iff there exists a subalgebra $\mathscr{A}'$ of $\mathscr{A}$ such that every H in the subset assigns the same probability to every member of $\mathscr{A}'$. In the absence of a prior distribution over Φ, the disjunction of a set of hypotheses H can be considered to exist only if the set is a disjunctive set. For only then can the disjunction itself be regarded as a simple hypothesis (about the experiment $(\mathscr{X}, \mathscr{A}')$).

I fear you may find this all too muddling. I'll send you a copy of my second Aarhus paper when I have written it out. Briefly, I am pointing to the fact that the disjunction of simple events can be regarded as a simple event in another experiment; but the disjunction of simple hypotheses can *not* in general be regarded as a simple hypothesis, because an arbitrary set of Φ will not necessarily be disjunctive.

Incidentally, I have referred to Renyi because I have it handy; there is a similar distinction made by Kolmogoroff, though I don't remember just how he does it.

Regarding the "uniform prior", I guess we should agree to differ. All our analyses of real situations are to some extent approximations. Whether such "complete ignorance" is a useful approximation in any situation will be to some extent a matter of taste.

Yours,
George

Paper received : May, 1973.

On the Elimination of Nuisance Parameters

DEBABRATA BASU

Eliminating nuisance parameters from a model is universally recognized as a major problem of statistics. A surprisingly large number of elimination methods have been proposed by various writers on the topic. In this article we propose to critically review two such elimination methods. We shall be concerned with some particular cases of the marginalizing and the conditioning methods. The origin of these methods may be traced to the work of Sir Ronald A. Fisher. The contents of the marginalization and the conditionality arguments are then reexamined from the Bayesian point of view. This article should be regarded as a sequel to the author's three-part essay (Basu 1975) on statistical information and likelihood.

KEY WORDS: Marginalization and conditionality arguments; Specific and partial sufficiency; Ancillary and S-ancillary statistics; Unrelated parameters.

1. THE ELIMINATION PROBLEM AND METHODS

The problem begins with an unknown state of nature represented by the parameter of interest θ. We have some information about θ to begin with—e.g., we know that θ is a member of some well-defined parameter space Θ—but we are seeking more. Toward this end, a statistical experiment $\mathcal{E}$ is planned and performed which generates the sample observation x. Further information about θ is then obtained by a careful analysis of the data $(\mathcal{E}, x)$ in the light of all our prior information about θ and in the context of the particular inference problem related to θ. For going through the rituals of the traditional sample-space analysis of data, we must begin with the invocation of a trinity of abstractions $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$, where $\mathfrak{X}$ is the sample space, $\mathcal{A}$ is a σ algebra of events (subsets of $\mathfrak{X}$), and $\mathcal{P}$ is a family of probability measures on $\mathcal{A}$. If the model $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$ is such that we can represent the family $\mathcal{P}$ as $\{P_\theta: \theta \in \Theta\}$, where the correspondence $\theta \to P_\theta$ is one-one and (preferably) smooth, then we go about analyzing the data according to our own light and are thankful for not having to contend with any nuisance parameters.

However, instances of statistical models with $\mathcal{P}$ indexed by θ alone are very rare. Typically, we have to work with a family $\mathcal{P}$ that is indexed as

$$\mathcal{P} = \{P_{\theta,\phi}: \theta \in \Theta, \phi \in \Phi\} ,$$

where ϕ is an additional unknown parameter. If the inference problem at hand relates only to θ and if information gained on ϕ is of no direct relevance to the problem, then we classify ϕ as the nuisance parameter.

The big question in statistics is: How can we eliminate the nuisance parameter from the argument? During the past seven decades an astonishingly large amount of effort and ingenuity has gone into the search for reasonable answers to this question. Broadly speaking, this collective endeavor of the community of statisticians may be classified into the following overlapping categories:

1. To plan the experiment ε in such a fashion that the model is related to the parameter of interest and is relatively free of disturbing nuisance parameters. In this article we are not concerned with the important problems of planning experiments. Our concern is with the problem of data analysis. However, a few elimination methods, such as randomization and sequential sampling which will be discussed in a sequel, may well be classified under this heading.
2. To justify a replacement of the basic model $(\mathfrak{X}, \mathcal{A}, \mathcal{P})$ by a related θ-oriented model $(\mathcal{T}, \mathcal{C}, \mathcal{Q})$, the family $\mathcal{Q}$ is indexed by θ alone. The marginalization and the conditionality arguments that we shall be examining in this article belong to this category.
3. To estimate the nuisance parameter away; that is, to substitute the unknown nuisance parameter ϕ by an estimated value $\hat{\phi}$. This classical method of elimination is used repeatedly in the large sample theory of statistics.
4. To Studentize in the manner of W.S. Gossett with the idea in mind to construct a reasonable looking pivotal quantity involving the sample x and the parameter of interest θ.
5. To invoke the invariance argument of Pitman-Stein-Lehmann. This particular marginalization argument will be examined in a subsequent article.
6. To delimit the argument to a small class of decision procedures, e.g., unbiased estimators, fixed size confidence intervals, similar tests, etc., whose average performance characteristics are, at least in part, free of the nuisance parameter. Mathematicians love this argument. See, e.g., Linnik (1965, 1968).
7. To eliminate the nuisance parameter from the risk function $r_\delta(\theta, \phi)$ of the decision procedure δ by the invocation of a so-called maximization (or minimax) principle. The recommendation for the choice of δ is then made on the basis of the eliminated risk function
$$R_\delta(\theta) = \sup_\phi r_\delta(\theta, \phi) .$$
In Lehmann (1959) we find this argument used quite frequently. For example, the size of a test is always understood as the maximum probability of committing an error of the first kind.
8. To invoke the fiducial argument of R.A. Fisher. With the departure of Sir Ronald from our midst, we seem to have lost our zest for this novel elimination argument.
9. To justify an elimination of the nuisance parameter directly from the likelihood function $L(\theta, \phi|x)$ generated by the particular data (ε, x). The idea is to construct a new scale $L_e(\theta, x)$ (the suffix e denotes the process of elimination of the nuisance parameter) for a direct comparison of the amount of support that the data lends to various values of θ. The maximization of likelihood with respect to ϕ is the classic example of this kind of elimination.

* Debabrata Basu is Professor, Department of Statistics, Florida State University, Tallahassee, FL 32306. This is a revised version of an earlier work with the same title that was presented at a symposium held at the Carleton University, Ottawa, Ontario, October 24–26, 1974. The earlier version appeared in mimeographed form in Proceedings of Symposium on Statistics and Related Topics, ed., Md. E. Saleh, Carleton Math. Lecture Notes No. 52, 1975.

© Journal of the American Statistical Association
June 1977, Volume 72, Number 358
Theory and Methods Section

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_26, © Springer Science+Business Media, LLC 2011

10. To act like a Bayesian; that is, to fix a prior, compute the posterior, integrate out the nuisance parameter from the posterior to arrive at the posterior marginal distribution of the parameter of interest, and then to let the statistical argument rest on the posterior marginal distribution.

    In addition, we have the choice of a fairly large number of specialized elimination methods: the two-stage sampling plan of Stein (1945), the randomization method of Durbin (1961), the characterization argument of Prohorov (1967), the partial sufficiency argument of Hájek (1967), the M-ancillarity argument of Barndorff-Nielsen (1973), etc.

After this introduction to the problem and methods of elimination, we plunge headlong into the depths of the marginalization and the conditionality arguments and try to sort out a number of ideas related to partial sufficiency and partial ancillarity.

2. MARGINALIZATION AND CONDITIONING

The marginalization method of elimination consists of: Choosing a suitable statistic T: $(\mathfrak{X}, \mathfrak{A}) \to (\mathcal{T}, \mathfrak{B})$, such that the family

$$\mathcal{P}_T = \{P_{\theta,\phi}T^{-1}: \theta \in \Theta, \phi \in \Phi\}$$

of probability measures on $(\mathcal{T}, \mathfrak{B})$ is θ-oriented, i.e., the family $\mathcal{P}_T$ is indexed by θ alone; and then recommending that the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ be given up in favor of the model $(\mathcal{T}, \mathfrak{B}, \mathcal{P}_T)$.

In effect, the method replaces the data $(\mathcal{E}, x)$ by its reduction $(\mathcal{E}_T, t)$, where $T(x) = t$. By $\mathcal{E}_T$ we mean the marginal experiment that may be operationally defined as "perform $\mathcal{E}$ but record only $T(x)$." It is not easy to justify data reduction of the above kind. A great deal of thought and mathematical expertise have gone into the many efforts made so far at such justification. Two distinct major lines of thought in this general direction are: (a) the invariance argument and (b) the partial sufficiency argument. In this article, we shall be concerned with the partial sufficiency argument only.

The conditioning method of elimination consists of: Choosing a suitable statistic Y: $(\mathfrak{X}, \mathfrak{A}) \to (\mathcal{Y}, \mathfrak{C})$ such that the conditional distribution of the sample x, given $Y = y$, is θ-oriented (it depends on (θ, ϕ) only through θ) for all $y \in \mathcal{Y}$; and recommending that the data $(\mathcal{E}, x)$ be analyzed by looking at the sample x, not as a random variable with the unconditional distribution model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ but as a random variable with the θ-oriented conditional distribution model that corresponds to the condition $Y = y$, where y is the observed value of the statistic Y. In effect, the method aims at replacing the data $(\mathcal{E}, x)$ by the conditioned data $(\mathcal{E}_y{}^Y, x)$, where $\mathcal{E}_y{}^Y$ is a conceptual conditional experiment that corresponds to the observed value y of a suitable statistic Y.

For the marginalization argument, the statistic T not only needs to be θ-oriented but also has to be one that, in some sense, summarizes in itself all the relevant and usable information about θ that is contained in the data. Similarly, for the conditionality argument, it is not enough to choose just any statistic Y that will do the elimination job. The static Y needs to be such that, in some meaningful sense, we can assert that referring the observed sample x to the reference set of all possible samples x' with $Y(x')$ fixed at the present observed value $y = Y(x)$ entails no loss of information on the parameter of interest θ. The statistical literature is strewn with logicians' nightmares of the above kind. Let us see what sense we can make of such nightmares.

3. PARTIAL SUFFICIENCY AND PARTIAL ANCILLARITY

In this section we put together a number of mathematical definitions.

Definition 1 (Model): By the model (or statistical structure) of an experiment $\mathcal{E}$ we mean the usual trinity of abstractions $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$.

We suppose that the family $\mathcal{P}$ is indexed as $\mathcal{P} = \{P_\omega: \omega \in \Omega\}$ and call ω the *universal parameter*. Let $\theta = \theta(\omega)$ be the parameter of interest. By a statistic T we mean a measurable map of $(\mathfrak{X}, \mathfrak{A})$ into another measurable space $(\mathcal{T}, \mathfrak{B})$.

Definition 2 (Ancillarity): The statistic T is ancillary if the marginal (or sampling) distribution of T is the same for all $\omega \in \Omega$—i.e., for all $B \in \mathfrak{B}$, the function $P_\omega(T^{-1}B)$ is a constant in ω.

Definition 3 (θ-Oriented Statistic): The statistic T is θ oriented if the marginal distribution of T depends on ω only through $\theta = \theta(\omega)$. That is, $\theta(\omega_1) = \theta(\omega_2)$ implies $P_{\omega_1}(T^{-1}B) = P_{\omega_2}(T^{-1}B)$ for all $B \in \mathfrak{B}$.

Observe that every ancillary statistic is θ oriented irrespective of what θ is

Example 1: Let $x = (x_1, x_2, \ldots, x_n)$, with n fixed in advance, be a sample of n independent observations on a $N(\mu, \sigma)$. Let $D = (x_2 - x_1, x_3 - x_1, \ldots, x_n - x_1)$ be the difference statistic. Clearly, D is σ oriented and, therefore, so is every measurable function $h(D)$ of D. That the class $\{h(D)\}$ of measurable functions of the difference statistic does not exhaust the family of σ-oriented statistics is seen as follows. Choose and fix two functions $h_1(D)$ and $h_2(D)$ that are identically distributed and also a Borel set E in R_1. Since $\bar{x}$ is stochastically independent of D for all (μ, σ), it now follows that the statistic T_E defined as

$$\begin{aligned} T_E(\bar{x}, D) &= h_1(D) \quad \text{if} \quad \bar{x} \in E \\ &= h_2(D) \quad \text{if} \quad \bar{x} \notin E \ , \end{aligned}$$

is σ oriented—indeed, T_E is identically distributed as $h_1(D)$ and $h_2(D)$. It is thus clear that D is not the maximum σ-oriented statistic. In fact no maximum σ-oriented statistic exists. (See Basu (1959) and (1967) for more information on this kind of problem.) In this case we have a plentiful supply of σ-oriented statistics. However, the notion of μ-orientedness is vacuous in the sense that no nontrivial (nonancillary) statistic can be μ oriented. This remark is generally true for the location parameter μ in a location-scale parameter setup.

Definition 4 (Variation Independence): The two functions $\omega \to a(\omega)$ and $\omega \to b(\omega)$ on the space Ω with respective ranges A and B are said to be variation independent if the range of the function $\omega \to (a(\omega), b(\omega))$ is the Cartesian product $A \times B$.

If the universal parameter ω can be represented as $\omega = (\theta, \phi)$, where θ and ϕ are variation independent in the preceding sense—that is, $\Omega = \Theta \times \Phi$ where Θ and Φ are the respective ranges of θ and ϕ—then we call ϕ a variation independent complement of θ. With θ as the parameter of interest, we may then call ϕ the nuisance parameter.

We have not come across a satisfactory definition of the notion of a nuisance parameter. It is only hoped that the above working definition will meet with little resistance. (See Barndorff-Nielsen (1973) for further details on the notion of variation independence.)

By a sufficient statistic, we mean a statistic that is sufficient in the usual sense with respect to the full model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$. The following definition of a specific sufficient statistic appears in Neyman and Pearson (1936). Let ϕ be a variation independent complement of θ.

Definition 5 (Specific Sufficiency): The statistic T is specific sufficient for θ if, for each fixed $\phi \in \Phi$, the statistic T is sufficient with respect to the restricted model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P}_\phi)$, where $\mathcal{P}_\phi = \{P_{\theta,\phi}: \theta \in \Theta, \phi \text{ fixed}\}$.

In Example 1, the sample mean $\bar{x}$ is specific sufficient for μ. In fact, $\bar{x}$ is a minimum specific sufficient statistic for μ. The sample standard deviation s is, however, not sufficient for σ for any specified value of μ. Indeed, a statistic can be specific sufficient for σ only if it is sufficient.

In the spirit of Definition 5, we then define the notion of specific ancillarity in the following terms. As before, let ϕ be a variation independent complement of θ.

Definition 6 (Specific Ancillarity): The statistic T is specific ancillary for θ if, for each fixed $\phi \in \Phi$, it is ancillary with respect to the restricted model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P}_\phi)$.

In other words, T is specific ancillary for θ if it is ϕ oriented, where ϕ is a variation independent complement of θ. It should be noted that the definition of θ-orientedness does not presuppose the existence of a variation independent complement ϕ, but the definitions of specific sufficiency and specific ancillarity (for θ) do.

In Example 1, with σ as the parameter of interest, it is tempting to marginalize to the statistic s. But can we logically justify such a marginalization? In what sense can we say that s summarizes in itself all the relevant and available information about σ in the absence of any information on μ? We shall return to the question later.

Suppose μ is the parameter of interest in Example 1. Marginalization to the statistic $\bar{x}$, which is specific sufficient for μ, will not eliminate σ as $\bar{x}$ is not μ oriented. We shall also lose valuable information on μ if we throw away the s-part of the sufficient statistic $(\bar{x}, s)$ and record only $\bar{x}$. For one thing, we shall no longer be able to speculate about the accuracy of $\bar{x}$ as a point estimate of μ. The marginalization method is of no use for the purpose of eliminating the scale parameter σ. As we have noted earlier, if T is μ oriented then it has to be an ancillary statistic. Surely, we do not want to marginalize to something that has nothing to do with μ! The conditionality argument is also of no use for eliminating σ. If conditioning with respect to Y eliminates σ, then Y has to be specific sufficient for σ. But, as we have stated earlier, every such Y has to be sufficient for (μ, σ). Hence, conditioning with respect to Y will eliminate μ as well! The problem of eliminating the scale parameter σ is not an easy one. Student's t-test and Stein's two-stage sampling plan are classical examples of statistical methodology that were developed to solve the problem in non-Bayesian terms.

The following definition of partial sufficiency is usually attributed to Fraser (1956). But we find the definition clearly laid out in Olshevsky (1940), who attributed it to Neyman (1935).

Definition 7 (p-Sufficiency): The statistic T is partially sufficient (denoted by p-sufficient) for θ if T is specific sufficient for θ and T is θ oriented. From this it is clear that the notion of p-sufficiency for θ presupposes the existence of a variation independent complement ϕ for θ. With the same presupposition, Sandved (1967) defined a notion of partial ancillarity in the following terms.

Definition 8 (S-Ancillarity): The statistic Y is a partial ancillary (S-ancillary) for θ if Y is specific ancillary for θ (Y is ϕ oriented) and Y is specific sufficient for ϕ. It should be noted that in Definitions 7 and 8, we are looking at the same concept but from two different angles. The statistic Y is S-ancillary for θ if and only if it is p-sufficient for ϕ.

The name S-ancillary (ancillary in the sense of Sandved) is due to Barndorff-Nielsen (1973) whose terminology for p-sufficiency is S-sufficiency. Barndorff-Nielsen's mathematical formalization of the twin notions of p-sufficiency and S-ancillarity as a "cut" may be defined as follows.

Definition 9 (Barndorff-Cut): A statistic T: $(\mathfrak{X}, \mathfrak{A}) \to (\mathcal{T}, \mathfrak{B})$ defines a Barndorff-cut of an experiment

$$\mathcal{E} = \{(\mathfrak{X}, \mathfrak{A}, P_\omega): \omega \in \Omega\} ,$$

if there exist two variation independent and complementary subparameters $\theta = \theta(\omega)$ and $\phi = \phi(\omega)$, such that the marginal experiment $\mathcal{E}_T = \{(\mathcal{T}, \mathfrak{B}, P_\omega T^{-1}): \omega \in \Omega\}$ is θ oriented ($P_\omega T^{-1}$ depends on ω only through $\theta(\omega)$) and that each one of the family $\{\mathcal{E}_t{}^T: t \in T\}$ of conditional experiments is ϕ oriented.

The statistic T is then p-sufficient for θ and S-ancillary for ϕ. Observe that every sufficient statistic defines a Barndorff-cut and so also does every ancillary statistic. In the former case $\theta(\omega) = \omega$ and $\phi(\omega)$ is a known constant, and in the latter case it is the other way around.

In Example 1 there exists no Barndorff-cut that separates μ and σ. The following are a few other examples where the definition yields something.

Example 2: Let the random variables $x_i (i = 1, 2, \ldots, m)$ be iid $N(\theta, 1)$, and let $y_j (j = 1, 2, \ldots, n)$ be an independent set of iid $N(\phi, 1)$. Clearly, $\bar{x}$ is p-sufficient and $\bar{y}$ is S-ancillary for θ.

Example 3: Let x and y be independent Poisson variables with means μ and ν, respectively. With the reparametrization $\theta = \mu/(\mu + \nu)$ and $\phi = \mu + \nu$, it can

be checked that $Y = x + y$ is S-ancillary for θ. There does not exist a statistic T that is p-sufficient for θ.

Example 4: Let $x = (y, z, w)$ have a multinomial distribution with $y + z + w = n$ and p, q, $r(p + q + r = 1)$ as probabilities (parameters). The parameters p and q are not variation independent. However, when we reparametrize as $\theta = p$, $\phi = q/(1 - p)$, it is easy to check that the statistic y becomes p-sufficient for θ (S-ancillary for ϕ).

Example 5: Let $0 < \theta < 1$ and $0 < \phi < \infty$. Let X be a random variable with pdf

$$p(x|\theta, \phi) = (1 - \theta)\phi e^{\phi x} \quad \text{for} \quad x \leq 0$$
$$= \theta\phi e^{-\phi x} \quad \text{for} \quad x > 0 \ .$$

Let $x_1, x_2, \ldots, x_n$ be n independent observations on X. Let T be the number of positive x_i's, and let $Y = \sum |x_i|$. Then T and Y are respectively p-sufficient and S-ancillary for the parameter θ.

Note the similarities between Examples 2 and 5. In either case, we have for the parameter of interest θ a statistic T that is p-sufficient and a statistic Y that is S-ancillary. In each case, however, the two statistics are stochastically independent for all possible values of the universal parameter. The fact that this is not generally true is going to bother us in due course.

It will be useful to review the various definitions in terms of the corresponding factorizations of the likelihood functions. To this end let us suppose that the family $\mathcal{P} = \{P_{\theta,\phi}: \theta \in \Theta, \phi \in \Phi\}$ is dominated by a σ-finite measure μ and let $\{p(\cdot|\theta, \phi)\}$ be the corresponding family of probability density functions. To fix our ideas and to avoid all measure-theoretic difficulties let us pretend for the time being that $\mathfrak{X}$ is a countable set and that μ is the counting measure on $\mathfrak{X}$. Corresponding to any statistic $T: \mathfrak{X} \to \mathcal{T}$ we have a factorization (of p) of the form

$$p(x|\theta, \phi) = g(T|\theta, \phi)f(x|T, \theta, \phi) \ ,$$

where g defines the marginal distribution of T and f defines the conditional distribution of x given T. (Our notations are admittedly rather sloppy, but there should be no difficulty in following our meaning.) Consider now the following particular cases of the above general factorization.

Case I: $p = g(T|\theta, \phi)f(x|T)$—this corresponds to the case where T is sufficient.

Case II: $p = g(T)f(x|T, \theta, \phi)$—the statistic T is ancillary.

Case III: $p = g(T|\theta)f(x|T, \theta, \phi)$—the statistic T is θ oriented. The case where T is ϕ oriented is similar.

In the situation where θ and ϕ are variation independent parameters, the notion of θ-orientedness is the same as the notion of specific ancillarity for ϕ. Case III, therefore, also corresponds to the case where T is specific ancillary for ϕ. With θ and ϕ variation independent, we have the next case.

Case IV: $p = g(T|\theta, \phi)f(x|T, \phi)$—the statistic T is specific sufficient for θ. The case where T is specific sufficient for ϕ is similar.

Case V: $p = g(T|\theta)f(x|T, \phi)$—the statistic T is p-sufficient for θ and is S-ancillary for ϕ.

Case Va: $p = g(T|\phi)f(x|T, \theta)$—the statistic T is S-ancillary for θ and is p-sufficient for ϕ.

Instead of looking at factorizations in terms of marginal and conditional frequencies, suppose we consider factorizations of the more general form

$$p(x|\theta, \phi) = G(x, \theta, \phi)F(x, \theta, \phi) \ .$$

The very familiar

Case VI: $p = G(T, \theta, \phi)F(x)$,

when proved equivalent to Case 1, constitutes the well-known factorization theorem for sufficiency. Similarly, the factorization

Case VII: $p = G(T, \theta, \phi)F(x, \phi)$

can be shown to be equivalent to Case IV (the case of specific sufficiency for θ). Now consider

Case VIII: $p = G(T, \theta)F(x, \phi)$.

Is Case VIII equivalent to Case V? It is important to recognize that the answer is in the negative. The examples in Section 9 will clarify the matter. Finally, we have factorizations of the form

Case IX: $p = G(x, \theta)F(x, \phi)$.

It will turn out later that we really should be after factorizations of this form. Clearly, p factors in the manner of Case IX whenever we have a Barndorff-cut separating θ from ϕ (as in Cases V or Va). That the converse is not true will be variously exemplified in Section 9.

4. GENERALIZED SUFFICIENCY AND CONDITIONALITY PRINCIPLES

To understand the logic of the generalized sufficiency and conditionality principles $\mathfrak{s}^*$ and $\mathfrak{C}^*$, it is useful to consider a few hypothetical situations. (For a comprehensive discussion on the sufficiency, conditionality, invariance, and the likelihood principles refer to Basu (1975).)

(i) We have two experimental setups ε and ε', where the former provides information only on the parameter of interest θ while the latter is informative about an unrelated parameter ϕ alone—the parameter ϕ is unrelated to θ in the sense that we do not recognize the relevance of any information on ϕ for the purpose of inference making on θ. Faced with data such as $\{(\varepsilon, x), (\varepsilon', x')\}$, it makes good statistical sense to ignore the second part of the data and concentrate our attention on the relevant part (ε, x).

(ii) Let ε be an experiment whose randomness (probabilistic) characteristics depend only on θ. Having obtained the data (ε, x), suppose we choose to perform a randomization exercise $\varepsilon_{(x)}$ thus arriving at the additional data $(\varepsilon_{(x)}, y)$. If all the randomness characteristics of $\varepsilon_{(x)}$ (possibly influenced by x) are known to us, then the secondary data $(\varepsilon_{(x)}, y)$ cannot give us any additional information on θ, or

on anything for that matter. It makes good statistical sense then to suggest that the analysis of the data $\{(\mathcal{E}, x), (\mathcal{E}_{(x)}, y)\}$ ought to proceed on a total nonrecognition of the randomization exercise $\mathcal{E}_{(x)}$ and the resulting outcome y. Indeed, this is one way of looking at the sufficiency principle $\mathcal{S}$ (see Basu 1975).

(iii) If in (ii) we find that the randomness characteristics of $\mathcal{E}_{(x)}$ are fully known except for a nuisance parameter ϕ that is unrelated to θ, then we are in a situation quite analogous to (i). Conforming to the statistical intuition that told us to ignore $(\mathcal{E}', x')$ in (i), the generalized sufficiency principle $\mathcal{S}^*$ tells us to ignore $(\mathcal{E}_{(x)}, y)$ in this situation.

(iv) We have a choice of k experiments $\mathcal{E}_{(1)}, \mathcal{E}_{(2)}, \ldots, \mathcal{E}_{(k)}$. The randomness structure of $\mathcal{E}_{(y)} (y = 1, 2, \ldots, k)$ is related only to the parameter of interest. Let $\mathcal{E}$ stand for a randomization exercise that selects one of the k experiments with known (predetermined) selection probabilities $\pi_1, \pi_2, \ldots, \pi_k$. The experiment $\mathcal{E}_{(y)}$ selected by $\mathcal{E}$ is then performed resulting in the outcome x. The full data is $\{(\mathcal{E}, y), (\mathcal{E}_{(y)}, x)\}$. Since the part $(\mathcal{E}, y)$ of the data is totally uninformative, it makes good statistical sense to disregard this part of the data and focus our attention on the relevant part, i.e., $(\mathcal{E}_{(y)}, x)$. This is a version of the conditionality principle.

(v) Now, suppose in (iv) above the selection probabilities $\pi_1, \pi_2, \ldots, \pi_k$ are not fully known but depend on (are functions of) an unrelated nuisance parameter ϕ. We are now in a situation that is very similar to (i). The generalized conditionality principle $\mathcal{C}^*$ tells us to analyze the data by concentrating our whole attention on that part of the data—namely $(\mathcal{E}_{(y)}, x)$—that is related to θ.

We are now ready to state formally the two generalized principles of sufficiency and conditionality.

Principle $\mathcal{S}^*$ *(Generalized Sufficiency Principle):* If, in terms of the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$ for the data $(\mathcal{E}, x)$, we recognize the statistic T as p-sufficient (partially sufficient in the sense of Definition 7) for the parameter of interest θ, then the data $(\mathcal{E}, x)$ should be reduced by marginalization to $(\mathcal{E}_T, t)$, where $\mathcal{E}_T$ is the marginal experiment corresponding to T and $t = T(x)$.

Principle $\mathcal{S}^*$ may be stated in a less severe form in the following terms.

Principle $\mathcal{S}^{**}$: If T is p-sufficient for θ, then $T(x') = T(x'')$ implies that the information content (the evidential meaning) of the data $(\mathcal{E}, x')$ and $(\mathcal{E}, x'')$ relative to the parameter θ are identical in all respects. In other words, the data $(\mathcal{E}, x')$ warrants the same inference on θ as does the data $(\mathcal{E}, x'')$.

Principle $\mathcal{C}^*$ *(Generalized Conditionality Principle):* If Y is an S-ancillary (Definition 8) for θ, then the data $(\mathcal{E}, x)$ should be analyzed by reinterpreting it as $(\mathcal{E}_{(y)}, x)$, where $\mathcal{E}_{(y)} (= \mathcal{E}_y{}^Y)$ is the conditional experiment that corresponds to the observed value $y = Y(x)$ of Y.

As we have said before, corresponding to any statistic T we can conceive of a decomposition of the experiment $\mathcal{E}$ into a two-stage experimental setup in which the marginal experiment $\mathcal{E}_T$ is followed by the conditional experiment $\mathcal{E}_t{}^T$ that corresponds to the observed value $t = T(x)$ of T. The original data $(\mathcal{E}, x)$ may then be viewed as $\{(\mathcal{E}_T, t), (\mathcal{E}_t{}^T, x)\}$. If T is p-sufficient for θ then, by definition, the experiment $\mathcal{E}_T$ is θ oriented, and the experiment $\mathcal{E}_t{}^T$ is ϕ oriented. So, in view of (i) and (iii), it makes good statistical sense to invoke principle $\mathcal{S}^*$ and marginalize the data to $(\mathcal{E}_T, t)$. Conversely, if T is S-ancillary for θ then, by definition, $\mathcal{E}_T$ is ϕ oriented and $\mathcal{E}_t{}^T$ is θ oriented. So, in view of (i) and (v), it appears logical that we ought to ignore the $(\mathcal{E}_T, t)$ part of the data and analyze it as $(\mathcal{E}_t{}^T, x)$. This is the generalized conditionality principle $\mathcal{C}^*$.

5. A CHOICE DILEMMA

In the writings of R.A. Fisher we find the conditionality argument used in three different ways: to recover the ancillary information in the data when it is found that the maximum likelihood estimator is not sufficient; to eliminate the nuisance parameter as in the case of the celebrated test of independence with a 2×2 multinomial data; and to generalize the fiducial argument as in the case of multiple observations on a random variable with a location parameter in its distribution.

In Basu (1964), while studying in depth Fisher's recovery of information argument, the author discovered a disturbing inherent difficulty in the conditionality argument. The difficulty flows from the fact that, in general, there does not exist a largest ancillary statistic in the sense of the usual partial order on statistics. Even in the simplest of situations we may have two ancillary statistics Y and U such that the statistic (Y, U) is not ancillary. Indeed, the pair (Y, U) may be fully informative, i.e., sufficient. In such a situation, the conflict between which of the two ancillaries to choose for the purpose of conditioning the data remains unresolved, despite some valiant efforts by Barnard and Sprott (1971) and Cox (1971), in non-Bayesian terms. The generalized conditionality (S-ancillarity) argument founders on the same non-uniqueness rock. We reproduce here an example from Basu (1964) that has attracted a lot of attention from non-Bayesians.

Example: Let X be a random variable with range $\{1, 2, 3, 4\}$ and probability distribution

$$X: \quad 1 \qquad 2 \qquad 3 \qquad 4$$
$$\text{Prob}: \quad (1-\theta)/6 \quad (1+\theta)/6 \quad (2-\theta)/6 \quad (2+\theta)/6 \ ,$$

where $0 < \theta < 1$. We have n independent observations on X. The cell frequencies $x = (n_1, n_2, n_3, n_4)$ constitute the minimum sufficient statistic. The likelihood function is

$$L(\theta) = (1-\theta)^{n_1}(1+\theta)^{n_2}(2-\theta)^{n_3}(2+\theta)^{n_4} \ .$$

Let us write Bin (n, p) for the Binomial distribution with parameters n and p. Observe that $Y = n_1 + n_2$ is an ancillary statistic with probability distribution Bin $(n, \frac{1}{3})$ and that $U = n_1 + n_4$ is another ancillary with distribution Bin $(n, \frac{1}{2})$. If we condition x by Y then we can look upon the data as a pair of independent random variables n_1 and n_3 that are distributed as

$$\text{Bin } (Y, (1-\theta)/2) \quad \text{and} \quad \text{Bin } (n - Y, (2-\theta)/4) \ ,$$

respectively. However, if we choose to condition x by the other ancillary U, then we simplify the data to two independent variables n_1 and n_3 distributed re-

spectively as

$$\text{Bin } (U, (1-\theta)/3) \quad \text{and} \quad \text{Bin } (n-U, (2-\theta)/3) \ .$$

In either case, the sample-space analysis of the conditioned data will be fairly easy and straightforward. But can anyone give a convincing argument for the choice of either Y or U as the conditioning ancillary?

We can easily introduce a nuisance parameter into the foregoing example by incorporating into the data, say, the result z of an independent coin-tossing experiment with an unknown bias ϕ in the coin. (George Barnard once remarked that when he retires he will go into business manufacturing biased coins and selling them to people like Basu!) In this case both (Y, z) and (U, z) will be S-ancillaries and we shall be back in the choice dilemma.

In contrast to the conditionality argument, the marginalization argument, in terms of the sufficiency or the generalized sufficiency principles, does not suffer from the above kind of a choice dilemma. With the kind of models that we work with in statistics, the existence of an essentially unique minimum sufficient statistic is always assured, and if the class of statistics that are p-sufficient for θ is not vacuous, then there will exist an essentially unique minimum such statistic.

6. A CONFLICT

The two elimination methods, namely, the one that marginalizes to a statistic T that is p-sufficient for θ and the one that conditions with respect to a statistic Y that is S-ancillary for θ, owe their origin to the same statistical intuition that guided us through (i) to (v) in Section 4. However, this does not mean that the two methods can co-exist in logical harmony. The possibility of a natural conflict between the methods was pointed out to the author by Philip Dawid (1975). We give below a simple example along the lines of the dilemma example of the previous section to highlight this conflict.

Example: Let T and Y be two random variables with the same range $\{1, 2, 3, 4\}$ and a joint distribution as described in the following table.

To simplify the argument let us suppose that we have only one observation $x = (t, y)$ on the pair (T, Y)—the general case where we have n observations on (T, Y) is very similar. Observe that the statistic T, defined as $T(x) = t$, is p-sufficient for θ and that the statistic Y, defined as $Y(x) = y$, is S-ancillary for θ. The trouble is that T and Y are not stochastically independent in this example. The marginal distribution of T is very different from its conditional distribution for any given value of Y.

It is thus clear that we can have a pair of stochastically dependent statistics T and Y such that (T, Y) is sufficient for (θ, ϕ), T is p-sufficient for θ, and Y is S-ancillary for θ. The nuisance parameter ϕ can be eliminated from the argument either by marginalizing to T or by conditioning (T, Y)—that is, T—by the S-ancillary Y. The two elimination methods cannot be reconciled in such cases.

What went wrong? Should we blame the statistical intuition that guided us through (i) to (v) in Section 4? The above conflict is only a manifestation of the difficulties that we have to face when we try to interpret data in some sample-space terms.

7. RAO-BLACKWELL TYPE THEOREMS

In Section 4, our case for the Sufficiency Principle $\mathcal{S}$, the Conditionality Principle $\mathcal{C}$ and their generalizations $\mathcal{S}^*$ and $\mathcal{C}^*$ rested on the highly nonmathematical phrase, "It makes good statistical sense." The author does not know how else to argue in non-Bayesian terms for these essentially Bayesian principles of data analysis. A large majority of the statisticians belonging to the Fisher-Neyman school of thought seem to agree wholeheartedly with $\mathcal{S}$ although most of them are quite wary of $\mathcal{C}$. This almost universal faith in $\mathcal{S}$ is there, partly because it makes good statistical sense, but mainly because of the widespread belief that principle $\mathcal{S}$ has been mathematically proved in the Rao-Blackwell theorem. On p. 17 of Basu (1975) we briefly examined this mathematical proof of a statistical principle. Now, let us turn the spotlight on a similar proof of $\mathcal{S}^*$ given by Fraser (1956).

Let $a(\theta)$ be a real valued function of θ. We are looking for a reasonable point estimate of $a(\theta)$ on the basis of the data $(\mathcal{E}, x)$. Let us suppose that the loss $W(t, \theta)$, when $a(\theta)$ is estimated by t, is convex in t for each θ. Let $\mathfrak{U}$ be the class of all estimators U of $a(\theta)$ such that the risk function

$$r_U(\theta) = r_U(\theta, \phi) = E[W(U, \theta) \mid \theta, \phi]$$

is well defined and θ oriented, that is, depends on $\omega = (\theta, \phi)$ only through θ.

Theorem (Fraser): If T is p-sufficient for θ then, for each $U \in \mathfrak{U}$, there exists an estimator $U_0 = U_0(T)$ such that $r_{U_0}(\theta) \leq r_U(\theta)$ for all $\theta \in \Theta$.

Joint Distribution of T and Y

	T				
Y	1	2	3	4	Total
1	$(1-\theta)(1-\phi)/12$	$(1+\theta)(1-\phi)/12$	0	0	$(1-\phi)/6$
2	$(1-\theta)(1+\phi)/12$	$(1+\theta)(1+\phi)/12$	0	0	$(1+\phi)/6$
3	0	0	$(2-\theta)(2-\phi)/24$	$(2+\theta)(2-\phi)/24$	$(2-\phi)/6$
4	0	0	$(2-\theta)(2+\phi)/24$	$(2+\theta)(2+\phi)/24$	$(2+\phi)/6$
Total	$(1-\theta)/6$	$(1+\theta)/6$	$(2-\theta)/6$	$(2+\theta)/6$	1

Proof: The statistic T is θ oriented by definition, so the risk function generated by any function of T, if well defined, must be θ oriented. Now, choose and fix $\phi_0 \in \Phi$ and define

$$U_0 = E(U \mid T, \theta, \phi_0) \ .$$

Since T, by definition, is sufficient for θ when ϕ is fixed, it follows that U_0 is well defined as an estimator; that is, the unknown θ does not enter into the definition of U_0. From Jensen's inequality it follows that, for all $\theta \in \Theta$,

$$r_{U_0}(\theta) = r_{U_0}(\theta, \phi_0) \leq r_U(\theta, \phi_0) = r_U(\theta) \ ,$$

and thus the theorem is proved. The Rao-Blackwell theorem clearly corresponds to the particular case where ϕ is a known constant ϕ_0.

The above theorem may be generalized further along the following lines suggested by Hájek (1965). Let $\mathfrak{U}'$ be the class of all estimators U such that the risk function $r_U(\theta, \phi)$ is well defined (but not necessarily θ oriented). Using the so-called minimax principle (see paragraph 7 of Section 1) let us define

$$R_U(\theta) = \sup_\phi r_U(\theta, \phi)$$

as the eliminated risk function associated with U, if $U \in \mathfrak{U}$ then $r_U(\theta) = r_U(\theta, \phi)$ is θ oriented and thus $R_U(\theta) = r_U(\theta)$. Now, if we define U_0 as in the Fraser theorem, then it follows (in view of the fact that U_0 is θ oriented) that $R_{U_0}(\theta) = r_{U_0}(\theta, \phi_0) \leq r_U(\theta, \phi_0) \leq R_U(\theta)$. This generalizes the Fraser theorem to the following result:

Theorem (Hájek): If T is p-sufficient for θ, then for each $U \in \mathfrak{U}'$ there exists an $U_0 = U_0(T)$ such that $R_{U_0}(\theta) \leq R_U(\theta)$ for all $\theta \in \Theta$.

The proofs of the preceding two theorems do not make full use of the supposition that T is p-sufficient for θ. They rest heavily on the supposition that T is θ oriented but require T to be sufficient for θ for just one specific value ϕ_0 of ϕ. This suggests the following generalization of the notion of partial sufficiency. For each $\theta \in \Theta$, let us define $\bar{\mathcal{P}}_\theta$ to be the convex hull of the family $\mathcal{P}_\theta = \{P_{\theta,\phi} : \theta \text{ fixed}, \phi \in \Phi\}$ of measures on $(\mathfrak{X}, \mathfrak{A})$. In other words, $\bar{\mathcal{P}}_\theta$ is the family of all measures Q of the form:

$$Q(A) = \int_\Phi P_{\theta,\phi}(A) d\xi(\phi) \quad \text{for all} \quad A \in \mathfrak{A} \ , \tag{7.1}$$

where ξ is an arbitrary probability measure on Φ. The following definition is due to Hájek (1965).

Definition (H-Sufficiency): The statistic T is H-sufficient (partially sufficient in the sense of Hájek) for θ if, for each $\theta \in \Theta$, there exists a choice of a measure $Q_\theta \in \bar{\mathcal{P}}_\theta$ such that, with $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$, T is sufficient in the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{Q})$, and T is θ oriented in the model $(\mathfrak{X}, \mathfrak{A}, \mathcal{P})$.

It should be noted that for the definition of H-sufficiency it is not necessary for θ and ϕ to be variation independent. Clearly, p-sufficiency implies H-sufficiency. We have only to choose and fix $\phi_0 \in \Phi$ and then define $Q_\theta = P_{\theta,\phi_0}$. Let us check now that the Fraser-Hájek theorems remain true even if we replace the requirement of p-sufficiency for the statistic T by the less stringent requirement of H-sufficiency. For any U, we define U_0 as $E(U \mid T, Q_\theta)$. Observe that U_0 is an estimator in view of the definition of H-sufficiency. We then invoke Jensen's inequality to prove that, for all $\theta \in \Theta$,

$$\int_{\mathfrak{X}} W(U_0, \theta) dQ_\theta \leq \int_{\mathfrak{X}} W(U, \theta) dQ_\theta \ .$$

Now, if we look back on the supposition that Q_θ is in the form (7.1) above, then, from the fact that U_0—being a function of T—is θ oriented, it follows at once that the left side of the above inequality is equal to

$$r_{U_0}(\theta) = R_{U_0}(\theta)$$

for all θ. Similarly, the right side is equal to $r_U(\theta)$ if $U \in \mathfrak{U}$ and is clearly not greater than $R_U(\theta)$ if $U \in \mathfrak{U}'$. Thus the two preceding theorems may be finally restated as:

Theorem (Fraser-Hájek): If T is H-sufficient for θ, then for any $U \in \mathfrak{U}$ there exists a $U_0 = U_0(T)$ such that $r_{U_0}(\theta) \leq r_U(\theta)$ for all θ. Furthermore, for any $U \in \mathfrak{U}'$ it is true that $R_{U_0}(\theta) \leq R_U(\theta)$ for all θ.

How much comfort can an advocate of the generalized sufficiency principle S* derive from the Fraser-Hájek theorem? Before answering this question, let us take a brief look at the question of how and where the notion of H-sufficiency fits into the ten-fold factorization scheme of the likelihood that we laid out in Section 4.

In order for T to be H-sufficient for θ it is necessary that T is θ-oriented; that is, we have a factorization of the form

$$p(x \mid \theta, \phi) = g(T \mid \theta) f(x \mid T, \theta, \phi) \ . \tag{7.2}$$

It is also necessary (in view of the sufficiency condition for T) that there exists a family $\{\xi_\theta : \theta \in \Theta\}$ of probability measures on Φ such that the "mixed" frequency function

$$q(x \mid \theta) = \int_\Phi p(x \mid \theta, \phi) d\xi_\theta(\phi)$$

factors as

$$q(x \mid \theta) = G(T, \theta) F(x) \ . \tag{7.3}$$

Let us look back at the classical problem where the sample $x = (x_1, x_2, \ldots, x_n)$ consists of n independent observations on an $N(\mu, \sigma)$. Clearly $\bar{x}$ is not H-sufficient for μ—indeed, no T can be H-sufficient for μ. But is $s^2 = \sum (x_i - \bar{x})^2$ H-sufficient for σ? Can we find a family $\{\xi_\sigma : 0 < \sigma < \infty\}$ of "mixing measures" on R_1 that will lead to a factorization of the type (7.3) above with $T = s^2$? Observe that

$$p(x \mid \mu, \sigma) = A(\sigma) \exp\left(-\frac{s^2}{2\sigma^2}\right) \exp\left[-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right],$$

where $A(\sigma) = ((2\pi)^{\frac{1}{2}}\sigma)^{-n}$.

We, therefore, need a family of mixing measures ξ_σ such that

$$\int_{-\infty}^{\infty} \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] d\xi_\sigma(\mu) = B(\bar{x})C(\sigma) \quad . \quad (7.4)$$

The above factorization clearly holds if we choose for ξ_σ the uniform distribution (the Lebesgue measure) over the whole real line. But, with such improper mixings, it is easily seen that the Fraser-Hájek theorem will fall to pieces. If the range of σ is the whole of the positive half line, then there cannot exist a family of proper mixing measures ξ_σ for which the factorization (7.4) will hold.

So how are we going to prove that we ought to marginalize to s when the parameter of interest is σ? Hájek (1965) came up with the following ingenious mathematical argument. In any particular situation, we should always be able to limit (on a priori considerations) the parameter σ to some finite interval $(0, k)$. With σ restricted to such a finite interval, the statistic s becomes H-sufficient for σ. Just check that the factorization (7.4) holds if we choose for ξ_σ the Normal measure with mean zero and variance $(k^2 - \sigma^2)/n$.

Hájek's definition of partial sufficiency is intriguing and full of mathematical possibilities. But, what are the statistical contents of Hájek's definition of partial sufficiency and his generalized Rao-Blackwell theorem? Hájek's 'proof,' that we should marginalize to s when we do not know μ, certainly does not scandalize our statistical intuition. In the language of R.A. Fisher, if we throw away $\bar{x}$ and marginalize to s, then our loss of information on σ has the measure of only one degree of freedom in the worst possible case (when μ is fully known). Of the total information available on σ, the fraction of information summarized in s is at least $(n-1)/n$. Let us now look at the following celebrated example due to Neyman and Scott (1948):

Example (Neyman & Scott): The data x consists of $2n$ observations $x_1, x_1', x_2, x_2', \ldots, x_n, x_n'$. The statistical model here corresponds to $2n$ independent normal variables with equal variances σ^2 and with x_i, x_i' having common mean $\mu_i (i = 1, 2, \ldots, n)$. The parameter of interest is σ, the nuisance parameter is the vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$.

With $S^2 = \sum (x_i - x_i')^2$, $\bar{x}_i = (x_i + x_i')/2$ and $A(\sigma)$ $((2\pi)^{\frac{1}{2}}\sigma)^{-2n}$, we then have

$$p(x \mid \boldsymbol{\mu}, \sigma) = A(\sigma) \exp\left(-\frac{S^2}{4\sigma^2}\right) \exp\left[-\frac{\sum (\bar{x}_i - \mu_i)^2}{\sigma^2}\right].$$

The statistic S^2 is clearly σ oriented. Is it H-sufficient for σ? Again the answer is no if σ is unrestricted, but it is yes if we restrict σ to a finite interval (0,k). For the mixing measure ξ_σ on R_n, we now choose the one for which $\mu_1, \mu_2, \ldots, \mu_n$ are iid normal variables with means zero and variances $(k^2 - \sigma^2)/2$.

Of course, we are prepared to assume that $0 < \sigma < k$ for some k. The Hájek proof notwithstanding, how secure do we really feel about marginalizing to S without taking a hard look at $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$? If $\boldsymbol{\mu}$ were known, then the sample would contain $2n$ units (degrees of freedom) of information on σ, out of which S summarizes in itself only n units. Are we really prepared to sacrifice n degrees of freedom at the altar of ignorance on $\boldsymbol{\mu}$? The issues raised in this example of Neyman and Scott are all very complex and we shall return to them again in a subsequent article.

We close this section with one more jab at the notion of H-sufficiency and the Rao-Blackwell type proof of our generalized sufficiency principle in sample-space terms.

Example:[1] Let $x = (x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_n)$ be $m + n$ independent normal variables all with unit variances. It is known that $Ex_i = \theta (i = 1, 2, \ldots, m)$ and $Ey_i = \theta\phi (j = 1, 2, \ldots, n)$, where $-\infty < \theta < \infty$ is the parameter of interest and $\phi (= 0 \text{ or } 1)$ is the nuisance parameter.

The likelihood function neatly factors as

$$p(x \mid \theta, \phi) = A(x) \exp\left[-m(\bar{x} - \theta)^2/2\right] \cdot \exp\left[-n(\bar{y} - \theta\phi)^2/2\right] .$$

Clearly, the pair $(\bar{x}, \bar{y})$ constitutes the minimal sufficient statistic. The statistic $\bar{x}$ is θ oriented. It is also sufficient (for θ) when ϕ is fixed at the value zero. Therefore, $\bar{x}$ is H-sufficient for θ and so the Fraser-Hájek theorems proved earlier recommend marginalization to $\bar{x}$. However, the reduction of the data from $(\bar{x}, \bar{y})$ to $\bar{x}$ will mean a substantial loss of information on θ in the event $\phi = 1$. From the full data we should be able to tell (with a reasonable amount of certainty if m and n are large) whether $\phi = 0$ or 1. (If E stands for the event $m(\bar{x} - \bar{y})^2 > (m+n)\bar{y}^2$, then it is easy to check that the maximum likelihood (ML) estimator $\hat{\phi}$ of ϕ is the indicator of E and that the ML estimator of θ is $\hat{\theta} = (1 - \hat{\phi})\bar{x} + \hat{\phi}(m\bar{x} + n\bar{y})/(m + n)$.)

This example does not contradict the good statistical sense that led us to the generalized (or partial) sufficiency principle S*, but only tells us not to be unduly impressed with Fraser's mathematical proof of the principle. The statistical literature is full of this kind of proof (see for instance Lehmann (1959)) where we start on the wrong foot either by delimiting the discussion to a conveniently small (and nice) class of decision procedures or by simplifying the hypothetical risk function by an ad hoc maximization process. The author is very skeptical about the relevance of this kind of statistical mathematics in theoretical statistics.

8. THE BAYESIAN WAY

After a long journey through a whole forest of confusing ideas and examples, we seem to have lost our way. Let us now see if our Bayesian guide can find a way out of this wilderness for us.

According to a Bayesian, the role of the data $(\mathcal{E}, x)$ is to act as an operator on the experimenter's prior

[1] A referee has pointed out that a similar example appears in Barndorff-Nielsen (1973).

opinion q (a probability measure on Ω) and to transform it into a posterior opinion $q_x{}^*$. This transformation is effected through a formal use of the Bayes Theorem and the likelihood function $L(\omega) = p(x|\omega)$ generated by the data.

With $\omega = (\theta, \phi)$, where θ is the parameter of interest and ϕ is the nuisance parameter, the Bayesian analysis of data is always firmly anchored to the posterior marginal distribution $q_x{}^\dagger$ on Θ defined as

$$q_x{}^\dagger(\theta) = \sum_\phi q_x{}^*(\theta, \phi) \ ,$$

where $q_x{}^*(\omega) = L(\omega)q(\omega)/\sum_\omega L(\omega)q(\omega)$. As we said in paragraph (10) of Section 1, the Bayesian way of eliminating the nuisance parameter from the argument is to integrate it out from the posterior distribution of (θ, ϕ).

In 1942, A.N. Kolmogorov defined the notion of a sufficient statistic in the following Bayesian terms:

Definition: The statistic T is sufficient if, for every prior q on Ω, the posterior $q_x{}^*$ depends on x only through T; that is, $T(x) = T(x')$ implies that $q_x{}^* = q_{x'}{}^*$.

In the discrete setup, there is no difficulty in proving the equivalence of the above definition and the classical Fisher definition of sufficiency. In the same 1942 paper, we find Kolmogorov suggesting the following definition of partial sufficiency.

Definition (K-Sufficiency): The statistic T is partially sufficient for θ if, for all prior q on Ω, the posterior marginal distribution $q_x{}^\dagger$ on Θ depends on x only through T. (Let us call such a statistic K-sufficient for θ.)

At last we seem to have something for which we have been looking for so long. However, it was demonstrated by Hájek (1965) that the definition of K-sufficiency is vacuous in the following sense:

Theorem (Hájek): If the parameter θ is not a constant in ω, then every T that is K-sufficient for θ is sufficient (in the usual sense).

Proof: Pretending as always that we are dealing with a discrete model, we first recall that if T is not sufficient then there must exist x, x' such that $T(x) = T(x')$, but the likelihood ratio $p(x|\omega)/p(x'|\omega)$ is not a constant in ω. Therefore, if T is not sufficient, then we must have x, x' and ω_1, ω_2 such that $T(x) = T(x')$ and

$$p(x|\omega_1)/p(x'|\omega_1) \neq p(x|\omega_2)/p(x'|\omega_2) \ . \qquad (8.1)$$

Let $\omega_1 = (\theta_1, \phi_1)$ and $\omega_2 = (\theta_2, \phi_2)$. There is no loss of generality in supposing that $\theta_1 \neq \theta_2$. (Otherwise, we choose $\omega_3 = (\theta_3, \phi_3)$, with $\theta_3 \neq \theta_1 = \theta_2$, and consider the ratio $p(x|\omega_3)/p(x'|\omega_3)$ along with any one of the two ratios in (8.1) that differs from it.) Now, consider the prior q whose entire mass is equally distributed over the two points ω_1 and ω_2. Observe that

$$q_x{}^\dagger(\theta_1) = q_x{}^*(\omega_1) = p(x|\omega_1)/\sum_{i=1}^{2} p(x|\omega_i) \ ,$$

and that a similar expression holds true for $q_{x'}{}^\dagger(\theta_1)$. In view of (8.1) it follows that

$$q_x{}^\dagger(\theta_1) \neq q_{x'}{}^\dagger(\theta_1) \text{ even though } T(x) = T(x') \ .$$

Thus, T not-sufficient implies T not K-sufficient. This proves the theorem. (Observe that we do not require θ and ϕ to be variation independent either in the definition of K-sufficiency or in the proof of the above theorem.)

The fault in Kolmogorov's definition of partial sufficiency is easily detected. We may try to correct this by restricting the discussion to a relatively small class Q of prior measures q on Ω. We find the following definition in Raiffa and Schlaifer (1961):

Definition (Q-Sufficiency): The statistic T is Q-sufficient for θ if, for all $q \in Q$, the posterior marginal distribution $q_x{}^\dagger$ on Θ depends on x only through T. (In the language of Raiffa and Schlaifer, such a T is called marginally sufficient with respect to Q.)

From the beginning, we have been concerned with the problem of eliminating a parameter ϕ that is "unrelated" to the parameter of interest θ. However, we have not as yet clearly stated what we mean by two unrelated parameters. Is it enough to say that θ and ϕ are unrelated if they are variation independent and if the loss depends only on the terminal decision and the parameter θ? Clearly not, but what else can a non-Bayesian say? Just ask a non-Bayesian what he means when he agrees that the unknown true height ϕ of Mount Everest is unrelated to the unknown number θ of civilians who lost their lives in the Vietnam war! A Bayesian has no problem in defining the term. He calls θ and ϕ unrelated parameters if, apart from the condition on the loss function, his prior q for $\omega = (\theta, \phi)$ is of the form

$$q(\theta, \phi) = q_1(\theta)q_2(\phi) \ .$$

Let Q_0 be the class of all (independent) priors q of the form $q(\theta, \phi) = q_1(\theta)q_2(\phi)$. When is a statistic T going to be Q_0-sufficient for θ in the sense of our modified Kolmogorov definition of partial sufficiency? We find the following result in Raiffa and Schlaifer (1961):

Theorem (Raiffa and Schlaifer): If, for all $x \in \mathfrak{X}$, the likelihood function factors as

$$p(x|\theta, \phi) = G(T, \theta)F(x, \phi) \ ,$$

then T is Q_0-sufficient.

Proof: If the prior distribution is $q(\theta, \phi) = q_1(\theta)q_2(\phi)$, then

$$\begin{aligned} q_x{}^\dagger(\theta) &= \sum_\phi q_x{}^*(\theta, \phi) \\ &= G(T, \theta)q_1(\theta)/\sum_\theta G(T, \theta)q_1(\theta) \end{aligned}$$

depends on x only through T.

The above theorem suggests the following definition.

Definition (L-Sufficiency): The statistic T is L-sufficient for θ if, for all $x \in \mathfrak{X}$, the likelihood factors as in the statement of the previous theorem.

We just proved that L-sufficiency implies Q_0-sufficiency. Is the converse true? The answer is, of course, no. If T is sufficient, in the sense of Fisher or Kolmogorov, then it is Q-sufficient for every Q and in particular for Q_0. Raiffa and Schlaifer's definition of Q_0-sufficiency for θ suffers from a defect very similar to that of Kolmogorov's definition of partial sufficiency (K-sufficiency). The definition is too wide and fails to pinpoint the exact notion of partial sufficiency we are after. Perhaps an example will make clear the point we are driving at.

Example: Let $x_1, x_2, \ldots, x_n$ be iid $N(\theta\phi, 1)$ where $0 < \theta < \infty$ is the parameter of interest and $\phi(=-1 \text{ or } 1)$ is the nuisance parameter. Just imagine $\mu(-\infty < \mu < \infty)$ to be the common mean and then let $\theta = |\mu|$ and $\phi = \text{Sgn } \mu$.

The statistic $T = |\bar{x}|$ is θ oriented. It is a reasonable estimator of θ, but is it, in some sense, partially sufficient for θ? Check that T is not Q_0-sufficient for θ. Indeed, the notion of Q_0-sufficiency leads us to $\bar{x}$ which is sufficient. If, however, we agree to restrict our discussion to the smaller class $Q_0' \subset Q_0$ of (independent) priors q of the form $q(\theta, \phi) = q_1(\theta)q_2(\phi)$ such that q_2 is the uniform prior on $\Phi = \{-1, 1\}$, then it is easy to check that $T = |\bar{x}|$ is Q_0'-sufficient for $\theta = |\mu|$.

If we look back on the proof of the one-way implication theorem above, then it will be clear that L-sufficiency takes us far beyond Q_0-sufficiency. If T is L-sufficient for θ then the posterior marginal $q_x^\dagger$ on Θ depends on the sample x only through T and on the prior $q = q_1q_2$ only through q_1. In Bayesian terms, we may redefine the notion of L-sufficiency as follows:

Definition (*B-Sufficiency*): The statistic T is B-sufficient (partially sufficient in a restricted Bayes sense) for θ if, for $q = q_1q_2 \in Q_0$ and $x \in \mathfrak{X}$, the posterior marginal $q_x^\dagger$ on Θ depends on x only through T and on q only through q_1. (Indeed, one may try to further generalize the above notion of partial sufficiency by restricting q to an arbitrary but fixed class Q of priors on $\Omega = \Theta \times \Phi$ and calling q_1 the prior marginal on Θ. In the present context we have, however, no use for such a generalization.) In the next section we develop the theme of B-sufficiency to its natural conclusion.

9. UNRELATED PARAMETERS

Let us consider a rather loosely formulated question: Under what circumstances can we recognize the nuisance parameter ϕ to be so *unrelated* to the parameter of interest θ that we can meaningfully isolate the whole of the relevant information about the parameter θ contained in the data $(\mathcal{E}, x)$?

This is a good test question with which we can try to classify a statistician into one or another of the numerous feuding groups (or mutual admiration societies) that divide the current community of statisticians. For instance, a pucca (fully baked) Bayesian will probably dismiss the question out of hand as naive, incompetent, and unnecessarily argumentative. This is because a pucca-Bayesian has no use for the notion of "information in the data." According to him the natural dwelling place for information is the head of a homo sapien, and he recognizes only two kinds of statistical information—prior and posterior. Being a pucca-Bayesian, he always knows his prior q as a well-defined probability measure on $\Omega = \Theta \times \Phi$. Given the data he can, therefore, compute the posterior information q_x^* and then isolate the information $q_x^\dagger$ on θ by integration.

In the pucca-Bayesian statistical theory of Bruno de Finetti and L.J. Savage, there is no room for a family Q of prior distributions. However, having examined the question from various angles, the author has come to recognize the merit of Kolmogorov's half-baked[2] Bayesian approach to the problem at hand. In the spirit of Kolmogorov, Raiffa, and Schlaifer, let us put down the following definition for unrelated parameters. Let $\theta \in \Theta$ and $\phi \in \Phi$ be two parameters that enter into the statistical structure or model of an experiment $\mathcal{E}$, and let Q_0 be the class of all product probability distributions $q = q_1q_2$ on $\Omega = \Theta \times \Phi$.

Definition (*Unrelatedness*): The parameters θ, ϕ are unrelated relative to a model of the experiment $\mathcal{E}$ if, for all prior $q \in Q_0$ and all sample outcomes x of $\mathcal{E}$, the posterior distributions q_x^* also belong to the class Q_0.

If the likelihood function $L(\theta, \phi|x) = p(x|\theta, \phi)$ factors as

$$p(x|\theta, \phi) = A(\theta, x)B(\phi, y) \ , \tag{9.1}$$

then, for any prior $q(\theta, \phi) = q_1(\theta)q_2(\phi)$, it is easily seen that the posterior factors as

$$q_x^*(\theta, \phi) = q_x^\dagger(\theta)q_x'(\phi) \ ,$$

where

$$q_x^\dagger(\theta) = A(\theta, x)q_1(\theta)/\sum_\theta A(\theta, x)q_1(\theta) \ ,$$

with a similar expression holding true for $q_x'(\phi)$. Conversely, if

$$q_x^*(\theta, \phi) = p(x|\theta, \phi)q_1(\theta)q_2(\phi)/\sum_{\theta,\phi} pq_1q_2$$

belongs to Q_0 then it is equally clear that $p(x|\theta, \phi)$ must factor in the manner of (9.1) above. We thus have the

Theorem: The parameters θ, ϕ are unrelated relative to a model of the experiment $\mathcal{E}$ if and only if the likelihood function factors in the manner of (9.1).

It is then easy to recognize whether the parameter of interest is unrelated (in the preceding sense) to the nuisance parameter or not. With such a recognition of unrelatedness, (and, of course, with the further condition that the nuisance parameter has nothing to do with the hazards of incorrect decisions) the Bayesian will not waste his time in figuring out his prior q_2 for ϕ as long as he is satisfied that his prior q for (θ, ϕ) must be in the class Q_0. He will carefully figure out his prior q_1 for θ and then work out his posterior for θ as

$$q_x^\dagger(\theta) = A(\theta, x)q_1(\theta)/\sum_\theta A(\theta, x)q_1(\theta) \ .$$

[2] The Hindi antonym of pucca is so hard to spell in English!

In Basu (1975) we examined in depth the question of information in the data. Our conclusion was that, relative to a particular statistical model for the experiment $\mathcal{E}$ in question, Fisher's notion of the "whole of the relevant information" about $\omega = (\theta, \phi)$ that is contained in the data $(\mathcal{E}, x)$ may be identified with the likelihood function

$$L(\theta, \phi \mid x) = p(x \mid \theta, \phi) \ .$$

What we are saying now is that when the likelihood comes factored as in (9.1), when, on a priori considerations, we are willing to regard θ and ϕ as independent entities, and when information gained on ϕ is of no direct relevance to the decision problem on hand (i.e., ϕ does not enter into the loss function), then we may regard the function

$$L^{\dagger}(\theta \mid x) = A(\theta, x)$$

as the "whole of the relevant information" on θ that is supplied by the data $(\mathcal{E}, x)$. This may be regarded as a generalized likelihood principle.

The generalized sufficiency principle $\mathcal{S}^*$ and the generalized conditionality principle $\mathcal{C}^*$ are in conformity with the above principle. The existence of a statistic T that is p-sufficient for θ or of a statistic Y that is S-ancillary for θ presupposes a factorization of the likelihood as in (9.1). The principles $\mathcal{S}^*$ and $\mathcal{C}^*$ are indirectly advising us to concern ourselves with the factor of $L(\theta, \phi \mid x)$ that involves only θ. This is precisely why the p-sufficiency and the S-ancillarity arguments do not lead us astray.

Also observe that we can have a statistic T that is L-sufficient (B-sufficient) for θ if and only if θ and ϕ are unrelated in the sense of the likelihood factoring as in (9.1). If and when the likelihood factors in the above manner, we can always fashion a statistic T that is minimal L-sufficient for θ and a statistic Y that is minimal L-sufficient for ϕ. For example, T will be defined in terms of the equivalence relation: $x' \sim x''$ if $A(\theta, x') = C(x', x'')A(\theta, x'')$ for all $\theta \in \Theta$. In general, such a T will fail to be θ-oriented; that is, T will not be p-sufficient for θ. Similarly, Y will, in general, fail to be S-ancillary for θ. Indeed, we shall give an example where T and Y are the same. In such an example the same statistic T is in some sense isolating all the relevant information about θ and also all the information about the unrelated parameter ϕ.

A major source of our confusion on the important question of when and how we can isolate the information on the parameter of interest, is the fact of our arguing (in the manner of Sir Ronald) in terms of statistics. The notion of a statistic as a measurable map has hardly any relevance at the data analysis stage. It was Sir Ronald who distorted the question "what is information?" to the question "what (statistic) has all the information?" He taught us that a statistic is sufficient if and only if it summarizes in itself all the relevant information in the data. In the same spirit, we have been looking for a statistic T that is partially sufficient for θ—a statistic that summarizes in itself all the relevant and usable information about θ in the event of ignorance about the nusiance parameter ϕ.

We end this marathon discussion with three examples of statistical models where the parameters come naturally separated in the manner of (9.1), and yet we cannot take advantage of the fact (and isolate the information on the parameter of interest) in terms of either the generalized sufficiency or the conditionallity principle.

Example 1: We have a multinomial distribution with three categories and with probabilities

$$\theta\phi, \ (1 - \theta)(1 + \phi)/2 \quad \text{and} \quad (1 + \theta)(1 - \phi)/2 \ ,$$

where $0 < \theta < 1$ and $0 < \phi < 1$. With n observations, the three frequencies (n_1, n_2, n_3) constitute the minimal sufficient statistic, and the likelihood factors as

$$2^{-(n_2+n_3)}[\theta^{n_1}(1 - \theta)^{n_2}(1 + \theta)^{n_3}][\phi^{n_1}(1 + \phi)^{n_2}(1 - \phi)^{n_3}] \ .$$

We do not have any statistic that is p-sufficient, H-sufficient or S-ancillary for θ. The statistic $T = (n_1, n_2, n_3)$ is minimal L-sufficient (B-sufficient) for θ and also for ϕ. The (likelihood) information in the data on the parameter of interest θ is crying to be isolated as

$$L^{\dagger}(\theta) = \theta^{n_1}(1 - \theta)^{n_2}(1 + \theta)^{n_3} \ .$$

If θ and ϕ are independent a priori and if ϕ does not enter into the loss function, then a Bayesian will analyze the data in the same manner as he would have done in the hypothetical case when ϕ were known to be equal to $\frac{1}{2}$, say. Can anyone suggest a reasonable sample-space analysis of the data?

Example 2: Let $0 < \theta < \infty$ and $0 < \phi < \infty$. Let X and Y be two random variables with probability density functions

$$\theta e^{-\theta(x-\phi)}I(x - \phi) \quad \text{and} \quad \phi e^{-\phi(y+\theta)}I(y + \theta) \ ,$$

respectively, where $I(\cdot)$ stands for the indicator of the positive half of the real line. The sample consists of n independent observations $x_1, x_2, \ldots, x_n$ on X together with an independent set $y_1, y_2, \ldots, y_n$ of n independent observations on Y. Observe that the likelihood neatly factors as

$$[\theta^n \exp(-n\theta\bar{x})I(y_{(1)} + \theta)] \cdot [\phi^n \exp(-n\phi\bar{y})I(x_{(1)} - \phi)] \ ,$$

where $x_{(1)} = \min x_i$ and $y_{(1)} = \min y_i$. Clearly, the two parameters θ, ϕ are unrelated relative to the model. The statistic $(\bar{x}, y_{(1)})$ is B-sufficient (L-sufficient) for θ. The Bayesian analysis of the data is very simple as ϕ gets eliminated almost by itself. Can anyone suggest how to deal with the nuisance parameter in non-Bayesian terms?

Anyone who would sneer at the last two examples, on the grounds that they are not apparently related to any real life problem, is advised to take a hard look at the next example.

Example 3: The experiment consists of the observation, for each of n week days in a large metropolitan area, of the number of accidents involving one or more auto-

mobiles and also the corresponding number of such accidents involving one or more fatalities. The parameter of interest is the proportion θ of automobile accidents that result in death. The mean number ϕ of auto accidents per working day is the nuisance parameter. The statistical model for our record,

$$x = \{(x_1, y_1), \ldots, (x_n, y_n)\} ,$$

of the number of accidents x_i and the corresponding number of fatal accidents y_i on the ith day ($i = 1, 2, \ldots, n$) is that we have a set of n independent observations on a pair of random variables (X, Y) such that X is a Poisson variable with mean ϕ and Y, given X, is a Binomial variable Bin (X, θ). Now with $N = \sum x_i$ and $T = \sum y_i$, the likelihood neatly factors as

$$p(x|\theta, \phi) = A(x)\{\phi^N \exp(-n\phi)\}\{\theta^T(1-\theta)^{N-T}\} . \quad (9.2)$$

If n were a preselected constant, then the statistic N, distributed as a Poisson variable with mean $n\phi$, would qualify as an S-ancillary for θ. In this case the generalized conditionality principle will eliminate ϕ and will permit us to argue in some sample-space terms. Sir Ronald would have advised us to reduce the data to the minimal sufficient statistic (N, T), hold the ancillary N as fixed (at its observed value), and then look upon T as an observation on a Binomial variable with parameters N (known) and θ (unknown).

What happens if we do not preselect n but let it be determined by the very system that was under observation? Suppose we continue our observations until $T = \sum y_i$ exceeds a preselected number, say 10. How should we analyze the data then? Observe that our stopping rule has no effect on the likelihood function which comes factored in the same form as (9.2) above. Now the triple (n, N, T) constitutes the minimal sufficient statistic—the statistic T is nearly a constant but not quite. The statistics (N, T) and (n, N) are B-sufficient (L-sufficient) for θ and ϕ, respectively, but the notions of p-sufficiency and S-ancillarity are vacuous in this instance.

In a subsequent article, we shall study in depth various conditionality and marginalization arguments which have been put forward for the purpose of eliminating a nuisance parameter that is *not* unrelated to the parameter of interest in the present sense of separated (factored) likelihood.

[*Received January 1976. Revised December 1976.*]

REFERENCES

Anderson, E.B. (1967), "On Partial Sufficiency and Partial Ancillarity," *Skandinavisk Aktuarietidskrift*, 50, 137–52.

Barnard, G.A., Jenkins, G.M., and Winsten, C.B. (1962), "Likelihood Inference and Time Series (with Discussions)," *Journal of the Royal Statistical Society*, Ser. A, 125, 321–75.

———, and Sprott, D.A. (1971), "A Note on Basu's Examples of Anomalous Ancillaries (with Discussions)," *Foundations of Statistical Inference*, Toronto: Holt, Rinehart & Winston, 163–76.

Barndorff-Nielsen, O. (1973), "Exponential Families and Conditioning," published Sc.D. thesis, Department of Mathematics, University of Copenhagen.

Basu, D. (1959), "The Family of Ancillary Statistics," *Sankhyā*, 21, 247–56.

——— (1964), "Recovery of Ancillary Information," *Sankhyā*, A 26, 3–16.

——— (1965), "Problems Relating to the Existence of Maximal and Minimal Elements in Some Families of Statistics (Subfields)," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 41–50.

——— (1975), "Statistical Information and Likelihood (with Discussions)," *Sankhyā*, A, 37, 1–71.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference (with Discussions)," *Journal of the American Statistical Association*, 57, 269–326.

Cox, D.R. (1958), "Some Problems Connected with Statistical Inference," *Annals of Mathematical Statistics*, 29, 357–72.

——— (1971), "The Choice Between Alternative Ancillary Statistics," *Journal of the Royal Statistical Society*, Ser. B, 33, 251–5.

Dawid, A.P. (1975), "On the Concepts of Sufficiency and Ancillarity in the Presence of Nuisance Parameters," *Journal of the Royal Statistical Society*, Ser. B, 37, 248–58.

Durbin, J. (1961), "Some Methods of Constructing Exact Tests," *Biometrika*, 48, 41–55.

Fisher, R.A. (1934), *Statistical Methods for Research Workers*, 5th ed., Edinburgh: Oliver and Boyd.

——— (1956), *Statistical Methods and Scientific Inference*, 1st ed., Edinburgh: Oliver and Boyd.

Fraser, D.A.S. (1956), "Sufficient Statistics with Nuisance Parameters," *Annals of Mathematical Statistics*, 27, 838–42.

Hájek, J. (1965), "On Basic Concepts of Statistics," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 139–62.

Kolmogorov, A.N. (1942), "Determination of the Center of Dispersion and Degree of Accuracy for a Limited Number of Observations," (in Russian) *Izvestija Akademii Nauk* SSSR, Ser. Mat. 6, 3–32.

Lehmann, E.L. (1959), *Testing Statistical Hypotheses*, New York: John Wiley & Sons.

Linnik, Yu.V. (1965), "On the Elimination of Nuisance Parameters in Statistical Problems," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 267–80.

——— (1968), *Statistical Problems with Nuisance Parameters*, Translations of Math. Monographs, Vol. 20, Providence, Rhode Island: American Mathematical Society.

Neyman, J. (1935), "On a Theorem Concerning the Concept of Sufficient Statistic," (in Italian), *Giornale dell' Instituto Italiano degli Atturai*, 6, 320–34.

———, and Pearson, E.S. (1936), "Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses," *Statistical Research Memoirs of the University of London*, 1, 133–37.

———, and Scott, E.L. (1948), "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

Olshevsky, L. (1940), "Two Properties of Sufficient Statistics," *Annals of Mathematical Statistics*, 11, 104–6.

Prohorov, Yu.V. (1965), "Some Characterization Problems in Statistics," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 341–50.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Cambridge: Harvard University Press.

Sandved, E. (1966), "A Principle for Conditioning on an Ancillary Statistic," *Skandinavisk Aktuarietidskrift*, 50, 39–47.

——— (1972), "Ancillary Statistics in Models Without and With Nuisance Parameters," *Skandinavisk Aktuarietidskrift*, 55, 81–91.

Stein, Charles (1945), "A Two-Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance," *Annals of Mathematical Statistics*, 16, 243–58.

Journal of Statistical Planning and Inference 2 (1978) 1-13.

ON PARTIAL SUFFICIENCY: A REVIEW*

D. BASU**
The Florida State University, Tallahassee, Fl, U.S.A.

Received 8 September 1976; revised manuscript received 12 September 1977
Recommended by O. Kempthorne

The notion of a sufficient statistic—a statistic that summarizes in itself all the relevant information in the sample x about the universal parameter ω—is acclaimed as one of the most significant discoveries of Sir Ronald A. Fisher. It is however not well-recognized that the related notion of a partially sufficient statistic—a statistic that isolates and exhausts all the relevant and usable information in the sample about a sub-parameter $\theta = \theta(\omega)$—can be very elusive if the question is posed in sample space terms. In this review article, the author tries to unravel the mystery that surrounds the notion of partial sufficiency. For mathematical details on some of the issues raised here one may refer to Basu (1977).

AMS subject classifications: 60A05, 62A15.

Key words:
Ancillary statistics, information, nuisance parameters, specific sufficiency, θ-oriented statistics.

1. Introduction

In the beginning we have a parameter of interest—an unknown state of nature θ. With a view to gaining additional information on θ, we plan and then perform a statistical experiment $\mathscr{E}$ and thus generate the sample x. The problem of data analysis is to extract all the relevant information in the data $(\mathscr{E}, x)$ about the parameter of interest θ.

The notion of partial sufficiency arises in the context where the statistical model

$$\{(\mathscr{X}, \mathscr{A}, P_\omega): \omega \in \Omega\}$$

of the experiment $\mathscr{E}$ involves the universal parameter ω and where $\theta = \theta(\omega)$ is a sub-parameter. In this case it is natural to ask:

Question A: What is the whole of the relevant information about θ that is available in the data $(\mathscr{E}, x)$?

*Based on an invited talk given by the author at the August, 1976 annual meeting of the Institute of Mathematical Statistics.

**Partially supported by NSF research grant no. MCS77 01661.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_27, © Springer Science+Business Media, LLC 2011

It is not easy for a non-Bayesian to face up to this question. Most of us would feel more at ease when the question is rephrased in the following familiar form:

Question B: What statistic T summarizes in itself the whole of the relevant information about θ that is available in the sample x?

Let us understand that the two questions A and B, though similarly phrased, are very different in their orientations. Question A is clearly addressed to the particular data $(\mathscr{E}, x)$. But in B we are searching for a principle of data reduction. We may rephrase B in the following nearly equivalent form:

Question B* Does there exist a statistic T such that, in some meaningful sense, there is no loss of information on θ in the reduction of the data $(\mathscr{E}, x)$ to $(\mathscr{E}_T, t)$, where $\mathscr{E}_T$ is the marginal experiment—perform $\mathscr{E}$ but record only $t = T(x)$—corresponding to the statistic T?

Questions B or B*, when asked in the context of the universal parameter ω, led Fisher to the important notion of a sufficient statistic. But the same question, when asked in the context of a sub-parameter θ, turns out to be surprisingly resistant to a neat solution. The notion of partial sufficiency is indeed shrouded in a lot of mystery.

It is interesting to note that Sir Ronald introduced the notion of sufficiency into statistical literature (Fisher, 1920) first in the context of partial sufficiency. With a sample $x = (x_1, x_2, \ldots, x_n)$ from a normal population with unknown μ and σ, Fisher (1920) was concerned with the relative precisions of the two estimators

$$s_1 = (\tfrac{1}{2}\pi)^{1/2} \sum |x_i - \bar{x}| / n, \qquad s = [\sum (x_i - \bar{x})^2 / n]^{1/2}$$

of the standard deviation σ. [Fisher had used the notations σ_1 and σ_2 for the above estimators, but we have opted for the more familiar s.] Introducing this paper in Fisher (1950), Sir Ronald described the main thrust of his 1920 argument in the following terms.

> "..., but the more general point is established that, for a given value of s, the (conditional) distribution of s_1 is independent of σ. Consequently, when s, the estimate based on the mean square is known, a value of s_1, the estimate based on the mean deviation, gives no additional information as to the true value (of σ). It is shown that the same proposition is true if any other estimate is substituted for s_1, and consequently the whole of the relevant information respecting the variance which a sample provides is summed up in the single estimate s".

[*Author's note*: The proposition stated in the final sentence of the above quoted paragraph was not proved in Fisher (1920). Indeed, the proposition is not true unless we limit the discussion to location invariant statistics.]

In Fisher (1922), p. 316 we find the first mention of the now famous:

Criterion of Sufficiency: That the statistic chosen should summarize the whole of the relevant information supplied by the sample.

On the same page we find it suggested that, in the case of a sample $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma)$, the statistic s fully satisfies the criterion of sufficiency. It is thus clear that from the very beginning Sir Ronald had been grappling with the notion of partial sufficiency.

In this article we shall be examining several definitions of partial sufficiency that have been proposed from time to time. In every case we shall look back on this original problem of Fisher and ask ourselves the question: "Does this definition make s partially sufficient for σ?"

[*Author's note*: The name "sufficient" is, of course, very misleading. We should never have allowed an expression like "T is sufficient for θ" to creep into any statistical text. It is less misleading to use expressions like "T is sufficient for the sample x" or "T isolates and exhausts all the information in x about θ". Perhaps we should agree to substitute the name "sufficient" by the more descriptive characterization "exhaustive", which also comes from Fisher. Having said all these, we are nevertheless going to use the expression "partially sufficient for θ" in the rest of this essay!]

2. Specific sufficient statistics

In Neyman and Pearson (1936) we find one of the earliest attempts at making some sense of the elusive notion of partial sufficiency. Let us suppose that the parameter of interest θ has a "variation independent" complement ϕ—that is, the universal parameter ω may be represented as $\omega = (\theta, \phi)$ with the domain of variation Ω of ω being the Cartesian product $\Theta \times \Phi$ of the respective domains of θ and ϕ. In this case, we have (from Neyman–Pearson) the following:

Definition (specific sufficiency). The statistic $T: \mathscr{X} \to \mathscr{T}$ is *specific sufficient* for the parameter θ if, for every fixed $\phi \in \Phi$, the statistic T is sufficient in the usual sense—that is, T is sufficient with respect to the restricted model

$$\{(\mathscr{X}, \mathscr{A}, P_{\theta,\phi}) : \theta \in \Theta, \phi \text{ fixed}\}$$

for the experiment $\mathscr{E}$.

With a sample $x = (x_1, x_2, \ldots x_n)$ of fixed size n from $N(\mu, \sigma)$, the sample mean $\bar{x}$ is specific sufficient for μ. The sample standard deviation s is, however, not specific sufficient for σ. Even though $\bar{x}$ is specific sufficient for μ, in no meaningful sense of the terms can we suggest that $\bar{x}$ exhaustively isolates all the relevant information in the sample x about the parameter μ. Surely, we also need to know s in order to be able to speculate about, say, the precision of $\bar{x}$ as an estimate of μ. Clearly, we are looking for something more than specific sufficiency.

The fact of T being specific sufficient for θ may be characterized in terms of the following factorization of the frequency (or density) function p on the sample space $\mathscr{X}$:

$$p(x \mid \theta, \phi) = G(T(x), \theta, \phi)\, H(x, \phi).$$

Alternatively, we may characterize the specific sufficiency of T (for θ) by saying that the conditional distribution of any other statistic T_1, given T and (θ, ϕ), depends on (θ, ϕ) only through ϕ.

Before going on to other notions of partial sufficiency, it will be useful to state the following:

Definition (θ-oriented statistics). The statistic $T: \mathscr{X} \to \mathscr{T}$ is *θ-oriented* if the marginal (or sampling) distribution of T—that is, the measure $P_\omega T^{-1}$ on $\mathscr{T}$—depends on ω only through $\theta = \theta(\omega)$. In other words, $\theta(\omega_1) = \theta(\omega_2)$ implies

$$P_{\omega_1}(T^{-1}B) = P_{\omega_2}(T^{-1}B)$$

for all 'measurable' sets $B \subset \mathscr{T}$.

It should be noted that the notion of θ-orientedness does not rest on the existence of a variation independent complementary parameter ϕ. In our basic example of a sample from $N(\mu, \sigma)$, observe that $\bar{x}$ is not μ-oriented but that s is σ-oriented.

3. Partial sufficiency

If we put together the two definitions of the previous section, then we have the following definition of partial sufficiency that is usually attributed to Fraser (1956).

Definition. The statistic T is *partially sufficient* for θ if it is specific sufficient for θ and is also θ-oriented.

See Basu (1977) for a number of examples of partially sufficient statistics. In the example of a sample $(x_1, x_2, \ldots, x_n)$ from $N(\mu, \sigma)$, the statistic $\bar{x}$ is not partially sufficient for μ as it is not μ-oriented and the statistic s is not partially sufficient for σ as it is not specific sufficient for σ. In view of the specific sufficiency part of the above definition, it is necessary that the parameter θ has a variation independent complement ϕ. The requirement of θ-orientedness brings in the unpleasant consequence that T may be partially sufficient for θ but a wider statistic T_1 need not be. Indeed, the whole sample x is never partially sufficient for θ.

The notion of partial sufficiency may be characterized in terms of the following factorization criterion:

$$p(x \mid \theta, \phi) = g(T \mid \theta)\, h(x \mid T, \phi)$$

where g and h denote respectively the marginal probability function of T and the conditional probability function of x given T. Note that the marginal distribution is θ-oriented and the conditional distribution is ϕ-oriented.

The interest in the Fraser definition of partial sufficiency stems from the following generalization (Fraser, 1956) of the Rao–Blackwell argument. Let $a(\theta)$ be an arbitrary real valued function of θ and let $W(y, \theta)$ denote the loss sustained when $a(\theta)$ is estimated by y. Let us suppose that, for each $\theta \in \Theta$, the loss function $W(y, \theta)$ is convex in y. Finally, let $\mathscr{U}$ be the class of all estimators U such that the risk function

$$r_U(\theta) = r_U(\theta, \phi) = \mathbf{E}[W(U, \theta) \mid \theta, \phi]$$

is finite and depends on (θ, ϕ) only through θ.

Theorem (Fraser). *If T is partially sufficient for θ, then for any $U \in \mathscr{U}$ there exists an estimator $U_0 = U_0(T) \in \mathscr{U}$ such that $r_{U_0}(\theta) \leqq r_U(\theta)$ for all $\theta \in \Theta$.*

The proof of the theorem consists of choosing and fixing a particular value ϕ_0 of ϕ and then considering the statistic $U_0 = U_0(T) = \mathbf{E}(U \mid T, \theta, \phi_0)$ as an estimator of $a(\theta)$. That U_0 does not involve the parameter θ follows from the supposition that T is sufficient for θ when ϕ is fixed at ϕ_0. That $U_0 \in \mathscr{U}$ follows from the supposition that T is θ-oriented. The rest follows at once from Jensen's inequality.

The above theorem may be generalized along the lines suggested by Hájek (1967). Let $\mathscr{U}'$ be the class of all estimators U for which the risk function $r_U(\theta, \phi)$ is finite (but not necessarily free of ϕ). Let $R_U(\theta) = \sup_\phi r_U(\theta, \phi)$ be the maximum risk associated with U for a particular θ.

Theorem (Hájek). *If T is partially sufficient for θ, then for any $U \in \mathscr{U}'$ there exists a $U_0 = U_0(T)$ such that $R_{U_0}(\theta) \leqq R_U(\theta)$ for all θ.*

The definition of U_0 is the same as in the previous theorem. The rest of the proof follows from the following chain of relations

$$R_U(\theta) \geqq r_U(\theta, \phi_0) \geqq r_{U_0}(\theta, \phi_0) = R_{U_0}(\theta).$$

If $U \in \mathscr{U}$, that is, if the risk function for U is free of ϕ, then $r_U(\theta) \equiv R_U(\theta)$ and so the above theorem is a generalization of the Fraser theorem.

Let us take note of the fact that the proofs of the previous two theorems rest heavily on the supposition that T is θ-oriented but make very little use of the supposition that T is specific sufficient for θ. What is needed is the sufficiency of T (for θ) for just one specified value ϕ_0 of ϕ. Consider the following example.

Example. Let $x = (x_1, x_2, \ldots, x_m; y_1, y_2, \ldots, y_n)$ be $m+n$ independent normal variables with unit variances and with $E(x_i) = \theta$ $(i = 1, 2, \ldots, m)$ and $E(y_j) = \theta\phi$ $(j = 1, 2, \ldots, n)$, where $\theta \in [a, b]$ is the parameter of interest and $\phi \in \{0, 1\}$ is the nuisance parameter. The likelihood function factors as

$$p(x \mid \theta, \phi) = A(x) \exp[-\tfrac{1}{2}m(\bar{x} - \theta)^2] \exp[-\tfrac{1}{2}n(\bar{y} - \theta\phi)^2].$$

The pair $(\bar{x}, \bar{y})$ constitute the minimal sufficient statistic. The statistic $\bar{x}$ is θ-oriented and is sufficient for θ when $\phi = 0$. Therefore, we can invoke either the Fraser or the Hájek complete class theorem and suggest a reduction of the data x to the statistic $\bar{x}$. However, such a data reduction will clearly result in a substantial loss of information in the event $\phi = 1$. Looking at the full data we should usually be able to make a good guess of the true value of ϕ. For instance, if $m = 2$, $n = 200$, $\bar{x} = 16.02$ and $\bar{y} = 17.45$ then we know for (almost) sure that $\phi = 1$ and should naturally rebel against the idea of reducing the data to $\bar{x}$.

This example highlights the inherent weakness of the Fraser–Hájek argument. Fraser limited his discussion to the class $\mathscr{U}$ of estimators U whose risk functions involve only θ. It is not at all clear why we have to limit our universe of discourse to such a limited class. [It is true that the statistical literature is so full of Fraser-type limited complete class theorems. Familiar examples of such theorems abound in the theories of

best unbiased estimates, best similar region tests, best invariant procedures, etc.] In this example, the class of estimators of θ that are functions of $(\bar{x}, \bar{y})$ is complete in the class $\mathscr{U}'$ of all estimators, provided the loss function is convex. But the only functions of $(\bar{x}, \bar{y})$ that belong to $\mathscr{U}$ are those that do not involve $\bar{y}$. Thus, Fraser's requirement that we limit the discussion to $\mathscr{U}$ sort of forces $\bar{y}$ out of the picture even though it contains a lot of information on θ.

Hájek considered the wider class $\mathscr{U}'$ but eliminated the nuisance parameter from the argument by redefining the risk function as

$$R_U(\theta) = \sup_{\phi} r_U(\theta, \phi).$$

This method of eliminating the nuisance parameter from the risk function has been made popular by Lehmann (1959) in his famous text on tests of statistical hypotheses. A generalized version of the Minimax Principle is being invoked in this elimination argument. The author is not at all clear in his mind about the statistical content of this generalized principle. The example of this section is clearly in conflict with the principle.

4. *H*-sufficiency

Hájek (1967) pushed Fraser's notion of partial sufficiency to its natural boundary in the following manner. For each $\theta \in \Theta$ let $\Omega_\theta = \{\omega : \theta(\omega) = \theta\}$ and let $\bar{\mathscr{P}}_\theta$ be the convex hull of the family $\mathscr{P}_\theta = \{P_\omega : \omega \in \Omega_\theta\}$ of the probability measures on the sample space $\mathscr{X}$. The class $\bar{\mathscr{P}}_\theta$ is the class of all probability measures Q_θ on $\mathscr{X}$ that has the representation

$$Q_\theta(A) = \int_{\Omega_\theta} P_\omega(A) \, d\xi_\theta(\omega)$$

for all measurable sets A, where ξ_θ is some 'mixing' probability measure on Ω_θ. [Note that we are riding slipshod over the usual measurability requirements.]

Definition (*H*-Sufficiency). The statistic T is *H-sufficient* (partially sufficient in the sense of Hájek) for θ, if, for each $\theta \in \Theta$, there exists a choice of a $Q_\theta \in \bar{\mathscr{P}}_\theta$ such that

(i) T is sufficient with respect to the model $\{(\mathscr{X}, \mathscr{A}, Q_\theta) : \theta \in \Theta\}$ and

(ii) T is θ-oriented in the model $\{(\mathscr{X}, \mathscr{A}, P_\omega) : \omega \in \Omega\}$.

Observe that the notion of *H*-sufficiency (unlike the Fraser definition of partial sufficiency) does not require θ to have a variation independent complement ϕ. If T is partially sufficient in the sense of Fraser, then it is also *H*-sufficient. In order to see this, we have only to choose and fix $\phi_0 \in \Phi$ and then take $Q_\theta = P_{\theta, \phi_0}$ which is a mixture probability corresponding to a degenerate mixing measure.

Also observe that the requirement of θ-orientedness in the definition of *H*-sufficiency has the same unfortunate consequence (as in the case of Fraser's definition) that T may be *H*-sufficient but a wider statistic (e.g., the whole sample x) need not be so. Hájek (1967) sought to remedy this fault in his definition by putting in the additional clause (almost as an afterthought) that any statistic T_1 wider than an *H*-sufficient T should be regarded as *H*-sufficient. But such a wide definition of partial sufficiency cannot be admitted when we are concerned with the problem of isolating the whole of the relevant information about a subparameter.

The two theorems of the previous section may now be consolidated in the following complete class theorem. [For a proof refer to p. 361 of Basu (1977).]

Theorem (Hájek). *If T is H-sufficient for θ, then, for any $U \in \mathscr{U}$, there exists a $U_0 = U_0(T)$ such that $r_{U_0}(\theta) \leqq r_U(\theta)$ for all θ. Furthermore for any $U \in \mathscr{U}'$ it is true that $R_{U_0}(\theta) \leqq R_U(\theta)$ for all θ.*

Let us look back on the classical problem of a sample $x = (x_1, x_2, \ldots, x_n)$ of fixed size n from $N(\mu, \sigma)$. No statistic T can be H-sufficient for μ. This is because T can be μ-oriented only if it is an ancillary statistic, in which case it cannot, of course, be partially sufficient for μ. [This remark holds true for a general location-scale parameter set-up with μ as the location parameter.] On the other hand the statistic s is σ-oriented. Let us examine whether s is H-sufficient for σ.

The density (or likelihood) function factors as

$$p(x \mid \mu, \sigma) = A(\sigma) \exp\left[-\frac{ns^2}{2\sigma^2}\right] \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right]$$

where $A(\sigma) = (\sqrt{2\pi}\,\sigma)^{-n}$.

For each $\sigma \in (0, \infty)$, let ξ_σ be our choice of the mixing measure on the range space R_1 of the nuisance parameter μ. The corresponding family $\{Q_\sigma : 0 < \sigma < \infty\}$ of mixture measures on the sample space R_n will have the density function

$$\begin{aligned} \bar{p}(x \mid \sigma) &= \int_{-\infty}^{\infty} p(x \mid \mu, \sigma)\, d\xi_\sigma(\mu) \\ &= A(\sigma) \exp\left[-\frac{ns^2}{2\sigma^2}\right] \int_{-\infty}^{\infty} \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] d\xi_\sigma(\mu). \end{aligned}$$

We shall recognize s as H-sufficient for σ provided we can find a family $\{\xi_\sigma\}$ of mixing measures such that

$$\int_{-\infty}^{\infty} \exp\left[-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right] d\xi_\sigma(\mu) = B(\bar{x})C(\sigma) \tag{1}$$

because in that case $\bar{p}(x \mid \sigma)$ will factor as

$$\bar{p}(x \mid \sigma) = A(\sigma) \exp\left[-\frac{ns^2}{2\sigma^2}\right] B(\bar{x})C(\sigma)$$

establishing condition (i) of the definition of H-sufficiency.

One way to ensure (1) is to choose for ξ_σ the uniform distribution over the whole of R_1. But, with a family $\{Q_\sigma\}$ of improper mixtures, the proof of the Hájek theorem will break down. If the range of σ is the whole of the positive half line, then it can be shown that the factorization (1) can be achieved with no proper mixing. However, if we are willing to set a finite upper bound K for the parameter σ—from a practical point of view this is hardly a restriction—then it is easy to check that the choice of ξ_σ as the normal

distribution with mean zero and variance $(K^2-\sigma^2)/n$ will achieve the desired factorization (1). The above argument of Hájek (1967) establishing the H-sufficiency of s for σ $(0<\sigma<K)$ is very intriguing. At this point we like to contrast the approaches of Fisher and Hájek to the question of partial sufficiency of s for σ. First, let us look at the question from the:

Fisher Angle: The pair $(\bar{x}, s)$, being jointly sufficient for (μ, σ), contains the whole of the available information on the parameter of interest σ. Furthermore, the two statistics $\bar{x}$ and s, being stochastically independent, yield independent (additive, that is) bits of information on σ. If μ were known, then we have n 'degrees of freedom' worth of information on σ. Of these, the statistic s summarizes in itself $n-1$ 'degrees of freedom' worth of information on σ. If the only (prior) information about μ that we have is $-\infty < \mu < \infty$, then there is no way that we can recover any part of the (at most one 'degree of freedom' worth of) information contained in $\bar{x}$ about μ. It is in this situation of no (prior) information on μ that Fisher would label s as exhaustive of all available and usable information on σ. And in the event of no (prior) information on σ either (other than $0<\sigma<\infty$) Fisher would invoke his celebrated fiducial argument to declare that the status of the parameter σ has been altered from that of an unknown constant to that of a random variable with (fiducial) probability distribution $\sqrt{n}s/\chi_{n-1}$. Observe that the fiducial distribution of σ depends on the sample only through the statistic s.

A sort of improper Bayesian justification for the Fisher intuition on the problem at hand can be given by suggesting that, for every prior $q(\mu, \sigma)$ for the parameter (μ, σ) that is of the form

$$q(\mu, \sigma)\,d\mu d\sigma = g(\sigma)\,d\mu d\sigma$$

[μ and σ are independent a-priori and μ has the (improper) uniform distribution over the whole real line], the posterior marginal distribution of σ depends on the sample x only through the statistic s. Furthermore, the fiducial distribution of σ corresponds to the case where $g(\sigma)=1/\sigma$ $(0<\sigma<\infty)$. Although Fisher never put his arguments in the above straightforward Bayesian framework, the fact remains that Fisher's thinking on the problem of inference had a distinct Bayesian orientation.

Hájek Angle: On the surface, Hájek's partial sufficiency argument carries a distinct Bayesian flavour. His mixing measure ξ_σ—normal with zero mean and $(K^2-\sigma^2)/n$ as variance for μ may be interpreted as the prior conditional distribution of μ given σ. With any prior $q(\mu, \sigma)$ of the form

$$q(\mu, \sigma)\,d\mu d\sigma = [d\xi_\sigma(\mu)]g(\sigma)\,d\sigma$$

the posterior marginal distribution of σ will depend on the sample x only through the statistic s. It will, however, be very hard to make any Bayesian interested in a prior $q(\mu, \sigma)$ of the above form. Apart from the fact that q depends on the sample size (which it should not), it is not possible to make any sense of q as a measure of prior belief pattern. The main thrust of the Hájek argument is, however, not Bayesian at all. He was using the Bayesian device (of averaging over the parameter space) only as a mathematical artifact to prove a complete class theorem in the fashion of Abraham Wald.

We have already pinpointed the flaw in Hájek's definition of partial sufficiency through our example of the previous section. In that example, $\bar{x}$ is H sufficient for θ even though marginalization to $\bar{x}$ will entail a substantial loss of information on θ in the case of the (easily discernable) event $\phi = 1$. Many such examples can be easily constructed. [See Barndorff–Nielsen (1973) and Basu (1977) for other such examples.]

5. Invariantly sufficient statistics

In this section we briefly review George Barnard's thoughts on the knotty question of partial sufficiency of s for σ. The following quotation is from p. 113 of Barnard (1963).

> "The definition of sufficiency which has become universally accepted required that the distribution of any function of the observations, conditional on a fixed value of the sufficient statistic, should be independent of the parameter in question, and there is no doubt that with, this definition, s fails to be sufficient for σ. However, as was usual for him, Fisher's definition of sufficiency was designed to embody a logical notion, that of providing the whole of the available relevant information for a given parameter and the definition just referred to does not altogether succeed in this object.
>
> The availability or otherwise of information is critically dependent on knowledge or lack of knowledge. Obviously if σ is already known, s provides us with no information whatsoever. The failure of s to satisfy the definition given above for sufficiency arises from the fact that the distribution of $\bar{x} - \mu$ (with the usual notations) depends also on σ. However,... μ is given as unknown, and so the information in $\bar{x} - \mu$ is unavailable.
>
> As already remarked, Fisher was very much concerned, up to the end of his life, with the difficulty of expressing in precise mathematical form, the notions corresponding to 'known' and 'unknown'. The present writer several times suggested to him, in connection with parameters such as μ in the case of the normal distribution, ..., that these parameters correspond to groups under which the problems considered are invariant, and the notion of ignorance of μ can be represented in terms of group invariance properties".

Barnard's thoughts on the problem are best understood in the context of the simple example of a sample $x = (x_1, x_2, \ldots x_n)$ of fixed size n from $N(\mu, \sigma)$. The group $G = \{g_a : a \in R_1\}$ of transformations

$$g_a(x_1, x_2, \ldots x_n) = (x_1 + a, x_2 + a, \ldots x_n + a)$$

of the sample space R_n onto itself is associated with the group $\bar{G} = \{\bar{g}_a : a \in R_1\}$ of transformations

$$\bar{g}_a(\mu, \sigma) = (\mu + a, \sigma)$$

of the parameter space onto itself. The group $\bar{G}$ leaves the parameter of interest σ invariant but acts transitively on (traces a single orbit on the domain of) the nuisance parameter μ.

The problem of estimating the parameter σ is invariant with respect to the group G of

transformations $g_a : \mathscr{X} \to \mathscr{X}$. The maximal invariant is the difference statistic

$$D = (x_2 - x_1, x_3 - x_1, \ldots x_n - x_1)$$

The statistic s is invariantly sufficient for σ in the sense that

(i) s is a function of D and is, therefore, σ-oriented, and

(ii) the conditional distribution of any other invariant statistics $s_1 = s_1(D)$, given s, is the same for all possible values of σ (and, of course, of μ as well).

[The notion of invariantly sufficient statistic is due to Charles Stein. See Hall, Wijsman and Ghosh (1965), and Basu (1969) for further discussion on the subject.]

We are now ready for the following

Question. What is the logical necessity for restricting our attention to only G-invariant estimators of σ?

The standard argument for restricting attention to only such T that satisfies the identity

$$T(x_1 + a, x_2 + a, \ldots x_n + a) = T(x_1, x_2, \ldots x_n)$$

for all samples $x \in R_n$ and all $a \in R_1$—that is, to measurable functions of the maximal invariant $D = (x_2 - x_1, x_3 - x_1, \ldots, x_n - x_1)$—runs along the following lines:

Argument. The sample $(x_1, x_2, \ldots, x_n)$ consist of n i.i.d. $N(\mu, \sigma)$'s with $\mu (-\infty < \mu < \infty)$ 'unknown' and with σ as the parameter of interest. If we shift the origin of measurement to $-a$, then the sample will take on the new look $(x_1 + a, x_2 + a, \ldots, x_n + a)$. The new model for the new-look sample will then correspond to n i.i.d. $N(\mu + a, \sigma)$'s.

Note that the new mean $\mu + a$ is 'equally unknown' as μ and that σ remains unaltered. The problem of estimating σ (with μ unknown), therefore, remains invariant with any shift in the origin of measurement. Now, an estimator T is a formula for arriving at an estimate $T(x_1, x_2, \ldots x_n)$ based on the sample $x = (x_1, x_2, \ldots, x_n)$. With the same sample represented differently as $(x_1 + a, x_2 + a, \ldots x_n + a)$, but with the problem (of estimating σ) unaltered, the same formula T will yield the estimate $T(x_1 + a, x_2 + a, \ldots x_n + a)$. Clearly, the formula T will look rather ridiculous if $T(x_1 + a, x_2 + a, \ldots x_n + a)$ is not equal to $T(x_1, x_2, \ldots x_n)$ for some x and a.

The above invariance argument of Pitman–Stein–Lehmann has been sold in many different packages to a vast community of statisticians. However, a close look at the present package will immediately reveal the fact that the argument does not really add up to anything that is logically compelling.

For one thing, the part of the argument that asserts that the problem remains invariant with any shift of the origin of measurement is questionable. The argument rests heavily on the supposition that $\mu + a$ is 'equally unknown' as μ. Only an improper Bayesian with uniform prior (over the whole real line) for μ can make a case for such a statement.

Secondly, implicit in the argument lies the supposition that the choice of the estimator (estimating formula) T as a function on the sample space may depend on the statistical model (which, in this case, does not change with any shift in the origin of measurement) and the kind of 'average performance characteristics' that we find satisfactory but must not (repeat not) depend on any pre-conceived notions that we may have on the parameters in the model. This, of course, is not a tenable supposition (as all Bayesians will readily agree).

Let T_q be a typical Bayes estimator of σ that corresponds to the prior distribution q for (μ, σ)—for the sake of this argument let us imagine $T_q(x)$ to be the posterior mean of σ for a given sample x and the prior q. In T_q we thus have a well-defined formula for estimating σ. Every such formula T_q is invariant for every shift in the origin of measurement. This is because when the origin is shifted to $-a$, the sample $(x_1, x_2, \ldots, x_n)$ shifts to $(x_1+a, x_2+a, \ldots, x_n+a)$, the parameters (μ, σ) move to $(\mu+a, \sigma)$ and the prior q changes itself to the corresponding prior q_a for $(\mu+a, \sigma)$. It is easy to see then that

$$T_q(x_1, x_2, \ldots x_n) = T_{q_a}(x_1+a, x_2+a, \ldots x_n+a)$$

for all q, x and a. Thus, no Bayes rule violates the essence of the invariance argument.

However, if for a particular q, we look upon $T_q(x)$ as a function on the sample space, then we shall find that the function will typically depend on x through both $\bar{x}$ and s. [As we have noted in the previous section, for all (improper) priors q of the form $q(\mu, \sigma)\,d\mu d\sigma = g(\sigma)\,d\mu d\sigma$ and also for some curious looking proper priors of the Hájek kind, the posterior marginal distribution of σ will depend on x only through s and so with such a choice of the prior q, the Bayes estimator $T_q(x)$ for σ will be G-invariant as a function on the sample space.]

There is no logical necessity for restricting our attention to only G-invariant estimators as long as we take care to avoid using estimating procedures that do not recognize the arbitrariness that is inherent in the choice of the origin of measurement, etc. As we have noted earlier, all Bayes estimation procedures are invariant in a sense.

6. Final remarks

Sir Ronald was deeply concerned with the notion of information (about a parameter) in the data, but never directly faced up to such basic questions as: What is information? How informative is this data? Have we obtained enough information on the parameter of interest? etc.

The mathematical definition of information that we got from Fisher is a most curious one. The definition does not relate to the concept of information in the data but is supposed to bring out the notion of information in (the statistical model of) an experiment and the associated family of marginal experiments. Even then, the Fisher information $I(\omega)$ can hardly be interpreted in terms of the average (or expected) amount of knowledge gained (or uncertainties removed) about the universal parameter ω when the experiment is performed. And we get no prescription from Sir Ronald about how to 'marginalize' his information function (or matrix) to a sub-parameter. We must reject the notion of Fisher information on the ground of irrelevance in the present context.

The Fisher criterion of sufficiency—that the statistic chosen should summarize the whole of the relevant information supplied by the data—should be looked upon only as a principle of data reduction relative to a particular statistical model of the experiment. The earliest thoughts of Fisher on the subject of sufficiency crystalized around the following two propositions that are stated here relative to a fixed experiment $\mathscr{E}$ that is already endowed with an assumed statistical model.

Proposition 1. *To reduct (or marginalize) the data x to the statistic $T = T(x)$ will entail a total loss of all available information on the (universal) parameter ω if the marginal distribution of T is the same for all possible values of ω. Any such statistic T may be regarded as 'marginally uninformative' about ω.*

Proposition 2. *To reduce the data x to the statistic T will entail no loss of available information on ω if the conditional distribution of every other statistic T_1 given T is the same for all possible values of ω. Such a statistic T may be called sufficient, fully informative, or exhaustive of all available information on ω.*

Is it not remarkable that we now have the notions of 'no information' and 'full information' (meaning, exhaustive of all available information) without ever mentioning what we mean by information?! If by information we mean the state of our knowledge about the parameter ω, then should we not speculate about it in terms of the parameter space Ω rather than in terms of the sample space $\mathcal{X}$?!

It so happens that Fisher's 'sample space' definition of sufficient (information-full, that is) statistic agrees with the following Bayesian definition of sufficiency due to A. N. Kolmogorov (1942):

Definition. The statistic T is *sufficient* if, for every prior $q(\cdot)$ on Ω, the posterior distribution $q(\cdot|x)$ on Ω depends on x only through $T(x)$.

It is to the lasting credit of Sir Ronald that, having discovered the 'sample space' definition of sufficiency, he was able to put the notion in the correct perspective by characterizing a sufficient statistic as that characteristic of the sample knowing which we can determine the likelihood function up to a multiplicative factor. Fisher recognized that, relative to a given model, the whole of the relevant information in the data is summarized in the corresponding likelihood function. This is only a short step away from the Bayesian insight on the knowledge business.

The 'sample space' definition of sufficiency for the universal parameter ω is all right. But the weakness and inadequacy of this approach becomes apparent when we try the sample space way to 'isolate' all the 'available' relevant information on a sub-parameter. Note that we now have to deal with the new term 'isolate' and that the term 'available' suddenly springs to life with a new meaning. Fraser, Hajek and Barnard all seem to have tacitly assumed that T can isolate information on θ only if it is θ-oriented. This sample space requirement of θ-orientedness for the partially sufficient T has been a major source of our trouble with the notion of partial sufficiency. The statistical insight that leads to θ-orientedness as a prime requirement for partial sufficiency, cannot be reconciled with any Bayesian insight on the subject. What if there are no non-trivial θ-oriented statistic? Can't we then isolate the information on θ? What is information on θ? How can we isolate something that we have not even cared to define?

Barnard (1963) said "..., the notion of ignorance on μ can be represented in terms of group invariance properties." What is ignorance? Lack of prior information? How can we talk about lack of information when we have not even attempted to define what we mean by information? In any case, how can we possibly characterize ignorance on μ in terms of group invariance properties of the model? Who is ignora ? The scientist or the model?!

In September 1967 the author had asked the late Professor Renyi the question: "Why

are you a Bayesian?" Promptly came back the answer: "Because I am interested in the notion of information. I can make sense of the notion in no other way".

Acknowledgement

The author wishes to thank Professor Oscar Kempthorne for carefully going over an earlier draft of the paper.

References

Barnard, G.A. (1963). Some logical aspects of the fiducial argument. *J. R. Statist. Soc.* B25, 111–114.

Barndorff, Nielsen O. (1973). Exponential families and conditioning. Sc.D. thesis, Dept. of Maths., Univ. of Copenhagen, Denmark.

Basu, D. (1969). On sufficiency and invariance. In: *Essays in Probability and Statistics.* Univ. of North Carolina and Indian Statistical Institute. 61–84.

Basu, D. (1975a). Statistical information and likelihood (with discussions). *Sankhyā* A37, 1–71.

Basu, D. (1977). On the elimination of nuisance parameters. *Jl. Am. Stat. Assoc.* 72, 355–366.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and the mean square error. *Monthly Notices of the Royal Astronomical Soc.* 80, 758–770. [Also reproduced in Fisher (1950).]

Fisher. R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London* A222, 309–368. [Also reproduced in Fisher (1950).]

Fisher, R.A. (1950). *Contributions to Mathematical Statistics.* Wiley, New York.

Fraser, D.A.S. (1956). Sufficient statistics with nuisance parameters. *Ann. Math. Statist.* 27, 838–842.

Hájek, J. (1967). On basic concepts of statistics. In: *Proc. Fifth Berkeley Symp.* 1, 139–162.

Hall, W.J., R.A. Wijsman and J.K. Ghosh (1965). The relationship between sufficiency and invariance. *Ann. Math. Statist.* 36, 575–614.

Kempthorne, Oscar and Leroy Folks (1971). *Probability Statistics and Data Analysis.* The Iowa State University Press, Ames, IA.

Kolmogorov, A.N. (1942). Determination of the centre of dispersion and degree of accuracy for a limited number of observations. *Izv. Akad. Nauk. USSR. Ser. Mat.* 6, 3–32. [In Russian.]

Neyman, J. (1935). On a theorem concerning the concept of sufficient statistic. *Giorn. Ist. Ital. Attuari* 6, 320–334. [In Italian.]

Neyman, J. and E.S. Pearson (1936). Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Stat. Res. Memoirs* 1, 133–137.

Randomization Analysis of Experimental Data: The Fisher Randomization Test

D. BASU*

R.A. Fisher's classic text on the design of experiments is the principal source of inspiration for a mode of data interpretation that is usually characterized as randomization analysis. In Chapter III of this text, Fisher briefly commented on how to make a randomization test on some data generated by a Darwin experiment. Two variants of this randomization test are discussed in this article. The variant that is discussed in Section 4 may be regarded as the forerunner of all nonparametric tests. The original variant of the test is discussed in Section 6. The author concludes that the Fisher randomization test is not logically viable.

KEY WORDS: Prerandomization; Postrandomization; Sufficiency principle; Level of significance; Reference set.

1. INTRODUCTION

Randomization is widely recognized as a basic principle of statistical experimentation. Yet we find no satisfactory answer to the question, Why randomize? In a previous paper (Basu 1978) the question was examined from the point of view of survey statistics. In this article we take an uninhibited frontal view of a part of the randomization methodology generally known as the Fisher randomization test.

R.A. Fisher's classic text *The Design of Experiments* (*DE*) is the principal source of inspiration for a mode of data interpretation that may be characterized as randomization analysis of data. In Chapter III of *DE*, while discussing Galton's analysis of a Darwin experiment with 15 pairs of self-fertilized and cross-fertilized seeds, Fisher cursorily mentioned how one can take advantage of the physical act of randomization to make a test of significance that needs no assumption of normality for the error terms. This idea of Fisher's was immediately generalized by Pitman (1937) and then pushed to its natural boundary by Kempthorne (1952) and many others. Two variants of the Fisher randomization test are discussed in this article. The variant that is discussed in Section 4 may be regarded as the forerunner of all nonparametric tests. The original variant that is discussed in Section 6 may be regarded as one of the two supporting pillars (the other one being the famous case of the "lady tasting tea") of the complex theory of randomization analysis of experimental data. In between the two sections, I have inserted a section entitled: "Did Fisher Change His Mind?" I speculate that in 1956 Fisher had lost a great deal of his early enthusiasm for randomization analysis.

Whether Fisher changed his mind is not the present issue. What I am asking is whether, in the specific instances discussed in this article, it makes sense to compute a significance level (P value) in the manner of the Fisher randomization test. Can any evidential meaning be attached to a P value so computed?

Let us postpone the debate on significance testing in general and nonparametric tests in particular. Let us keep the issue sharply in focus and ask, Can the Fisher randomization test pass the test of common sense?

2. RANDOMIZATION

Let us define randomization as the incorporation of a fully controlled bit of randomness in the process of data generation. Randomization is usually carried out in the manner of items 1 and 2.

1. *Prerandomization.* This is the most common form of randomization. As the name suggests, the data-generation process begins with a fully controlled randomization exercise that determines the actual experimental (or observational) layout. Typical examples are random allocation of treatments in experimental designs and random selection of units in survey sampling. Along with replication and local control (blocking), prerandomization was characterized by Fisher (1960) as one of the three basic principles of statistical experimentation.

2. *Postrandomization.* Abraham Wald (1950) was one of the earliest to consider this kind of randomization as a statistical tool. After data x has been obtained, postrandomization is the generation of a further random entity y whose randomness characteristics may depend on x but are completely known to the randomizer. The statistician's conclusions or decisions are then based on the extended data (x, y). The average performance characteristics of a postrandomized decision rule δ are evaluated by taking into account all possible values of (x, y). With postrandomization, the statistician has a wider choice of attainable risk functions.

3. *Unrecorded randomization.* Occasionally, randomization is allowed to enter into the experimental process in a form quite different from the forms 1 and 2 discussed.

* D. Basu is Professor, Department of Statistics, Florida State University, Tallahassee, FL 32306. This article is based on an invited talk given at the SREB Summer Research Conference in Statistics at Arkadelphia, Arkansas, on June 15, 1978. This research was supported by National Science Foundation Grant MCS 79-04693.

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_28, © Springer Science+Business Media, LLC 2011

For instance, in a randomized-response survey the subjects may be instructed to respond to the question "Did you truthfully report your gross income in your 1977 tax return?" in the following manner. Each subject tosses a supposedly unbiased coin twice and then answers the question with a "Yes" if the coin yields two heads, with a "No" if the coin yields two tails, or with a truthful "Yes" or "No" if the coin yields a head and a tail. In this data-generation process the statistician may prerandomize to choose his or her sample subjects but has no control over the response randomizations done by the subjects. The statistician can only speculate about the outcomes of the response randomizations but cannot observe them. It may be argued that response randomization need not be classified as a form of experimental randomization. We shall not discuss this kind of randomization in this article. Warner (1965) proposed this kind of survey technique for eliminating evasive-answer bias.

3. TWO FISHER PRINCIPLES

As we said in the introduction, our primary concern is the so-called randomization analysis of data generated by a statistical experiment that has a large measure of prerandomization incorporated in it. It will, however, be useful to clear the deck with a short discussion of postrandomization and the two sides of the sufficiency principle.

Postrandomization injects into the data an element whose randomness characteristics are fully controlled by the experimenter. Let x be the initial data (sample) and let y be the postrandomized variable whose probability distribution, given x, depends only on x. In terms of the extended sample (x, y), the statistic x is sufficient and, as Fisher would put it, summarizes in itself the whole of the relevant information available in (x, y). To incorporate y in the inference-making process will be a violation of

The sufficiency principle: If T is a sufficient statistic then any conclusion that can be validly drawn from a statistical analysis of the data ought to depend on the data only through the statistic T.

In accordance with the sufficiency principle the data should be reduced to the minimal sufficient statistic. Not to reduce the data to the minimal sufficient statistic is to keep open the possibility of being influenced by irrelevant data characteristics such as, say, a postrandomization variable. In this connection it is interesting to read Fisher's (1956, pp. 96–98) comments on a postrandomization test proposed by Bartlett.

According to Fisher, a principal difference between the deductive and the inductive modes of inference is that in the former case valid conclusions (theorems) can be drawn from a partial use of the data (the primary postulates), whereas in the latter case no conclusion can be validly drawn from an examination of only a part of the relevant information core of the data. Fisher was quite concerned with the fact that the maximum likelihood estimator is not always a sufficient statistic. This led him to the conditionality principle and the celebrated recovery-of-ancillary-information method. The Fisher concern about using the whole of the relevant information in the data may be loosely stated as

The insufficiency principle: If the statistic T_1 is not sufficient then an inference-making procedure that depends on the data only through T_1 is insufficient, that is, lacking in substance.

It is not at all surprising, therefore, that Fisher took a rather dim view of nonparametric methods, especially those that make use of only the rank-order statistics. We shall revert to this theme with a Fisher quotation in Section 5.

4. THE FISHER RANDOMIZATION TEST

In Chapter III (Sec. 21) of *DE*, Fisher introduced his randomization test in the following terms: "In these discussions it seems to have escaped recognition that the physical act of randomization, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied." Fisher then gave a brief description of his randomization test as an alternative to the Student's t test. In this section we consider a popular variant of the test that may be regarded as the original permutation test. This is how the test is described in Kempthorne and Folks (1971, p. 342).

Let $x_1, x_2, \ldots, x_n$ be n independent observations on a random variable x. The problem is to test the null hypothesis H_0 that $E(x) = 0$. Under the parametric model that x is normally distributed, the test is usually carried out in terms of the studentized sum $T = \sum x_i$. Under the wider hypothesis (nonparametric model) that the distribution of x is continuous and is symmetric about its mean, the null hypothesis H_0 may be tested in terms of the criterion $T = \sum x_i$ as follows:

Write $\delta_i = \operatorname{sgn} x_i$, $i = 1, 2, \ldots, n$; that is, δ_i is -1 or 1 according as x_i is negative or positive. Note that

$$T = \sum x_i = \sum |x_i| \, \delta_i$$

and that the sample $(x_1, x_2, \ldots, x_n)$ may be split into the two parts

$$(|x_1|, |x_2|, \ldots, |x_n|) \quad \text{and} \quad (\delta_1, \delta_2, \ldots, \delta_n) .$$

Making the standard pretense that we are dealing with random variables and not particular observations, we recognize at once that the $|x_i|$'s are iid and that so also are the δ_i's. Under the null hypothesis, the two parts of the sample are stochastically independent and each δ_i is uniformly distributed over the two-point set $\{-1, 1\}$.

The distribution of the test criterion T is not well defined even under the null hypothesis. If we fix the $|x_i|$'s at their observed values and regard the δ_i's as random variables, however, then the conditional null distribution of T gets well defined. Although the actual computation may become somewhat tedious, the conditional probability

$$\Pr(T \geq t \mid H_0, |x_1|, |x_2|, \ldots, |x_n|)$$

can be worked out. Thus, we can carry out one-sided or two-sided tail area tests in terms of the conditional null distribution of T.

The conditional test just described bears the distinctive hallmark of Sir Ronald. It was Fisher who amazed and mystified the statistical world with his sensational 2×2 conditional test of independence, and it was he who taught us how to set up a conditional test for the equality of two Poisson means.

In order to find the attained significance level of data vis à vis a null hypothesis H_0, we have to search for an appropriate test criterion T and then refer it to an appropriate sample space (the reference set) for determining the tail-area probability under the null hypothesis. In the present case the criterion is the sample total T and the reference set is the set of all samples of the type

$$(\pm|x_1|, \pm|x_2|, \ldots, \pm|x_n|) \ ,$$

where $|x_1|, |x_2|, \ldots, |x_n|$ are fixed at their observed values. Before we turn the searchlight of careful scrutiny on this mystifying conditional test, it will be useful to compare it with two familiar nonparametric tests of the null hypothesis $\mu = 0$. (See Kempthorne and Folks 1971, pp. 340–345.)

The sign test: Choose as the test criterion the number S of positive signs among $\delta_i = \operatorname{sgn} x_i$, $i = 1, 2, \ldots, n$. The null distribution of S is bin $(n, \frac{1}{2})$. One-sided or two-sided tests can then be made in terms of S.

The Wilcoxon signed-rank test: Instead of the sample total $T = \sum |x_1|\delta_i$, choose the statistic $W = \sum r_i\delta_i$ as the test criterion, where r_i is the rank of $|x_i|$ among $|x_1|, |x_2|, \ldots, |x_n|$. Observe that the range of variation of W is the set of alternate integers in the interval $[-n(n+1)/2,\ n(n+1)/2]$. Under the null hypothesis H_0, the two vectors $(r_1, r_2, \ldots, r_n)$ and $(\delta_1, \delta_2, \ldots, \delta_n)$ are stochastically independent and the δ_i's are iid ± 1 variables with equal probabilities. The conditional distribution of W, given $(r_1, r_2, \ldots, r_n)$, can, therefore, be easily worked out under H_0. Since $(r_1, r_2, \ldots, r_n)$ is always a permutation of $(1, 2, \ldots, n)$, it is clear that the null distribution of W is the same for all possible realizations of $(r_1, r_2, \ldots, r_n)$; in other words, the Wilcoxon statistic W is stochastically independent of the rank vector if the null hypothesis is true. Thus, the Wilcoxon test is not a conditional test in the sense the Fisher randomization test is. The sign test and the Wilcoxon test are typical examples of nonparametric, distribution-free, marginal tests.

Commenting on the three tests, Kempthorne and Folks (1971, p. 344) wrote: "Since the sign test uses only the signs of x_i, the Wilcoxon test uses only the signs and the ranks of $|x_i|$, and the Fisher test uses the x_i without condensation, the Fisher test is superior as a significance test." Thus, it seems that Kempthorne and Folks are giving poorer ratings to the sign test and the Wilcoxon test on the score that they violate the insufficiency principle to a greater extent than does the Fisher test. But how to measure the extent of such violations! Do any of these tests violate the sufficiency principle? Let us examine the question.

In the context of our nonparametric statistical model, the set of order statistics $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is minimal sufficient. Each of the three test criteria T, S, and W can be written as a function of the order statistics, for example

$$T = \sum x_{(i)} = \sum |x_{(i)}|\delta_{(i)} \ ,$$
$$S = \tfrac{1}{2}(\sum \delta_{(i)} + n) \ ,$$
$$W = \sum r_{(i)}\delta_{(i)} \ ,$$

where $\delta_{(i)} = \operatorname{sgn} x_{(i)}$ and $r_{(i)}$ is the rank of $|x_{(i)}|$ among $|x_{(1)}|, |x_{(2)}|, \ldots, |x_{(n)}|$. Since the sign and the Wilcoxon tests are based on the marginal distributions of S and W, respectively (and, of course, on their observed values), there is no violation of the sufficiency principle in these cases—only the insufficiency principle is at stake.

In the case of the Fisher test, it may appear on the surface that the sufficiency principle has been violated in view of the fact that the conditioning statistic $(|x_1|, |x_2|, \ldots, |x_n|)$ is not a function of the minimal sufficient statistic $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$. If we carefully examine the conditional distribution of T given $(|x_1|, |x_2|, \ldots, |x_n|)$, then it will be clear that the conditional distribution depends on the sample $(x_1, x_2, \ldots, x_n)$ only through the order statistics. The sufficiency principle is sometimes interpreted as a requirement that the data ought to be first reduced to the minimal sufficient statistic (thus sieving out all the postrandomization impurities) and then the reduced data interpreted in terms of the marginal distribution of the minimal sufficient statistic. To satisfy the statistical intuition of such a purist we have only to point out that the Fisher test will remain unaltered if the conditioning statistic is chosen to be the ordered rearrangement of $|x_{(1)}|, |x_{(2)}|, \ldots, |x_{(n)}|$. The Fisher test does not violate the sufficiency principle.

The choice of the test criterion $T = \sum x_i$ and the choice of the conditioning statistic $(|x_1|, |x_2|, \ldots, |x_n|)$ are arbitrary elements in the Fisher test. For instance, we may want to condition T with respect to the statistic $(|x_1 + x_2 + \ldots + x_k|, |x_{k+1}|, \ldots, |x_n|)$ for some chosen k $(1 \leq k \leq n)$. It is easily seen that any such conditioning will make the null distribution of T dis-

tribution free. For instance, if $k = n$, then the conditional null distribution of T, given $|x_1 + \ldots + x_n| = d$, is uniform over the two-point set $\{-d, d\}$. With such a conditioning the one-sided test of the null hypothesis will result in a significance level of $\frac{1}{2}$ whenever the observed value of T is positive! Suppose we somehow convince ourselves that the only reasonable choice of a conditioning statistic is the one that corresponds to $k = 1$, the Fisher choice. (If $1 < k < n$, then the test procedure violates the sufficiency principle. The case $k = n$ is too ridiculous to deserve any serious consideration. And so on for any other conditioning statistic that one can think of.) Even then the question about the choice of the test criterion remains. Instead of $T = n\bar{x}$, why not choose the sample median $\tilde{x}$ as the test criterion? With the Fisher conditioning with respect to $(|x_1|, |x_2|, \ldots, |x_n|)$, the null distribution of $\tilde{x}$ is also distribution free. In our nonparametric setup, the sample median $\tilde{x}$ seems to be as reasonable a choice (as a test criterion) as the sample mean. Now, let us try to evaluate the significance level attained by the sample

$$(x_1, x_2, x_3, x_4, x_5) = (4, 7, 2, 3, 1).$$

The Fisher reference set consists of the 32 points

$$(\pm 4, \pm 7, \pm 2, \pm 3, \pm 1).$$

The observed sample mean is 3.4 and the median is 3. In the reference set there is only one point (viz., the sample itself) whose mean is at least as high as 3.4. In the same reference set, however, there are four points, namely, $(4, 7, \pm 2, 3, \pm 1)$ with median as high as 3. Therefore, with $\bar{x}$ as the test criterion the significance level (SL) of the data will be evaluated as 1/32, whereas with $\tilde{x}$ as the test criterion the data will be deemed to have attained SL = $\frac{1}{8}$. Note that every sample of five positive observations, irrespective of how far out or how scattered they are on the positive half-line, will be judged as significant (SL = 1/32) if $\bar{x}$ is the test criterion and not significant (SL = 1/8) if $\tilde{x}$ is the test criterion. With a sample of seven positive observations the SL will be 1/128 or 1/16 depending on whether $\bar{x}$ or $\tilde{x}$ is chosen as the test criterion. Consider the two samples $(-5, -4, -1, 6, 7, 8, 9)$ and (62, 63, 64, 65, 66, 67, 68). Does it make any sense to say that, with respect to the null hypothesis $\mu = 0$ with one-sided alternatives, the two samples are equally significant with SL = 1/16? But that is exactly what the Fisher test will do if $\tilde{x}$ is chosen to be the test criterion.

Let us take a short break from this ruthless cross-examination of the Fisher test with some speculation on Sir Ronald's later thoughts on the subject. The cross-examination will continue in Section 6.

5. DID FISHER CHANGE HIS MIND?

In all fairness to Sir Ronald we have to admit that, apart from making a passing reference to the randomization test method in Chapter III of *DE*, Fisher did not have much else to do with this kind of test procedure. Twenty-one years later, when Fisher came out with his last testament on statistics—*Statistical Methods and Scientific Inference* (*SI*)—he had apparently forgotten all about his randomization test method. In the winter of 1954–55, Fisher visited the Indian Statistical Institute for a couple of months and gave an extensive series of lectures based on the manuscript of *SI*. Those lectures profoundly influenced my own thinking on statistics. In *SI*, Fisher discussed the logic of inductive inference, his new outlook on significance testing, fiducial inference methods, likelihood methods of inference, and conditioning and recovery of ancillary information, but nowhere do we find any mention of randomization analysis of data. Randomization as an ingredient of statistical designs was mentioned only once, and that appeared in the following passage (Fisher 1956, p. 98):

> . . . whereas in the Theory of Games a deliberately randomized decision (1934) may often be useful to give an unpredictable element to the strategy of play; and whereas planned randomization (1935–53) is widely recognized as essential in the selection and allocation of experimental material, it has no useful part to play in the formation of opinion, and consequently in the tests of significance designed to aid the formation of opinion in Natural Sciences. [Note: The year 1934 refers to a Fisher article on randomization in card play and 1935–53 refers to *DE*.]

On the suggestion of an associate editor of *JASA* this passage is quoted in full so that the readers of the article can make up their own minds on the following.

Questions: Isn't it surprising that Fisher had no more to say about randomization in 1956? Was Fisher disassociating himself from the randomization test by not mentioning the method in *SI*? Does the remark about "formation of opinion" refer to postrandomization only?

It should be recognized that Fisher's views on significance testing underwent a major change during the period 1935–1956. On p. 77 of *SI* he made a clear distinction between tests of significance as used in natural sciences with tests for acceptance as in quality-control theory. According to him the dissimilarities (between the two methods) lie in the population, or reference set, available for making statements of probability. Let us quote Fisher (*SI*, p. 77) on this point:

> Confusion under this head has on several occasions led to erroneous numerical values; for, where acceptance procedures are appropriate the population of lots of one or more items, which could be chosen for examination, is unequivocally defined. The source of supply has an objective empirical reality. Whereas, the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination . . . . The demand was first made, I believe, in connection with Behrens' test of . . . significance . . . , that the level of significance should be determined by repeated sampling from the same population, evidently with no clear realization that the population in question is hypothetical, that it could be defined in many ways . . .; or, that an understanding, of what the information is which the test is to supply, is needed before an appropriate population, *if indeed we must express ourselves in this way*, can be specified. (Italics ours)

Again, on p. 91 of *SI*, Fisher quoted himself (from a 1945 *Sankhyā* article dealing with the fiducial argument) as follows:

> In recent times one often repeated exposition of the tests of significance, . . . , seems liable to lead mathematical readers astray, through laying down axiomatically, what is not agreed or generally true, that the level of significance must be equal to the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by hypothesis. This intrusive axiom, which is foreign to the reasoning on which tests of significance were in fact based seems to be a real bar to progress. . . .

It seems clear to me that, in 1956, Fisher's views on significance testing were somewhat close to the Bayesian position that the evidential content of data cannot be judged in sample space terms. Indeed, the 1945 quotation from Fisher might very well have been written by De Finetti himself. I am, therefore, not surprised at all that in *SI* Fisher mentioned neither the randomization test nor the lady-tasting-tea–type data analysis. For these are very extreme types of nonparametric data analysis in which the evidential meaning of the data is sought to be evaluated by referring it to a sample space that is formed by the statistician in his or her mind by imagining all the possible outcomes of the planned randomization input of the experiment. This will be made clearer in the next section.

Many of our contemporary statisticians are unaware of the fact that in the seventh edition of *DE* (1960), Fisher added what looks like a disclaimer in the form of a short section (Sec. 21.1; "Nonparametric" Tests) at the end of Chapter III. We quote this section in full.

> In recent years, tests using the physical act of randomization to supply (on the Null Hypothesis) a frequency distribution, have been largely advocated under the name of Nonparametric tests. Somewhat extravagant claims have often been made on their behalf. The example of this section, published in 1935, was by many years the first of its class. The reader will realize that it was in no sense put forward to supersede the common and expeditious tests based on the Gaussian theory of errors. The utility of such nonparametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly it is suspected that the simpler tests have been appreciably injured by departures from normality.
>
> They assume less knowledge, or more ignorance, of the experimental material than do the standard tests, and this has been an attraction to some mathematicians who often discuss experimentation without personal knowledge of the material. In inductive logic, however, an erroneous assumption of ignorance is not innocuous; it often leads to manifest absurdities. Experimenters should remember that they and their colleagues usually know more about the kind of material they are dealing with than do authors of textbooks written without such personal experience, and that a more complex, or less intelligible, test is not likely to serve their purpose better, in any sense, than those of proven value in their own subject.

Note Fisher's use of the phrase "physical act of randomization." The same phrase appears in the Fisher quotation in the opening paragraph of the previous section. Where is the physical act of randomization in the Fisher randomization test? The random entities $\delta_1, \delta_2, \ldots, \delta_n$ can hardly be called randomization variables. It is only under the null hypothesis that the δ_i's can be regarded as iid uniform ± 1 variables. The nonnull distribution of the δ_i's depends on the parameter of interest μ in a rather complex fashion. We should recognize the fact that in Section 21 of *DE* (1935) Fisher was not really concerned with the particular test situation that we have discussed in the previous section. He was talking about the problem of comparing two treatment effects under a wider hypothesis and was suggesting a (nonparametric) randomization analysis of data generated by paired comparisons on the basis of a physical act of randomization. In the next section we discuss this matter in some detail.

6. RANDOMIZATION AND PAIRED COMPARISONS

A scientist wants to test whether a so-called improved diet (treatment) is in effect superior to the standard diet (control). The scientist has 30 animals (subjects) with which to experiment. The scientist carefully pairs (blocks) the subjects into 15 homogeneous pairs. Let $\{(s_{1i}, s_{2i}): i = 1, 2, \ldots, 15\}$ be the set of 15 subject pairs. The subjects in each pair are of the same sex, come from the same litter, and so on. From each pair the scientist selects one subject for the treatment and the other one for control. The 30 responses (weight gain in so many weeks) are laid out as $\{(t_i, c_i): i = 1, 2, \ldots, 15\}$, where t_i and c_i are the responses of the treated subject and the control subject, respectively.

The scientist observes that

$$T = \sum t_i - \sum c_i$$

is a large positive number and also notes that

$$d_i = t_i - c_i > 0 \text{ for all } i \ .$$

The scientist, therefore, concludes that he or she has obtained very strong evidence in favor of the hypothesis H_1 that the improved diet is really superior to the standard diet. For measuring the strength of the evidence the scientist consults a statistician.

The statistician decides to make a one-sided test of significance of the null hypothesis H_0 (that the two diets are the same in their short-term weight-gain effects) on the basis of the scientist's data. The statistician also thinks that $T = \sum d_i$ is an appropriate test criterion in this case. For finding the significance level of the observed value of T, the statistician has to find the null distribution of T. So what the statistician needs now is a nice reference set.

The response difference d_i between the ith pair of subjects can be explained in terms of a possible treatment difference and other possible nuisance factors like subject differences (which the scientist tried his or her best to control by blocking), virus infection, loss of appetite, and many such uncontrollable factors that may have acted differently on the two subjects in the ith pair. If hypothesis H_0 is true, then there is no treatment difference; so the response difference d_i must be presumed to be caused by the previously mentioned nuisance factors. As Fisher explained in *DE*, randomiza-

tion enables the statistician to eliminate all these nuisance factors from the statistical argument. Let us see how this elimination is achieved in the present case.

Suppose the scientist had made 15 independent random decisions (on the basis of 15 tosses of a fair coin) as to which subject in the ith pair gets the improved diet ($i = 1, 2, \ldots, 15$). Having recorded the 15 response differences $d_1, d_2, \ldots, d_{15}$ and having computed $T = \sum d_i$, the scientist can speculate about a hypothetical rerun of the experiment in which all but one of the experimental factors (controllable or uncontrollable) are supposedly held fixed at the level of the last experiment—the same 30 animals exactly as they were at the commencement of the last trial, paired the same way into 15 blocks, exactly the same set of animals coming down with the same kind of virus infections with the same effects on them, and so forth. The only thing that is allowed free play in the hypothetical rerun of the experiment is the random allocation of treatment—the fair coin has to be tossed again 15 times. If H_0 is true, then the response difference $d_i = t_i - c_i$ for the ith pair must have been caused by the nuisance factors (subject differences, virus infection, etc). In the hypothetical rerun of the experiment all such nuisance factors are supposedly held fixed at the past level. Therefore, in the new experiment the response difference for the ith pair can take only two values d_i or $-d_i$, depending on whether the treatment allocation for the ith pair is the same as in the past experiment or is different. If we denote the response differences for the hypothetical experiment by $(d'_1, d'_2, \ldots, d'_{15})$ then it is clear that, under the null hypothesis H_0, the sample space (for the response differences) is the set R of 2^{15} points (vectors)

$$R = \{(\pm d_1, \pm d_2, \ldots, \pm d_{15})\} ,$$

with all the points equally probable. This is the reference set that the statistician was looking for. Let $T' = \sum d'_i$. The significance level of the data

$$\text{SL} = \Pr(T' \geq T | H_0)$$

is now computed as follows.

The statistician looks back on the data and notes that $d_i > 0$ for all i. Therefore, $T' \geq T$ if and only if $d'_i = d_i$ for all i. Hence, $\text{SL} = \frac{1}{2}^{15}$. This is randomization analysis of data in its classical form.

The rest of this section is devoted to an evaluation of this particular data analysis. The evaluation is laid out in the form of a hypothetical sequence of remarks and counterremarks by the statistician, the scientist, and the author.

Statistician: Observe that the randomization test argument does not depend on any probabilistic assumptions. The randomization probabilities are fully understood and are completely under control. I do not have to assume that the treatment-allocation process was like a sequence of 15 Bernoulli trials with $p = \frac{1}{2}$. Surely, I can regard that as demonstrably correct. In this argument there is no mention of a population. The experimental animals do not have to be regarded as a random sample from a population of animals. This test is an ultimate nonparametric test. Not only do we not have to deal with model parameters, we do not have to contend with even a statistical model. There is no mention of a sample space X equipped with a σ-field A of events and a family P of probability measures, no measurement errors, no mention of a sequence of iid random variables with an unknown distribution function.

Scientist: I am greatly puzzled by your data analysis. Your analysis seems to depend only on the randomization probabilities and the observed fact that $d_i > 0$ for all i. The fact that the test criterion $T = \sum d_i$ attained a rather large value in this case does not seem to enter into the probability evaluation of $\frac{1}{2}^{15}$.

Author: Suppose we choose the median of $d_1, d_2, \ldots, d_{15}$ to be the test criterion instead of $T = \sum d_i$. The significance level of the data will then be evaluated as $\frac{1}{2}^8$. How can we explain the big difference between $\frac{1}{2}^{15}$ and $\frac{1}{2}^8$?

Scientist: I do not understand the relevance of the randomization probabilities. Why is it so crucial that the coin with which I made the treatment allocation be a fair coin? Suppose I had used a biased coin with $p = \frac{1}{4}$. Suppose for the ith pair (s_{1i}, s_{2i}) of experimental animals my treatment allocation was (t, c) or (c, t) depending on whether the ith toss of the biased coin resulted in a head or a tail. How significant would my present data have been then?

Author: Let me answer the question. The hypothetical rerun of the experiment will be defined as before, but this time the biased coin will define the randomization scheme. The reference set for $(d'_1, d'_2, \ldots, d'_{15})$ will still be the same set R. Note that in this case $\Pr(d'_i = d_i) = \frac{1}{4}$ or $\frac{3}{4}$ depending on whether, in the original experiment, the response difference d_i was associated with the (t, c) or the (c, t) treatment allocation. Therefore, the level of significance will be evaluated as

$$\text{SL} = \Pr(T' \geq T | H_0) = (\tfrac{1}{4})^m (\tfrac{3}{4})^{15-m} ,$$

where m is the number of (t, c) allocations in the original experiment. The larger the value of m is, the more significant are the data.

Scientist: This is patently absurd. How can the SL depend so largely on such an irrelevant data characteristic as m? It is relevant to know that the 30 animals have been paired into 15 homogeneous blocks. The manner of my labeling the two animals in the ith block as (s_{1i}, s_{2i}) does not seem to be of much relevance. The number m of treatment allocations of the type (t, c) seems to be of no consequence at all. I have not been asked about all the background information that I have on the problem. For instance, I happen to have made a nutrition analysis of the two diets. I know that the improved diet has a much higher protein content and is very rich in vitamins C and D. I know the results

of several past experiments on the same set of animals when they were fed the standard diet. I know that six animals came down with virus infections during the experiment and that five of them were fed the improved diet. I am amazed to find that a statistical analysis of my data can be made without reference to these relevant bits of information.

Statistician: You are trying to make a joke out of an excellent statistical method of proven value, a method that originated in the mind of one of the two (Fisher and Einstein) really outstanding men of genius that the world has seen in this century. Your criticisms are based on an extreme example and then on a misunderstanding of the very nature of tests of significance. Tests of significance do not lead to probabilities of hypotheses. I do not believe in "belief probabilities." I do not believe that any useful purpose can be served by trying to quantify your knowledge in the form of a belief probability. Go to a Bayesian if you wish to make any input of your subjective beliefs in the data analysis process. It does not make much sense to set up a statistical model for the purpose of analyzing experimental data. The randomization analysis of data is so simple, so free of unnecessary assumptions that I fail to understand how anyone can raise any objection against the method. In the case of the present experiment you have in effect tossed a fair coin 15 times, have you not? So why confuse the issue by bringing in the case of an absurdly biased coin with $p = \frac{1}{4}$? Note that the probability of $\frac{1}{2}^{15}$ that I have computed for you is a gambler's probability, a frequency probability, a propensity measure of a well-defined physical system. A belief probability it is not.

Scientist: Your probability of $\frac{1}{2}^{15}$ is defined in terms of a hypothetical experiment, a rerun of the original experiment with everything (repeat everything) but the randomization part fixed at the level of the original experiment. But how can you even think of such an utterly impossible experiment? My experimental animals have changed—one of them died last week—the weather has changed, the virus epidemic is gone. I do not see how you can claim any objective reality for the randomization probability of $\frac{1}{2}^{15}$. In any case, I knew all along that the null hypothesis could not possibly be true. So any probability computed under the supposition that the null hypothesis is true cannot have much of an objective reality.

Author: The computation SL = $\frac{1}{2}^{15}$ was based on the supposition that in the hypothetical rerun of the experiment all the 2^{15} treatment-allocation patterns are equally probable. It is not clear from the argument that the scientist had to make all the 2^{15} possible allocations equally probable in the original experiment.

Scientist: This is a good time for me to confess that in fact I did not randomize over the full set of 2^{15} possible allocations. As a scientist I have been trained to put as much control into the experimental setup as I am capable of, to balance out the nuisance factors as far as possible. After carefully blocking the 30 subjects into 15 nearly homogeneous pairs, I could still detect differences within the subject pairs. There were differences in weight, height, some relevant blood characteristics, and a few other relevant features. I wanted the set of 15 treated subjects to be nearly equal to the set of 15 control subjects in some group characteristics like average weight, average height, and so on. I worked very hard on the project of striking a perfect balance between the treatment and the control groups. Finally, I found two such complementary groups and then decided on the treatment/control allocation to the two groups by a mental process that may be likened to the toss of a fair coin. I wonder what the significance level of my data is going to be in the light of this confession.

Statistician: Had I known about this before, I would not have touched your data with a long pole. Now the reference set for $(d'_1, d'_2, \ldots, d'_{15})$ consists of only the two points

$$(d_1, d_2, \ldots, d_{15}) \quad \text{and} \quad (-d_1, -d_2, \ldots, -d_{15}) ,$$

and the significance level

$$\text{SL} = \Pr(T' \geq T | H_0)$$

works out to be $\frac{1}{2}$ if $T > 0$ and 1 if $T \leq 0$. Your data is not significant at all.

Scientist (utterly flabbergasted): But my experiment was better planned than a fully randomized experiment, was it not? With my group control (in addition to the usual local control) I made it much harder for $T = \sum t_i - \sum c_i$ to be large in the absence of any treatment difference. In spite of this careful global control, I found that T is a large positive number and that every $t_i > c_i$. And you are telling me that, under the null hypothesis, it is as easy to get a result as significant as mine as it is to get a head from a single toss of a symmetric coin!

Statistician: My good man, you must realize that your experiment is no good. The prerandomization that you had carried out was not wide enough; the randomization sample space has only two points in it with a uniform probability distribution under the null hypothesis. Thus, the only attainable significance levels are $\frac{1}{2}$ and 1. Your experiment is not informative enough. I wish you had consulted me before planning your experiment. It appears that you do not have a clear understanding of the role of randomization in statistical experiments.

7. CONCLUDING REMARKS

So the randomization argument foundered on the rocks of restricted and unequal probability randomization. The statistician had the last word but lost the argument. The statistician was clearly wrong in characterizing the scientist's one-toss randomized experiment as uninformative. During the last 15 years, I have heard three very eminent statisticians characterizing the one-

toss experiment as uninformative on the score that the sample space has only two points in it. That this cannot be so is easily seen as follows.

An urn contains two balls that are either both white or both black. The draw of a single ball from the urn is then fully informative, although the sample space has only two points in it. If this example seems to be too artificial, then consider the case of an urn in which the proportion of white balls is either $\frac{1}{4}$ or $\frac{3}{4}$. Consider the sequential sampling plan that requires drawing of balls one at a time and with replacements until the likelihood ratio either exceeds 100 or falls below 1/100. Suppose the outcome of this experiment is recorded as "below 1/100" or "above 100." This is a highly informative experiment with only two sample points in it.

It should be noted that the sample space of the experiment performed by the scientist had a huge number of points in it. The statistician took a thin cross-section of the sample space (after holding fixed all the relevant factors like subjects, treatment effects, recognizable nuisance factors, and error terms) and then found only two points in it. No wonder the scientist failed to understand the argument.

The scientist was correct in questioning the relevance of randomization at the data analysis stage. Prerandomization injects an element of uncertainty about the actual experimental layout. But that uncertainty is removed once the scientist goes through the randomization ritual early in the game. At the data analysis stage, why is it still necessary to find out about the details of the actual randomization process? The randomization exercise cannot generate any information on its own. The outcome of the exercise is an ancillary statistic. Fisher advised us to hold the ancillary statistic fixed, did he not?

Our statistician is a most ardent admirer of R.A. Fisher. But he does not like the postfiducial (1936–62) Fisher. During the last 27 years of his astonishing career, we find Sir Ronald entertaining such counterrevolutionary thoughts as the conditionality and the likelihood principle and toying with the half-baked Bayesian idea of fiducial probability distribution.

We have noted earlier how the sufficiency principle rejects postrandomization analysis of data. Similarly, the conditionality principle (see Basu 1975 for more on this) rejects prerandomization analysis of data. In view of Fisher's postfiducial rethinking on statistical inference, it was almost inevitable for him finally to insert that astonishing short section on nonparametric tests in the seventh edition of *The Design of Experiments.*

[*Received March 1979. Revised October 1979.*]

REFERENCES

Basu, D. (1975), "Statistical Information and Likelihood" (with discussions), *Sankhyā*, Ser. A, 37, 1–71.
——— (1977), "On the Elimination of Nuisance Parameters," *Journal of the American Statistical Association*, 72, 355–366.
——— (1978), "On the Relevance of Randomization in Data Analysis" (with discussion), in *Survey Sampling and Measurement*, ed. N.K. Namboodiri, New York: Academic Press, 267–339.
Fisher, R.A. (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.
——— (1960), *The Design of Experiments* (7th ed.), Edinburgh: Oliver and Boyd.
Hodges, J.L., Jr., and Lehmann, E.L. (1973), "Wilcoxon and *t*-Test for Matched Pairs of Typed Subjects," *Journal of the American Statistical Association*, 68, 151–158.
Kempthorne, O. (1952), *The Design and Analysis of Experiments*, New York: John Wiley & Sons.
——— (1955), "The Randomization Theory of Experimental Inference," *Journal of the American Statistical Association*, 50, 946–967.
——— (1966), "Some Aspects of Experimental Inference," *Journal of the American Statistical Association*, 61, 11–34.
——— (1974), "Sampling Inference, Experimental Inference and Observations Inference," Paper presented at the Mahalanobis Memorial Symposium on Recent Trends of Research in Statistics, Calcutta, India.
——— (1975), "Inference for Experiments and Randomization," in *A Survey of Statistical Designs and Linear Models*, ed. J.N. Srivastava, Amsterdam: North Holland Publishing Co., 303–331.
——— (1977), "Why Randomize?" *Journal of Statistical Planning and Inference*, 1, 1–25.
Kempthorne, O., and Folks, J.L. (1971), *Probability Statistics and Data Analysis*, Ames: Iowa State University Press.
Pitman, E.J.G. (1937), "Significance Tests Which Can Be Applied to Samples From Any Population III. The Analysis of Variance Test," *Biometrika*, 29, 322–335.
Wald, A. (1950), *Statistical Decision Functions*, New York: John Wiley & Sons.
Warner, S.L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60, 66–69.

Comment

DAVID V. HINKLEY*

Basu has provided us with an interesting and provocative critique of significance tests related to randomized experiments. It does seem to be true that there is not a unified Fisherian mathematical theory of significance tests. This should not be surprising, however, since Fisher was wont to warn of the dangers of

* David V. Hinkley is Professor and Chairman, Department of Applied Statistics, and Professor, Department of Theoretical Statistics, University of Minnesota, Minneapolis, MN 55455.

September 1980, Volume 75, Number 371
Theory and Methods Section

routine, formal application of mathematical statistics without very careful regard for scientific context and operational meaning. Indeed, one might view Basu's paper as an illustration of Fisher's warning.

In terms of Fisher and randomization, the first four sections of the paper require little comment, since they deal with the separate topic of permutation and rank tests. Nevertheless, it is important to point out a fallacy in Basu's criticism of nonunique significance level (SL) in Section 4: The data as such do not possess an SL, which instead attaches to a particular statistic. Moreover, it is important to recognize that in Section 4 there is only an abstract null probability model—not a general model—so that the statistician has no basis for choice of statistic: The scientist must specify the relevant statistical measure. The role of the statistician here is to ensure that a valid, operational interpretation of the chosen statistic can be, and is, made.

After confessing to a "ruthless cross-examination" of the wrong topic—the non-Fisherian nonparametric tests of Section 4—Basu suggests that Fisher's silence in 1956 may be used to condemn the randomization test. This speculation seems unwarranted on two counts. First, I do not think that Fisher ever did recommend the randomization test for analysis of data, but rather that he introduced it as a device for demonstrating that randomization validates the usual normal-theory methods of analysis. This notion seems clear in Yates (1933), for example. Unfortunately, Basu has not chosen to discuss the connection between randomization and the validity of mathematical models. The second point is that Fisher's views on randomization would have been so widely known and accepted after 25 years that it would not have seemed necessary to repeat them in a book on statistical inference for parametric probability models. Fisher did repeatedly assert that randomization could guarantee the relevance of such abstract models—including the seemingly innocuous model in Section 4—but he realized that randomization was "sufficient" (CP 204)[1] rather than necessary, since often "Nature has done the randomization for us" (CP 212). These last remarks should be borne in mind when reading the amusing developments in Section 6.

What are we to make of the Statistician and the Scientist? They are certainly an entertaining addition to the literature, but hardly enlightening or enlightened. The first serious issue seems to be that of the biased-coin design, for which the author provides the SL. Surely the SL given is an appropriately cautious evaluation in the worst case in which the experimenter knowingly takes advantage of the bias—cheats, that is. But apparently the Scientist did not cheat ("my labeling . . . does not seem to be of much relevance"), so that in effect the treatments were allocated at random within each pair: Nature has done the randomization for us. Thus the usual analysis is presumably valid, and if a randomization SL were computed it would still be $\frac{1}{2}^{15}$. The Statistician has apparently mistaken Fisher's "sufficient" for "necessary."

What follows after the biased-coin episode is a series of irrational remarks and misunderstandings. The Statistician's dogmatic attitude is hardly characteristic of the statistician who inspired the Rothamsted song "Why! Fisher can always allow for it" (Box 1978, pp. 138–139).

What was Fisher's position on randomization and the induced distribution of statistics? While this is not entirely clear down to the last detail, I think it clear enough to suggest that Basu has missed the point. For a brief introduction to the relevant parts of Fisher's work, see the lectures by Holschuh, Picard, and Wallace in Fienberg and Hinkley (1980). Highly informative and balanced accounts of the issues may be found in Yates (1970), particularly the 1965 Berkeley Symposium paper reviewing experimental design. As I see it, the purpose of randomization in the design of agricultural field experiments was to help ensure the validity of normal-theory analysis. Nature was not in the habit of doing the randomization. Studies by Tedin (1931) and others on uniformity trial data showed that for systematic designs (such as that finally described by Basu's Scientist) the usual properties of t and F tests did not hold in an operational sense. Thus standard significance tests were invalidated. "Student," among others, correctly pointed out that effects could be more precisely estimated from carefully chosen systematic designs. But, said Fisher, this was of no use if the estimated precision were too high, higher even than the valid estimates obtained from randomized experiments. Thus, for some systematic designs, the computed normal-theory SL corresponding to a theoretically precise effect was in fact appreciably larger than the "real" SL. This is exactly what could happen in the case of the two-point design of Basu's Scientist, although it probably did not if nature has randomized. With the limited information given us by Basu, we cannot give a reliable standard error for the Scientist's accurate estimate $\bar{d}$, at least not one with a clear operational meaning. The apparently silly SL values ($\frac{1}{2}$ and 1) are a warning of possible difficulty, surely, nothing more.

The empirical evidence confronting Fisher certainly suggested the necessity of randomization in most field experiments, if the standard methods of analysis were to be used. In recent years it has become apparent that relatively simple spatial models can often account for some of the effects that randomization was designed to overcome; see Bartlett (1978), for example. Complete or partial failure to randomize can have adverse effects in other areas too, for example, in survey sampling in which systematic grids are randomly positioned on a sampling frame. In such a case systematic effects can accidentally (or purposely) change the variation, as I have seen myself. Cochran (1977, Ch. 8) discussed

[1] Fisher's papers are referred to by their numbering in the *Collected Papers* (Fisher 1974).

this problem in detail. For informative accounts of the importance of randomization in medical and public-policy studies, respectively, see Chapters 9, 10, and 11 of Bunker, Barnes, and Mosteller (1977) and Gilbert, Light, and Mosteller (1977). In all areas, the randomization distribution literally induced by the experimental randomization is of value in assessing the validity of a standard analysis. This, I think, is Fisher's message.

The final substantial issue of Basu's paper is that of the ancillarity of the design outcome. Technically Basu is quite correct, if the randomization has validated a parametric model—the design outcome is then ancillary *by design!* It would, however, be as well not to forget the purpose of an ancillary statistic, since otherwise we are merely playing with abstract mathematical definitions. An ancillary statistic indicates the set of comparable cases against which to judge the observed sample and the statistical summary thereof. Usually "comparable cases" is taken to mean "equally informative samples" in some appropriate sense, as in Fisher's brief comments on the 2×2 table (CP 205). Admittedly this is not mathematically precise, but it seems to have the merit of common sense. It is often unnecessary, and sometimes plain foolish, to take an infinitesimal slice through an abstract space as the set of comparable cases, although the mathematical definition of ancillarity would require this. Lest Basu think that Fisher has been caught with his conditional pants down here, let me suggest that Fisher implicity invoked conditionality in his criticism of Knut-Vik Squares, which in Tedin's (1931) analyses correspond to an ancillary set of real import.

For a more constructive use of design ancillarity, consider a randomized block design (RBD) with four replicates of four treatments. Suppose that in the particular physical layout the selected design coincides with a 4×4 Latin Square and that the accidental block structure corresponds to a noticeable effect not due to the treatments. Here my qualitative notion of ancillarity would suggest that we analyze the experimental data as coming from a Latin Square design, that is, treat the 4×4 Latin Squares as that subset of RBD's that constitute the set of comparable cases. The Latin Square analysis would be exactly equivalent to using a covariate-adjusted RBD analysis with accidental block totals as covariates. (There is of course a question as to whether randomization validates the latter analysis.)

This short discussion has necessarily focused on my major misgivings with Basu's interesting paper. As to whether randomization tests are *logically* viable, I think Basu has not made a case. There may be no case in logic if, with John Clerk Maxwell, we believe that "the true logic for this world is the calculus of Probabilities." What we need to know is: Which probabilities?

[*Received December 1979.*]

REFERENCES

Bartlett, M.S. (1978), "Nearest Neighbour Models in the Analysis of Field Experiments" (with discussion), *Journal Royal of the Statistical Society*, Ser. B, 40, 147–174.

Box, J.F. (1978), *R.A. Fisher. The Life of a Scientist*, New York: John Wiley & Sons.

Bunker, J.P., Barnes, B.A., and Mosteller, F. (1977), *Costs, Risks, and Benefits of Surgery*, New York: Oxford University Press.

Cochran, W.G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley & Sons.

Fienberg, S.E., and Hinkley, D.V. (1980), *R.A. Fisher: An Appreciation, Lecture Notes in Statistics*, New York: Springer-Verlag.

Fisher, R.A. (1974), *Collected Papers of R.A. Fisher*, Adelaide, Australia: University of Adelaide. (Papers are referred to by number.)

Gilbert, J.P., Light, R.J., and Mosteller, F. (1977), "Assessing Social Innovations: An Empirical Base for Policy," in *Statistics and Public Policy*, ed. W.B. Fairley and F. Mosteller, Menlo Park, Calif.: Addison-Wesley.

Tedin, O. (1931), "The Influence of Systematic Plot Arrangements Upon the Estimate of Error in Field Experiments," *Journal of Agricultural Science, Cambridge*, 21, 191–208.

Yates, F. (1933), "The Formation of Latin Squares for Use in Field Experiments," *Empire Journal of Experimental Agriculture*, 1, 235–244. (Also reprinted in Yates 1970.)

——— (1970), *Experimental Design. Selected Papers of Frank Yates, C.B.E., F.R.S.*, London: Griffin.

Comment

OSCAR KEMPTHORNE*

Basu states that we have no satisfactory answer to the question, "Why randomize?" Various workers have attempted to give "satisfying" partial answers to this question. Surely, Fisher (7th ed., 1960) did so, with extensive exposition in his *The Design of Experiments*. Then we can examine various writings of the 1930's.

We can cite Greenberg (1951) with the question as the title, as also was the title of the cited Kempthorne

* Oscar Kempthorne is Professor of Statistics and Distinguished Professor in Sciences and Humanities, Iowa State University, Ames, IA 50011.

September 1980, Volume 75, Number 371
Theory and Methods Section

(1977). It would be useful to have an attempted exhaustive bibliography of the topic.

It is obvious that expositions that some regard as carrying *some* real force are not so regarded by Basu and, for example, by Harville (1975). Is there any possibility that discussion will resolve the disagreement? I believe not. But I do believe that discussion is useful. All of us surely subscribe to the absolute necessity of critical examination of our ideas.

My discussion consists of two parts: (a) reactions of Basu's essay and (b) a few comments on the nature and role of randomization.

Basu writes entertainingly, perhaps, but not informatively. He poses the question, "Can the Fisher randomization test pass the test of common sense?" We must, I suggest, force Basu to be explicit and clear. What is this "common sense" that Basu refers to? Presumably, it is Basu's "common sense." The philosophy of statistics is plagued with writers who talk about "the probability" only to tell us that they mean "my probability"; now we have "common sense," but it is "Basu's common sense."

Section 2 given us prerandomization, postrandomization, and unrecorded randomization. This discussion is irrelevant. But it is useful, perhaps, to make a remark. It is ludicrous that Basu, a keen bridge player, does not, it seems, give a role to postrandomization. Let Basu play poker with significant (to him) payoff. Then if he does not use some sort of postrandomization, maybe very informal, he will be "cleaned out." I speak from past personal experience of playing social, but nontrivial (it now bores me!) poker. I regret that I surmise that many of those who write about gambling do not practice it. Section 2 is a red herring.

Section 3 discusses the sufficiency principle. As he has written, and others too numerous to cite, this is a data-reduction principle. It was used by Fisher in his (inadequate) formulation of tests of significance and reached its summit for Fisher in his fiducial inversion (which I shall not discuss). Problems with both of these led to the also inadequate formulation of use of ancillaries. The Fisher prescription, "The (Basu-titled) Insufficiency Principle," reached its summit only with fiducial inference, which was, in fact, the only real inference that Fisher espoused. It is in these terms, I suggest, that the later Fisher must be examined.

Section 4 discusses the Fisher randomization test on the basis of the cited Kempthorne-Folks presentation. My only regret here is that Basu did not examine, it seems, an article by Kempthorne and Doerfler (1969). I found Basu's remarks on the test not violating the sufficiency principle interesting and possibly a justification of the quoted remark of Kempthorne and Folks about condensation. Also, it was comforting that Basu seems to conclude that $k = 1$ gives "the only reasonable choice of a conditioning statistic." On the choice of test criterion, the naive idea I had was that the alternative is a uniform shift so that the sample total contains, perhaps, the maximum total shift. Here, there surely is a question. To this I add that my own use of the randomization test is in the experimental setting in which the alternative hypothesis is that a treatment adds some quantity Δ to each and every unit to which it is assigned. Also on a technological level, a question of interest is whether the treatment gives a gain when applied to *all* the experimental units.

Section 5 discusses whether Fisher changed his mind. This has perplexed others, including me (Kempthorne 1966, 1974, 1975). I have commented (1975) on what appeared to me to be outright inconsistencies in Fisher's whole output. It serves no useful purpose to try to psychoanalyze this phenomenon, I suggest. I do suggest, however, that we frankly admit the occurrence of these inconsistencies. Clearly, Fisher (1956) was writing in part a polemic against acceptance procedures. In connection with one quotation, it is obvious that the population in a randomization test of a randomized experiment is "the product of the statistician's imagination." (Why Fisher should include "exclusively," I do not know.) *It is not clear* to me that in 1956 Fisher had the position that "evidential content of data cannot be judged in sample space terms." On the matter of the "lady tasting tea" and randomized experiments, Fisher (1956) is, I judge, entirely silent, and that is surely a mystery. I have to state my opinion that I do not find Basu's psychoanalysis clear or convincing. On the Fisher (1960) quotation, Fisher is merely polemic, for some unknown reasons. In fact one can use Fisher's words against Fisher. One does *not* have knowledge of distribution. If one did, one would not be involved in transformation search, for instance. Any supposed statistician who believes he or she knows the model, for example, of normality and independence, is not a real statistician; that is surely obvious. The only interpretation of this is that Fisher was polemicizing, against what we can guess, but with no profit.

With respect to Basu's writing on "the physical act of randomization," I believe Basu is merely *plain wrong*. In a randomized experiment, the δ_i's have a known distribution whether or not the null distribution holds.

Section 6 gives in the first part what is, or should be, routinely taught on the randomized pair trial. This has been known for decades and an elementary substantively oriented exposition is that of Kempthorne (1961).

In the second part Basu gives a hypothetical interchange of a statistician and a scientist and the author. I suggest that this serves *no* useful purpose. Comments on sentences of this interchange follow.

1. *Scientist:* "The fact that . . . $T = \sum d_i$ attained a large value . . . does not seem to enter."

Comment: Precisely! When is T large? With reference to what is T large? If one has external information on the possible magnitude of T, then one will have an idea of what values of T are large. If one is a Bayesian, one *claims to know* the possible distribution of T. Clearly,

if one is in this situation, one does not randomize. I believe Basu has no familiarity with the problems of evaluating drugs for human illnesses, of evaluating diets on humans or mice or whatever. He does not see the variability among humans "treated alike." Why, is a mystery to me. Basu seems to say: If you are an expert on cancer, then you know a probability model. Otherwise, you should withdraw from the field. I regret that I find the lack of knowledge that underlies Basu's thesis rather surprising, incongruous, and deplorable. I would like Basu to take up a "very small" branch of investigative science, learn all the available background, and then design and conduct, with aid, of course, his own research program. Because Basu is highly competent, I believe, in the game of bridge, I would like him to make a comparative trial of two bridge systems. How would he do this? He surely has as good a background as any bridge player or writer. It is the absence of any effort on a problem outside the WFFing (constructing well-formed-formulas) of mathematical statistics that concerns me. As I have said before, we must be skeptical of individuals who write books on cooking but have never made a meal in a kitchen.

2. *Scientist:* "Why is it crucial that the coin . . . be a fair coin?"

Comment: From Basu's viewpoint, this is obviously irrelevant. From the viewpoint of the investigator who is not a Bayesian, the situation is different. If one does not regard experimentation as a process of investigation, with the value of the process being determined, partly at least, by its operating characteristics, the question is irrelevant. But a scientist who does not care about the operating characteristics of his or her observation procedure is a "pretty poor" scientist. This is my opinion, of course, but one that is shared by the great bulk of scientists, I am totally sure. Indeed, the question with respect to any substantive experimental outcome is whether other scientists can duplicate the results. This is all very elementary, and I will not waste valuable journal space discussing it. The question is irrelevant to Basu, because one should not use any coin. But for the person who accepts the idea that operating characteristics of a procedure are important and who regards significance tests as evidential, the answer is obvious. If one uses a collection of plans, of which one plan has probability .99, then the significance level regardless of outcome and regardless of whether there is a real treatment difference will be equal or exceed .99 with probability .99. One then has to discard the idea of significance tests—at least as they are used at present. If there are M plans, then using these with equal probability gives the possibility with huge treatment effects of obtaining a significance level of $1/M$. So, for the significance tester, there is value to equal probabilities, or the "fair coin." From another point of view, the use of equal probabilities gives estimates with nice properties, an analysis of variance with nice properties, and, surely, a valid use of the central limit theorem for the distribution of the test criterion in comparative trials of reasonable size. The reply of the author does not consider, I think, operating characteristics.

3. *Scientist:* "This patently absurd . . . ."

Comment: I can say, equally as Basu, that his writings on randomization are "patently absurd"—but, of course, this does not lead to improved understanding. Basu credits the scientist with all sorts of "background information." The scientist "knows etc." The scientist and Basu are entitled to "be amazed to find that a statistical analysis of my data can be made without reference to these relevant bits of information." Why? Because if the scientist really has these "bits of information" a decent statistician will attempt to take them into account. *No one claims that the randomization test of significance is the beginning, middle, and end of statistical analysis.* Finally, I must ask the question, "Has Basu worked intimately with any scientists with a real problem (as opposed to a circus trainer with 10 elephants)?"

4. *Scientist:* "But how can you even think of such an utterly impossible experiment?"

Comment: I, Kempthorne, can! A very simple answer! I follow this with a question to Basu, related to the very interesting arcane mathematics he sometimes does. "How can you, Basu, even think of observing a real number exactly?" Each of us has a mental problem. Let us rest the matter there.

5. *Scientist:* ". . . I did not randomize over the full set . . . ."

Comment: For me as a randomizing significance tester, that presents no problems. Tell me what your randomization frame was, and I can proceed. I may well find that to interpret your results a repeated sampling principle is useless. I, or rather you, have to supply a prior and a probability distribution. This is, perhaps, no problem for you. But it is for me, because now I have to assess for myself how much belief to hold with respect to your opinion. That is the rub!

6. *Statistician:* "Your data are not significant at all."

Comment: Right on! Your data are not significant to me. They may be to you, of course.

7. *Basu:* "So the randomization argument founders on the rocks of restricted and unequal probability randomization."

Comment: I do not see the claimed foundering.

8. *Basu:* "The one toss experiment (is) uninformative . . . ."

Comment: It is uninformative to me in the absence of forcing relevant and supported prior or external information. With such information, the actual ex-

periment is only a part of the total information. What is there to argue about?

9. *Basu:* "The outcome of the (randomization) exercise is an ancillary statistic."

Comment: Yes and no. This outcome does not depend on the probability model, but if one does not know the probability model, one cannot (or should not) characterize the randomization outcome as ancillary. Furthermore, Basu's own work (not cited, but very well known) shows that there are huge difficulties in strict formulation of ancillary statistics.

I have one final comment about Fisher (1956). It is clear from Chapter III that Fisher envisaged various forms of inference from tests of significance to distributions on unknown parameters. He did *not*, then, reject tests of significance in 1956. Furthermore, there is *no* evidence that he rejected his "lady tasting tea" example.

I close with the statement (which will be unknown to most readers of this journal) that Basu and I are very deep friends. The argumentation and the comments I make in this article must be interpreted with that background.

[*Received December 1979.*]

REFERENCES

Fisher, R.A. (1956), *Statistical Methods and Scientific Inference*, Edinburgh: Oliver and Boyd.

——— (1960), *The Design of Experiments* (7th ed.), Edinburgh: Oliver and Boyd.

Greenberg, B.G. (1951), "Why Randomize?" *Biometrics*, 7, 309–322.

Harville, D.A. (1975), "Experimental Randomization: Who Needs It?" *The American Statistician*, 29, 27–31.

Kempthorne, O. (1961), "The Design and Analysis of Experiments With Some Reference to Educational Research," in *Research Designs and Analysis, Second Annual Phi Delta Kappa Symposium on Educational Research*, 97–126.

——— (1966), "Some Aspects of Experimental Inference," *Journal of the American Statistical Association*, 61, 11–34.

——— (1974), "Sampling Inference, Experimental Inference and Observations Inference," Paper presented at the Mahalanobis Memorial Symposium on Recent Trends of Research in Statistics, Calcutta, India.

——— (1975), "Inference for Experiments and Randomization," in *A Survey of Statistical Designs and Linear Models*, ed. J.N. Srivastava, Amsterdam: North-Holland Publishing Co., 303–331.

Kempthorne, O., and Doerfler, T.E. (1969), "The Behaviour of Some Significance Tests Under Experimental Randomization," *Biometrika*, 56, 231–247.

Comment

DAVID A. LANE*

The scientist's experimental results contain evidence bearing on the superiority of the improved diet. He asks the statistician to evaluate this evidence. The statistician answers by computing a significance probability, $\frac{1}{2}^{15}$, by means of Fisher's randomization test. The scientist is baffled:

How can the evidence in his results be measured by a computation that ignores so much relevant information: the magnitude of the difference in weight gain between the two groups of animals, the ingredients of the two diets, previous experience with the standard diet, knowledge of the experimental animals gathered before and during the experiment, the mechanisms of the growth process, and so forth?

The statistician's computation refers to a biologically irrelevant feature of the experiment, the physical properties of the device determining the assignment of animals to diet; how can such a computation connect to the biological problem of the superiority of the improved diet?

The answer to the first question is clear: The statistician's significance probability cannot summarize completely the evidence in the scientist's experiment. The scientist cannot get something for nothing. If the scientist wants to assess what his experimental results imply about the effects of his improved diet and the nature of the growth process, he must analyze them in terms of a statistical model that describes as much as possible of what he knows about the biology of the experiment. But there are rhetorical as well as inferential issues involved in discussing an experiment. One of the scientist's goals is to obtain public confirmation for the superiority of the improved diet. If this can be accomplished with a minimum of fuss and assumption, preliminary to the detailed, model-based analysis, and without contradicting explicitly or implicitly the results of that analysis, so much the better. Here, the randomization test may be of use.

The randomization test addresses the question: Might the two diets really be equally effective and the apparent superiority of the improved diet be attributed to chance variability? The success of the test depends on the relevance of the interpretation it requires for the notions of "equally effective diets" and "chance

* David A. Lane is Assistant Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455.

September 1980, Volume 75, Number 371
Theory and Methods Section

variability." According to the Fisherian foundation of the test, two treatments can be considered equally effective only if they would each elicit exactly the same response from each experimental unit. The experimenter may, however, be interested in a weaker notion of equality between two treatments: Their distributions for the responses over the experimental units (or over all potential recipients of the treatments) should coincide. For example, a physician may not believe that each cancer patient faces the same prospect for a cure from radiotherapy as from chemotherapy, but he or she still might want to entertain the hypothesis that the overall success rates of the two treatments might be the same. The randomization test would be of no help to the physician.

The way in which "chance variability" enters into his experiment should be carefully explicated by the scientist when he constructs the statistical model he will use for analyzing his results. The randomization test ignores this model and substitutes an alternative relation between chance and the experiment, based on a frequency distribution induced by the physical act that assigns animals to diets. The logical foundation of this relation is challenged by the scientist's second question.

Basu presses this challenge home and denies Fisher's dictum that the physical act of randomization validates the randomization test. I find his argument convincing, and yet it seems to me that the significance probability of $\frac{1}{2}^{15}$ can possess a rhetorical force that tells for the superiority of the improved diet, without reference to the distribution induced by the physical randomization. To explain this, I need to describe certain thoughts that the scientist might have about his experiment.

For each of his 30 animals, the scientist has ideas at the beginning of the experiment about what the animal would weigh at the end, were it fed the standard diet. Although it undoubtedly implies more precision than the scientist could readily supply, think of these ideas as generating 30 standard-diet predictive distributions. One hypothesis about the diets that the scientist might entertain—although we know he does not believe it!—is H_0: Each animal would end up weighing the same under the standard diet as it would under the improved diet. In particular, if H_0 were true, the 30 standard-diet predictive distributions also describe the scientist's ideas about what the animals would weigh if they were fed the improved diet. Now suppose the scientist has paired his 30 animals so successfully that the predictive distributions for the two animals in each pair coincide. Moreover, suppose also that there are no patterns of covariation among his animals such that, if H_0 were true and he knew the outcome of the experiment for some group of pairs, the scientist's conditional predictive distributions for the two animals in each of the remaining pairs would differ. Call this state of knowledge—or lack of it!—about the experimental animals *null neutrality*.

Under H_0 and null neutrality, the scientist's predictive probability that the 15 animals on the improved diet all end up heavier than their partners is $\frac{1}{2}^{15}$. In fact, the joint predictive distributions under H_0 and null neutrality induce the uniform distribution on the set $S = \{(\pm 1, \ldots, \pm 1)\}$, where the ith coordinate of a point in S is $+1$ if d_i is positive and -1 otherwise. So the usual null distribution of the sign test derives from the scientist's predictive distributions under H_0 and null neutrality, without regard to the method of assignment of animal to diet. The null distribution for the Fisher randomization test can also be derived, with somewhat more tedious assumptions about conditional predictive distributions, in terms of the scientist's prior beliefs about the experimental outcome under H_0. Since the scientist's real beliefs about the superiority of the improved diet imply predictive distributions weighted toward large positive values for the d_i's, small values of the significance probability from the randomization test indicate small posterior probability near H_0, if the scientist had assessed null neutrality and fully probabilized the problem—hence the rhetorical if not inferential force of the $\frac{1}{2}^{15}$.

What about the assessment of null neutrality? It is to be regarded as a rough approximation at best; if the scientist is willing to think hard enough, he can of course recognize differences between any pair of animals. Still, if null neutrality holds approximately, so do the conclusions that follow from assuming it, which serve only as guidelines anyway. In this regard, it is not much different from assuming normality in measurement situations. Yet, just as with normality, it is an assessment not to make lightly—and, as I shall argue later, the physical act of randomization can play a role in deciding whether the assessment is appropriate.

To see whether this or the Fisherian interpretation of the statistician's significance probability provides the sounder guidance for the scientist, it is useful to consider some extreme cases. The issue should not be whether these cases occur in practice, but whether the logic that you claim to follow in practice guides you rightly or wrongly when pressed into extremity.

Example 1 (a variant of Basu's biased coin): The scientist achieves a successful pairing—null neutrality seems reasonable. He generates 15 random numbers on the university computer, associates each of these numbers with a distinct pair of animals, and assigns the first animal (first, relative to a list of the animals' cage addresses) in each pair to the improved diet, if the pair's random number is even. It turns out that, in each pair, the animal that received the improved diet ends up heavier.

Just as the scientist is about to write up his results, however, the computer center informs him that because of a faulty program, only about 40 percent of the random digits the generator produced during the experiment were even.

Example 2: Same story, but the scientist knew about the generator's quirk before he chose the numbers.

Example 3: The scientist is not so lucky as in example 1, or perhaps he knows more about his experimental animals: In each pair, he can identify one animal that seems to have more growth potential than the other. This time, the random-number generator is working fine. Surprisingly, all the animals that the scientist judged to have higher growth potential get assigned to the improved diet. And they all end up heavier.

Basu carefully—and, as far as I can see, successfully—argues that Fisher's logic leads to a significance probability different from $\frac{1}{2}^{15}$ for example 2. The same argument must apply to example 1, since Fisher's logic allows the scientist's knowledge of the randomizing mechanism to enter into the analysis only when he writes down a probabilistic model for it—and since this model attempts to represent the mechanism's physical properties, he must use whatever he knows when he analyzes the experiment, not what he thought he knew when he generated the random numbers. The significance probability derived from the Fisherian logic, as discussed by Basu, is singularly unattractive as a measure of evidence, depending as it does on an artifact of the method of listing cages.

Fisher's logic, tied to the random-number generator and imaginary repetitions, cannot fault the calculation of a significance probability of $\frac{1}{2}^{15}$ in example 3. The design of the experiment may be at fault here, and the experiment itself quite uninformative scientifically, but this does not seem to stand in the way—in the Fisherian framework—of analyzing its evidential content by the $\frac{1}{2}^{15}$ significance probability.

Interpreting the statistician's significance probability in terms of the scientist's predictive probability distribution changes this analysis completely. In example 1, the significance probability of $\frac{1}{2}^{15}$ is unchanged by the computer center's information, since the probability refers to the scientist's thoughts at the commencement of the experiment, to which the information is irrelevant. At first sight, the same holds in example 2, since the probability does not refer to the method of assignment of animal to diet. But the scientist is interested in sharing his assessment of null neutrality: He wants his readers to feel the force of his argument, and so his assessment must be theirs. From this point of view, using a biased or arbitrary mode of assignment is to invite suspicion of loading the experiment in the scientist's favor—perhaps unconsciously, as in the famous Lanarkshire milk experiment ("Student" 1931). Randomly assigning animals to diets with public probability $\frac{1}{2}$ is a way of guaranteeing the honesty—to the public and the scientist himself—of his subjective assessment that both animals in a pair had the same standard-diet predictive distribution.

In example 3, the scientist cannot assess null neutrality, and so the significance probability of $\frac{1}{2}^{15}$ does not apply. Here, he can block his pairs according to his predictive distributions for the d_i's, to ensure as informative an experiment as possible. Again, he can employ one or more physical acts of randomization as a check and guarantee of his subjective assessments of these distributions. The experiment can of course be analyzed, but the null distribution for the randomization test will no longer follow Fisher's frequency distribution and will necessarily be somewhat less open to general agreement.

[*Received December 1979.*]

REFERENCE

Student (1931), "The Lanarkshire Milk Experiment," *Biometrika*, 23, 398.

Comment

D.V. LINDLEY*

What is one to do with this paper but applaud it? Another incoherent procedure has its nature clearly displayed. Here is an encore that I have used in class to suggest that the randomization test does not "pass the test of common sense."

The example is artificial in that the experiment is very small, but this has the virtue of simplifying the arithmetic. The same principle holds for a larger and more realistic experiment at the expense of computations that might obscure the essential ideas. Two scientists are to conduct an experiment to compare a treatment, T, thought to improve the yield, with a control, C. Four units are to be used, two each for T and C. The six possible assignments of T and C to the units are

* D.V. Lindley was formerly Head, Department of Statistics and Computer Science, University College London. He is now retired and lives at 2 Periton Lane, Minehead TA24 8AQ, England.

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

listed in the first column of the tabulation appearing two paragraphs after this one. The first scientist, A, decides to select one of the six designs at random. The second scientist, B, feels that the first and last designs would be unsatisfactory, because all the treatments and all the controls come together, and therefore selects a design at random from the four remaining. (In practice, as mentioned before, larger sets of designs would be used.) Both A and B carry out their respective randomizations and both come up with the design *TCTC*, in the second row of the tabulation. On implementing the design, both scientists obtain the results 5, 4, 3, 2 shown in the final row of the tabulation. The total for the treated units is 8, that for the control 6, and the effect is measured by the difference, 2. So far the scientists agree, but now see what happens if they use the randomization argument for analysis.

Had the observed values arisen from any other of the designs that might have been used, the differences would have been those listed in the second column of the tabulation. Consider scientist B first. Scientist B excluded the first and last designs, and so the possible differences are (2, 0, 0, −2), of which the first, the one actually obtained, is the largest. Hence the result is significant at 25 percent, because all designs had the same 25 percent chance of being used. Scientist A, however, included the first and last designs in the randomization so must include the differences 4 and −4 that could have arisen by use of them. Of all six differences, 4 is the largest and 2, the one actually observed, the next largest. Hence the chance of the observed difference, or more extreme differences, is 2 out of 6 and the result is significant at 33⅓ percent.

There, then, are two scientists who have performed exactly the same experiment, *TCTC*, obtained exactly the same result, and yet one is quoting a significance level substantially in excess of the other. And the reason for this difference in level is that A contemplated doing experiments that B did not (viz., those in the first and last rows of the tabulation), although, in fact, A did not perform one of these experiments. Expressed slightly differently, the analysis of the results of the experiment depended on what might have been done, but in fact was not done. Certainly in this context, in which the only probability ideas leading to the level are the equal probabilities involved in the random assignment, the argument seems unsatisfactory.

		Designs	Differences
A		TTCC	4
	B	TCTC	2
		TCCT	0
		CTTC	0
		CTCT	−2
		CCTT	−4
Results		5 4 3 2	

The whole concept of A and B reaching substantially different conclusions seems so absurd that the randomization-analysis argument has to be dismissed. There are two defenses: first, that in practice substantial differences (like 25 and 33⅓ percent) are not observed and that the results are typically the same as normal theory. In that case, why not use normal theory? The second defense is that A and B ought to argue differently because B thought that the first and last experiments might be unsatisfactory, whereas A did not. In other words, both scientists had different ideas before the experiment; is it not reasonable that the two scientists should have different ideas afterwards? This argument violates the claim often made for significance tests—that they allow the data to speak for themselves and are not affected by considerations outside the data—and if admitted plays straight into the Bayesian camp, where the ideas of prior information are considered explicitly.

A minor comment is that it is perhaps a little unfair to say that there is "no satisfactory answer to the question: Why randomize?" The work of Rubin (1978) has at least made a substantial contribution to the answer. The answer for me is tied up with what we mean by *random*. (Basu's definition of randomization, in the first section of Section 2, is in terms of randomness.) I suggest that X is random, given H, if X is independent of any A, given H; that is, if $p(A|X, H) = p(A|H)$. The idea is that the generation of X, whether by a random mechanism, or by pseudorandom numbers, is unconnected with anything else. It is thus a subjective notion, in that what you consider random, I might not; though, in practice, we observe a lot of agreement among people. The value of randomization in design may then be illustrated by an experiment to test the efficacy of treatment T in aiding the recovery R of a patient. We require the probability of a patient's recovery were the patient to be given a treatment, $p(R|T, D)$, using data D from a planned experiment. This may differ from $p(R|T, D, A)$, where A is some factor unrecognized by us. (Had it been recognized it could have been planned for in the acquisition of D.) In order to make reasonably sure that our design does not confound the effects of T and A, we may assign treatments at random, that is, independent of A. This does not ensure lack of confounding but reduces its possibility to an acceptable level. Thus prerandomization has a place in coherent analysis: Basu shows that postrandomization is incoherent.

[*Received December 1979.*]

REFERENCE

Rubin, Donald B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6, 34–58.

Comment

DONALD B. RUBIN*

Basu's article on Fisher's randomization test for experimental data (FRTED) is certainly entertaining. Although much of the paper is devoted to the thesis that Fisher changed his views on FRTED, apparently the primary point of the paper is to argue that FRTED is "not logically viable." Admittedly, FRTED is not the ultimate statistical weapon, even in randomized experiments, but calling it illogical is rather bizarre.

Basu criticizes FRTED through two primary arguments. His first line of criticism follows from his attack on a nonparametric test labeled in Section 4 as "Fisher's randomization test." But this test was not proposed by Fisher and is not a logical variant of FRTED; consequently, these criticisms are not of FRTED. I believe that Basu agrees with this contention because in concluding this first criticism he states, "Where is the physical act of randomization in the Fisher randomization test? . . . We should recognize the fact that in Section 21 of *Design of Experiments* (1935) Fisher was not really concerned with the particular test situation that we have discussed in the previous section." Basu's second line of criticism of FRTED takes the form of a discussion between a statistician and a scientist; I find this discussion so confused that it is easier for me to challenge the argument indirectly by clearly describing FRTED than directly by correcting particular misconceptions.

In the paired comparison experiment, let Y_{ij} be the response of the ith unit ($i = 1, \ldots, 2n$) if exposed to treatment j ($j = 1, 2$), where $Y = \{Y_{ij}\}$ is the $2n \times 2$ matrix of values of Y_{ij}. The assumption that such a representation is adequate may be called *the stable unit-treatment value assumption:* If unit i is exposed to treatment j, the observed value of Y will be Y_{ij}; that is, there is no interference between units (Cox 1958, p. 19) leading to different outcomes depending on the treatments other units received and there are no versions of treatments leading to "technical errors" (Neyman 1935). If Y were entirely observed, we could simply calculate the effect of the treatments for these $2n$ units; for example, $Y_{i1} - Y_{i2}$ would be an obvious measure of the effect of treatment 1 versus treatment 2 for the ith unit, and the average value of $Y_{i1} - Y_{i2}$ would be a common measure of the typical effect of treatment 1 versus treatment 2 for these $2n$ units. Because each unit can be exposed to only one treatment, we cannot observe both Y_{i1} and Y_{i2}, and so we will have to draw inferences about the unknown values of Y from observed values of Y.

Let $T = (T_1, \ldots, T_{2n})$ be the indicator for treatment received: $T_i = 1$ if the ith unit received treatment 1 and $T_i = 2$ if the ith unit received treatment 2; if $T_i = 1$, Y_{i1} is observed and Y_{i2} is missing, whereas if $T_i = 2$, Y_{i2} is observed and Y_{i1} is missing. In order to avoid confusion about the inferential content of indices, suppose that the unit indices i are simply a random permutation of $(1, \ldots, 2n)$. The pairing of the units in the paired comparison experiment will be represented by X, where $X_i = 1$ for the two units in the first pair, ..., and $X_i = n$ for the two units in the nth pair. Other characteristics of units can be coded in other variables, but for simplicity assume for now that only values of Y, X, and T will be used for drawing inferences, where Y is partially observed and both X and T are fully observed.

Both randomization and Bayesian inferences for unobserved Y values require a specification for the conditional distribution of T given (Y, X), say $\Pr(T|Y, X)$. The physical act of randomization in the experiment (e.g., the physical act of haphazardly pointing to a starting place in a table of random numbers) is designed to ensure that all scientists will accept the specification $\Pr(T|Y, X) = \Pr(T|X)$. In the paired comparison experiment,

$$\Pr(T|X) = \begin{cases} 0 & \text{if } T_i = T_j \text{ for any } i \neq j \text{ s.t. } X_i = X_j \\ 2^{-n} & \text{otherwise.} \end{cases} \quad (1)$$

If treatments are assigned using characteristics Z of the units that are correlated with Y (the scientist's confessed experiment at the end of Sec. 5), then $\Pr(T|Y, X) = \Pr(T|X)$ would generally not be acceptable. For example, if treatment assignments are determined by tossing biased coins where the bias favors the first unit in each pair receiving treatment 1 (Z = order of unit in pair), then whether $\Pr(T|Y, X) = \Pr(T|X)$ is generally acceptable depends on the scientific view of the partial correlation between Z and Y given X; if the order "does not seem to have much relevance," then $\Pr(T|X, Y) = \Pr(T|X)$ may be plausible with (1) as the accepted specification for $\Pr(T|Y, X)$. Of course, even if unit order is randomly assigned within pairs,

* Donald B. Rubin is Senior Statistical Research Adviser, Educational Testing Service, Princeton, NJ 08541.

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

one could decide to record its values and use $\Pr(T|X, Z)$ to draw inferences; this is analogous to recording the random numbers used to assign treatments and observing that given them no randomization took place (i.e., $\Pr(T|X, Z) = 1$ for one value of T and 0 for all other values of T). In order to make sensible use of FRTED, we cannot condition on numbers accepted a priori to be unrelated to Y.

Suppose that we wish to consider the hypothesis H_0 that $Y_{i1} = Y_{i2}$ for all i, or any other sharp null hypothesis such that given H_0 and the observed values in Y, all values of Y are known. Under H_0 and accepting specification (1), the difference in observed averages $\bar{y}_d = \sum Y_{i1}(2 - T_i)/n - \sum Y_{i2}(T_i - 1)/n$, or any other statistic, has a conditional distribution given Y and X consisting of 2^n equally likely known values. Because the expectation of $\bar{y}_d$ over this distribution is zero, values of $\bar{y}_d$ far from zero are a priori considered to be more extreme than values near zero. The proportion of possible values as extreme or more extreme than the observed value of $\bar{y}_d$, that is, the significance level of FRTED is not a property solely of the data and the null hypothesis but also of the statistic and the definition of extremeness of the statistic. If the observed value of $\bar{y}_d$ is extreme (e.g., if the significance level is less than 1 in 20), then we must believe that

1. H_0 is false with the result that the treatments have an effect; or
2. $\Pr(T|Y, X) = \Pr(T|X)$ is false with the result that the 2^n values of $\bar{y}_d$ are not a priori equally likely; or
3. An a priori unusual (extreme) event took place.

The physical act of randomization is designed to rule out option 2 and consequently leave us believing either that an a priori unusual event has taken place or that H_0 is false.

I see nothing illogical about the FRTED; it is relevant for those rare situations when a purely confirmatory test of an a priori sharp hypothesis is to be made using an a priori defined statistic having an associated a priori definition of extremeness. On this point, I find myself in total agreement with the following statement of Brillinger, Jones, and Tukey (1978, p. F–1):

> If we are content to ask about the simplest null hypothesis, that our treatment ("seeding") has absolutely no effect in any instance, then the randomization, that must form part of our design, provides the justification for a randomization analysis of our observed result. We need only choose a measure of extremeness of result, and learn enough about the distribution of this result
> - for the observed results held fixed
> - for re-randomizations varying as is permitted by the specification of the designed process of randomization.
>
> If $p\%$ of the values obtained by calculating as if a random re-randomization had been made are more extreme than (or equally extreme as) the value associated with the actual randomization, then $p\%$ is an appropriate measure of the unlikeliness of the actual result.
>
> Under this very tight hypothesis, this calculation is obviously logically sound.

Of course, there are limitations of FRTED of which Fisher was well aware. For example, the null hypothesis that $Y_{i1} = Y_{i2}$ for all i may not be very realistic; when Neyman (1935) criticized the FRTED for Latin Squares, Fisher (1935a) replied:

> [The null hypothesis that "the treatments were wholly without effect"] may be foolish, but that is what the Z-test [FRTED] was designed for, and the only purpose for which it has been used . . . Dr. Neyman thinks that another test would be more important [one for the average treatment effect being zero]. I am not going to argue that point. It may be that the question which Dr. Neyman thinks should be answered is more important than the one I have proposed and attempted to answer . . . I hope he will invent a test of significance, and a method of experimentation, which will be as accurate for questions he considers to be important as the Latin Square is for the purpose for which it was designed.

More complicated questions, such as those arising from the need to adjust for covariates brought to attention after the conduct of the experiment, simultaneously estimate many effects, or generalize results to other units, require statistical tools more flexible than FRTED. Such tools are essentially based on a specification for $\Pr(Y|X, Z)$, where now Y refers to outcome variables in general, X refers to blocking and design variables, and Z refers to covariates. Fisher (1935a) was certainly willing to specify particular distributional forms for data in experiments, and I believe that he was simply advocating such an attack whenever justified in his "astonishing short section on nonparametric tests in the seventh edition of *DE*." This desire to condition on all relevant information is obviously very Bayesian.

I believe (Rubin 1978) that Bayesian thinking, which requires specifications for both $\Pr(T|Y, X, Z)$ and $\Pr(Y|X, Z)$ and draws inferences conditional on all observed values, provides, in principle, the most effective framework for inference about causal effects. Other statisticians view the specification $\Pr(Y|X, Z)$ as something to be avoided in principle: "For crucial comparisons . . . the appropriate role for the classical kind of parametric analysis would seem to be confined to assistance in the selection of the test statistics to be used . . . in a randomization analysis" (Brillinger, Jones, and Tukey 1978, p. F–5). Using the test statistic (in conjunction with the null hypothesis and definition of extremeness) to summarize all scientific knowledge relevant for data analysis seems to be unduly restrictive. Although much care is needed in applying Bayesian principles because of the sensitivity of inference to the specification $\Pr(Y|X, Z)$, the increased flexibility and directness of the resulting inferences make the Bayesian approach scientifically more satisfying.

On this point, perhaps Basu and I are actually in substantial agreement. FRTED cannot adequately handle the full variety of real data problems that practicing statisticians face when drawing causal infer-

ences, and for this reason it might be illogical to try to rely solely on it in practice.

[*Received December 1979.*]

REFERENCES

Brillinger, D.R., Jones, L.V., and Tukey, J.W. (1978), "The Role of Statistics in Weather Resources Management," Report of the Statistical Task Force to the Weather Modification Advisory Board.

Cox, D.R. (1958), *Planning of Experiments*, New York: John Wiley & Sons.

Fisher, R.A. (1935a) (7th ed. 1960), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

——— (1935b), Discussion of "Statistical Problems in Agricultural Experimentation" by J. Neyman, *Journal of the Royal Statistical Society*, II, 2, 154–180.

Neyman, J. (1935), "Statistical Problems in Agricultural Experimentation," *Journal of the Royal Statistical Society*, II, 2, 107–154.

Rubin, D.B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6, 34–58.

Rejoinder

D. BASU

Let me begin by thanking Hinkley, Lane, Lindley, Rubin, and my good friend Kemp for their many interesting comments. I also offer my apologies to them for my inability, because of an eye condition needing surgical treatment, to read the discussions for myself. They were read out to me, and so I may have missed out on some of the many issues raised. I thank Carlos Pereira for his help in putting together this reply.

Rubin wonders about the relevance of the material discussed in Section 4. Let me explain why I challenged the Fisher nonparametric test—the first nonparametric test by many years, as Fisher (*DE* 1960) put it. The logic of the test is essentially the same as that of the paired-comparison test discussed in Section 6. Both are conditional tests of a very extreme kind. In the nonparametric test, the statistic $(|x_1|, |x_2|, \ldots, |x_n|)$ is held fixed; the δ_i's define the reference set. In the randomization test of Section 6, everything but the design outcome is held fixed. Kempthorne and Folks (1971) labeled the nonparametric test as the Fisher randomization test even though, as I explained at the end of Section 5, the δ_i's cannot really be likened to a set of randomization variables. (Kemp disputes this, but then he disputes almost everything I said.) Each of my difficulties with the nonparametric test also persists with the randomization test. For instance, why must we choose $\bar{x}$ (in Sec. 6, $\bar{d}$) as the test criterion and not the median $\tilde{x}$? With $n = 7$ and each $x_i > 0$, the significance level (SL) works out as 1/128 with $\bar{x}$ as the criterion and as 1/16 with $\tilde{x}$ as the criterion. Neither Kemp or Hinkley answers my question. At one place Kemp mumbles about the central limit theorem, but that is hardly relevant for my sample size. Hinkley makes the curious suggestion that the choice of the test criterion is not a statistical problem. How to justify holding $|x_1|, |x_2|, \ldots, |x_n|$ fixed in the nonparametric test? Why not hold $|\bar{x}|$ fixed instead? In the latter case, the SL is either $\frac{1}{2}$ or 1. In Section 6, when the scientist admitted that he had made a one-toss restricted randomization, the statistician declared the experiment to be uniformative because, for every possible outcome of the experiment, the SL is either $\frac{1}{2}$ or 1. Kemp agrees with the statistician. But Kemp, why? Should we not treat such value-loaded terms like significant or informative with greater respect?

When I said that the Fisher randomization test is not logically viable—Rubin calls the characterization "bizarre" and Kemp, in classical debating style, queries my system of logic—I only meant that the logic of the test procedure is not viable. How else can you characterize a test procedure that falls to pieces when confronted with the slightly altered circumstances of a restricted or unequal probability randomization? I am happy to note that Lane and Lindley agree with me on this point.

My working definition of a Bayesian fellow traveler is one who has trouble in understanding a P value as the level of significance attained by the particular data. Rubin, who claims to be a Bayesian, seems to be quite at home with significance testing. George Box is another notable exception to my working definition.

Let us try to make some sense—please Kemp, do not ask me to define *sense*—of the P value of 2^{-15} in Section 6. Suppose each of the 15 subject pairs is indistinguishable to the scientist. Also suppose that the scientist believes that there is no treatment difference. No doubt then the scientist will be surprised if, at the end of the experiment, he finds that each of the 15 treated subjects gains more weight than the corresponding control subjects. The SL of 2^{-15} may be regarded as

© Journal of the American Statistical Association
September 1980, Volume 75, Number 371
Theory and Methods Section

a measure of this element of surprise. It is a probability (measure of doubt) that existed in the mind of the scientist before the experiment and under the assumed circumstances. As Lane observes, this probability does not depend on the nature of prerandomization. But Kemp, the frequentist, refuses to interpret the SL in terms of such nonexistent belief probabilities.

If the scientist cannot truly distinguish between the subjects in each block, then "Nature has done the randomization for us," says Hinkley, and so he cannot understand the point in all the fuss that I am making. But our scientist, like most scientists, can distinguish between the subjects in each block—one subject is heavier, the other one is older and so on. Mother Nature is asking for a helping hand, and so the scientist must randomize! But the scientist can still distinguish between the subjects in each pair. How can we evaluate his surprise index? So we very sternly tell the scientist, "Randomize and close your eyes!" The scientist randomizes, closes his eyes, but still refuses to be greatly surprised in the end. Because, he says, he knew all along that the improved diet is superior to the standard diet. At this point Kemp will perhaps say, "I am surprised that you can write so much on *surprise* without even defining the term."

Many of my esteemed colleagues believe that postrandomization is a useful statistical device. I know my friend Kemp well enough to say that he is not one among them. He agrees with Fisher, Lindley, and me that postrandomization has no place in scientific thinking. But, today, fighting for every inch of the ground, Kemp is trying to prove me wrong even on this issue. Perhaps one can play better poker by wearing a mask, making hand signals instead of using one's vocal chords, and carrying a randomizer hidden in one's pocket. But does Kemp really think that our scientist is engaged in something like a poker game against Mother Nature? Why does he not advise the scientist also to wear a mask?!

I have no objection to prerandomization as such. Indeed, I think that the scientist ought to prerandomize and have the physical act of randomization properly witnessed and notarized. In this crooked world, how else can he avoid the charge of doctoring his own data? In order to make the device a superior cosmetic agent it may be necessary to make the extent of prerandomization sufficiently wide. In Basu (1978) I have mentioned a few noncosmetic uses of the prerandomization device.

Lindley agrees wholeheartedly with my criticisms of the Fisher randomization test. But, disagreeing with me on what he calls a "minor point," he suggests that there may be a place for randomization in a subjective Bayesian theory of statistics. All I know is that L.J. Savage had similar thoughts but he never spelt them out for us. I may have something to say on the Rubin (1978) thesis on another occasion.

Hinkley and Rubin quote from the prefiducial Fisher to dispute me on the randomization test. In the thirties, Fisher knew that the unrestricted, equal probability randomization test closely parallels the traditional test based on the Gaussian law. So Lindley is asking, "Why not use normal theory?" I remember having seen a Fisher quotation (from the prefiducial time) saying that the randomization test provides a logical justification for the parametric tests based on the normal theory. So Kemp is asking us to discard the normal theory and use the randomization logic instead. In Section 21(a) of *DE* (1960), we find Fisher summarily discarding the Kempthorne thesis on experimental designs. Kemp says that no useful purpose can be served by trying to "psychoanalyze" the mind of Fisher. But what purpose does it serve to dismiss much of Fisher's later writings as mere polemics?

I cannot understand what Hinkley is trying to communicate with his comments on the ancillarity of the design outcome. Is it "plain foolish" to regard the design outcome as an experimental constant? Since there are only a finite number of design outcomes, how can one get an "infinitesimal slice" of the sample space by holding the ancillary statistic fixed? As I pointed out, it is the randomization-test argument that rests on an infinitesimal slice of the sample space by holding fixed everything but the design outcome. The Bayesian recommendation is to hold the data fixed and to speculate about the still-variable parameters. When you push the Fisher conditionality argument to the limit, you become a Bayesian.

On the ancillarity issue, Kemp adopts the proverbial Chinese philosophy of seeing no evil. He is in effect saying, "How can there be an ancillary statistic when there is no probabilistic statistical model and, therefore, no parameters?" I have no difficulty in recognizing the 60 parameters $\omega = \{(x_i, y_i): i = 1, 2, \ldots, 30\}$ in the scientist's diet problem—x_i and y_i are, respectively, the would-be treatment and control responses of subject i at the planning stage of the experiment. Let us suppose that the scientist's parameter of interest is $\theta = \bar{x} - \bar{y}$. Consistent with his prior opinion ξ on ω, the scientist has a prior opinion η on θ. After the experiment, the scientist, having observed 15 of the x_i's, and the complementary set of 15 y_i's, must have drastically revised his prior opinion ξ to a new opinion ξ^*. Consistent with ξ^*, the scientist has then an opinion η^* on the parameter of interest θ.

According to DeFinetti, probability, like beauty, exists only in the mind; it is a formal representation of opinion on parameters. The subjective Bayesian thesis on statistics deals with the process of opinion changes in the very limited context of what we may call statistical parameters. The Bayesian thesis appears to me to be coherent and pertinent to the real issues of scientific inference. That the Bayesian paradigm is useful is slowly gaining recognition. Fuller recognition will take time. But by then it will perhaps be time for us to move on to a more useful paradigm.

When it comes to changing one's opinion on a scientific paradigm, the mind of a stubborn scientist—for that matter, the minds of a whole community of trained scientists—certainly does not, perhaps cannot, follow any logic. In his *Scientific Autobiography and Other Papers* (1949, pp. 33–34) Max Planck wrote, "A new scientific truth does not triumph by convincing the opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." It rarely happens that Saul becomes Paul.

[*Received March 1979. Revised March 1980.*]

REFERENCE

Planck, Max (1949), *Scientific Autobiography and Other Papers*, New York: Greenwood Press.

CHAPTER IX

ANCILLARY STATISTICS, PIVOTAL QUANTITIES AND CONFIDENCE STATEMENTS

1. Introduction

The most commonly used expression in Statistics is information; yet, we have no agreement on the definition or usage of this concept. However, in the particular situation where the problem is to predict a future value of a random variable X with a known probability distribution p(.), we all seem to agree that the information on the yet unobserved future value of X may be characterized by the function p(.) itself. And if we have another variable Y such that the conditional distribution $p(.|Y)$ of X, given Y, is also known then, having observed Y, we can claim that the information on X has shifted from p(.) to $p(.|Y)$. [To avoid a multiplicity of notations, we do not distinguish between a random variable X, an observed value of X and a typical point in the sample space of X.] If $p(.|Y)$is the same for all values of Y, then X is stochastically independent of Y. In this case Y is said to have no information on X. And we know how to prove then that X has no information on Y.

A problem of statistical inference is somewhat similar. We have a parameter of interest θ and an observable random variable X. The argument begins with a choice of a model $\{p(.|\theta)\}$ for X, with the interpretation that $p(.|\theta)$ is the conditional probability distribution of X given θ. [In this article we do not consider the case where a nuisance parameter exists.] Typically, the distributions $p(.|\theta)$ are different for different values of the parameter θ. So θ has information on X. The converse proposition that X has information on θ is a reasonable one. However this cannot be proved in probabilistic terms unless we take the Bayesian route and regard the parameter θ as a random variable. For five decades R. A. Fisher tried to set up a non-Bayesian theory of information in the data. This gave rise to a set of novel ideas like sufficiency, likelihood, information function, ancillarity, reference sets, conditionality argument, pivotal quantities and fiducial probability. Many learned discussions by contemporary statisticians and philosophers on Fisher's theory have illuminated as well as clouded the statistical literature. In the hope of dispelling some of the still lingering clouds, I propose to take yet another look at the twin concepts of ancillary statistics and pivotal quantities and the related issue of confidence statements.

2. Ancillary Statistics

A statistic T is ancillary if the probability distribution of T, given θ, is the same for all θ. If the problem is to predict a future value of T, then the parameter θ has no information on T. Is it reasonable then to say that T has no information on θ? Can we generate any information on an unknown State of Nature by, say,

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_29, © Springer Science+Business Media, LLC 2011

rolling a fair coin a number of times? When we are arguing within the framework of a particular model for an observable X, and T = T(X) is ancillary, then the act of recording only the T-value of the sample X seems to be quite as useless a statistical exercise as that of rolling the fair coin. R. A. Fisher must have been guided by a reasoning of the above kind to arrive at the conclusion that an ancillary statistic by itself cannot possibly have any information on θ . To a Bayesian the proposition is almost self-evident, because, in the context of a prior opinion q(.) on θ , whatever q might be, the posterior distribution of θ , given T = t, will work out as

$$\begin{aligned} q(\theta|\ t) &= q(\theta)\ p_T(t|\theta) / \sum_{\theta} q(\theta)\ p_T(t|\theta) \\ &= q(\theta) / \sum_{\theta} q(\theta), \text{ since } p_T(t|\theta) \text{ does not involve } \theta\ , \\ &= q(\theta) \end{aligned}$$

for all θ and t. Since an observation on T in isolation cannot change any opinion on θ , the Bayesian must regard T as uninformative in itself. Strange as it may seem, this hardcore statistical intuition on ancillarity has been challenged by V. P. Godambe (1979a,1980) with his repeated assertions that there are situations where an ancillary statistic by itself can yield a quantum of information on the parameter of interest. This Godambe intuition on ancillary information will be examined in the final section of the essay.

The notion of sufficiency is closely related to the above notion of ancillarity. A statistic S = S(X) is sufficient if the sample X becomes ancillary when it is conditioned by S. In other words, S is sufficient if the conditional distribution p(. | S,θ) of X, given S and θ , depends on (S, θ) only through S. Thus, once we know the S-value of X, the parameter θ has no further information on the sample X. The Fisher intuition that "a sufficient statistic exhausts (summarises) all the relevant information on θ in X" was perhaps based on an unconscious reversal of the roles of θ and X in the previous sentence. Kolmogorov (1942) presented us with the correct Bayesian perspective on sufficiency with the following

Definition : The statistic S is sufficient if, for every prior q(.) for θ , the posterior distribution q(.| X) of θ , given X, depends on X only through S = S(X).

In other words, S is sufficient if, for every prior q, the variables X and θ are conditionally independent given S. Likewise T is ancillary if, for every prior q, the variables T and θ are independent of each other.

Following Neyman-Pearson (1936), we call an X-event A (a measurable subset of the sample space) **similar** (a similar region) if p(A | θ) is a constant, say α , in θ. Despite the fact that the sample space is endowed (by the model) with a multiplicity

of probability distributions, a similar event A with $P(A \mid \theta) \equiv \alpha$, like the impossible and the sure event, seems to be endowed with an absolute (unconditional, that is) probability α. As before, a Bayesian characterization of a similar event is that A is similar if it is independent of θ for every prior q. [It should be understood that we are always arguing in the context of a fixed model.]

The class $\mathcal{A}$ of similar regions is endowed with the following closure properties. It (i) contains the whole space $\mathcal{X}$, (ii) is closed for differences (if $A \subset B$ and both belong to $\mathcal{A}$ then so also does (B - A) and (iii) is a monotone class (closed for monotone limits). It is what Dynkin (1965) calls a **λ-system.** A λ-system that is closed for intersection is a Borel field. The class $\mathcal{A}$ is typically not a Borel field. A statistic T is ancillary if and only if every T-event (a subset of X defined in terms of T(X)) belongs to $\mathcal{A}$. Whenever $\mathcal{A}$ is not a Borel field, we can find two ancillary statistics T_1 and T_2 such that the pair (T_1, T_2) is not ancillary. In such a situation we cannot have an ancillary statistic T_0 that is a maximum ancillary in the sense that every other ancillary statistic is a function of T_0.

Example **1:** Let $\underline{X} = (X_1\ X_2)$ be a pair of i.i.d. random variables whose common distribution on the real line is known excepting for a location parameter θ. Clearly, $T = X_1 - X_2$ is ancillary and, therefore, so also is every function h(T) of T. Contrary to popular impression T is not the maximum ancillary. Indeed, a maximum ancillary can never exist in a situation like this. That the family $\mathcal{A}$ of similar regions is not a Borel field is seen as follows. For any $\underline{X}$-event A let A^c denote its complement and A* the event obtained from A by interchanging X_1, X_2 in its definition. We call A symmetric in the co-ordinates if A* = A. Clearly (A*)* = A and (AB)* = A*B* for all A,B. Since the distribution of $\underline{X} = (X_1, X_2)$ is symmetric in the co-ordinates for each θ, we have $p(A^* \mid \theta) = p(A \mid \theta)$ for all θ and A. Now, let A be an arbitrary T-event that is not symmetric, e.g., $A = \{(x_1, x_2) : x_1 - x_2 > 1\}$ and let B be an arbitrary symmetric event that is not a T-event, e.g., $B = \{(x_1, x_2): x_1^2 + x_2^2 < 1\}$. By definition A and $A^* = \{(x_1, x_2) : x_2 - x_1 > 1\}$ are similar regions. Consider $E = AB \cup A^*B^c$. Since $(A^*B^c)^* = A^{**}B^c = AB^c$, we have $p(A^*B^c \mid \theta) \equiv p(AB^c \mid \theta)$ and so

$$\begin{aligned} p(E \mid \theta) &= p(AB \mid \theta) + p(A^*B^c \mid \theta) \\ &= p(AB \mid \theta) + p(AB^c \mid \theta) \\ &= p(A \mid \theta) \end{aligned}$$

for all θ. That is, E is a similar region even though it is not a T-event.

Observe that in this particular instance
$AE = AB = \{(x_1, x_2) : x_1^2 + x_2^2 < 1,\ x_1 - x_2 > 1\}$ and that this, being a bounded set,

cannot be a similar region for a location parameter family. This is, $\mathcal{A}$ is not closed for intersection. The family of ancillary statistics is very large.

Example 2 : Let X be uniformly distributed over the interval $(\theta, \theta+1)$. With a single observation on X, can we have a nontrivial ancillary statistic? Let [X] be the integer part of X and $\Psi(X) = X - [X]$, the fractional part. It is then easy to verify that $\Psi(X)$ is uniformly distributed over the interval (0, 1) for all θ and so it is an ancillary statistic. In this case $\Psi(X)$ is the maximum ancillary. If $X_1, X_2, \ldots, X_n$ are n independent observations on X, then the statistic $T = (\Psi(X_1), \Psi(X_2), \ldots, \Psi(X_n))$ is ancillary and so also is the difference statistic $D = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$.

3. Ancillary Information

An ancillary statistic T by itself carries no information on the parameter θ, but in conjunction with another statistic Y (which, as we shall see later, may even be ancillary itself) may become fully informative (sufficient). Fisher's controversial theory of recovery of ancillary information (Basu, 1964) is based on a recognition of the fact that an ancillary statistic can carry a lot of potent information. A simple example is given to elucidate the Fisher method.

Example 3 : Consider a threefold multinomial model with cell probabilities $\theta/2$, $(1 - \theta)/2$ and 1/2. Let the observed cell frequencies be n_1, n_2 and $n - n_1 - n_2$. Now the statistic $S = (n_1, n_2)$ is sufficient and $T = n_1 + n_2$ being distributed as Bin(n,1/2), is ancillary. The statistic T by itself does not tell us anything about θ that we did not know already, but it tells us how informative the sample is. For instance, if T is zero then we know that the likelihood function generated by the data is the constant $1/2^n$ and so the sample is void of any information on θ. In a sense it seems clear that larger the observed value of T the more informative the sample is. The maximum likelihood (ML) estimate $\hat{\theta}$ of θ is n_1/T if $T \neq 0$ and is undefined if $T = 0$. The ML estimator $\hat{\theta}$, assuming that we have suitably defined it for the case $T = 0$, is not a sufficient statistic and so, according to Fisher, there will be a loss of information if we evaluate $\hat{\theta}$ in terms of its (marginal) distribution $p_{\hat{\theta}}(. \mid \theta)$. The statistic T is an ancillary complement to $\hat{\theta}$ in the sense that the pair $(\hat{\theta}, T)$, being equivalent to $S = (n_1, n_2)$, is fully informative. This is the kind of situation where Fisher (1935c) wants us to evaluate the ML estimator $\hat{\theta}$ by conditioning it by the ancillary statistic T. The conditional distribution of n_1, given T, θ, is Bin (T, θ), when $T \neq 0$. Thus $E(\hat{\theta} \mid T, \theta) = \theta$, $V(\hat{\theta} \mid T, \theta) = \theta(1 - \theta)/T$ and the Fisher information content of $\hat{\theta}$, conditionally on T, is $T/\theta(1 - \theta)$. Fisher would regard $\hat{\theta}$ to be an unbiased estimate of θ with small variance and large information if T is large. The smaller the observed value of T, the less informative $\hat{\theta}$ will be as an estimate of θ. In the extreme case when T = 0, there is no information at all.

This example brings out the Fisher dilemma. He recognized that as a rule not all samples are equally informative and so a sample space analysis of the data may not be quite appropriate. He realized that the information in the data obtained is fully summarized in the corresponding likelihood function but he did not quite know how to analyze the likelihood function in a non-Bayesian fashion. [In Fisher (1956) we do find some half-hearted advice on how to analyze the likelihood function. This has been fully scrutinized (and rejected) by the author in Basu (1973).] So he had to go for a compromise solution : Analyze the data in sample space terms, but suitably restrict the sample space by clustering together all sample points that are, in some qualitative sense, as informative as the sample actually observed.

In the present case the Fisher solution seems pretty attractive. With n =15, the two samples (1, 2, 12) and (12, 2, 1) are qualitatively quite different. It seems reasonable enough to say that the data (1, 2, 12) should be interpreted as 1 success in a succession of 3 Bernoulli trials with θ as the probability of success. Similarly the data (12, 2, 1) should be looked upon as 12 successes in 14 Bernoulli trials. In the first case 1/3 is an unbiased estimate of θ with variance $\theta(1 - \theta)/3$, whereas in the second case 12/14 is an unbiased estimate of θ with variance $\theta(1 - \theta)/14$.

Admittedly, it is not easy to make an unconditional sample space analysis of the data $X = (n_1, n_2, n_3)$. Our heart reaches out for the ML estimator $n_1/(n_1+n_2)$ but we worry about it being undefined when $T = n_1 + n_2 = 0$. We do not relish the idea of figuring out the bias, variance and the information content of the estimator. We recognize many unbiased estimators like $2n_1/n$, $1 - 2n_2/n$ and $1/2 + (n_1 - n_2)/n$ but we like none of them as we recognize that they can take values outside the parameter space (0, 1). The minimum sufficient statistic $S = (n_1, n_2)$ is not complete, there does not exist a minimum variance unbiased estimator of θ, no unbiased estimator can be admissible and so on. All the difficulties seem to be stemming from the fact that T is ancillary. So why not cut the Gordian knot by holding fixed the observed value of $T = n_1 + n_2$ just as we do for the sample size $n = n_1 + n_2 + n_3$?

The sample size analogy for an ancillary statistic comes from Fisher (1935c) who wrote : "It is shown that some, or sometimes all of the lost information may be recovered by calculating what I call ancillary statistics, which themselves tell us nothing about the value of the parameter, but instead, tell us how good an estimate we have made of it. Their function is, in fact, analogous to the part which the size of the sample is always expected to play, in telling us what reliance to place on the result".

Example 4 : An experiment consists of n Bernoulli trials where n = 5 or 100 depending on the flip of a fair coin. The sample is X = (Y, n), where Y is the number

of successes in n trials. The ML estimate of θ is $\hat{\theta} = Y/n$. The sample size n is an ancillary statistic now. How to analyze the data X = (3, 5)? Clearly, the data is qualitatively very different from the possible data (60,100), although the ML estimate is the same in either case. As Fisher would have put it, the ML estimate is insufficient (in this case) to identify the full information content of the data, namely, the likelihood function. In terms of the conditional model $p(.|n,\theta)$ for the data, the ML estimator $\hat{\theta}$ is fully sufficient.

Suppose we slightly alter the experiment in Example 4 by determining the sample size n, not by the flip of a fair coin, but by the outcome of the first trial – the sample size is 5 or 100 according as the first trial yields a success or a failure. Now, n is no longer an ancillary statistic even though the quality of the sample depends very much on n. The likelihood function generated by an observed sequence of n successes and failures is $\theta^{y}(1-\theta)^{n-y}$, where y is the number of successes in the sequence. Therefore, (y, n) is the minimum sufficient statistic and $\hat{\theta} = y/n$ is still the ML estimate of θ. To evaluate $\hat{\theta}$ in terms of its full (unconditional) model $p_{\theta}(.|\theta)$ will clearly entail a substantial loss of information. The recovery of information argument does not work here in view of the fact that we do not have an ancillary complement to $\hat{\theta}$.

In Basu (1964) there are other illustrations of how the argument can fail. For instance, there can be a multiplicity of ancillary statistics with no clear cut choice for one to condition the data with. There can also be situations where conditioning X with an ancillary statistic T reduces it to a degenerate random variable with the point of degeneracy depending on T and θ. In Example 2 we have such a situation because the conditional distribution of X, given $\psi(X)$ and θ, is degenerate. In Section 6 we discuss a similar situation. A traditional sample space analysis of data with a degenerate sample space is unheard of.

With his recovery of ancillary information argument, Fisher was seeking for a via media between the Bayesian and the Neyman-Pearson way. As a compromise solution it failed as most compromises do. From the point of view of history, it is more important to recognize what Fisher was attempting to do than the fact of his failure to do so. Clearly, he was trying to cut down the sample space to size. Why? Because he realized that not all sample points in the same sample space are equally informative. With a data of poor quality in hand, e.g. X = (3, 5) in Example 4, it makes little sense to derive some comfort from the thought of a might have been excellent data, e.g., X = (60, 100) or any other sample with n = 100. In this situation, why not evaluate the data (3, 5) in the context of the reduced (conditional) sample space (reference set) consisting of only the six points (0, 5), (1, 5), ... ,(5,.5)?

Fisher (1936) wrote : "The function which this ancillary information is required to perform is to distinguish among samples of the same size those from which more or less accurate estimates can be made, or, in general, to distinguish among samples having different likelihood functions, even though they may be maximized at the same value". Clearly, Fisher though that two samples, even though they may be of the same size, may be different in their information content in that a more (or less) accurate estimate of the parameter can be made from the one than from the other. It is not at all clear why he thought that his ancillary statistic can "distinguish among samples with different likelihood functions". The minimum sufficient statistic S is the one that distinguishes between samples with different likelihood functions. [Two likelihood functions are equivalent if they differ only by a multiplicative constant. The minimum sufficient statistic S partitions the sample space into sets of points with equivalent likelihood functions.] The M.L. estimator $\hat{\theta}$ is a function of S. In Example 4, the ancillary statistic T complements $\hat{\theta}$, that is, $(\hat{\theta}, T)$ is a one-one function of S. Thus, $(\hat{\theta}, T)$ can distinguish among samples with different likelihood functions. If $\hat{\theta}(x_1) = \hat{\theta}(x_2)$, then x_1 and x_2 generate different likelihood functions if and only if $T(x_1) \neq T(x_2)$. Of course, T by itself cannot distinguish among samples with different likelihood functions, for, if it could, it would have been the minimum sufficient statistic.

In the thirties, when Fisher came up with the recovery of information argument, he was still trying to justify the method of maximum likelihood in some sample space terms. In the mid-fifties, however, we find Fisher (1956, pp 66-73) recognize particular situations where he thought that the data ought to be interpreted only in terms of the particular likelihood function generated by it.

Apart from recovering ancillary information, Fisher found two other uses for the conditionality argument. In the next section we briefly discuss how he extended the scope of his fiducial argument by conditioning pivotal quantities that were not based on what he called "sufficient estimates". Fisher's conditioning method for elimination of nuisance parameters has been discussed by me at some length in Basu (1977).

4. Pivotal Quantities

The notion of a pivotal quantity, perhaps the most innovative idea that came from Sir Ronald, is an extension of the concept of an ancillary statistic. A quantity $Q = Q(\theta, X)$ is a measurable function of the parameter $\theta \in \Theta$ and the sample $X \in \mathcal{X}$, that is, Q is a map of the product space $\Omega = \Theta \times \mathcal{X}$ into a range space R that is usually taken to be the real line. A statistic is a quantity that is constant in θ.

Definition : $Q : \Omega \to R$ is a **pivotal quantity** in short, a **p-quantity**, if the conditional distribution of $Q(\theta, X)$, given θ, is the same for all θ.

Thus, an ancillary statistic is the extreme case of a p-quantity that is constant in θ. If $Q : \Omega \to R$ is a p-quantity then so also is $h(Q)$ for every measurable function h on R. In the location parameter models of Examples 1 and 2, typical examples of p-quantities are $X_1 - \theta$, $\bar{X} - \theta$, $\tilde{X} - \theta$, etc., where $\tilde{X}$ is an equivariant statistic like the mean, median or the mid-range. A more complex example of a p-quantity involving a location parameter is as follows :

Example 5 : Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common c.d.f. $F(x - \mu)$ where about F we only know that it is continuous at the origin and that $F(0) = 1/2$. With μ as the parameter of interest and $X = (X_1, X_2, \ldots, X_n)$, define the quantity $Q(\mu, X)$ as the number of i's for which $X_i - \mu > 0$, $i = 1,2, \ldots, n$. Then Q is a p-quantity since the conditional distribution of Q, given $\theta = (\mu, F)$, is Bin(n, 1/2).

A Bayesian definition of a p-quantity would run along the following lines:

Definition : $Q(\theta, X)$ is a pivotal quantity with respect to a model $\{p(.|\theta) : \theta \in \Theta\}$ for the sample X, if, for every prior distribution q of θ, the quantity Q and the parameter θ are stochastically independent.

Equivalently, Q is a p-quantity if its predictive distribution depends on the prior q and the model $\{p(.|\theta)$ only through the later.

With n i.i.d. observations $X_1, X_2, \ldots, X_n$ on $N(\theta, 1)$, Fisher derived the fiducial distribution of θ as $N(\bar{X}, 1/n)$ by using the p-quantity $\bar{X} - \theta$ as a pivot. It is not the purpose of this article to examine the fiducial argument once again. During the past five decades the argument has been thoroughly examined many times and has been declared, with a few exceptions, as logically invalid. We propose to examine here the confidence statement argument that Neyman-Pearson in the early thirties, synthesized from the fiducial argument. But before we get into that it will be useful to examine why Fisher chose the particular p-quantity $\bar{X} - \theta$ as the pivot.

Fisher regarded the p-quantity $\bar{X} - \theta$ as the correct pivot for the fiducial argument because i) $\bar{X}$ is a natural estimate of θ, ii) $\bar{X}$ is a (minimum) sufficient statistic and iii) the range of variation of $\bar{X}$ is the same as that of the parameter θ, namely, the whole real line. Fisher's fiducial logic would not permit us to use $X_1 - \theta$ as a pivot because X_1 is not sufficient.

The weak p-quantity $X_1 - \theta$ can, however, be made strong by proper conditioning! The difference statistic $D = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$ is an ancillary complement of X_1. Consider the conditional distribution of $X_1 - \theta$ for fixed D and θ. Since $X_1 - \theta = \bar{X} - \theta + (X_1 - \bar{X})$, $X_1 - \bar{X}$ is a function of D, and $\bar{X}$ is independent of D for fixed θ, it follows that $X_1 - \theta$, given D and θ, is distributed as $N(X_1 - \bar{X}, 1/n)$. Thus, $X_1 - \theta$ remains a p-quantity even when it is conditioned by D. If the fiducial argument is based on the conditioned p-quantity $X_1 - \theta | D$ then we arrive at the correct fiducial distribution $N(\bar{X}, 1/n)$ for θ.

Fisher generalized the above conditional pivotal quantity argument to the case of a location parameter model as follows. Let $X = (X_1, X_2, \ldots, X_n)$ be the sample with density function $f(x_1 - \theta, x_2 - \theta, \ldots, x_n - \theta)$. There are many p-quantities like $X_1 - \theta$, $\bar{X} - \theta$, $\tilde{X} - \theta$, etc., where $\tilde{X}$ is any equivariant estimator of θ. On which one of these shall we base the fiducial argument? It does not matter as long as we condition the chosen p-quantity by the difference statistic! Suppose we choose the sickly p-quantity $X_1 - \theta$, condition it by D and θ, thus arriving at a density function, say, p(t). Invoking the fiducial argument we then arrive at the fiducial density function $p(X_1 - t)$ for the parameter θ. But had we started off with a stronger looking p-quantity $\tilde{X} - \theta$, where $\tilde{X}$ is an equivariant estimate of θ like the mean or the median, then, in view of the fact that $\tilde{X} - X_1$ is a function of D, the conditional density function for $\tilde{X} - \theta$, given D and θ, would have been $p(t - \tilde{X} + X_1)$. Therefore, the fiducial distribution of θ, with $\tilde{X} - \theta$ held as the pivot, would also have worked out as $p(X_1 - t)$.

Many of our contemporary statisticians — Pitman, Barnard, Kempthorne, Fraser, to name only a few — have been greatly (and very diversely) influenced by the above conditional pivot argument of Fisher. For one thing, the extensive study of invariance as a data reduction principle originated in the fiducial distribution $p(X_1 - t)$ for the location parameter θ. Note that the mean

$$\begin{aligned}
\hat{\theta} &= \int tp(X_1 - t)\, dt \\
&= \int (X_1 - t)\, p(t)\, dt \\
&= X_1 - \int tp(t)\, dt \\
&= X_1 - E(X_1 - \theta | D, \theta) \\
&= X_1 - E(X_1 | D, \theta = 0)
\end{aligned}$$

of the fiducial distribution is the Pitman estimator (the best equivariant estimator with squared error loss function) of θ. Similarly, the median of the fiducial distribution is the best equivariant estimator when the absolute error is taken as the

loss function. If the proof of the pudding is in the eating, then one may argue that Fisher's fiducial argument is plausible at least in the case of a location parameter.

I regard the following proposition as an empirically established statistical metatheorem:

No inferential argument in statistics has anything going for it unless a sensible Bayesian interpretation can be found for it.

It so happens that the fiducial distribution $p(X_1 - t)$ for the location parameter θ may be interpreted as the Bayes posterior with the (improper) uniform prior over the real line. So it is possible to condition the sample X all the way down to its observed value and derive a distribution for θ directly from the likelihood function. However, the uniform prior over the unbounded real line can hardly be regarded as a sensible representation of anyone's ignorance about the parameter θ.

Sir Ronald is no longer with us, but his pivotal qunatities are very much alive. As we shall see in the next section, all confidence statements are based on p-quantities. Fisher tried hard but failed to establish a coherent theory for the choice of an appropriate p-quantity. Currently there appears to be no law and order regarding the choice of the pivot. Often a sizeable part of the data is ignored to arrive at a particular pivot. Stein's (1945) classical work on fixed width confidence interval is an example of this kind. Sometimes post-randomization variables are deliberately introduced in the data so as to create a p-quantity in terms of the extended data. For instance, if the data generated by a sequence of n Bernoulli trials be enhanced by a randomization variable uniformly distributed over (0, 1), then we can construct a 95% confidence interval for the probability parameter p. Finally, there is the curious (Fisher inspired) method of holding a substantial part of the data as fixed and then recognizing that a quantity $Q(\theta, X)$ becomes a p-quantity in terms of the restricted reference set. Godambe and Thompson (1971) and Seheult (1980) are only two of many such licentious use of the fiducial argument that Fisher himself thought to be of rather limited coverage.

5. Confidence Statements

A quantity Q was defined earlier as a measurable function of (θ, X). A measurable subset of $\Omega=\Theta \times \mathcal{X}$ will be called a **quantal.** Let $Q:\Omega \to R$ be an arbitrary pivotal quantity(p-quantity) and let B be a measurable subset of R. Consider the quantal $E = \{(\theta, X) : Q(\theta, X) \in B\}$. For each $\theta \in \Theta$, let $E^{\theta} = \{X: (\theta, X) \in E\}$ denote the θ-section of E. For a given θ, the statements $Q(\theta, X) \in B$ and $X \in E^{\theta}$ are identical. Since Q

is a p-quantity, the conditional probability of the former statement, given θ , is a constant in θ, and therefore, so also is Prob($X \in E^{\theta} | \theta$). This motivates the following

Definition : In the context of a model $\{p(.|\theta)\}$ for the sample X, a subset E of $\Omega = \Theta \times \mathcal{X}$ is called a **p-quantal** if $p(E^{\theta} | \theta)$ is a constant in θ . The constant value of $p(E^{\theta} | \theta)$ is called the size of the p-quantal E.

The empty set and the whole space Ω are trivial examples of p-quantals of size 0 and 1 respectively. If A is a similar region of size α then $\Theta \times A$ is a p-quantal of size α . Like the family $\mathcal{A}$ of similar regions, the family $\mathcal{E}$ of p-quantals is a λ -system of sets. The indicator function of a p-quantal is a p-quantity. A function $Q(\theta, X)$ is a p-quantity if and only if every subset of $\Theta \times \mathcal{X}$ defined in terms of Q is a p-quantal. Analogous to our Bayesian definition of a p-quantity, the notion of a p-quantal may be redefined as

Definition : The set $E \subset \Omega$ is a p-quantal if, irrespective of the scientist's prior distribution (or belief) on Θ , the event E is independent of the random variable θ .

Equivalently, E is a p-quantal if its probability is well defined in terms of the model $\{p(.|\theta)\}$ alone.

R.A. Fisher's fiducial argument was severely restrictive in that only a few simplistic statistical models could cope with his stringent requirements for the right pivotal quantity. Jerzy Neyman and E.S. Pearson rejected the fiducial logic but they nevertheless accepted a part of it and generalized it to the limit. Neyman-Pearson's confidence statement argument is based not on the limited stock of the Fisherian p-quantities but on the plentiful supply of p-quantals. Any p-quantal E can be the basis of a confidence statement as described below.

Let $E_X = \{\theta : (\theta, X) \in E\}$ denote the X-section of a p-quantal E. Clearly, the three statements $(\theta, X) \in E$, $X \in E^{\theta}$ and $\theta \in E_X$ are logically equivalent. If we regard E_X as a random set determined by the random variable X, then, prior to the observation of X, the probability of $\theta \in E_X$ is well defined and is independent of any prior opinion (or lack of it) that the scientist may have on the parameter θ . Suppose the p-quantal E is of size 0.95. Then, prior to the observation of X, the scientist is 95% sure (ordinary probability) that the true θ will belong to the set E_X that is going to be determined by the observance of X. Once X is observed and the particular E_X determined, then is it still reasonable for the scientist to assert that he or she is still 95% sure that $\theta \in E_X$? Fisher would have answered the question with a firm negative if he found that the p-quantal E has not been defined in the right manner in terms of the right pivotal quantity. Neyman-Pearson's

theory of confidence statements suffers from no such inhibition. Any p-quantal of size 0.95 is a generator of a 95% confidence set estimator $\{ E_X : X \in \mathcal{X}\}$ of θ. The converse proposition that every confidence set estimator corresponds to a p-quantal is easily seen to be also true.

If we consider the hypothetical population of a sequence of observations $X_1, X_2, \ldots$ on X with θ fixed at its true value, then is it not correct to say that 95% of the sets $E_{X_1}, E_{X_2}, \ldots$ will cover θ and only 5% will not?

The frequency interpretation of the confidence statement $\theta \in E_X$ that is implicit in the above hypothetical question is the cornerstone of the Neyman-Pearson argument. Fisher always maintained that the argument was a logical error. I vividly recall an occasion (Winter of 1955, Indian Statistical Institute, Calcutta) when Professor Fisher bluntly asserted that a confidence interval, unless it coincides with a fiducial interval, cannot be interpreted in frequency probability terms. At the end of the Fisher seminar, Professor Mahalanobis, who was chairing the session, invited me to comment on the proposition. So, shaking in my shoes, I rose to defend Professor Neyman and gave the standard frequency interpretation of confidence statements. Professor Fisher summarily dismissed my explanation as yet another example of the "acceptance test" type argument.

It is important for us to understand the Fisher point of view. According to Fisher (1956, p. 77) the population (an hypothetical sequence of repeated trials with certain elements held fixed) to which a certain inferential statement is referred to for probabilistic interpretation is a figment of the mind. Only in some "acceptance test" type situations, Fisher would concede that the population (reference set) is well-defined. In problems of scientific inference it should be recognized that the data can be differently interpreted in terms of different reference sets. Of course, Fisher alone knew how to artfully choose the reference set to suit an individual scientific problem!

Let us go back to a situation that we have already discussed in Section 3. Suppose with a sample $X = (X_1, X_2, \ldots, X_n)$ of n i.i.d. $N(\theta, 1)$'s, a misguided scientist chooses to make a confidence statement based on the first observation, X_1 alone. Since the p-quantity $X_1 - \theta \sim N(0,1)$, the 95% confidence interval is $I = (X_1 - 1.96, X_1 + 1.96)$. Following Neyman the scientist will claim that the random interval I covers the true θ with 95% probability. But following Fisher the scientist may recognize quite a different probability for I covering θ. Since X_1 is not a sufficient statistic, the scientist will have to look for an ancillary complement to X_1. Suppose $d = \bar{X} - X_1$ is chosen as the ancillary complement to X_1. [Note

that $\bar{X} - X_1$ is an ancillary statistic and that $(X_1, \bar{X} - X_1)$ is a sufficient statistic.] The conditional probability

$$\Pr(|X_1 - \theta| < 1.96 | d, \theta)$$
$$= \Pr(|\bar{X} - \theta - d| < 1.96 | d, \theta)$$
$$= \Pr[|N(d, 1/n)| < 1.96]$$

and this is the same as Fisher's fiducial probability that θ is in I. That the choice of the reference set (or the conditioning statistic) can dramatically affect the confidence co-efficient is evident from this example.

Most of my statistical colleagues can never cease to admire the sheer elegance and simplicity of Neyman-Pearson's confidence statement argument. And then it is so wonderfully easy to construct a p-quantal E. Choose and fix a number α in (0, 1); for each $\theta \in \Theta$, choose and fix a subset E^θ of $\mathcal{X}$ such that $p(E^\theta | \theta) = \alpha$; and then define E as

$$E = \{(\theta, X) : \theta \in \Theta, X \in E^\theta\}.$$

By construction E is a p-quantal of size α and so the family $\{E_X : X \in \mathcal{X}\}$ of X-sections of E is a 100α % confidence set estimator of θ. With E_X as the confidence set corresponding to the observed sample X, can any **evidential meaning** be attached to the assertion $\theta \in E_X$? Suppose on the basis of sample X one can construct a 95% confidence interval estimator for the parameter θ, then does it mean that (the random variable) X has **information** on θ in some sense?

Anyone who fails to answer both the questions quickly and firmly in the negative is invited to take a look at the following simple example.

Example 6 : The parameter θ lies somewhere in the unit interval (0, 1) and the sample X is a pure randomization variable having the θ-free uniform distribution over (0, 1). Surely X has no information on θ. No evidential meaning can be attributed to any inference about θ based on X. That there is a plentiful supply of 95% confidence intervals for θ is seen as follows. Choose and fix any subset B of (0,1) and then define the quantal E as the union of the two sets $B \times (0, .95)$ and $B^c \times (.05, 1)$. For each θ the section E^θ is either (0, .95) or (.05, 1), and so E is a p-quantal of size 0.95. The 95% confidence intervals $\{E_X\}$ based on the p-quantal E are

$$E_X = \begin{cases} B & \text{if } 0 < X \leq .05 \\ (0, 1) & \text{if } .05 < X < .95 \\ B^c & \text{if } .95 \leq X < 1. \end{cases}$$

Of course it does not make any sense to say that 95% confidence can be placed on the statement $\theta \in E_X$ irrespective of what X turns out to be.

As I said in the last paragraph of the previous section, the statistical literature is full of many singular kinds of confidence statements. But has anyone ever dared to base confidence statements on no data, that is, on an uninformative part of the data? Surprisingly, the answer is, yes. In the next section we briefly consider the Godambe (1979a,1980) contention that the label part of the survey data, despite being an ancillary statistic, is informative by itself.

6. Ancillarity in Survey Sampling

In current survey literature a lot of confusion and controversy surround the notion of the **label set,** that is, the set of names or label identities of the surveyed units. Let us denote the population to be surveyed as {1, 2, . . . , N} of unit labels. The universal parameter is $\theta = (Y_1, Y_2, \ldots, Y_N)$, where Y_j is an unknown characteristic of unit j. The survey design (sampling plan) selects a subset s of the population lebels {1, 2, . . ., N} with a known selection probability p(s). The survey fieldwork determines the set $y = \{ Y_i : i \in s \}$ of Y-values of the units in the label set s. We write x = (s, y) to denote the sample generated by the survey.

For a typical survey design the probability p(s) of the label set s does not depend on the parameter θ , and so s is an ancillary statistic. Can we then discard s and marginalize the data to the set of observed Y-values? Since s is ancillary, should we not condition the data x by holding s fixed? If we factor the likelihood function $L(\theta)$ as

$$L(\theta) = Pr(s, y|\theta) = p(s)Pr(y| s, \theta)$$

and discard the θ -free factor p(s) from the likelihood, then it becomes clear [see Basu (1969) for more on this] that the sampling plan is not a determinant of the likelihood. Invoking the **Likelihood Principle** should we not declare then that at the data analysis stage we need not concern ourselves with the details of the particular survey sampling plan? These are some of the hotly debated issues in survey theory.

It is useful to note the close similarity between the survey setup and Example 2, where the sample X is uniformity distributed over the interval $(\theta, \theta + 1)$ and $-\infty < \theta < \infty$ is the parameter.The label set s of the survey sample x corresponds to the fractional part $\varphi(X)$ of the sample X in Example 2, both being "very large" ancillary statistics in the following sense. The conditional distribution of the sample X, given the ancillary statistic $\varphi(X)$ and the parameter θ is degenerate in Example 2, that

is, X is a function of $\varphi(X)$ and θ . In precisely the same sense, the sample x is a function of the label set s and the universal parameter θ of our survey setup.

The fact that the survey sample x has a degenerate conditional distribution, given s and θ , has the simple consequence that no unbiasedly estimable parametric function (unless it is a constant) can have a uniformly minimum variance unbiased estimator (UMVUE). This is seen as follows.

Let T be an unbiased estimator of $g(\theta)$. Choose and fix a particular parameter value θ_0, and consider the estimator

$$T_0 = T - E(T|\, s, \theta_0) + g(\theta_0).$$

The second term on the right hand side, being a function of the ancillary s, has a constant (θ -free) mean. Considering the particular case $\theta = \theta_0$, it is then clear that the constant mean is $g(\theta_0)$. Hence T_0 is an unbiased estimator of $g(\theta)$. Since the distribution of T, given s and θ, is degenerate, the first two terms on the right hand side are the same when $\theta = \theta_0$. We have thus established that for each unbiasedly estimable parametric function $g(\theta)$ and for any prefixed parameter value θ_0, there exists an unbiased estimator T_0 for $g(\theta)$ with zero variance at $\theta = \theta_0$. As in Example 2, we cannot talk of UMVUE's in survey theory.

In Example 2, the likelihood function is flat over the interval (X - 1, X) and is zero outside. [With n observations on X, the likelihood is flat over the interval (M - 1, m), where m and M are, respectively, the minimum and the maximum sample.] Exactly the same thing is true for the survey setup. It is rather curious that the flat likelihood for the survey parameter is sometimes characterised as "uninformative". I have not met anyone yet who would regard the equally flat likelihood in Example 2 as uninformative.

Survey theory is full of all kinds of confusing ideas. We end this essay with a good look at a mystifying Godambe proposition that was stated at the end of the previous section. It will be useful to consider a simplified version of the Godambe example first.

Example 7 : The population consists of 100 individuals labeled as 1,2, ..., 100. It is known that only one of the 100 individuals is black, the rest being all white. It is also known that the black individual is either 1 or 2. Writing 1 for black and 0 for white, the universal parameter then takes one of the two values $\theta_1 = (1, 0, 0, 0, \ldots, 0)$ and $\theta_2 = (0, 1, 0, 0, \ldots, 0)$. An individual is selected at random (equal probabilities) and the sample is x = (s, y), where s and y are, respectively, the label index and the colour value of the chosen individual. In this case both s and y are ancillary

statistics, an example of how two ancillaries can jointly be sufficient. As we noted earlier, the statistic **y** gets fully determined in terms of the parameter θ and the label **s**. In other words, the ancillary statistic **y** may be represented as a p-quantity $Q(\theta, s)$. Since $Pr(y = 0 \mid \theta) = 0.99$ for both θ, the set $E=\{(\theta, s) : Q(\theta, s) = 0\}$ is a p-quantal of size 0.99. Defining E_s as the **s**-section of E,we then have in $\{E_s\}$ a 99% confidence set estimator of θ in the sense of Neyman-Pearson. Note that $E_1 = \{\theta_2\}$, $E_2 = \{\theta_1\}$ and $E_s = \{\theta_1, \theta_2\}$ for all other **s**. Does it make any sense to say that when **s** = 1, we should be 99% confident that $\theta = \theta_2$? How confident should we be in the proposition $\theta \in E_s$ when **s** = 3 ?!

Example 8 (V. P. Godambe) : The population consists of four individuals 1, 2, 3, 4, the universal parameter θ is either

$$\theta_1 = (-1, 1, -1, 1) \text{ or } \theta_2 = (-1, 1, 1, -1).$$

With a simple random sample of size 2, let $\mathbf{s} = (s_1, s_2)$ be the label set and let $\mathbf{y} = (y_1, y_2)$ be the sample Y-values. It is then easy to see that $T = |y_1 + y_2|$ is an ancillary statistic taking the two values 0 and 2 with probabilities 4/6 and 2/6 respectively. As in the previous example, we represent the ancillary statistic T as a pivotal quantity

$$Q(\theta, \mathbf{s}) = \left| \sum_{i \in s} Y_i \right| .$$

The set $E = \{(\theta, \mathbf{s}) : Q(\theta, \mathbf{s}) = 0\}$ is then a p-quantal of size 4/6 and the family $\{E_s\}$ of **s**-sections of E constitute a confidence set estimator of θ with a confidence level of 4/6. Note that $E_{(1,3)} = \{\theta_2\}$. According to Godambe, the partial data **s**= (1, 3), despite being ancillary in the sense of Fisher, is informative (in itself) about the parameter in the sense that it makes the parameter value θ_2 "more plausible" than the value θ_1.

Example 6 of the previous section and the two examples of this section are really the same. They all demonstrate how it is possible to construct sizeable confidence statements with meaningless data. We must recognize that the confidence statement argument when presented in its classical p-quantal form is a mistake. Sir Ronald was certainly more right on this controversial issue than Professor Neyman.

Like a blind man in a dark room groping for a black cat that is perhaps not there, we statisticians are still seeking for the true meaning of statistical information. This article was written in the hope of sharing with my colleagues my imperfect intuition on (and limited understanding of) the subject. If in doing so I have hurt anyone's feelings, I am sorry.

Internat. J. Math. & Math. Sci.
Vol. 4 No. 3 (1981) 571-582

A NOTE ON SUFFICIENCY IN COHERENT MODELS

D. BASU

Department of Statistics
Florida State University
Tallahassee, Florida 32306
and

S.C. CHENG

Math Department, Creighton University
Omaha, Nebraska 68178

(Paper Received February 18, 1980)

ABSTRACT. Partly of an expository nature, this article brings together a number of notions related to sufficiency in an abstract measure theoretic setting. The notion of a coherent statistical model, as introduced by Hasegawa and Perlman [6], is studied in some details. A few results are generalized and their earlier proofs simplified. Among other things, it is shown that a coherent model can be connected in the sense of Basu [2] if and only if no splitting set (Koehn and Thomas, [7]) exists.

KEY WORDS AND PHRASES. Sufficiency, coherent model, splitting set.

1980 MATHEMATICS SUBJECT CLASSIFICATION CODES. Primary 62B05.

1. INTRODUCTION.

This article is partly of an expository nature and is written mainly for its pedagogical interest. Given a mathematical model $(X, \mathcal{A}, P)$ for a statistical experiment, it is of some theoretical interest to inquire whether the family of sufficient sub-σ-fields (subfields) C has a minimum element in some sense. It is now known that if the model is coherent in the sense of Hasegawa and Perlman [6] then such a minimum sufficient subfield exists. The case of a coherent model is

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_30,

studied in some details in this article. Among other things it is demonstrated that for a coherent model the notion of connectedness (Basu, [2]) and that of the nonexistence of a splitting set (Koehn and Thomas, [7]) coincide.

2. NOTATION AND DEFINITIONS.

The basic statistical model is denoted by (X, A, P), where X is the sample space, A a σ-field of subsets of X, and $P = \{P_\theta: \theta \in \Theta\}$ a family of probability measures on A. By an A-measurable function we mean a measurable map of (X, A) into $(R_1, \mathcal{B}_1)$. Any sub-σ-field C of A will be referred to as a subfield. The function f is C-measurable if $f^{-1}\mathcal{B}_1$ is contained in C. Given a family $\{A_t: t \in T\}$ of measurable sets, we write $\sigma\{A_t: t \in T\}$ for the subfield generated by the family of sets. Likewise, $\sigma\{f_t: t \in T\}$ will stand for the smallest subfield C such that each f_t is C-measurable. By $C \vee \mathcal{D}$ we denote $\sigma(C \cup \mathcal{D})$, that is, the smallest subfield containing both C and $\mathcal{D}$.

A set N in A is P-null if $P_\theta(N) = 0$ for all $\theta \in \Theta$. Let N denote the class of all P-null sets. For $A, B \in A$, the statement "$A = B$ $[P]$" means that the symmetric difference $A \Delta B$ is P-null. Similarly, for any two A-measurable functions f and g, we write $f = g$ $[P]$ to indicate that $\{x: f(x) \neq g(x)\} \in N$.

The completion $\overline{C}$ of a subfield C is defined as $\overline{C} = C \vee \sigma(N)$. Accordingly, a subfield C is called complete if $N \subset C$. Let C and $\mathcal{D}$ be two subfields. We write $C \subset \mathcal{D}[P]$ if $C \subset \overline{\mathcal{D}}$, that is, corresponding to each set C in C, there exists a set $D \in \mathcal{D}$ such that $C = D[P]$. If $C \subset \mathcal{D}[P]$ and $\mathcal{D} \subset C[P]$, then $C = \mathcal{D}[P]$. By a P-essentially $\overline{C}$-measurable function we mean a function f such that $f^{-1}\mathcal{B}_1 \subset C$ $[P]$, i.e., f is $\overline{C}$-measurable. This is also equivalent to the statement that there exists a C-measurable function g such that $f = g$ $[P]$. An A-measurable function is P-integrable if $\int_X |f| \, d P_\theta < \infty$ for all $\theta \in \Theta$.

DEFINITION 1. (Halmos and Savage, [5]). The statistical model (X, A, P) is called dominated if every probability measure in P is absolutely continuous with respect to a fixed σ-finite measure λ on A.

In this case, we say that the family is dominated by λ and write $P \ll \lambda$.

DEFINITION 2. (Basu and Ghosh, [3]). The statistical model (X, A, P) is called discrete if

(i) each P_θ is a discrete probability measure,

(ii) A is the class of all subsets of X, and

(iii) for each $x \in X$, there exists a $\theta \in \Theta$ such that $P_\theta(\{x\}) > 0$.

Condition (iii) implies that the empty set is the only P-null set. A discrete model with a countable sample space X is clearly dominated. If P is countable, then the model is dominated. We assume that both X and P are uncountable. In this case, the model will be undominated. As we shall see in the next section, dominated and discrete models are particular cases of what Hasegawa and Perlman [6] called a coherent model.

Finally, let us state the notion of sufficiency as follows. A subfield C is sufficient with respect to the model (X, A, P) if, corresponding to each A in A, there exists a C-measurable function $I_A^\dagger$ such that $I_A^\dagger = E_\theta(I_A|C)$ $[P_\theta]$ for all $\theta \in \Theta$. A subfield C is pairwise sufficient with respect to (X, A, P) if, for each A in A and each pair $\theta_1, \theta_2 \in \Theta$, there exists a C-measurable function I_A^* such that $I_A^* = E_{\theta_i}(I_A^*|C)$ $[P_{\theta_i}]$ for $i = 1, 2$. (The function I_A^* may depend on θ_1 and θ_2.)

3. COHERENT STATISTICAL MODEL.

Let F denote the class of all measurable functions $f: X \to [0, 1]$ and let

$$s = \{f_\theta: \ f_\theta \in F, \ \theta \in \Theta\}$$

be a collection of members of F that is indexed by θ. Let $S = \{s\}$ be the family of all such collections s.

DEFINITION 3. A member $\{f_\theta\}$ of S is said to be pairwise coherent if, for every pair θ_1, θ_2 in Θ, there exists a function f_{12} in F such that $f_{\theta_i} = f_{12}$ $[P_{\theta_i}]$ for $i = 1, 2$.

DEFINITION 4. A member $\{f_\theta\}$ of S is said to be countably coherent if, for every countable subfamily $\Theta_0 = \{\theta_1, \theta_2, \ldots\}$ of Θ, there exists a function f_0 in F such that $f_{\theta_i} = f_0 [P_{\theta_i}]$ for all $i = 1, 2, \ldots$.

DEFINITION 5. A member $\{f_\theta\}$ of S is said to be coherent if there exists a function f in F such that $f_\theta = f[P_\theta]$ for all $\theta \in \Theta$.

DEFINITION 6. (Hasegawa and Perlman, [6]). The statistical model (X, A, P) is said to be coherent if every countably coherent member of S is coherent.

In the following lemma, we show that the notions of pairwise coherence and countable coherence do coincide.

LEMMA 1. If $\{f_\theta\}$ is a pairwise coherent member of S, then it is countably coherent.

PROFF. Choose and fix a countable subfamily $\Theta_0 = \{\theta_1, \theta_2, \ldots\}$ of Θ. For each pair θ_i, θ_j in Θ_0, there exists a function f_{ij} in F such that

$$f_{ij} = f_{\theta_i}\ [P_{\theta_i}] \text{ and } f_{ij} = f_{\theta_j}\ [P_{\theta_j}].$$

Let

$$g_i = \sup_j f_{ij},\ h_j = \inf_i f_{ij}, \text{ and } f = \inf_i \sup_j f_{ij}.$$

For each fixed i, the functions $f_{i1}, f_{i2}, \ldots$ are P_{θ_i}-equivalent to f_{θ_i}. The supremum of a countable number of P_{θ_i}-equivalent functions is also P_{θ_i}-equivalent to those functions. Thus, $g_i = f_{\theta_i}[P_{\theta_i}]$. Likewise, for each fixed j, we have $h_j = f_{\theta_j}[P_{\theta_j}]$. Therefore, it follows that $g_n = h_n = f_{\theta_n}[P_{\theta_n}]$ for $n = 1, 2, \ldots$. Observe that $h_n \leq f \leq g_n$ for all n. It then follows that

$$f = g_n = h_n = f_{\theta_n}\ [P_{\theta_n}] \text{ for } n = 1, 2, \ldots .$$

Hence, $\{f_\theta\}$ is countably coherent.

In general, the notions of pairwise coherence and coherence do not coincide as the following example shows.

EXAMPLE 1. (Pitcher, [10]). Let X be the unit interval $[0, 1]$, A the σ-field of Borel subsets of X, and P the family of all probability measures on A which are either degenerate at a single point of X or else are absolutely continuous with respect to the Lebesgue measure. For each P_θ in P and each $x \in X$, define $f_\theta(x) = P_\theta(\{x\})$. Then $\{f_\theta\}$ is pairwise coherent, but not coherent.

In this model, no proper subfield of A can be sufficient. To see this, let C be an arbitrary sufficient subfield and let P_x be the probability measure degenerate at x. Then it follows from the sufficiency of C that, for all A in

A, there exists a C-measurable function $I_A^\dagger$ such that

$$I_A^\dagger(x) = \int_X I_A^\dagger \, dP_x = \int_X I_A dP_x = I_A(x).$$

Let $C = \{x: \ I_A^\dagger(x) = 1\}$. Then $A = C \in \mathcal{C}$. Hence, $\mathcal{C} = \mathcal{A}$.

We now show that the model $(X, \mathcal{A}, \mathcal{P})$ is coherent if it is either dominated or discrete.

LEMMA 2. (Hasegawa and Perlman, [6]). If $(X, \mathcal{A}, \mathcal{P})$ is dominated, then it is coherent.

PROOF. Since $\mathcal{P}$ is dominated, it follows that there is a countable subfamily $\Theta_0 = \{\theta_1, \theta_2, \ldots\}$ of Θ such that $\mathcal{P}_0 = \{P_\theta: \ \theta \in \Theta_0\}$ is equivalent to $\mathcal{P}$. Suppose that $\{f_\theta\}$ is countably coherent. Then there exists a function $f_0 \in \mathcal{F}$ such that $f_0 = f_{\theta_n}[P_{\theta_n}]$ for $n = 1, 2, \ldots$. We now show that $f_0 = f_\theta[P_\theta]$ for all $\theta \in \Theta$, so $\{f_\theta\}$ is coherent. For each $\theta \in \Theta$, consider $\Theta_1 = \Theta_0 \bigcup \{\theta\}$. Then there exists a function $f_1 \in \mathcal{F}$ such that $f_1 = f_{\theta_n}[P_{\theta_n}]$ for $n = 1, 2, \ldots$. Thus, $f_0 = f_1[\mathcal{P}_0]$ and so $f_0 = f_1[\mathcal{P}]$. Hence, $f_0 = f_1 = f_\theta[P_\theta]$ as required.

LEMMA 3. If $(X, \mathcal{A}, \mathcal{P})$ is discrete, then it is coherent.

PROOF. Let $\{f_\theta\}$ be pairwise coherent. We must show that it is also coherent. For each $\theta \in \Theta$, let $S_\theta = \{x \in X: \ P_\theta(\{x\}) > 0\}$ denote the countable support of P_θ. For each pair θ_1, θ_2 in Θ, there exists a function f_{12} in $\mathcal{F}$ such that $f_{12} = f_{\theta_i}[P_{\theta_i}]$ for $i = 1, 2$. Thus, we have

$$f_{\theta_i}(x) = f_{12}(x) \text{ for all } x \in S_{\theta_i}, \ i = 1, 2. \tag{1}$$

Hence, $f_{\theta_1}(x) = f_{\theta_2}(x)$ for all $x \in S_{\theta_1} \bigcap S_{\theta_2}$. Now, choose and fix $x \in X$. Let $\Theta_x = \{\theta \in \Theta: \ x \in S_\theta\}$. Let θ_0 be a member of Θ_x. Note that $x \in S_{\theta_0} \bigcap S_\theta$ for all $\theta \in \Theta_x$. In view of (1), we have $f_{\theta_0}(x) = f_\theta(x)$ for all $\theta \in \Theta_x$, that is, $f_\theta(x)$ is constant in $\theta \in \Theta_x$ (for this prefixed x). Let c_x be the common value of f_θ for $\theta \in \Theta_x$, evaluated at x, that is, $f_\theta(x) = c_x$ for all $x \in X$. Since x is arbitrary, define a function f by $f(x) = c_x$ for all $x \in X$. Clearly, $f = f_\theta[P_\theta]$ for all $\theta \in \Theta$ as required.

That the coherent case is not exhausted by the dominated and discrete cases is shown in the following example.

EXAMPLE 2. Let (X_1, A_1, P_1) be a non-discrete dominated model, let (X_2, A_2, P_2) be an undominated discrete model, where X_1 and X_2 are disjoint, $P_1 = \{P_\theta: \theta \in \Theta_1\}$, and $P_2 = \{P_\theta: \theta \in \Theta_2\}$. Let $X = X_1 \bigcup X_2$, $A = \{A \subset X: AX_i \in A_i \text{ for } i = 1, 2\}$, and extend P_1 and P_2 to A by defining

$$P_\theta(A) = P_\theta(AX_i) \text{ for all } \theta \in \Theta_i, \ i = 1, 2.$$

It follows from Lemmas 2 and 3 that both (X_1, A_1, P_1) and (X_2, A_2, P_2) are coherent. Now we claim that (X, A, P) is also coherent, where $P = P_1 \bigcup P_2 = \{P_\theta: \theta \in \Theta_1 \bigcup \Theta_2 = \Theta\}$. Suppose that $\{f_\theta: \theta \in \Theta\}$ is pairwise coherent with respect to (X, A, P). Each f_θ, being an A-measurable function, can be written as

$$f_\theta = \begin{cases} f_\theta^1 & \text{on } X_1 \\ f_\theta^2 & \text{on } X_2, \end{cases}$$

where f_θ^i is A_i-measurable for $i = 1, 2$. Since $\{f_\theta^i: \theta \in \Theta_i\}$, being pairwise coherent with respect to (X_i, A_i, P_i), is coherent, it follows that there exists an A_i-measurable function $0 \le f_i \le 1$ such that

$$f_i = f_\theta^i \ [P_\theta] \text{ for all } \theta \in \Theta_i.$$

Define

$$f = \begin{cases} f_1 & \text{on } X_1 \\ f_2 & \text{on } X_2 \end{cases}$$

Clearly, f is A-measurable. Observe that $f_\theta = f_\theta^i [P_\theta]$ for all $\theta \in \Theta_i$, $i = 1, 2$. Therefore, $f = f_\theta [P_\theta]$ for all $\theta \in \Theta$. However, it should be noted that (X, A, P) is neither dominated nor discrete.

4. SUFFICIENCY IN THE COHERENT CASE.

For each $\theta \in \Theta$, let N_θ denote the class of all P_θ-null sets, that is, $N_\theta = \{N \in A: P_\theta(N) = 0\}$. For each subfield C of A, define

$$\check{C} = \bigcap_{\theta \in \Theta} [C \vee \sigma(N_\theta)].$$

Then $\check{C}$ is also a subfield of A. It is easy to see that a function f is $\check{C}$-measurable

if and only if, for each $\theta \in \Theta$, there exists a C-measurable function f_θ such that $f = f_\theta[P_\theta]$. Clearly, $C \subset \overline{C} \subset \tilde{C}$.

Let (X, A, P) be a coherent model. Then, so is (X, A, P_0) for any subfamily P_0 of P. However, it is not true that (X, C, P) is coherent for every subfield C of A. Such an example is explicitly included in Pitcher's [9] example.

EXAMPLE 3. Let X be the real line and let $\mathcal{B}$ be the σ-field of all Borel subsets of X. Choose and fix a non-empty, non-Borel set E that excludes the origin but is symmetric about the origin, i.e., $E = -E = \{x: -x \in E\}$. Let Θ be also the real line, and define a family $P = \{P_\theta: \theta \in \Theta\}$ of probability measures as follows: If $\theta \in E$, then P_θ is the discrete measure allotting probabilities 1/2 and 1/2 to the two points $-\theta$ and θ. If $\theta \notin E$, then P_θ is degenerate at θ. We claim that $(X, \mathcal{B}, P)$ cannot be coherent. To see this, for each $\theta \in \Theta$ and each $x \in X$, define

$$f_\theta(x) = P_\theta(\{x\}).$$

Then $\{f_\theta: \theta \in \Theta\}$ is pairwise coherent with respect to $(X, \mathcal{B}, P)$. Suppose, on the contrary, that $\{f_\theta: \theta \in \Theta\}$ is coherent. Then there exists a $\mathcal{B}$-measurable function $0 \leq f \leq 1$ such that $f = f_\theta[P_\theta]$ for all $\theta \in \Theta$. This implies that

$$f(x) = \begin{cases} 1/2 & \text{if } x \in E \\ 1 & \text{if } x \notin E, \end{cases}$$

which obviously contradicts the initial supposition that $E \notin \mathcal{B}$. However, if we consider the class A of all subsets of X, then (X, A, P) is a discrete model and hence is coherent.

In the following lemma, however, we show that if (X, A, P) is a coherent model, then so is $(X, \tilde{C}, P)$ for any subfield C of A.

LEMMA 4. If (X, A, P) is coherent and C is a subfield of A, then $(X, \check{C}, P)$ is coherent.

PROOF. Let $\{f_\theta: \theta \in \Theta\}$ be pairwise coherent with respect to $(X, \check{C}, P)$. Then it is also pairwise coherent with respect to (X, A, P). Since (X, A, P) is coherent, there exists an A-measurable function $0 \leq f \leq 1$ such that $f = f_\theta[P_\theta]$ for all $\theta \in \Theta$. Since each f_θ is $\tilde{C}$-measurable, it is seen that f must be

measurable with respect to $\tilde{\tilde{C}} = \tilde{C}$. This proves that $(X, \tilde{C}, P)$ is coherent.

Let $\tilde{P}$ denote the convex hull of P, that is, $\tilde{P} = \{Q: Q = \Sigma a_i P_{\theta_i}, a_i \geq 0, \Sigma a_i = 1, \theta_i \in \Theta\}$.

LEMMA 5. (X, A, P) is coherent if and only if $(X, A, \tilde{P})$ is coherent.

PROOF. The "only if" part is what needs to be proved. Let $\{f_Q: Q \in \tilde{P}\}$ be pairwise coherent with repsect to $(X, A, \tilde{P})$. Then the subset $\{f_\theta: \theta \in \Theta\}$ of $\{f_Q: Q \in \tilde{P}\}$ is pairwise coherent with respect to (X, A, P). Since (X, A, P) is coherent, there exists an A-measurable function $0 \leq f \leq 1$ such that $f = f_\theta[P_\theta]$ for all $\theta \in \Theta$. We now claim that $f = f_Q[Q]$ for all $Q \in \tilde{P}$. Choose and fix $Q \in \tilde{P}$. Then $Q = \Sigma a_i P_{\theta_i}$, where $\theta_i \in \Theta$, $a_i \geq 0$, and $\Sigma a_i = 1$. For each pair Q, P_{θ_i}, there exists an A-measurable function $0 \leq f_i \leq 1$ such that

$$f_i = f_Q[Q]$$

and

$$f_i = f_{\theta_i} = f[P_{\theta_i}].$$

Since $\{P_{\theta_i}: i = 1, 2, \ldots\} \equiv Q$, it follows that

$$f = f_i = f_Q[P_{\theta_i}] \text{ for } i = 1, 2, \ldots .$$

Hence, $f = f_Q[Q]$ as required.

LEMMA 6. Let C be a subfield of A such that (X, C, P) is coherent. If P is closed for countable convex combinations (i.e., $P = \tilde{P}$), then $\overline{C} = \tilde{C}$.

PROOF. Since $\overline{C} \subset \tilde{C}$, it suffices to show that $\tilde{C} \subset \overline{C}$. Let f be a $\tilde{C}$-measurable function such that $0 \leq f \leq 1$. Then, for each $\theta \in \Theta$, there exists a C-measurable function $0 \leq f_\theta \leq 1$ such that $f = f_\theta[P_\theta]$. For each pair $\theta_1, \theta_2 \in \Theta$, let $Q = (P_{\theta_1} + P_{\theta_2})/2$. Then $Q \in \tilde{P} = P$ and so $f = f_Q[Q]$. Since $\{P_{\theta_1}, P_{\theta_2}\} \equiv Q$, we have $f_Q = f = f_{\theta_i}[P_{\theta_i}]$ for $i = 1, 2$. Thus, we have shown that $\{f_\theta: \theta \in \Theta\}$ is pairwise coherent with respect to (X, C, P). Therefore, there exists a C-measurable function $0 \leq f_0 \leq 1$ such that $f_0 = f_\theta[P_\theta]$ for all $\theta \in \Theta$, and hence $f = f_0[P]$. This shows that $\tilde{C} \subset \overline{C}$.

LEMMA 7. (X, C, P) is coherent if and only if $(X, \overline{C}, P)$ is coherent.

PROOF. Let us first prove the "only if" part. Let $\{f_\theta: \theta \in \Theta\}$ be pairwise coherent with respect to $(X, \overline{C}, P)$. For each $\theta \in \Theta$, there exists a C-measurable function $0 \leq g_\theta \leq 1$ such that $f_\theta = g_\theta[P_\theta]$. For each pair $\theta_1, \theta_2 \in \Theta$, there exists a $\overline{C}$-measurable function $0 \leq f_{12} \leq 1$ such that $f_{12} = f_{\theta_i}[P_{\theta_i}]$ for $i = 1, 2$. Since there is a C-measurable function $0 \leq g_{12} \leq 1$ such that $f_{12} = g_{12}[P]$, it follows that $g_{12} = g_{\theta_i}[P_{\theta_i}]$ for $i = 1, 2$. Hence, $\{g_\theta: \theta \in \Theta\}$ is pairwise coherent with respect to (X, C, P). Since (X, C, P) is coherent, there exists a C-measurable function $0 \leq f \leq 1$ such that $f = g_\theta = f_\theta[P_\theta]$ for all $\theta \in \Theta$. Therefore, $(X, \overline{C}, P)$ is coherent.

To prove the "if" part, let $\{f_\theta: \theta \in \Theta\}$ be pairwise coherent with respect to (X, C, P). Then, it is also pairwise coherent with respect to $(X, \overline{C}, P)$. Thus, there exists a $\overline{C}$-measurable function $0 \leq \overline{f} \leq 1$ such that $\overline{f} = f_\theta[P_\theta]$ for all $\theta \in \Theta$. Since there is a C-measurable function $0 \leq f \leq 1$ such that $f = \overline{f}[P]$, we have $f = f_\theta[P_\theta]$ for all $\theta \in \Theta$ as required.

Under the assumption of coherence, we now prove a number of results on sufficiency.

PROPOSITION 1. Suppose that (X, C, P) is coherent. If C is pairwise sufficient, then C is sufficient.

PROOF. Chose and fix $A \in \mathcal{A}$. For each $\theta \in \Theta$, let $0 \leq f_\theta \leq 1$ be a version of $E_\theta(I_A|C)$. Since C is pairwise sufficient, for each pair $\theta_1, \theta_2 \in \Theta$, there exists a C-measurable function $0 \leq f_{12} \leq 1$ such that $f_{12} = f_{\theta_i}[P_{\theta_i}]$ for $i = 1, 2$. Thus, $\{f_\theta: \theta \in \Theta\}$ is pairwise coherent with respect to (X, C, P). Since (X, C, P) is coherent, there exists a C-measurable function $0 \leq f \leq 1$ such that $f = f_\theta[P_\theta]$ for all $\theta \in \Theta$. That is, C is sufficient.

COROLLARY 1. Suppose that $(X, \mathcal{A}, P)$ is coherent and $\overline{C} = \hat{C}$. If C is pairwise sufficient, then C is sufficient.

PROOF. Since $(X, \mathcal{A}, P)$ is coherent, it follows from Lemma 4 that $(X, \hat{C}, P)$ is coherent. Since $\overline{C} = \hat{C}$, $(X, \overline{C}, P)$ is coherent. In view of Lemma 7, (X, C, P) is coherent. Consequently, the result follows from Proposition 1.

COROLLARY 2. Let $\mathcal{C} \subset \mathcal{D}[P_{\theta_1}, P_{\theta_2}]$ for all pairs $\theta_1, \theta_2 \in \Theta$. If $\mathcal{C}$ is pairwise sufficient and $(X, \mathcal{D}, P)$ is coherent, then $\mathcal{D}$ is sufficient.

PROOF. Since $\mathcal{C}$ is pairwise sufficient and $\mathcal{C} \subset \mathcal{D}[P_{\theta_1}, P_{\theta_2}]$ for all $\theta_1, \theta_2 \in \Theta$, it is easy to verify that $\mathcal{D}$ is also pairwise sufficient. Since $(X, \mathcal{D}, P)$ is coherent, it follows from Proposition 1 that $\mathcal{D}$ is sufficient.

COROLLARY 3. Suppose that $(X, \mathcal{A}, P)$ is coherent and $\mathcal{C}$ is sufficient. If $\mathcal{C} \subset \mathcal{D}[P_{\theta_1}, P_{\theta_2}]$ for all $\theta_1, \theta_2 \in \Theta$ and $\overline{\mathcal{D}} = \tilde{\mathcal{D}}$, then $\mathcal{D}$ is sufficient.

PROOF. Since $(X, \mathcal{A}, P)$ is coherent and $\overline{\mathcal{D}} = \tilde{\mathcal{D}}$, it follows from Lemmas 4 and 7 that $(X, \mathcal{D}, P)$ is coherent. In view of Corollary 2, we therefore conclude that $\mathcal{D}$ is sufficient.

We now show that if $(X, \mathcal{A}, P)$ is coherent, then so is $(X, \mathcal{C}, P)$ for any sufficient subfield $\mathcal{C}$. To this end, we shall need the following lemma.

LEMMA 8. (Pitcher, 1965). If $\mathcal{C}$ is sufficient, then $\overline{\mathcal{C}} = \tilde{\mathcal{C}}$.

PROOF. Let $A \in \tilde{\mathcal{C}}$. Since $\mathcal{C}$ is sufficient, there exists a $\mathcal{C}$-measurable function $I_A^\dagger$ such that $I_A^\dagger = E_\theta(I_A|\mathcal{C})$ $[P_\theta]$ for all $\theta \in \Theta$. Note that $I_A - I_A^\dagger$ is $\tilde{\mathcal{C}}$-measurable. Thus, for each $\theta \in \Theta$, there exists a $\mathcal{C}$-measurable function f_θ such that $I_A - I_A^\dagger = f_\theta[P_\theta]$ and so

$$\begin{aligned}
\int_X (I_A - I_A^\dagger)^2 dP_\theta &= \int_X (I_A - I_A^\dagger) f_\theta dP_\theta \\
&= \int_X E_\theta[(I_A - I_A^\dagger) f_\theta | \mathcal{C}] dP_\theta \\
&= \int_X f_\theta E_\theta(I_A - I_A^\dagger | \mathcal{C}) dP \\
&= 0.
\end{aligned}$$

Hence, $I_A = I_A^\dagger[P_\theta]$ for all $\theta \in \Theta$. Since $I_A^\dagger$ is $\mathcal{C}$-measurable, it follows that $A \in \overline{\mathcal{C}}$. Therefore, $\overline{\mathcal{C}} = \tilde{\mathcal{C}}$.

PROPOSITION 2. If $(X, \mathcal{A}, P)$ is coherent and $\mathcal{C}$ is sufficient, then $(X, \mathcal{C}, P)$ is coherent.

PROOF. Since $\mathcal{C}$ is sufficient, $\overline{\mathcal{C}} = \tilde{\mathcal{C}}$ in view of Lemma 8. Since $(X, \mathcal{A}, P)$ is coherent, it follows from Lemma 4 and 7 that $(X, \mathcal{C}, P)$ is coherent.

5. BASU'S THEOREM.

As before, $(X, \mathcal{A}, P)$ is our basic model, where $P = \{P_\theta: \theta \in \Theta\}$. Two

probability measures P_{θ_1} and P_{θ_2} in $\mathcal{P}$ are said to be overlapping if, for any set A in $\mathcal{A}$, $P_{\theta_1}(A) = 1$ implies that $P_{\theta_2}(A) > 0$. We write $\theta_1 <=> \theta_2$ if P_{θ_1} and P_{θ_2} overlap. If there exists a finite number of parameter points $\theta_1, \theta_2, \ldots, \theta_k$ such that

$$\theta_1 <=> \theta_2 <=> \ldots <=> \theta_k <=> \theta',$$

then we say P_θ and $P_{\theta'}$ are connected. The family $\mathcal{P}$ is called connected if every pair of probability measures in the family are connected.

THEOREM 1. (Basu, [2]). If T is a sufficient statistic, $\mathcal{P}$ is connected, and V is a statistic which is independent of T for all $\theta \in \Theta$, then the distribution of V does not depend on θ.

A set A in $\mathcal{A}$ is called splitting set if there is a partition Θ_0, Θ_1, of Θ such that

$$P_\theta(A) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ 1 & \text{if } \theta \in \Theta_1 \end{cases}$$

The above theorem has been generalized by Koehn and Thomas [7] as follows:

THEOREM 2. Let T be a sufficient statistic. There exists a statistic V, independent of T for all $\theta \in \Theta$, whose distribtuion depends on θ if and only if there exists a splitting set.

We now demonstrate that the two notions of connectedness of $\mathcal{P}$ and the nonexistence of a splitting set are equivalent in the coherent case.

LEMMA 9. Let $(X, \mathcal{A}, \mathcal{P})$ be coherent. Then $\mathcal{P}$ is connected if and only if there exists no splitting set.

PROOF. Clearly, the connectedness of $\mathcal{P}$ implies the nonexistence of a splitting set. For each $\theta \in \Theta$, let A_θ be a set such that $P_\theta(A_\theta) = 1$ and $P_\theta(B) > 0$ if $\phi \neq B \subset A_\theta$. Suppose that $\mathcal{P}$ is not connected. Choose and fix $\theta_0 \in \Theta$. Let $\Theta_0 = \{\theta \in \Theta: P_\theta \text{ and } P_{\theta_0} \text{ are connected}\}$. Since $\mathcal{P}$ is not connected, $\Theta_1 = \Theta \setminus \Theta_0$ is not empty. For each $\theta \in \Theta$, let $f_\theta = I_{A_\theta}$. It is easy to show that $\{f_\theta: \theta \in \Theta\}$ is pairwise coherent. Since $(X, \mathcal{A}, \mathcal{P})$ is coherent, there exists an $\mathcal{A}$-measurable

function $0 \leq f \leq 1$ such that $I_{A_\theta} = f[P_\theta]$ for all $\theta \in \Theta$. Let $A = \bigcup_{\theta \in \Theta_0} A_\theta$. Since A_θ is the support of P_θ, it is easily seen that

$$P_\theta(A) = P_\theta(A_\theta) = 1 \text{ for all } \theta \in \Theta_0.$$

Note that $A \bigcap A_\theta = \phi$ if $\theta \in \Theta_1$. Thus,

$$P_\theta(A) = 0 \text{ for all } \theta \in \Theta_1.$$

That nonexistence of a splitting set is weaker than the connectedness property is seen from the following example.

EXAMPLE 4. Let $\mathcal{P}$ be a family consisting of all two-point distributions and the standard normal distribution. Then $\mathcal{P}$ is not connected and this does not possess a splitting set.

ACKNOWLEDGMENTS. Work of the first author is partially support by NSF Grant No. 79-04693.

REFERENCES

1. Basu, D. (1955). On statistics independent of a complete sufficient statistic. Sankhya A, 15, 337-380.

2. Basu, D. (1958). On statistics independent of sufficient statistic. Sankhya A, 20, 223-226.

3. Basu, D. and Ghosh, J.K. (1967). Sufficient statistics in sampling from a finite universe. Proc. 36th Session Internat. Statist. (In ISI Bulletin), 850-859.

4. Cheng, S.C. (1978). A Mathematical Study of Sufficiency and Adequacy in Statistical Theory. Ph.D. Thesis, submitted to the Florida State University.

5. Halmos, P. R. and Savage, L. J. (1949). Applications of the Radon-Nikodym theorem to the theory of sufficient statistics. Ann. Math. Statist. 20, 225-241.

6. Hasegawa, M. and Perlman, M. D. (1974). On the existence of a minimal sufficient subfield. Ann. Statist. 2, 1049-1055.

7. Koehn, V. and Thomas, L. D. (1975). On statistics independent of a sufficient statistic: Basu's lemma. American Statistician, 29, 40-43.

8. Pathak, P. K. (1975). Note on Basu's lemma. Technical Report No. 308. The University of New Mexico.

9. Pitcher, T. S. (1957). Sets of measures not admitting necessary and sufficient statistics or subfield. Ann. Math. Statist. 28, 267-268.

10. Pitcher, T. S. (1965). A more general property than domination for sets of probability measures. Pacific J. Math. 15, 597-611.

G. Kallianpur, P.R. Krishnaiah, J.K. Ghosh, eds., *Statistics and Probability: Essays in Honor of C.R. Rao*

A NOTE ON THE DIRICHLET PROCESS

D. BASU and R. C. TIWARI
The Florida State University, Tallahassee, FL, U.S.A.

Written mainly for its pedagogical interest, this expository note is concerned primarily with the question of existence of a Dirichlet process. A random probability measure on a measurable space $(\mathcal{X}, \mathcal{A})$ is a stochastic process $\{P(A): A \in \mathcal{A}\}$ – a collection of random variables indexed by the measurable sets in $\mathcal{X}$ – such that almost every realization of the process is a probability measure on $(\mathcal{X}, \mathcal{A})$. Given a finite measure α on $(\mathcal{X}, \mathcal{A})$, a Dirichlet process D^{α} is a random probability measure on $(\mathcal{X}, \mathcal{A})$ such that, for every partition $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$ into a finite number of measurable sets, the joint distribution of the random variables $(P(A_1), P(A_2), \ldots, P(A_k))$ is a singular Dirichlet distribution with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$.

Part one of this article deals with the familiar case where $\mathcal{X}$ is a finite set. Properties of the k-dimensional Dirichlet distribution are so exposited as to motivate Blackwell's (1973) constructive definition of the Dirichlet process. In part two, the case where $(\mathcal{X}, \mathcal{A})$ is a Borel space is discussed in some detail.

1. Introduction

This report is concerned primarily with the question of existence of a Dirichlet process on a Borel space $(\mathcal{X}, \mathcal{A})$. Part one of this report deals with the familiar case when $\mathcal{X}$ is a finite set. Properties of the k-dimensional Dirichlet distribution are so exposited as to motivate Blackwell's (1973) constructive definition of the Dirichlet process. In part two, the case where $(\mathcal{X}, \mathcal{A})$ is a Borel space is discussed in some detail.

Section 2 deals with Bayesian (parametric) inference and the family of natural conjugate priors. Section 3 is devoted to some characterizations of the Dirichlet distribution and elucidations of its useful properties. In Section 4, the results of Section 3 are extended and it is shown that there exists a Dirichlet process on $(\mathcal{X}, \mathcal{A})$, when $\mathcal{X}$ is a finite or a countably infinite set.

In Section 5, some preliminary material on the Dirichlet process is presented. Some useful results on a Borel space are given in Section 6. Blackwell's (1973) construction of a Dirichlet process on a Borel space $(\mathcal{X}, \mathcal{A})$ is discussed in Section 7. Section 8 is devoted to the study of some properties of a Dirichlet process.

PART I: THE DIRICHLET DISTRIBUTION

2. Bayesian inference and natural conjugate priors

Let X be an observable random variable (r.v.) with a statistical model that is characterized by a probability density function (p.d.f.) $f(\cdot|\theta)$, where $\theta \in \Theta$ is the

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_31,

unknown parameter of interest. Before any data are collected, a Bayesian represents his prior opinion about θ by a distribution on Θ, called a prior distribution. If he observes X n-times and denotes by $\boldsymbol{D}_1^n=(x_1, x_2, \ldots, x_n)$ the data so obtained, then his opinion about θ is represented by the distribution of θ given $\boldsymbol{D}_1^n$, called the posterior distribution.

Let q be the prior p.d.f. of θ. The posterior p.d.f. of θ given $\boldsymbol{D}_1^n$, namely $q(\cdot|\boldsymbol{D}_1^n)$, is expressed by the relation

$$q(\theta|\boldsymbol{D}_1^n)\propto q(\theta)L(\theta|\boldsymbol{D}_1^n). \tag{2.1}$$

Here, $L(\theta|\boldsymbol{D}_1^n)$ is the likelihood function of θ at point $\boldsymbol{D}_1^n$, and the proportionality symbol, $\propto$, is used to indicate that the posterior p.d.f. of θ given $\boldsymbol{D}_1^n$ is equal to the right side of (2.1) divided by the factor $\int_\Theta L(\theta|\boldsymbol{D}_1^n)q(\theta)\mathrm{d}\theta$ which does not involve θ.

This is how new knowledge, obtained through data, may be combined with prior knowledge. The Bayesian continually updates his knowledge as more observations are taken. Clearly,

$$\begin{aligned} q(\theta|\boldsymbol{D}_1^{n+m}) &\propto q(\theta)L(\theta|\boldsymbol{D}_1^{n+m}) \\ &\propto q(\theta)L(\theta|\boldsymbol{D}_1^n)L(\theta|\boldsymbol{D}_{n+1}^{n+m}) \\ &\propto q(\theta|\boldsymbol{D}_1^n)L(\theta|\boldsymbol{D}_{n+1}^{n+m}). \end{aligned}$$

Thus, the opinion $q(\theta|\boldsymbol{D}_1^{n+m})$ based on data $\boldsymbol{D}_1^{n+m}$ may be regarded as the posterior based on data $\boldsymbol{D}_{n+1}^{n+m}$ and prior $q(\theta|\boldsymbol{D}_1^n)$. This process of updating opinion may go through many stages.

It is clear from the relation (2.1) that the change of opinion about θ after the data are obtained is effected through the likelihood function. In the context of a chosen statistical model, a Bayesian will regard the likelihood function as the sole reservoir of all the relevant information about the parameter that is contained in the data. This is usually stated as: *The Likelihood Principle*. Two sets of data generating equivalent likelihood functions contain the same relevant information about the parameter. Two likelihood functions are said to be equivalent if one of them is a constant multiple of the other, where the constant may depend on the data. [See Basu (1975) for more on the likelihood principle.]

In many situations, it is convenient to access the prior within a family, C, of distributions. The class C should be large enough to accommodate various shades of opinion about the parameter. Further, if $q\in C$ is a prior p.d.f. of θ, then the posterior p.d.f. $q(\theta|\boldsymbol{D}_1^n)$ of θ given the data $\boldsymbol{D}_1^n$ ought to be in a simple computable form. If $q(\theta|\boldsymbol{D}_1^n)\in C$ for all $q\in C$ and data $\boldsymbol{D}_1^n$, then C is called a *conjugate family of priors*.

It frequently happens that a conjugate family of priors naturally coexists with a given statistical model for the observable r.v. X. Suppose the model is such that there exists an $n_0>0$ with the property that for all data $\boldsymbol{D}_1^n$ with $n\geqslant n_0$ the induced likelihood function $L(\theta|\boldsymbol{D}_1^n)$ is integrable (with respect to some integrating measure μ) over the parameter space Θ. Consider then the family C_0 of p.d.f.'s $q(\theta)$ of the

form:

$$q(\theta)=\frac{L(\theta|\boldsymbol{D}_1^n)}{\int_\theta L(\theta|\boldsymbol{D}_1^n)\,\mathrm{d}\mu(\theta)}.$$

If the prior p.d.f. $q(\theta)$ corresponds to a so-called prior data, $\boldsymbol{D}_1^n=(y_1, y_2,\ldots, y_n)$, then with the current data $\boldsymbol{D}_1^m=(x_1, x_2,\ldots, x_m)$. The posterior p.d.f. $q(\theta|\boldsymbol{D}_1^m)$ will correspond to the likelihood function $L(\theta|\boldsymbol{D}_1^{n+m})$, where $\boldsymbol{D}_1^{n+m}$ is the extended data $(y_1, y_2,\ldots, y_n, x_1, x_2,\ldots, x_m)$. Thus for each prior $q\in C_0$, the posterior $q(\cdot|\boldsymbol{D}_1^m)$ belongs to C_0 for all possible current data $\boldsymbol{D}_1^m$.

The natural conjugate family C_0 of prior distributions takes on a simple form when, irrespective of the sample size n, there exists a sufficient statistic $T=(T_1, T_2,\ldots, T_k)$ of fixed and small dimension $k(k\geqslant 1)$. Then,

$$L(\theta|\boldsymbol{D}_1^n)\propto H_n(\theta, T_1, T_2,\ldots, T_k), \tag{2.2}$$

where $T_1, T_2,\ldots, T_k$ are functions of $\boldsymbol{D}_1^n$. In this case, the natural conjugate family C_0 of prior distributions is characterized by $k+1$ superparameters, namely, particular values of $T_1, T_2,\ldots, T_k$ and n.

For example, suppose each observation on an observable r.v. X belongs to one of the $k+1$ mutually exclusive and collectively exhaustive categories. Let $p_i(0<p_i<1)$ be the probability that an observation belongs to the ith category, $i=1,2,\ldots, k+1$, where $\sum_{i=1}^{k+1}p_i=1$. We may regard $(p_1, p_2,\ldots, p_k)$ as the model parameters. Suppose X is observed n times and let $\boldsymbol{D}_1^n$ be the data $(x_1, x_2,\ldots, x_n)$ collected. Furthermore, let n_i denote the number of x's that belong to the ith category, $i=1,2,\ldots, k+1$. Then each n_i is a non-negative integer and $\sum_{i=1}^{k+1}n_i=n$. Also, since $\sum_{i=1}^{k+1}n_i=n$, we may regard $T=(n_1, n_2,\ldots, n_k)$ as the k-dimensional sufficient statistic. Before the data are collected, $(n_1, n_2,\ldots, n_k)$ are r.v.'s having a multinomial distribution with parameters n and $(p_1, p_2,\ldots, p_k)$. The likelihood function $L(p_1, p_2,\ldots, p_k|\boldsymbol{D}_1^n)\propto \prod_{i=1}^k p_i^{n_i}(1-\sum_{i=1}^k p_i)^{n-\sum_{i=1}^k n_i}$, which is of the form (2.2) with $\theta=(p_1, p_2,\ldots, p_k)$ and $T=(n_1, n_2,\ldots, n_k)$.

The natural conjugate family of prior distributions for the parameters $(p_1, p_2,\ldots, p_k)$ is then the family $\mathcal{C}_0$ of distributions with p.d.f.'s of the form

$$q(p_1, p_2,\ldots, p_k)\propto \prod_{i=1}^k p_i^{a_i-1}\left(1-\sum_{i=1}^k p_i\right)^{a_{k+1}-1};\quad p_i>0,\quad i=1,2,\ldots, k,$$

$\sum_{i=1}^k p_i<1$, and each a_i is a positive integer.

Clearly, for any prior p.d.f. $q(p_1, p_2,\ldots, p_k)\in\mathcal{C}_0$ and any data $\boldsymbol{D}_1^n$ the posterior p.d.f.,

$$q(p_1, p_2,\ldots, p_k|\boldsymbol{D}_1^n)\propto \prod_{i=1}^k p_i^{a_i+n_i-1}\left(1-\sum_{i=1}^k p_i\right)^{a_{k+1}+(n-\sum_{i=1}^k n_i)-1}$$

and so $q(p_1, p_2,\ldots, p_k|\boldsymbol{D}_1^n)\in\mathcal{C}_0$.

The natural conjugate family $\mathcal{C}_0$ of prior distributions for the parameters $(p_1, p_2, \ldots, p_k)$ is a subfamily of the family of the Dirichlet distributions, defined in the next section.

3. The Dirichlet distribution

This section is devoted to the study of the family of the Dirichlet distributions, as the natural conjugate family for the parameters of a multinomial distribution, and its characterizations. The Dirichlet distribution is defined as follows.

Definition 3.1. Let $\alpha_i > 0, i = 1, 2, \ldots, k+1$. The r.v.'s $(Y_1, Y_2, \ldots, Y_k)$ are said to have a Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, denoted by $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, if the joint distribution of $(Y_1, Y_2, \ldots, Y_k)$ has the p.d.f. $f(y_1, y_2, \ldots, y_k) = \text{const}\, y_1^{\alpha_1 - 1} y_2^{\alpha_2 - 1}, \ldots, y_k^{\alpha_k - 1}(1 - y_1 - \cdots - y_k)^{\alpha_{k+1} - 1}$, over the k-dimensional simplex S_k defined by the inequalities $y_i > 0, i = 1, 2, \ldots, k, \sum_{i=1}^{k} y_i < 1$.

More generally, in the above definition one may assume $\alpha_i \geq 0$ for each i, and $\sum_{i=1}^{k+1} \alpha_i > 0$. However, if $\alpha_i = 0$ for some i, then the corresponding $Y_i = 0$ with probability one.

For $k = 1$, the Dirichlet distribution $D(\alpha_1, \alpha_2)$ for Y_1 is the familiar Beta distribution with parameters α_1 and α_2, Beta(α_1, α_2). The proof of the following basic proposition has already been outlined in the previous section.

Proposition 3.1. *Let $D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$ be the prior probability model for the parameters $(p_1, p_2, \ldots, p_k)$ in the statistical model of a $k+1$ valued r.v. X. Then, with n independent observations on X, giving rise to the sample frequencies $n_1, n_2, \ldots, n_{k+1}$ for the $k+1$ values of the r.v. X, the posterior distribution of $(p_1, p_2, \ldots, p_k)$ will be $D(\alpha_1 + n_1, \alpha_2 + n_2, \ldots, \alpha_{k+1} + n_{k+1})$.*

The rest of this section is devoted to some characterizations of the Dirichlet distribution and elucidations of some of its more useful properties.

First of all, note that if we define Y_{k+1} as $1 - \sum_{i=1}^{k} Y_i$ then the joint distribution of $(Y_1, Y_2, \ldots, Y_{k+1})$ is singular with respect to the $k+1$ dimensional Lebesgue measure λ_{k+1} on R_{k+1}. The support of this singular distribution is the k-dimensional simplex E_{k+1} defined by the inequalities $y_i > 0, i = 1, 2, \ldots, k+1, \sum_{i=1}^{k+1} y_i = 1$. The joint p.d.f. (with respect to the k-dimensional Lebesgue measure on E_{k+1}) of the $k+1$ variables may be neatly represented as const $\prod_{i=1}^{k+1} y_i^{\alpha_i - 1}$.

The following result follows immediately.

Proposition 3.2. *If $i_1, i_2, \ldots, i_k$ is any sequence of distinct integers from the set $\mathcal{X} - \{1, 2, \ldots, k+1\}$ then $(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_k}) \sim D(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_{k+1}})$.*

A characterization of the Dirichlet distribution in terms of mutually independent Beta r.v.'s is given by

Proposition 3.3. *Let* $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$. *Let* $U_1 = Y_1$, *and* $U_i = Y_i/(1 - Y_1 - \ldots - Y_{i-1})$, $i = 2,3,\ldots,k$. *Then* $U_i \sim Beta(\alpha_i, \sum_{j=i+1}^{k+1} \alpha_j)$, $i = 1,2,\ldots,k$, *and* $U_1, U_2, \ldots, U_k$ *are mutually independent.*

Proof. The joint p.d.f. of $(Y_1, Y_2, \ldots, Y_k)$ is $f(y_1, y_2, \ldots, y_k) = \text{const} \prod_{i=1}^{k} y_i^{\alpha_i - 1}(1 - \sum_{i=1}^{k} y_i)^{\alpha_{k+1} - 1}; (y_1, y_2, \ldots, y_k) \in S_k$. Consider the one–one transformation of S_k onto the k-dimensional cube $(0,1)^k$ given by the relation $y_1 = u_1$, $y_i = u_i \prod_{j=1}^{i-1}(1 - u_j)$, $i = 2,3,\ldots,k$, the Jacobian of transformation being $\prod_{i=1}^{k-1}(1 - u_i)^{k-i}$. It follows then that the joint p.d.f. of $(U_1, U_2, \ldots, U_k)$ is $g(u_1, u_2, \ldots, u_k) = \text{const} \prod_{i=1}^{k} u_i^{\alpha_i - 1}(1 - u_i)^{\alpha_{i+1} + \alpha_{i+2} + \cdots + \alpha_{k+1} - 1}$.

That the converse of Proposition 3.3 is true, is established by reversing the above chain of arguments.

Remark 3.1. As a by-product of the above proposition we immediately have that the r.v. $Y_1 \sim \text{Beta}(\alpha_1, \alpha - \alpha_1)$ where $\alpha = \sum_{i=1}^{k+1} \alpha_i$. This fact together with Proposition 3.2 then gives:

Corollary 3.1. *If* $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, *then* $Y_i \sim Beta(\alpha_i, \alpha - \alpha_i)$, $i = 1,2,\ldots,k$.

This corollary may be generalized by using the converse of Proposition 3.3 to:

Corollary 3.2. *If* $(Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, *then* $(Y_1, Y_2, \ldots, Y_r) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_r, \alpha - \sum_{i=1}^{r} \alpha_i)$.

A more general extension is then given by Proposition 3.2 and the converse of Proposition 3.3 as

Corollary 3.3. *For any subset* $\{i_1, i_2, \ldots, i_r\}$ *of* $\mathfrak{X}$, $(Y_{i_1}, Y_{i_2}, \ldots, Y_{i_r}) \sim D(\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_r}, \alpha - \sum_{j=1}^{r} \alpha_{i_j})$.

The following proposition follows immediately from Proposition 3.3 and its converse.

Proposition 3.4. *Let* $(Y_1, Y_2, \cdots, Y_k) \sim D(\alpha_1, \alpha_2, \cdots, \alpha_{k+1})$. *Then, for any integer* r *such that* $2 \leqslant r \leqslant k$,

$$\left(\frac{Y_r}{1 - Y_1 - \cdots - Y_{r-1}}, \frac{Y_{r+1}}{1 - Y_1 - \cdots - Y_{r-1}}, \ldots, \frac{Y_k}{1 - Y_1 - \cdots - Y_{r-1}} \right)$$

is independent of $(Y_1, Y_2, \ldots, Y_{r-1})$. *Also,*

$$\left(\frac{Y_r}{1 - Y_1 - \cdots - Y_{r-1}}, \frac{Y_{r+1}}{1 - Y_1 - \cdots - Y_{r-1}}, \ldots, \frac{Y_k}{1 - Y_1 - \cdots - Y_{r-1}} \right) \sim D(\alpha_r, \alpha_{r+1}, \ldots, \alpha_{k+1}).$$

The Dirichlet distribution can also be characterized in terms of mutually independent Gamma r.v.'s. This is given by the following

Proposition 3.5. *Let $Z_1, Z_2, \ldots, Z_{k+1}$ be mutually independent Gamma r.v.'s with the common scale parameter $\beta>0$ and possibly different shape parameters $\alpha_i>0, i=1,2,\ldots,k+1$. Let $Z=\Sigma_i Z_i$ and $Y_i=Z_i/Z, i=1,2,\ldots,k$. Then $(Y_1, Y_2, \ldots, Y_k)\sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$. Also, $(Y_1, Y_2, \ldots, Y_k)$ is independent of Z.*

Proof. The joint p.d.f. of the Z_i's is

$$f(z_1, z_2, \ldots, z_{k+1}) \propto \exp\left\{-\beta \sum_i z_i\right\} \Pi_i z_i^{\alpha_i - 1}, \quad z_i>0, i=1,2,\ldots,k+1.$$

Consider the transformation $z=\Sigma_i z_i$, $y_i=z_i/z, i=1,2,\ldots,k$, the reverse transformation being $z_i=zy_i, i=1,2,\ldots,k$, and $z_{k+1}=z(1-\Sigma_{i=1}^k y_i)$. The Jacobian of transformation is z^k. It then follows that the joint p.d.f. of $(Z, Y_1, Y_2, \ldots, Y_k)$ is $g(z, y_1, y_2, \ldots, y_k) \propto \exp\{-\beta z\} z^{\alpha-1} \Pi_{i=1}^k y_i^{\alpha_i - 1}(1-\Sigma_{i=1}^k y_i)^{\alpha_{k+1}-1}$.

That $(Y_1, Y_2, \ldots, Y_k)$ is independent of Z in the above proposition can also be seen as follows. Regard the parameters $(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$, where each $\alpha_i>0$, as known constants and $\beta>0$ as the unknown parameter. With $(Z_1, Z_2, \ldots, Z_{k+1})$ as the sample, $Z=\Sigma_i Z_i$ is a complete sufficient statistic. The vector valued statistic $((Z_1/Z),(Z_2/Z),\ldots,(Z_k/Z))$ is scale invariant. Since $\beta>0$ is a scale parameter, it follows that $((Z_1/Z),(Z_2/Z),\ldots,(Z_k/Z))$ is ancillary statistic. Hence it follows (Basu (1955)) that $((Z_1/Z),(Z_2/Z),\ldots,(Z_k/Z))=(Y_1, Y_2, \ldots, Y_k)$ is independent of Z.

The converse of Proposition 3.5 may be stated as follows. The proof is omitted.

Proposition 3.6. *If Z is a r.v. having a Gamma distribution with shape parameter $\alpha_i>0$ and scale parameter $\beta>0$, denoted by $Z\sim G(\alpha, \beta)$, and if Z is independent of $(Y_1, Y_2, \ldots, Y_k)$, where $(Y_1, Y_2, \ldots, Y_k)\sim D(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$; then the r.v.'s $Z_i=ZY_i$, $i=1,2,\ldots,k$, and $Z_{k+1}=Z(1-\Sigma_{i=1}^k Y_i)$ are mutually independent with $Z_i\sim G(\alpha_i, \beta)$, $i=1,2,\ldots,k+1$.*

Remark 3.2. Proposition 3.4 can also be proved using the Basu theorem (Basu (1955)) and the Gamma characterization of the Dirichlet distribution.

For more on the Dirichlet distribution, see Wilks (1962).

4. Further properties of the Dirichlet distribution

Suppose A is any subset of the set $\mathcal{X}=\{1,2,\ldots,k+1\}$ and $(y_1, y_2, \ldots, y_{k+1})$ is any given point in the simplex E_{k+1}. Define $P(A)=\Sigma_{i\in A} y_i$. Then, P is a probability measure on $\mathcal{X}$, and P is identified by the point $(y_1, y_2, \ldots, y_{k+1})$ in E_{k+1}. Thus, E_{k+1} represents a class of probability measures on $\mathcal{X}$. If $(Y_1, Y_2, \ldots, Y_{k+1})$ is a

random point in E_{k+1} then $P(A)=\Sigma_{i\in A}Y_i$ is random probability measure of the set A, and P is a random probability measure on $\mathfrak{X}$. A random probability measure on $\mathfrak{X}$ can, therefore, be viewed as a probability measure on E_{k+1}.

From now on, we consider the particular case where the r.v.'s $(Y_1, Y_2, \ldots, Y_{k+1})$ have a singular Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \ldots, \alpha_{k+1})$. To simplify matters, we introduce $k+1$ mutually independent Gamma r.v.'s $Z_1, Z_2, \ldots, Z_{k+1}$ with $Z_i \sim G(\alpha_i, \beta)$, $i=1,2,\ldots,k+1$, and regard $Y_i=Z_i/Z$, where $Z=Z(\mathfrak{X})=\Sigma_i Z_i$. For any subset A of $\mathfrak{X}$ we write $Z(A)=\Sigma_{i\in A}Z_i$, and $\alpha(A)=\Sigma_{i\in A}\alpha_i$. Then $P(A)=\Sigma_{i\in A}Y_i=(Z(A))/(Z(\mathfrak{X}))$.

For any subsets A and B of $\mathfrak{X}$, let $P(A|B)$ be the (random) conditional probability of A given B defined as

$$P(A|B)=\begin{cases}P(AB)/P(B), & \text{if } P(B)>0\\ 0, & \text{if } P(B)=0.\end{cases}$$

Note that for any collection $(A_1, A_2, \ldots, A_n)$ of disjoint subsets of $\mathfrak{X}$, the r.v.'s $Z(A_1), Z(A_2), \ldots, Z(A_n)$ are mutually independent and $Z(A_i)\sim G(\alpha(A_i), \beta)$, $i=1,2,\ldots,n$.

The following is a general property of the Dirichlet distribution.

Proposition 4.1. *Let $\mathfrak{X}$ be partitioned into non-empty subsets $A_1, A_2, \ldots, A_{m+1}$, $1\leqslant m\leqslant k$. Then, $(P(A_1), P(A_2), \ldots, P(A_m))\sim D(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_{m+1}))$. Also, $(P(A_1), P(A_2), \ldots, P(A_m))$ is independent of $Z(\mathfrak{X})$.*

This follows immediately from Proposition 3.5.

For $m=1$, the above proposition can be stated as: For any two disjoint subsets A_1 and A_2 of $\mathfrak{X}$, $(Z(A_1))/(Z(A_1\cup A_2))$ is independent of $Z(A_1\cup A_2)$. Also, $(Z(A_1))/(Z(A_1\cup A_2))\sim \text{Beta}(\alpha(A_1), \alpha(A_2))$, and $Z(A_1\cup A_2)\sim G(\alpha(A_1\cup A_2), \beta)$. As a direct consequence of Proposition 4.1, we have the following:

Corollary 4.1. *The marginal distribution of the sum of any r, $1\leqslant r\leqslant k$, r.v.'s $Y_{i_1}, Y_{i_2}, \ldots, Y_{i_r}$ is Beta$(\alpha_{i_1}+\alpha_{i_2}+\cdots+\alpha_{i_r}, \alpha-\Sigma_{j=1}^{r}\alpha_{i_j})$, where $i_1, i_2, \ldots, i_r$ is any sequence of r distinct integers from $\mathfrak{X}=\{1,2,\ldots,k+1\}$.*

For any subset B of $\mathfrak{X}$ we denote by $\bar{B}$ the complement of B. The next result is a preliminary to Proposition 4.2.

Lemma 4.1. *Let B_1 and B_2 be any two subsets of $\mathfrak{X}$. Then, the r.v.'s $P(B_1)$, $P(B_2|B_1)$, $P(B_2|\bar{B}_1)$ are mutually independent.*

Proof. It suffices to show that the r.v.'s $(Z(B_1))/(Z(\mathfrak{X}))$, $(Z(B_1B_2))/(Z(B_1))$, $(Z(\bar{B}_1B_2))/(Z(\bar{B}_1))$ are mutually independent. Since the r.v.'s $Z(B_1B_2)$, $Z(B_1\bar{B}_2)$, $Z(\bar{B}_1B_2)$ and $Z(\bar{B}_1\bar{B}_2)$ are mutually independent, the pairs $(Z(B_1B_2), Z(B_1\bar{B}_2))$ and $(Z(\bar{B}_1B_2), Z(\bar{B}_1\bar{B}_2))$ of r.v.'s are independent both "within and between". Applying Proposition 4.1 to each of the two pairs we have then that the pairs $((Z(B_1B_2))/(Z(B_1)), Z(B_1))$ and $((Z(\bar{B}_1B_2))/(Z(\bar{B}_1)), Z(\bar{B}_1))$ of r.v.'s are

independent both "within and between". Thus, the r.v.'s $(Z(B_1B_2))/(Z(B_1))$, $(Z(\bar{B}_1B_2))/(Z(\bar{B}_1))$, $Z(B_1)$ and $Z(\bar{B}_1)$ are mutually independent. Applying Proposition 4.1 to the last two r.v.'s we finally conclude that the r.v.'s $(Z(B_1B_2))/(Z(B_1))$,$(Z(\bar{B}_1B_2))/(Z(\bar{B}_1))$, $(Z(B_1))/(Z(\mathcal{X}))$, and $Z(\mathcal{X})$ are mutually independent.

For any subset B of $\mathcal{X}$, define B^t as B when $t=1$ and as $\bar{B}$ when $t=0$, $i=1,2,\ldots,n$. We now state Proposition 4.2.

Proposition 4.2. *For any collection $B_1, B_2, \ldots, B_n$ of subsets of $\mathcal{X}$, the (2^n-1) r.v.'s $P(B_1), \{P(B_2|B_1^{t_1})\}, \ldots, \{P(B_n|B_1^{t_1}B_2^{t_2}\ldots B_{n-1}^{t_{n-1}})\}$ are mutually independent with $P(B_1) \sim Beta(\alpha(B_1), \alpha(\bar{B}_1))$ and $P(B_{r+1}|B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}) \sim Beta(\alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}B_{r+1}), \alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}\bar{B}_{r+1}))$, $r=1,2,\ldots,n-1$.*

Proof. For $n=2$ the proof is established in Lemma 4.1. The rest follows by induction.

Proposition 4.3. *If for any collection $B_1, B_2, \ldots, B_n$ of subsets of $\mathcal{X}$ the (2^n-1) r.v.'s $P(B_1), \{P(B_2|B_1^{t_1})\}, \ldots, \{P(B_n|B_1^{t_1}B_2^{t_2}\ldots B_{n-1}^{t_{n-1}})\}$ are mutually independent with $P(B_1) \sim Beta(\alpha(B_1), \alpha(\bar{B}_1))$ and $P(B_{r+1}|B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}) \sim Beta(\alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}B_{r+1}), \alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}\bar{B}_{r+1}))$, $r=1,2,\ldots,n-1$; then the joint distribution of 2^n r.v.'s $\{P(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$ is singular Dirichlet with parameters $\{\alpha(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$.*

Proof. Let $\{Y_{t_1t_2\ldots t_n}\}$ be a collection of 2^n r.v.'s having a singular Dirichlet distribution with parameters $\{\alpha(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$. Let $\mathfrak{y}=\{(t_1t_2\ldots t_n)\}$ be the set consisting of 2^n points. Define $C_i=\{(t_1t_2\ldots t_n): t_i=1\}$, $i=1,2,\ldots,n$, and for any subset C of $\mathfrak{y}$ write

$$Q(C)=\sum_{(t_1t_2,\ldots,t_n)\in C} Y_{t_1t_2\ldots t_n}.$$

Then Q is a random probability measure on $\mathfrak{y}$, and the joint distribution of 2^n r.v.'s $\{Q(C_1^{t_1}, C_2^{t_2}, \ldots, C_n^{t_n})\}$ is singular Dirichlet with parameters $\{\alpha(B_1^{t_1}, B_2^{t_2}, \ldots, B_n^{t_n})\}$. Furthermore, it follows from Proposition 4.2 that the (2^n-1) r.v.'s $Q(C_1), \{Q(C_2,|C_1^{t_1})\}, \ldots, \{Q(C_n|C_1^{t_1}C_2^{t_2}\ldots C_{n-1}^{t_{n-1}}\}$ are mutually independent with $Q(C_1) \sim \text{Beta}(\alpha(B_1), \alpha(\bar{B}_1))$ and $Q(C_{r+1}|C_1^{t_1}C_2^{t_2}\ldots C_r^{t_r}) \sim \text{Beta}(\alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}B_{r+1}), \alpha(B_1^{t_1}B_2^{t_2}\ldots B_r^{t_r}\bar{B}_{r+1}))$, $r=1,2,\ldots,n-1$. Thus, the joint distribution of (2^n-1) r.v.'s $\{P(B_1), P(B_2|B_1), P(B_2|\bar{B}_1), \ldots\}$ is the same as the joint distribution of (2^n-1) r.v.'s $\{Q(C_1), Q(C_2|C_1), Q(C_2|\bar{C}_1), \ldots\}$. It then follows that the joint distribution of 2^n r.v.'s $\{P(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$ is the same as the joint distribution of 2^n r.v.'s $\{Q(C_1^{t_1}C_2^{t_2}\ldots C_n^{t_n})\}$.

Remark 4.1. If the collection $\{(B_1^{t_1}B_2^{t_2}\ldots B_n^{t_n})\}$ is such that every single point subset of $\mathcal{X}=\{1,2,\ldots,k+1\}$ appears in the collection (in other words, $B_1, B_2, \ldots, B_n$ is a separating sequence), then the random probability measure P on $\mathcal{X}$ is a Dirichlet process (see Definition 5.1).

Up to this stage, only finite dimensional Dirichlet distributions were considered. A more general Dirichlet distribution may be defined as follows.

Let $\{\alpha_n\}$ be a sequence of numbers satisfying $\alpha_i > 0$ for each i and $\Sigma\alpha_i < \infty$. A sequence $\{Y_n\}$ of r.v's such that $0 < Y_i < 1$ for each i and $\Sigma Y_i = 1$ is said to have a Dirichlet distribution with parameters $\{\alpha_n\}$ if for each $k, (Y_1, Y_2, \ldots, Y_k) \sim D(\alpha_1, \alpha_2, \ldots, \alpha_k, \Sigma_{i=k+1}^{\infty}\alpha_i)$.

In the above definition one may assume $\alpha_i \geq 0$ for each i and $0 < \Sigma\alpha_i < \infty$. However, if $\alpha_i = 0$ for some i, then the corresponding $Y_i = 0$ with probability one.

The Dirichlet process on a countable infinite set $\mathcal{X} = \{1, 2, \ldots\}$ may now be defined as follows. Let $\{\alpha_n\}$ be a convergent sequence of non negative numbers. Consider the separating sequence $\{B_n\}$ of sets in $\mathcal{X}$ where $B_n = \{n\}$. Consider a sequence of mutually independent Beta r.v.'s $\{P(B_1), P(B_2|B_1), P(B_2|\bar{B}_1), \ldots\}$, where $P(B_1) \sim \text{Beta}(\alpha(B_1), \alpha(\bar{B}_1))$, $P(B_2|B_1) \sim \text{Beta}(\alpha(B_1B_2), \alpha(B_1\bar{B}_2))$, $P(B_2|\bar{B}_1) \sim \text{Beta}(\alpha(\bar{B}_1B_2), \alpha(\bar{B}_1\bar{B}_2))$, and so on. Then from Remark 4.1 it defines a Dirichlet distribution on $\{1, 2, \ldots, n\}$ for every n in a consistent manner.

In Part II we demonstrate how this constructive approach to the Dirichlet process also works in the case where $\mathcal{X}$ is a Borel space.

PART II: THE DIRICHLET PROCESS

5. Dirichlet process preliminaries

In the Bayesian analysis of non-parametric problems there is a sequence $\{X_n\}$ of independent identically distributed (i.i.d.) random variables with a common unknown distribution P, that is, given $P = \mathrm{P}$ the X_n's are i.i.d. P. Here P is regarded as the parameter and belongs to $\mathcal{P}$, the class of all probability measures on a given space $(\mathcal{X}, \mathcal{A})$. A prior for P is a probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$, where $\sigma(\mathcal{P})$ is the smallest σ-field of subsets of $\mathcal{P}$ such that the map $\mathrm{P} \to \mathrm{P}(A)$ from $\mathcal{P}$ into $[0, 1]$ is $\sigma(\mathcal{P})$-measurable $\forall A \in \mathcal{A}$. This prior may be viewed as a stochastic process $\{P(A): A \in \mathcal{A}\}$ whose sample functions are probability measures on $(\mathcal{X}, \mathcal{A})$. As in the parametric case, a class of processes satisfying the following properties is desired:

(I) It is wide enough to accommodate various shades of opinion about P.

(II) If a prior is selected from this class, then the posterior distribution given a sample of observations from P is manageable analytically, and it belongs to the class, i.e. the class is closed under "the Bayesian operation".

The class of Dirichlet process introduced by Ferguson (1973) is especially convenient since it satisfies the properties (I) and (II).

Let us look back at the definition of a random probability measure as given in the abstract of the paper. A random probability measure P on an arbitrary measurable

space $(\mathcal{X}, \mathcal{A})$ may be viewed as a measurable map from a probability space $(\Omega, \mathcal{F}, \mu)$ to the space $(\mathcal{P}, \sigma(\mathcal{P}))$. It may also be regarded as a transition function from $(\Omega, \mathcal{F})$ into $(\mathcal{X}, \mathcal{A})$. In otherwords, $P(\cdot, \cdot)$ is a measurable map from $\Omega \times \mathcal{A}$ into $[0,1]$ such that (i) for every ω in Ω, $P(\omega, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{A})$, and (ii) for every set A in $\mathcal{A}$, $P(\cdot, A)$ is a measurable function on $(\Omega, \mathcal{F})$, i.e. $P(\cdot, A)$ is a random variable with values in $[0,1]$. The distribution of P, namely μP^{-1}, is the prior probability measure on $(\mathcal{P}, \sigma(\mathcal{P}))$. Therefore, this paper can be thought of as dealing with a class of random probability measures, with a class of stochastic processes, or with a class of prior probabilities. The Dirichlet process is defined as follows.

Definition 5.1. Let α be a finite measure on $(\mathcal{X}, \mathcal{A})$. A Dirichlet process D^{α} is a random probability measure on $(\mathcal{X}, \mathcal{A})$ such that, for every partition $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$ into a finite number of measurable sets, the joint distribution of random variables $(P(A_1), P(A_2), \ldots, P(A_k))$ is singular Dirichlet with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$.

Ferguson (1973) shows through the Kolmogorov extension theorem that there exists a probability measure on $([0,1]^{\mathcal{A}}, \sigma([0,1]^{\mathcal{A}}))$ yielding the above finite dimensional Dirichlet distributions. Here $[0,1]^{\mathcal{A}}$ is the product space having for each of its factors the closed unit interval $[0,1]$, there being as many factors as elements of $\mathcal{A}$. Also, $\sigma([0,1]^{\mathcal{A}})$ is the product σ-field for $[0,1]^{\mathcal{A}}$, the σ-field generated by the measurable cylinders having a finite base. Viewing $[0,1]^{\mathcal{A}}$ as a class of set functions defined on $\mathcal{A}$ with values in $[0,1]$, each set in $\sigma([0,1]^{\mathcal{A}})$ may be defined by restrictions on a countable collection $\{p(A_n); n = 1, 2, \ldots\}$, where $\{A_n\}$ is a given countable subset of $\mathcal{A}$ and p denotes an element of $[0,1]^{\mathcal{A}}$. Observe that with $\mathcal{A}$ uncountable the single point sets in $[0,1]^{\mathcal{A}}$ are not in $\sigma([0,1]^{\mathcal{A}})$. Also, the class $\mathcal{P}$ of all probability measures on $(\mathcal{X}, \mathcal{A})$ does not belong to $\sigma([0,1]^{\mathcal{A}})$; it is not determined by a countable number of restrictions when $\mathcal{A}$ is uncountable. Thus a statement like "a Dirichlet process gives probability one to the class $\mathcal{P}$" is not meaningful.

Berk and Savage (1979) discuss other technical problems relating to measurability in addition to some fundamental difficulties with Ferguson's definition of a Dirichlet process. However, as proved by Blackwell (1973), none of these difficulties arise when $(X, \mathcal{A})$ is a Borel space. A full discussion on Blackwell's construction is given in Section 7. Some basic results on a Borel space is given in the next section.

6. Some useful results on a Borel space

Let $(\mathcal{X}, \mathcal{A})$ be a Borel space — a complete separable metric space, $\mathcal{X}$, with $\mathcal{A}$ being the σ-field generated by the open subsets of $\mathcal{X}$. Since $\mathcal{X}$ is a separable metric space, $\mathcal{A}$ is countably generated. We may, therefore, assume that there exists a countable field $\mathcal{B} = \{B_1, B_2, \ldots\}$ such that its Borel extension is $\mathcal{A}$. The family $\mathcal{B}$ forms a separating sequence, i.e. for any distinct points x_1 and x_2 in $\mathcal{X}$ there exists a set $B_n \in \mathcal{B}$ which contains either x_1 or x_2 but not both.

Consider all sequences $t = (t_1, t_2, \ldots)$ such that each $t_i = 0$ or 1. Let $T = \{t\}$ be the class of all such sequences and $\mathcal{T}$ be the σ-field for T — the σ-field generated by the

cylinders having a finite base, the so called Kolmogorov's sets. Then $(T, \mathcal{T})$ is a Borel space.

Consider the map $\xi: \mathcal{X} \to T$ defined by $\xi(x)=(\xi_1(x), \xi_2(x), \ldots)$, where ξ_i is the indicator of B_i, $i=1,2, \ldots$. Notice that ξ is a measurable map since each coordinate is measurable. Also, ξ is one–one since any two x's that agree on all ξ_i's are the same. Then we have the following

Lemma 6.1. *$\xi(\mathcal{X})$ is a Borel subset of T.*

The proof of the above lemma is a direct consequence of *The Kuratowski Theorem* (Parthasarathy (1967), Theorem 3.9). If ρ is a one–one measurable map from a Borel subset E_1 of a complete separable metric space into another complete separable metric space with $\rho(E_1)=E_2$, then E_2 is a Borel set. Also, the map ρ from E_1 onto E_2 is one–one and bimeasurable.

Let $[0,1]^\infty$ be the set of all sequences $(w_1, w_2, \ldots)$ with $0 \leqslant w_n \leqslant 1$ for each n. Note that $([0,1]^\infty, \sigma([0,1]^\infty))$ is a Borel space. Consider the map $\eta: \mathcal{P} \to [0,1]^\infty$ defined as $\eta(\mathrm{P})=\{\mathrm{P}(B_1), \mathrm{P}(B_2), \ldots\}$. The map is one–one since $\mathcal{B}$ is a field. And it is measurable since each of its coordinates is a measurable map of $\mathcal{P}$ into $[0,1]$. Let $\mathcal{S}$ be the range of the map η. From Kuratowski's theorem it then follows:

Lemma 6.2. *$\mathcal{S}$ is a Borel subset of $[0,1]^\infty$, and the map η from $\mathcal{P}$ onto $\mathcal{S}$ is one–one and bimeasurable.*

For the remainder of this section we need only to assume $\mathcal{A}$ is countably generated and contains the single point sets. Let $\mathcal{B}$ be a countable field that generates $\mathcal{A}$.

The following result is a preliminary to Proposition 6.1.

Lemma 6.3. *Let* P *be a probability measure on $(\mathcal{X}, \mathcal{A})$. Then, for every $x \in \mathcal{X}$ we have*

$$\inf_{n:\, x \in B_n} \mathrm{P}(B_n)=\mathrm{P}(\{x\}).$$

Proof. Let $C_1, C_2, \ldots$ be an enumeration of sets in $\mathcal{B}$ that contain x. Then, $\bigcap_{n=1}^{\infty} C_n=\{x\}$. Defining $D_n=C_1 C_2 \ldots C_n$, we have $\mathrm{P}(D_n) \downarrow \mathrm{P}(\{x\})$.

Consider the set of all pairs (P, x) such that $\mathrm{P} \in \mathcal{P}$ and $x \in \mathcal{X}$, that is, the product space $\mathcal{P} \times \mathcal{X}$. Equip this with product σ-field $\sigma(\mathcal{P}) \times \mathcal{A}$. Let $E=\{(\mathrm{P}, x): \mathrm{P}(\{x\})>0\}$. The following result is useful.

Proposition 6.1. $E \in \sigma(\mathcal{P}) \times \mathcal{A}$.

Proof. It suffices to show that the map $(\mathrm{P}, x) \to \mathrm{P}(\{x\})$ from $\mathcal{P} \times \mathcal{X}$ into $[0,1]$ is $\sigma(\mathcal{P}) \times \mathcal{A}$-measurable. Consider the map $H: R_2 \to R_1$ defined as

$$H(a, b)=\begin{cases} a, & \text{if } b \neq 0, \\ 1, & \text{if } b=0. \end{cases}$$

Now, the map $P \to P(B_n)$ is $\sigma(\mathcal{P})$-measurable $\forall n$, and the map $x \to \xi_n(x)$ is $\mathcal{A}$-measurable $\forall n$. Also, observe that H is a measurable map from R_2 into R_1. Therefore, the map $(P, x) \to H(P(B_n), \xi_n(x))$ is $\sigma(\mathcal{P}) \times \mathcal{A}$-measurable $\forall n$, and so is $\inf_n H(P(B_n), \xi_n(x))$. Note that $\inf_n H(P(B_n), \xi_n(x)) = \inf_{n: x \in B_n} P(B_n) = P(\{x\})$, where the last equality follows from Lemma 6.3. Thus the map $(P, x) \to P(\{x\})$ is $\sigma(\mathcal{P}) \times \mathcal{A}$-measurable.

For $P \in \mathcal{P}$, let $E_P = \{x : P(\{x\}) > 0\}$ be the P-section of E. E_P is the discrete mass points of P. Also, if P is a discrete probability measure, then E_P is the support of P. We have the following:

Proposition 6.2. *The map $\psi : \mathcal{P} \to [0, 1]$ defined as $\psi(P) = P(E_P)$, the discrete mass of P, is $\sigma(\mathcal{P})$-measurable.*

Note that the maps $P \to P_d$, the discrete part of P, and $P_d \to P_d(\mathcal{X})$ are measurable. Thus the map $P \to P_d(\mathcal{X}) = P(E_P)$ is $\sigma(\mathcal{P})$-measurable.

Corollary 6.1. *The class $\mathcal{P}_0$ of all discrete probability measures on $(\mathcal{X}, \mathcal{A})$ is $\sigma(\mathcal{P})$-measurable.*

Observe that $\mathcal{P}_0 = \{P \in \mathcal{P} : \psi(\mathcal{P}) = 1\} \in \sigma(\mathcal{P})$. For further details refer to Dubins and Freedman (1964).

7. Existence of a Dirichlet process

We proceed to prove the existence of a Dirichlet process D^α on a Borel space $(\mathcal{X}, \mathcal{A})$ corresponding to any finite measure α on $\mathcal{A}$. Choose and fix a countable field $\mathcal{B} = \{B_1, B_2, \dots\}$ of sets in $\mathcal{X}$ such that $\mathcal{B}$ is a generator of the σ-field $\mathcal{A}$. The map $x \to \xi(x) = \{\xi_1(x), \xi_2(x), \dots\}$, where ξ_n is the indicator of B_n, is then a one–one bimeasurable map of $(\mathcal{X}, \mathcal{A})$ into $(T, \mathcal{T})$. A probability measure Q on $(T, \mathcal{T})$ defines a probability measure $P = Q\xi$ on $(\mathcal{X}, \mathcal{A})$ provided $Q[\xi(\mathcal{X})] = 1$.

To simplify our notations we denote a typical point $(t_1, t_2, \dots, t_n)$ of the product space $T_n = \{0, 1\}^n$ by s_n. By $s_n 0$ we denote the point in T_{n+1} that is obtained by augmenting s_n by 0, that is, $s_n 0 = (t_1, t_2, \dots, t_n, 0)$, and similarly for $s_n 1$. Finally, we denote by $[s_n]$ the cylinder set of all points in T whose first n coordinates form the vector s_n. For example, [0] is the set of all $t \in T$ such that $t_1 = 0$.

It is easily seen that a probability measure Q on $(T, \mathcal{T})$ is uniquely defined by a sequence of blocks ω of numbers in the closed unit interval [0, 1]:

$$\omega = \{u, (u_0, u_1), (u_{00}, u_{01}, u_{10}, u_{11}), \dots\}, \tag{7.1}$$

where $u = Q([1])$, $u_0 = Q([0, 1]|[0])$, $u_1 = Q([1, 1]|[1])$, $u_{00} = Q([0, 0, 1]|[0, 0])$ and so on, a typical term of the $(n+1)$th block of the sequence of blocks being

$$u_{s_n} = Q([s_n 1]|[s_n]), \quad s_n \in T_n.$$

Let Ω denote the space of all sequence of blocks ω with its coordinates lying in [0, 1].

The probability measure on $(T, \mathcal{T})$ that coexists with each $\omega \in \Omega$ is denoted by Q_ω. If Ω is equipped with the product σ-field $\sigma(\Omega)$, then the map $\omega \to Q_\omega$ defines a transition function from $(\Omega, \sigma(\Omega))$ to $(T, \mathcal{T})$. If $(\Omega, \sigma(\Omega))$ is equipped with a probability measure μ, then we have a random probability measure on $(T, \mathcal{T})$ which we denote by Q_μ. How do we choose μ so that $Q_\mu \xi$ is a Dirichlet process on $(\mathcal{X}, \mathcal{A})$ with parameter α?

For an arbitrary but fixed n, consider the partition $\{B^{s_n}: s_n \in T_n\}$ of $\mathcal{X}$, where by B^{s_n} we denote the set $B_1^{t_1} B_2^{t_2} \dots B_n^{t_n}$. If P_μ is D^α on $(\mathcal{X}, \mathcal{A})$, then the joint distribution of the 2^n r.v.'s $\{P_\mu(B^{s_n}): s_n \in T_n\}$ is singular Dirichlet with parameters $\{\alpha(B^{s_n}): s_n \in T_n\}$. Invoking Proposition 4.2 we then have

$$P_\mu(B_1), P_\mu(B_2|\bar{B}_1), P_\mu(B_2|B_1), P_\mu(B_3|\bar{B}_1\bar{B}_2), \dots \tag{7.2}$$

are mutually independent random variables with

$$P_\mu(B_1) \sim \text{Beta}\big(\alpha(B_1), \alpha(\bar{B}_1)\big), \quad \text{and}$$

$$P_\mu(B_{m+1}|B^{s_m}) \sim \text{Beta}\big(\alpha(B^{s_m 1}), \alpha(B^{s_m 0})\big), \quad s_m \in T_m, \quad m = 1, 2, \dots n. \tag{7.3}$$

Observe that the map $\xi: \mathcal{X} \to T$ transforms B_1 to [1], B^{s_m} to $[s_m]$, $P_\mu(B_1)$ to $Q_\mu([1])$, and so on. It is, therefore, clear that if under μ the coordinates of ω are mutually independent and are distributed as

$$u \sim \text{Beta}\big(\alpha(B_1), \alpha(\bar{B}_1)\big), \quad \text{and}$$

$$u_{s_n} \sim \text{Beta}\big(\alpha(B^{s_n 1}), \alpha(B^{s_n 0})\big), \quad n = 1, 2, \dots, \tag{7.4}$$

then (7.2) and (7.3) hold true for all n.

Theorem 7.1 (Blackwell (1973)). *If under μ coordinates of ω are mutually independent and* (7.4) *holds, then P_μ is D^α on $(\mathcal{X}, \mathcal{A})$.*

Proof. Since (7.2) and (7.3) hold, it follows (Remark 4.1) that $P_\mu(B_n) \sim \text{Beta}(\alpha(B_n), \alpha(\bar{B}_n))$ for all n. Since $\mathcal{B}$ is a field, $\mathcal{X} = B_n$ some n. Therefore, $P_\mu(\mathcal{X}) = 1$ a.s. $[\mu]$. This proves that P_μ is a random probability measure on $(\mathcal{X}, \mathcal{A})$.

To prove that $P_\mu(A) \sim \text{Beta}(\alpha(A), \alpha(\bar{A}))$ for each $A \in \mathcal{A}$ we proceed as follows: The map $\mathrm{P} \to (\mathrm{P}(B_1), \mathrm{P}(B_2), \dots)$ from $\mathcal{P}$ to $\mathcal{S}$ is one–one and bimeasurable (Lemma 6.2). For any $A \in \mathcal{A}$, the map $\mathrm{P} \to \mathrm{P}(A)$ from $\mathcal{P}$ to $[0, 1]$ is measurable. Hence there exists a measurable map $h_A: \mathcal{S} \to [0, 1]$ such that $h_A(\mathrm{P}(B_1), \mathrm{P}(B_2), \dots) = \mathrm{P}(A)$ for all $\mathrm{P} \in \mathcal{P}$. For each n, the joint distribution of $P_\mu(B_1), P_\mu(B_2), \dots, P_\mu(B_n)$ is well defined in terms μ. And for different n these joint distributions are mutually consistent. The Kolmogorov extension theorem, therefore, guarantees that the joint distribution of the whole sequence $P_\mu(B_1), P_\mu(B_2), \dots$, is well defined. If we denote this joint distribution by Π_μ, then $P_\mu(A) \sim \Pi_\mu h_A^{-1}$.

Consider now the hypothetical situation where we might have started with $\mathcal{B}^* = \{A, B_1, B_2, \dots\}$ as generator of $\mathcal{A}$. Proceeding as before, we would then have defined a random probability measure P_μ^* on $(\mathcal{X}, \mathcal{A})$. Under μ, the random probability measures P_μ and P_μ^* on $(\mathcal{X}, \mathcal{A})$ are the same, and therefore the joint

distribution of $(P_\mu(A), P_\mu(B_1), \ldots)$ is the same as the joint distribution of $(P_\mu^*(A), P_\mu(B_1), \ldots)$. For the random probability measure P_μ^* it is clear that $P_\mu^*(A) \sim$ Beta$(\alpha(A), \alpha(\bar{A}))$ and $(P_\mu^*(B_1), P_\mu^*(B_2), \ldots) \sim \Pi_\mu$. Therefore, $\Pi_\mu h_A^{-1}$ is Beta$(\alpha(A), \alpha(\bar{A}))$. This proves that $P_\mu(A) \sim$ Beta$(\alpha(A), \alpha(\bar{A}))$ for all $A \in \mathcal{A}$.

The above argument goes through, word for word, for an arbitrary measurable partiation $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$ leading to the conclusion that the r.v.'s $(P_\mu(A_1), P_\mu(A_2), \ldots, P_\mu(A_k))$ have a singular Dirichlet distribution with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$. This proves that P_μ is D^α on $(\mathcal{X}, \mathcal{A})$.

8. Support of the Dirichlet process

The existence theorem of the previous section may be restated as:

Theorem 8.1. *If $(\mathcal{X}, \mathcal{A})$ is a Borel space, then, for each finite measure α on $(\mathcal{X}, \mathcal{A})$, there exists a probability measure D^α on $(\mathcal{P}, \sigma(\mathcal{P}))$ such that with $P \sim D^\alpha$, the r.v.'s $(P(A_1), P(A_2), \ldots, P(A_k))$ have singular Dirichlet distribution with parameters $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$ for any measurable partition $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$.*

Let $\mathcal{P}_0$ be the family of discrete probability measure on $(\mathcal{X}, \mathcal{A})$. That $\mathcal{P}_0$ belongs to $\sigma(\mathcal{P})$ has been noted in Corollary 6.1.

Theorem 8.2. *If $\mathrm{P} \sim D^\alpha$, then almost every realization of P is a discrete probability measure on $(\mathcal{X}, \mathcal{A})$, that is,*

$$D^\alpha(\mathcal{P}_0) = 1.$$

Historical Note: Ferguson (1973) gave a rather involved argument to prove this result. Blackwell (1973) and Blackwell and MacQueen (1973) gave alternative arguments for the same result. The proof given here is a streamlined version of an ingenious argument given by Berk and Savage (1979).

Consider the pair (P, X) of random entities such that (i) $P \sim D^\alpha$ and (ii) $X|P \sim P$, that is, conditional on $P = \mathrm{P}$, the probability distribution of X on $(\mathcal{X}, \mathcal{A})$ is P. Let Δ^α denote the joint distribution of (P, X) on the product space $(\mathcal{P} \times \mathcal{X}, \sigma(\mathcal{P}) \times \mathcal{A})$. The marginal distribution of X is then easily verified to be the normalized measure $\tilde{\alpha} = \alpha/\alpha(\mathcal{X})$. It is well known (see Ferguson (1973)) that $P|X = x \sim D^{\alpha + \delta_x}$, where δ_x denotes the degenerate probability measure with its whole mass concentrated at x.

We have noted earlier (Proposition 6.1) that $E = \{(\mathrm{P}, x) : \mathrm{P}(\{x\}) > 0\}$ belongs to $\sigma(\mathcal{P}) \times \mathcal{A}$. The following proposition is a preliminary to the proof of Theorem 8.2.

Proposition 8.1. $\Delta^\alpha(E) = 1$.

Proof. Writing E^x for the x-section of E, we have

$$\Delta^\alpha(E) = \int D^{\alpha + \delta_x}(E^x) \, \mathrm{d}\tilde{\alpha}(x).$$

Now, E^x is the set of all $\mathrm{P} \in \mathcal{P}$ such that $\mathrm{P}(\{x\}) > 0$. Under the distribution $D^{\alpha + \delta_x}$, the random variable $P(\{x\})$ is positive with probability one— this is because the $\alpha + \delta_x$ measure of the set $\{x\}$ is positive. Therefore, $D^{\alpha + \delta_x}\{\mathrm{P} : \mathrm{P}\{x\} > 0\} = 1$ for all x. In other words, $\Delta^\alpha(E) = 1$.

Proof of Theorem 8.2. Consider now the P-section $E_{\mathrm{P}}=\{x:\mathrm{P}(\{x\})>0\}$ of the set E. Since X, given $P=\mathrm{P}$, is distributed as P we have

$$\Delta^{\alpha}(E)=\int\psi(\mathrm{P})\,\mathrm{d}D^{\alpha}(\mathrm{P}),$$

where $\psi(\mathrm{P})=\mathrm{P}(E_{\mathrm{P}})$ is the discrete mass of P. Since $\Delta^{\alpha}(E)=1$, we at once have $\psi(P)=1$ a.s. $[\mathrm{D}^{\alpha}]$. But $\{\mathrm{P}:\psi(\mathrm{P})=1\}=\mathcal{P}_0$.

Let $\mathcal{V}=\{V\}$ be the collection of all open sets in $\mathcal{X}$. Since $\mathcal{X}$ is a separable metric space, there exists a countable subcollection $\{V_1,V_2,\dots\}$ of open sets such that every V contains some V_n. Let $\mathcal{P}'$ be the collection of all $\mathrm{P}\in\mathcal{P}$ such that $\mathrm{P}(V)>0$ for all $V\in\mathcal{V}$. Similarly, let $\mathcal{P}_n=\{\mathrm{P}:\mathrm{P}(V_n)>0\}$. It is then clear that $\mathcal{P}'=\bigcap_{n=1}^{\infty}\mathcal{P}_n$.

Theorem 8.3. *If $\alpha(V)>0$ for all $V\in\mathcal{V}$, then $D^{\alpha}(\mathcal{P}')=1$.*

Proof. Since $P(V_n)\sim\mathrm{Beta}(\alpha(V_n),\alpha(\overline{V}_n))$ and $\alpha(V_n)>0$ it follows that $P(V_n)>0$ a.s. $[D^{\alpha}]$, that is, $D^{\alpha}(\mathcal{P}_n)=1$. Therefore, $D^{\alpha}(\mathcal{P}')=1$.

The set $\mathcal{P}_0\cap\mathcal{P}'$ is the collection of all discrete probability measure P on $(\mathcal{X},\mathcal{A})$ such that the mass points of P are everywhere dense in $\mathcal{X}$. Putting Theorems (8.2) and (8.3) together we finally have:

Theorem 8.4. *If $P\sim D^{\alpha}$ and the α-measure of every open subset of $\mathcal{X}$ is positive, then for almost every realization* P *of P it is true that* P *is discrete with its mass points everywhere dense in $\mathcal{X}$.*

Further properties of the Dirichlet process will be discussed in a forthcoming note.

Acknowledgement

The authors are greatly indebted to Professor David Blackwell for the benefit of some prolonged discussions and consultations during the Fall and Winter Quarters of 1978–1979 when he was visiting the Florida State University. The authors are also indebted to Professor J. Sethuraman for his useful discussions.

References

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā A*, **15**, 337–380.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā A* **37**, 1–71.

Berk, R.H. and Savage, I.R. (1979). Dirichlet process produce discrete measures: An elementary proof. *Contributions to Statistics. Jaroslav Hájek Memorial Volume*, pp. 25–31. Academia, North-Holland, Prague.

Blackwell, D. (1973). Discreteness of Ferguson Selections. *Ann. Statist.* **1**, 356–358.

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–355.

Dubins, L. and Freedman, D. (1964). Measurable sets of measures. *Pacific J. Math.* **14**, 1211–1222.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Parthasarathy, K.R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.

Wilks, S.S. (1962). *Mathematical Statistics*. Wiley, New York.

Sankhyā : The Indian Journal of Statistics
1983, Volume 45, Series A, Pt. 3, pp. 324–337.

CONDITIONAL INDEPENDENCE IN STATISTICS

By D. BASU[1]
The Florida State University
and
CARLOS A. B. PEREIRA[2]
Universidade de São Paulo

SUMMARY. The theory of conditional independence is explained and the relations between ancillarity, sufficiency, and statistical independence are discussed in depth. Some related concepts like specific sufficiency, bounded completeness, and splitting sets are also studied in some details by using the language of conditional independence.

1. Introduction

The notion of conditional independence is a central theme in statistics. In a series of articles Dawid (1979a, 1979b; 1980), Florens and Mouchart (1977), and Mouchart and Rolin (1978) have explained at length the grammar of Conditional Independence as a language of statistics. This article is a further elucidation on the subject and is in part of an expository nature.

The statistical perspective of this article is that of a Bayesian. A problem begins with a parameter (State of Nature) θ with its prior probability model $(\theta, \mathcal{B}, \xi)$ that exists only in the mind of the investigator. There is an observable X with an associated statistics model $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$. Writing $\omega = (\theta, X)$, $(\Omega, \mathcal{F}) = (\Theta \times \mathcal{X}, \mathcal{B} \times \mathcal{A})$, and Π for the joint distribution of (θ, X), there then exists a subjective model $(\Omega, \mathcal{F}, \Pi)$ for ω. Hidden behind the wings of the Bayesian probability model $(\Omega, \mathcal{F}, \Pi)$ are the four models

(i) the prior model $(\Theta, \mathcal{B}, \xi)$,

(ii) the statistical model $(\mathcal{X}, \mathcal{A}, \{P_\theta : \theta \in \Theta\})$,

(iii) the posterior model $(\Theta, \mathcal{B}, \{\xi_x : x \in \mathcal{X}\})$,

and (iv) the predictive model $(\mathcal{X}, \mathcal{A}, P)$ where P is the marginal or predictive distribution of X.

[1]Research partially supported by NSF Grant No. 79-04693.

[2]Research supported by CNPq, CAPES and USP—Brazil.

Key Words : Conditional independence, ancillarity, sufficiency, Markov property, (strong) identification, splitting sets, measurable separability, specific sufficiency, variation independence.

AMS classification : Primary 62B05, Secondary 62B20.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_32,

In statistics the phenomenon of conditional independence manifests itself in a natural fashion. The statistical model that is most commonly in use is that of a sequence $\boldsymbol{X} = (X_1, X_2, \ldots)$ of observables that are independently and identically distributed (i.i.d) for each given value of θ. It was DeFinnetti (1937) who emphasized that, in view of the fact that θ is not fully known, it is appropriate to regard the sequence of X_i's not as i.i.d random variables but as an exchangeable process. The fact that the X_i's are conditionally i.i.d implies that they are positively dependent in the sense that the covariance (when exists) between any pair is non-negative. More specifically, $\text{cov}(X_i, X_j) = \text{var}(E(X_1 \mid \theta))$.

Consider for example the particular case where $X_1, X_2, \ldots, X_n$ are i.i.d with common distribution $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$ not fully known. In almost every text book of statistics it is proved that $\bar{X} = \frac{1}{n}\Sigma X_i$ is stochastically independent of $s^2 = \frac{1}{n}\Sigma(X_i - \bar{X})^2$. Does it mean that $\bar{X}$ when observed, carries no information about s^2 ? That the answer cannot be "yes" is easily seen as follows. Suppose that the sample size $n = 25$ and that our partial knowledge about $\theta = (\mu, \sigma^2)$ is as follows, $\mu = 0$ or 1 and $\sigma^2 = 1$ or 100, (that is $\Theta = \{(0, 1), (0, 100), (1, 1), (1, 100)\}$). Suppose now that $\bar{X}$ is observed and is equal to 2.1. This observation generates the four likelihoods $L(0, 1)$, $L(1, 1)$, $L(0, 100)$ and $L(1, 100)$ where $L(0, 1) = \frac{5}{\sqrt{2\pi}} \exp\left\{-\frac{25}{2}(2\,1)^2\right\}$ and so on. The relative likelihoods work out roughly as 10^{-17}, 1, $2(10)^5$ and $3(10)^5$ respectively. Thus, it is intuitive that the observation $\bar{X} = 2.1$ almost categorically rules out the points $(0, 1)$ and $(1, 1)$. Then $\bar{X}$ and s^2, even though they are conditionally independent given θ, are in effect highly dependent.

The three entities $\theta = (\mu, \sigma^2)$, $T = (\bar{X}, s^2)$ and $\boldsymbol{X} = (X_1, \ldots, X_n)$ in this order, have the Markov property in the sense that, given θ and T, the conditional distribution of $\boldsymbol{X}$ depends on (θ, T) only through T. This is the sufficiency property of T as recognized by Fisher (1920, 1922). Kolomogrov (1942) gave a Bayesian characterization of the notion of sufficiency by noting that irrespective of the choice of the prior distribution ξ for the parameter θ, the posterior distribution $\xi_{\boldsymbol{X}}$ of θ depends on $\boldsymbol{X}$ only through T. Note that the Fisher characterization of sufficiency is made only in terms of the statistical model for $\boldsymbol{X}$ whereas the Kolmogorov characterization is made in terms of a large family of Bayesian models $(\Omega, \mathcal{F}, \Pi)$ for $\omega = (\theta, X)$. (See Basu, 1977 and Cheng, 1978 for further details on these characterizations).

Fisher regarded a sufficient statistic T as one that summarizes in itself all the available relevant information in the sample $\boldsymbol{X}$ about the parameter θ. He called a statistic $Y = Y(X)$ ancillary if the conditional distribution of Y given θ does not involve θ (is the same for all values of θ). For example, the statistic $\Sigma \frac{(X_i - \bar{X})^4}{s^4}$ is ancillary. In a series of articles Basu (1955, 1958, 1959, 1964, 1967) studied the phenomena of sufficiency, ancillarity and conditional independence from various angles. In these articles, Basu's viewpoint was non-Bayesian in the sense that he did not introduce a prior distribution ξ for θ. Mouchart and Rolin (1978) studied in depth the familiar Basu theorems on sufficiency, ancillarity and conditional independence from the view point of the Bayesian model.

In this paper we too review Basu's results and also the two-parameter problem from the Bayesian perspective. Many results are stated without proof since the proofs involve standard measure theoretic arguments and can be found for instance in Loeve (1977).

2. Notation and preliminaries

Let $(\Omega, \mathscr{F}, \pi)$ be the basic probability space. By a "random object" X we mean a measurable map $\omega \to X(\omega)$ of $(\Omega, \mathscr{F})$ into another measurable space $(\mathscr{X}, \mathscr{A})$. The sub σ-algebra (to be called subfield) of X events $\{X^{-1}(A); A \in \mathscr{A}\}$ will be denoted by $\mathscr{F}_X$. The two probability spaces $(\Omega, \mathscr{F}_X, \Pi)$ and $(\mathscr{X}, \mathscr{A}, \Pi^{-1})$ are indistinguishable in a certain sense, and so we shall, as a rule, identify a random object X with the induced subfield $\mathscr{F}_X$ of $\mathscr{F}$. In that way, one could say that random objects are generators of subfields. Examples of random objects include random variables, random vectors etc.

For any two subfields $\mathscr{F}'$ and $\mathscr{F}''$ of $\mathscr{F}$, $\mathscr{F}' \vee \mathscr{F}''$ denotes the smallest subfield of $\mathscr{F}$ that contains both $\mathscr{F}'$ and $\mathscr{F}''$. The smallest subfield of $\mathscr{F}$ that contains all null sets of $\mathscr{F}$ (a set N is null if $\pi(N) = 0$) is denoted by $\bar{\mathscr{F}}_0$, and write $\mathscr{F}_0 = \{\varphi, \Omega\}$, the trivial subfield.

A subfield of $\mathscr{F}$ is said to be completed if it contains $\bar{\mathscr{F}}_0$. For any subfield $\mathscr{F}'$ of $\mathscr{F}$ its completion $\bar{\mathscr{F}}'$ is defined by

$$\bar{\mathscr{F}}' = \mathscr{F}' \vee \bar{\mathscr{F}}_0.$$

For a random object X, the notation $X \in \mathscr{F}'$ indicates that $\mathscr{F}_X \subseteq \mathscr{F}'$ and X is said to be essentially $\mathscr{F}'$ measurable. A random variable is a random object with range $(R_1, \mathscr{B}_1)$ where R_1 is the real line and $\mathscr{B}_1$ is the Borel

σ-algebra. A random variable f is said to be bounded if there exists $a \in R_1$ such that $\pi\{\omega : |f(\omega)| \leqslant a\} = 1$. In the sequel, all random variables will be regarded as bounded unless stated otherwise and the use of small letters shall be restricted to their representation. The notation $f \subset X$ indicates that the random variable f is ess—$\mathcal{F}_X$ measurable. In the same spirit, for two random objects X and Y, we write $X \subset Y$ to indicate that $\bar{\mathcal{F}}_X \subset \bar{\mathcal{F}}_Y$. The class of all bounded random variables on $(\Omega, \mathcal{F}, \Pi)$ is denoted by L_∞ and $L_\infty(X)$ denotes the class of all ess—$\mathcal{F}_X$ measurable random variables. Here and for the rest of this article, equality of two random variable means essential equality; that is $f = g$ means $\{\omega : f(\omega) \neq g(\omega)\}$ is a null set.

The conditional expectation of f, given a random object X, is a random variable $f^{*X} \in L_\infty(X)$ such that

$$\int f g d\pi = \int f^{*X} g d\pi \quad \forall g \in L_\infty(X)$$

Another notation for f^{*X} is $E(f \mid X)$. When the conditioning random object X is implicit in the context, f^* is substituted for f^{*X}. The map $f \to f^*$ of L_∞ to $L_\infty(X)$ is linear, constant preserving, idempotent and is a contraction in the L_p norm if $p \geqslant 1$.

3. Conditional independence : definition, properties and the drop/add principles

In this section the definition and properties of conditional independence are briefly discussed.

Three random objects X, Y and Z are being considered and in this section * stands for $^*Z$ operator.

Definition 1. (Intuition) : The random objects X and Y are conditionally independent given Z (in symbols $X \perp\!\!\!\perp Y \mid Z$) if for any $f \in L_\infty(X)$

$$E(f \mid Y, Z) = f^{*YZ} = f^*$$

Note that to say $X \perp\!\!\!\perp Y \mid Z$ is equivalent to say that $X \mid (Y, Z)$ has the same conditional disribution as $X \mid Z$. This is the intuition behind the definition. In the case where Z is essentially a generator of $\mathcal{F}_0$, we obtain the independence of X and Y in the usual sense. In this case the notation is $X \perp\!\!\!\perp Y$.

Definition 1*a*. (Symmetric) : The random objects X and Y are conditionally independent given Z if for any $f \in L_\infty(X)$ and $g \in L_\infty(Y)$

$$(fg)^* = f^* g^*.$$

The following well known theorem gives the equivalence of the two definitions showing that $X \amalg Y \mid Z$ implies $Y \amalg X \mid Z$ which is not clear by looking at definition 1.

Theorem 1 : *Definitions* 1 *and* 1*a are equivalent.*

The concept of conditional independence (c.i.) gives rise to many questions. Among them are questions involving the drop and add (Drop/Add) principles. Suppose that X, Y, Z, W, X_1, Z_1, are random objects such that $X \amalg Y \mid Z$, $X_1 \subset X$, $Z_1 \subset Z$. What can be said about the relationship $\amalg$ if X_1 is substituted for X, Z_1 for Z, (Y, W) for Y or (Z, W) for Z ? In other words, can $\mathscr{F}_X$, $\mathscr{F}_Y$ or $\mathscr{F}_Z$ be essentially reduced or enlarged without destroying the c.i. relation ? In general the answer is no. However for certain kinds of reductions and enlargements, the relationship will be preserved. To indicate that the relationship $\amalg$ does not hold we write not $\amalg$.

It is not difficult to find examples showing that arbitrary reductions or enlargements of $\mathscr{F}_Z$ = the conditioning subfield, may destroy the c.i. property. With the example of the normal distribution presented in the introduction, we have $\bar{X} \amalg s^2 \mid \theta$ but not $\bar{X} \amalg s^2$. In yet another statistical context, suppose that θ is the unknown (real) parameter of interest and let X and Y be two i.i.d. random variables with common uniform distribution on the interval $(\theta-\frac{1}{2}, \theta+\frac{1}{2})$. Since θ is unknown, we can only say that $X \in \mathscr{R}$, $Y \in \mathscr{R}$. However, after X has been observed equal to x, we would for sure know that $x-1 \leqslant Y \leqslant x+1$. This shows clearly that $X \amalg Y \mid \theta$ but not $X \amalg Y$. To show that we can have $X \amalg Y$ and X not $\amalg Y \mid Z$, let W and Z be two i.i.d. $N(0, 1)$ variables and take $X = Z-W$ and $Y = Z+W$.

The above examples may be viewed as cases of Simpson's paradox (see Dawid, 1979a). The paradox, however, is much stronger. For instance, let W and Z be two independent normal variables with zero means. As before define $X = Z-W$ and $Y = Z+W$. The correlation between X and Y is given by $\rho(X, Y) = \dfrac{1-\delta}{1+\delta}$ where $\delta = \dfrac{\text{var}(W)}{\text{var}(Z)}$. Given Z, the conditional correlation is clearly equal to -1. On the other hand, δ may be taken very small to make $\rho(X, Y)$ close to 1. This shows that we can have cases where X and Y are strongly positive (negative) dependent but, when Z is given, X and Y turn to be strongly negative (positive) dependent. The problem of dependence inversion is discussed in depth in Lindley and Novick (1981). The following example may be of relevance for applied statisticians.

Example : Suppose that an urn contains θ (unknown) white balls in a total of N (known) balls. A sample, without replacement, of n balls is selected

from this urn. Let X be the number of white balls in the sample which implies that $Y = \theta - X$ is the number of white balls that remain in the urn. Since $\rho(X, Y \mid \theta) = -1$, we have X not $\amalg\, Y \mid \theta$. It can be proved (see Whitt, 1979) that $X \amalg Y$ iff θ has binomal (prior) distributions with fixed parameters N and $p \in (0, 1)$. On the other hand, if a priori, $\Pr\{\theta = 0\} = \Pr\{\theta = N\} = \frac{1}{2}$, then $\rho(X, Y) = 1$ showing an extreme inverted dependence.

The essence of Drop/Add principles for c.i. is contained in the following propositions.

Proposition I : If $X \amalg Y \mid Z$, then for every $X' \subset X$, we have :

(i) $X' \amalg Y \mid Z$

(ii) $X \amalg Y \mid (Z, X')$

(iii) $(X, Z) \amalg (Y, Z) \mid Z$.

By way of explanation, if $X \amalg Y \mid Z$, then the relation $\amalg$ is preserved when (i) X and Y are increased (Add) by any essential part of Z, (ii) Z is increased (Add) by any essential part of X or Y, and (iii) X and Y are arbitratily reduced (Drop).

To end this section we present an extreme case of Drop/Add principles for the conditioning random object. It appears in Dawid (1980) and it was originally introduced by G. Udny Yule in terms of collapsibility of contingency tables. It must clarify the problems with Simpson's paradox for contingency tables.

Proposition 2 : Let X, Y and Z be three random objects such that $\mathcal{F}_Z = \{\phi, \Omega, A, A^C\}$ with $0 < \Pi(A) < 1$. If $X \amalg Y$ and $X \amalg Y \mid Z$, then either $X \amalg Z$ or $Y \amalg Z$.

The proof becomes simple when we recognize the following result.

Lemma : *If* $X \amalg Y$ *and* $X \amalg Y \mid Z$, *then for every atom* A *of* Z *with* $\Pi(A) > 0$, *we have*

$$E(I_A \mid (X, Y)) = \frac{1}{\Pi(A)} E(I_A \mid X) E(I_A \mid Y).$$

Proof of lemma : Let B, C, be two sets such that $I_B \subset X$, $I_C \subset Y$

$$\begin{aligned}\int_{BC} E(I_A \mid X)E(I_A \mid Y)d\Pi &= \int E(I_{AB} \mid X)E(I_{AC} \mid Y)d\Pi \\ &= \int E(I_{AB} \mid X)d\,\Pi \int E(I_{AC} \mid Y)d\Pi = \Pi(AB)\Pi(AC) \\ &= [\Pi(A)]^2\Pi(B \mid A)\Pi(C \mid A) = [\Pi(A)]^2\Pi(BC \mid A) \\ &= \Pi(A)\Pi(ABC) = \Pi(A) \int_{BC} E(I_A \mid (X, Y))d\Pi.\end{aligned}$$

▲ 3–10

Since sets of the form BC generate $\mathcal{F}_{XVY}$ a standard argument completes the proof.

Proof of proposition : Let $p = \Pi(A)$. From lemma we have

$$E(I_A|(X, Y)) = \frac{E(I_A|X)E(I_A|Y)}{p}$$

and

$$E(I_{A^C}|(X, Y)) = \frac{[1-E(I_A|X)][1-E(I_A|Y)]}{1-p}$$

and consequently

$$\left(1-\frac{E(I_A|X)}{p}\right)\left(1-\frac{E(I_A|Y)}{p}\right) = 0$$

Since $X \amalg Y$, this equation holds only if $\frac{E(I_A|X)}{p} = 1$ or $\frac{E(I_A|Y)}{p} = 1$.

4. Bayesian inference : sufficiency, ancillarity and independence

As discussed in Dawid (1979a, 1980) many of the important Statistical Concepts are simply manifestations of the concept of conditional independence. In this section we use the framework of conditional independence to describe the Bayesian version of those statistical concepts and their properties. First we review some of the structures involved.

Let $(\mathcal{X}, \mathcal{A})$ be the usual Sample Space and $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures on $(\mathcal{X}, \mathcal{A})$ where Θ is the usual parameter "Space". In addition the Bayesian considers a (prior) probability space $(\Theta, \mathcal{B}, \xi)$ where $\mathcal{B}$ is a σ-algebra of subsets of Θ such that $P_\theta(A)$ is a $\mathcal{B}$-measurable function for every fixed $A \in \mathcal{A}$. Clearly, the choice of the prior model is not completely arbitrary, since it has to match the statistical structure on the $\mathcal{B}$-measurability of $P_\theta(A)$.

We then consider the probability space $(\Omega, \mathcal{F}, \Pi)$, where now $\Omega = \Theta \times \mathcal{X}$, $\mathcal{F} = \mathcal{B} \times \mathcal{A}$ and Π is defined by

$$\Pi(F) = \int_\Theta P_\theta(F^\theta) d\xi(d\theta)$$

where $F^\theta = \{x : (\theta, x) \in F\}$. The marginal on $\mathcal{X}$ is defined by

$$P(A) = \Pi(\Theta \times A) \text{ for every } A \in \mathcal{A}.$$

Let X and Y be two random objects on $(\Omega, \mathcal{F})$. We say that X represents the sample and Y represents the parameter if

$$\mathcal{F}_X = \{\Theta \times A, A \in \mathcal{A}\}$$

and

$$\mathcal{F}_Y = \{B \times \mathcal{X}, B \in \mathcal{B}\}.$$

In addition to X and Y defined above, consider two random objects X_1, and X_2 such that $(X_1, X_2) \subseteq X$. The Bayesian version of the concepts of sufficiency and ancillarity is contained in the following.

Definition 2 :

(a) If $X \amalg Y | X_1$ we say X_1 is sufficient for X with respect to Y.

(b) If $X_2 \amalg Y$ we sav that X_2 is ancillary with respect to Y.

The classical concept of statistical independence between X_1 and X_2 nas its Bayesian version as

(c) $X_1 \amalg X_2 | Y$.

Basu (1955, 1958) speculates under what conditions two of the three relation (a), (b) and (c) imply the third. In this section we study Basu's theorems under the Bayesian framework. The next result which is Basu's first conjecture presents conditions to have (b) and (c) implying (a).

Proposition 3 : If in addition to $X_2 \amalg Y$ and $X_1 \amalg X_2 | Y$ we have $X \amalg Y | (X_1, X_2)$ then $X \amalg Y | X_1$.

Proof: Note that $X_2 \amalg Y$ and $X_1 \amalg X_2 | Y$ implies $X_2 \amalg Y | X_1$, also $X_2 \amalg Y | X_1$ and $X \amalg Y | (X_1, X_2)$ implies $X \amalg Y | X_1$.

Arguing similarly it is easy to see that if $X_1 \amalg X_2$, then (a) implies (b) and (c). The meaning of $X_1 \amalg X_2$ in classical statistics however is void.

Note that Proposition 3 gives conditions for reducing (Drop) the conditioning random object. Actually all of Basu's theorems are cases of Drop/Add principles.

Basu (1955) stated that any statistic independent of a sufficient statistic is ancillary. Later on Basu (1958) presented a counter example and recognized the necessity of an additional condition (connectedness) on the family $\{P_\theta : \theta \in \Theta\}$ of probability measures. Koehn and Thomas (1975) strengthened this result by introducing a necessary and sufficient condition on the family. More recently Basu and Cheng (1979), generalizing results of Pathak (1975) showed the equivalence between these two conditions in coherent models.

The following theorem is a Bayesian version of the result of Koehn and Thomas (1975).

Theorem 2 : *Let* $X_1 \subset X$ *be a sufficient random object (i.e.* $X \amalg Y \mid X_1$*). The random object* $Y \wedge X_1$ *is essentially a constant (i.e.* $F_{Y \wedge X_1} = F_0$*) iff* $X_2 \amalg Y$ *whenever* $X_2 \subset X$ *and* $X_1 \amalg X_2 \mid Y$ *(i.e.* X_2 *is ancillary if* X_1 *and* X_2 *are statistically independent).*

Proof : $E(I_A \mid Y) = E(I_A \mid X_1, Y) = E(I_A \mid X_1)$ by $X \amalg Y \mid X_1$. Now since, $X_1 \wedge Y = F_0$ it follows that $E(I_A \mid Y)$ is a constant. Take X_2 such that $X_2 \equiv Y \wedge X_1$. Since $X_2 \subset Y$, $X_1 \amalg X_2 \mid Y$. Then by hypotheses $X_2 \amalg Y$, which implies that $X_2 \amalg X_2$ since $X_2 \subset Y$; that is $X_2 \equiv Y \wedge X_1$ is essentially a constant.

Remarks : The condition introduced by Koehn and Thomas (1975) is the non existence of a splitting set. A set A in the sample space is a 'splitting' set if $P_\theta(A) = 0$ or 1 for all θ, and at least for a pair $\{\theta_1, \theta_2\} \epsilon \Theta$, $P_{\theta_1}(A) = P_{\theta_2}(A^c) = 1$. In the Bayesian framework, an analogous definition is as follows: A set A such that $I_A \subset X$ is a splitting set if $0 < \Pi(A) < 1$ and $E(I_A \mid Y) = E^2(I_A \mid Y)$. It is easy to see that if A is a splitting set then $I_A \subset Y \wedge X$. We conclude that the non existence of a splitting set is equivalent to $Y \wedge X$ being essentially a constant.

Basu (1955) proved that any ancillary statistic is statistically independent of any bounded complete sufficient statistic. The Bayesian analogue of the concept of boundedly completeness is the concept of strong identifiability (Dawid, 1980 and Mouchart and Rolin, 1978). The main objective of this section is to study this concept and present Basu's result under the Bayesian framework.

Definition 3 : The random objects X and Y are said to be measurably separated conditionally on Z if $(X, Z) \wedge (Y, Z) \equiv Z$. When Z is essentially a constant we simply say that X and Y are measurably separated.

A large list of results related with this concept appears in Mouchart and Rolin (1978).

Let X and Y be two random objects. We shall study some aspects of the linear maps $L_x(Y) \xrightarrow{*} L_\infty(X)$ and $L_\infty(X) \xrightarrow{+} L_\infty(Y)$ where $*$ is for $E(\cdot \mid X)$ and $+$ for $E(\cdot \mid Y)$.

Definition 4 : We say that X is strongly identified by Y and write $X \lll Y$ if the map $L_\infty(X) \xrightarrow{+} L_\infty(Y)$ is essentially one-one.

Proposition 4 : If the map $L_\infty(Y) \xrightarrow{*} L(X)$ is essentially onto then $X \ll Y$.

Proof : Let $(f, h) \subset X$ and $f^+ = 0$. Since $*$ is essentially onto $\exists\, g \subset Y$ such that $g^* = h$. Then

$$E(fh) = E(fg^*) = E(f^+g) = 0$$

since h is arbitrary $f = 0$.

Let $X_{[Y]}$ be the random object that generates the smallest subfield that contains all functions g^* where $g \subset Y$. The following result shows that $X_{[Y]}$ may be viewed as a Bayesian minimal sufficient statistic.

Proposition 5 : (i) $X \amalg Y \mid X_{[Y]}$

(ii) If $X_1 \subset X$ is such that $X \amalg Y \mid X_1$ then $X_{[Y]} \subset X_1$.

Proof : The proof is easy and hence omitted.

Remark : From Proposition 5 it is easy to see that a Burkholder type theorem on intersection of sufficient subfields is true in the Bayesian framework. Precisely, if $X \amalg Y \mid X_1$ and $X \amalg Y \mid X_2$ then $X \amalg Y \mid X_1 \wedge X_2$.

When $X_{[Y]} \equiv X$, X is said to be identified by Y (Dawid 1980, and Mouchart and Rolin 1978). The name strong identification was motivated by the following result.

Proposition 6 : If $X \ll Y$ then $X_{[Y]} \equiv X$.

Proof : Note that $X \amalg Y \mid X_{[Y]}$. Thus $\forall f \subset X$

$$E\{E(f \mid Y, X_{[Y]}) \mid Y\} = E\{E(f \mid X_{[Y]}) \mid Y\}.$$

For $f^+ = E(f \mid X_{[Y]})$. Since $X \ll Y$ we have that

$$E\{(f - f^+) \mid Y\} = 0 \rightarrow f = f^+.$$

Then $\forall f \subset X$, $f \subset X_{[Y]}$ and $X = X_{[Y]}$.

The Bayesian version of Basu's theorem is contained in the result below.

Theorem 3 : *Let X, Y and Z be three random objects. If $X \amalg Y$, $X \amalg Y \mid Z$ and $Z \ll Y$ then $X \amalg Z \mid Y$.*

Proof : Since $X \amalg Y \mid Z$ we have, for any $f \subset X$

$$E(f \mid Y, Z) = E(f \mid Z)$$

and since

$$X \amalg Y,\ E(f \mid Y) = E(f)$$

Therefore

$$E[\{E(f \mid Z) - E(f)\} \mid Y] = 0.$$

Now since $Z \ll Y$ $E(f \mid Z) = E(f)$ we thus have $E(f \mid Y, Z) = E(f)$.

Note that to obtain Basu's theorem, we consider X as the sample, Y as the parameter, and X_0 and X_1 two random objects such that $(X_0, X_1) \subset X$, $X_0 \amalg Y$, $X \amalg Y \mid X_1$ and $X_1 \ll Y$. Clearly $X_0 \amalg Y \mid X_1$ and the result $X_0 \amalg X_1 \mid Y$ follows.

Lehman and Scheffe (1950) proved that if a sufficient statistic is boundedly complete, then it is a minimal sufficient statistic. The proposition below is a Bayesian version of this result.

Proposition 7 : Suppose $X_1 \subset X$ and $X \amalg Y \mid X_1$. If $X_1 \ll Y$ then $X_1 = X_{[Y]}$.

Proof : From Proposition 6 $X_{[Y]} \subset X_1$ and $X \amalg Y \mid X_{[Y]}$.

Let $f \subset X_1$ then $E(f \mid Y) = E[E(f \mid X_{[Y]}) \mid Y]$ or

$$E(f \mid Y) = E[E(f \mid X_{[Y]}, Y) \mid Y]$$
$$= E(E(f \mid X_{[Y]}) \mid Y)$$

since $X_1 \ll Y$ we conclude that $f = E(f \mid X_{[Y]}) \subset X_{[Y]})$.

Remark : The concept of strong identifiability may be generalized as follows. X is strongly identified by Y conditionally on Z $(X \ll Y \mid Z)$ if for every $f \subset (X, Z)$, $E(f \mid Y, Z) = 0$ implies $f = 0$. Analogously, X is identified by Y conditionally on Z if

$$(X, Z)_{[Y, Z]} = (X, Z).$$

All the results of this section may be easily generalized by introducing a conditioning random object Z to each relation stated. For our future work we intend to relate these general results with the work of Dawid (1979c), Ferreira (1980) and Godambe (1980).

5. The two parameter problem

We now briefly discuss sufficiency in the presence of a nuisance parameter.

Suppose that the parameter Y is such that $Y \equiv (Y_1, Y_2)$. Let X represent the sample, $X_1 \subset X$ be specific sufficient with respect to Y_2, and $X_2 \subset X$ be specific sufficient with respect to Y_1. That is, $X \amalg Y_2 \mid (X_1, Y_1)$ and $X \amalg Y_1 \mid (X_2, Y_2)$. (See Basu, 1978 for details on the notion of specific sufficiency). The question here is under what conditions does the specific sufficiency of (X_1, X_2) imply the sufficiency of (X_1, X_2) ?

Proposition 8 : If $(X_1, Y_1) \wedge (X_2, Y_2) \subset (X_1, X_2)$ then $X \amalg Y_2 \mid (X_1, Y_1)$ and $X \amalg Y_1 \mid (X_2, Y_2)$ imply $X \amalg Y \mid (X_1, X_2)$.

Proof: We have

$$X \amalg Y \mid (X_1, Y_1)$$

and

$$X \amalg Y \mid (X_2, Y_2)$$

and a simple argument yields

$$X \amalg Y \mid (X_1, Y_1) \wedge (X_2, Y_2).$$

The following related result may also be of interest.

Proposition 9: If $X \amalg Y_2 \mid (X_1, Y_1)$ and $X \amalg Y_1 \mid (X_2, Y_2)$ then

$$X \amalg Y \mid (X_1, X_2) \text{ if and only if } X \amalg Y_1 \mid (X_1, X_2).$$

Note the condition $X \amalg Y_1 \mid (X_1, X_2)$ does not have an interpretation in classical statistics since distributions depend on both parameters Y_1 and Y_2.

The following example is again relevant. Note that the parameter space θ is variation independent (if the parameter space is the cartesian product of the domain Y_1 by the domain of Y_2; see Basu, 1977 and Barndorff-Nielsen, 1978).

Take

$$\Theta = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

Then $\Theta = \Theta_1 \times \Theta_2$ where $\Theta_1 = \Theta_2 = \{0, 1\}$

$$X = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

$$P_\theta = \delta_\theta \text{ the point mass at } \theta.$$

Then $T = I_{\{(0,1),(1,0)\}}(x)$ is specific sufficient for θ_1 and for θ_2 but is not sufficient. This example Shows that variation independence on (θ_1, θ_2) and specific sufficiency of X_1, and X_2 does not imply that (X_1, X_2) being sufficient.

Acknowledgement. Thanks are due to the referee for helping in improving presentation of the paper.

References

Ash, R. B. (1972): *Real Analysis and Probability*, Academic Press, New York.

Bahadur, R. R. (1955): Measurable subspaces and subalgebras. *Proc. Amer. Math. Soc.*, 6, 565–70.

Barndorff-Nielsen, O. (1978): *Information and Exponential Families in Statistical Theory*, John Wiley, New York.

Basu, D. (1955): On statistics independent of a complete sufficient statistics. *Sankhyā*, A, 15, 377–80

——— (1958): On statistics independent of a sufficient statistics. *Sankhyā*, A, 20, 223–26.

——— (1959): The family of ancillary statistics. *Sankhyā*, A, 21, 247–56

——— (1964): Recovery of ancillary information. *Sankhyā*, A, 26, 3–16.

——— (1967): Problems relating to the existence of maximal and minimal elements in some families of statistics (Subfields). *Proc. Fifth Berkeley Sym. Math. Statist. Prob.*, **1**, 41–50.

——— (1977): On the elimination of nuisance parameters. *Jour. Amer. Statist. Assoc.*, **72**, 355–66.

——— (1978): On partial sufficiency: A review. *J. Statist. Plan. Inf.*, **2**, 1–13.

BASU, D. and CHENG, S. C. (1979): A note on sufficiency in coherent models. *Int. J. Math. Math. Sci.* To appear.

BURKHOLDER, D. L. (1961): Sufficiency in the undominated case. *Ann. Math. Statist.*, **32**, 1191–200.

CHENG, S. C. (1978): A mathematical study of sufficiency and adequacy in statistical theory. Ph.D. Dissertation, FSU, Florida.

DAWID, A. P. (1979a): Conditional independence in statistical theory. *J. Roy. Statist. Soc.*, B, **41**, 1–31.

——— (1979b): Some misleading arguments involving conditional independence. *J. Roy. Statist. Soc.*, B, **41**, 249–52.

——— (1979c): A Bayesian look at nuisance parameters. *Trabajos de Estadistica*, To appear.

——— (1980): Conditional independence for statistical operations. *Ann. Statist.*, **8**, 598–617.

DE FINETTI, B. (1937): Foresight: Its logical laws, its subjective sources. Translated edition 1964 in *Studies in Subjective Probability* (H. E. Kyburg and H. E. Smokler, editors). John Wiely, New York.

——— (1970): *Theory of Probability*, Vols. 1 and 2. Translated edition, 1974. John Wiley, London.

DOOB, J. L. (1953): *Stochastic Processes*, John Wiley, New York.

FERREIRA, P. E. (1980): Comments on Berkson's Paper "In dispraise of ....". Unpublished report.

FISHER, R. A. (1920): A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. *Mon. Not. Roy. Ast. Soc.*, **80**, 758–70.

——— (1922): On the mathematical foundations of theoretical statistics. *Phil. Trans.*, **A**, **222**, 309–68.

FLORENS, J. P. and MOUCHART, M. (1977): Reduction of Bayesian experiments. *CORE*, Discussion Paper 7737.

GODAMBE, V. P. (1979): On sufficiency and ancillarity in presence of nuisance parameter. Unpublished report.

HALL, W. J., WIJSMAN, R. A. and GHOSH, J. K. (1965): The relationship between sufficiency and invariance. *Ann. Math. Statist.*, **36**, 375–614.

KOEHN, U. and THOMAS, D. L. (1975): On statistics independent of a sufficient statistics: Basu's lemma. *Amer. Statist.*, **29**, 40–2.

KOLMOGOROV, A. N. (1942): Determination of the center of dispersion and degree of accuracy for a limited number of observations (in Russian). *Izvestija Akademii Nauk*, Ser. Mat., **6**, 3–32.

LEHMAN, E. L. and ACHEFFE, H. (1950) : Completeness, similar regions and unbiased estimation, Part I. *Sankhyā*, A, **10**, 305–40.

LINDLEY, D. V. and NOVICK, M. R. (1981) : The role of exchangeability in inference. *Ann. Statist.*, **9**, 45–58.

LOEVE, M. (1977) : *Probability Theory*, 4th ed. Springer-Verlag, NY.

MOUCHART, M. and ROLIN, J. M. (1978) : A note on conditional independence. Unpublished report.

MOY, S. T. C. (1954) : Characterization of conditional expectation as a transformation on function spaces. *Pacific J. Math.*, **4**, 47–63.

PATHAK, P. K. (1975) : Note on Basu's lemma. Unpublished report.

PICCI, G. (1977) : Some connections between the theory of sufficient statistics and the identifiability problem. *SIAM J. Appl. Math.*, **33**, 383–98.

WHITT, W. (1979) A note on the influence of the sample on the posterior distribution. *Jour. Amer. Statist. Assoc.* **74**, 424–426.

Paper received : *March*, 1981.
Revised : *January*, 1983.

Sankhyā : The Indian Journal of Statistics
1983, Volume 45, Series A, Pt. 1, pp. 99–104.

A NOTE ON BLACKWELL SUFFICIENCY AND A SKIBINSKY CHARACTERIZATION OF DISTRIBUTIONS

By D. BASU* and CARLOS A. B. PEREIRA**

The Florida State University, Tallahassee

SUMMARY. A Skibinsky (1970) characterization of the family of hypergeometric distributions is re-examined from the point of view of sufficient experiments and a number of other distributions similarly characterized.

1. Introduction

Consider an urn containing N balls x of which are white. If a simple random sample of n ($n \leqslant N$) balls is drawn from the urn, then the number of white balls in the sample has the hypergeometric distribution with parameters N, n, and x [denoted by $h(N, n, x)$]. Skibinsky (1970) introduced the following characterization of $h(N, n, x)$:

"A family of $N+1$ probability distributions (indexed say by $x = 0, 1, \ldots, N$), each supported on a subset of $\{0, 1, \ldots, n\}$ is the hypergeometric family having population and sample size parameters N and n respectively (the remaining parameter of the x-th member being x), if and only if for each number θ, $0 < \theta < 1$, the mixture of the family with binomial (N, θ) mixing distribution is the binomial (n, θ) distribution."

Writing $b(N, \theta)$ for the binomial distribution over $\{0, 1, \ldots, N\}$ and the symbol $\sim$ for "distributed as", we may restate Skibinsky's characterization as follows :

Let $X \sim b(N, \theta)$, $0 < \theta < 1$, and let $\{\tau_x : x = 0, 1, \ldots, N\}$ be a family of probability distributions on $\{0, 1, \ldots, n\}$, where $n \leqslant N$. Consider the random variable Y such that the conditional probability distribution of Y given $\{X = x\}$ is τ_x for all x (i.e., $Y \mid X = x \sim \tau_x$). Then $Y \sim b(n, \theta)$ for all θ in $(0, 1)$ if and only if τ_x is $h(N, n, x)$ for all x.

Skibinsky (1970) proved the above result in several interesting ways, but somehow the perspective of Blackwell sufficiency eluded him. Written for its pedagogical interest, this note is an elucidation on the notion of Blackwell sufficiency and an unification of a number of results analogous to Skibinsky's characterization of the Hypergeometric distribution.

*Research partially supported by NSF Grant No. 79-04693.

**Research supported by CAPES and USP—Brazil.

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_33, © Springer Science+Business Media, LLC 2011

2. Blackwell sufficiency

A statistical experiment related to a parameter $\theta \in \Theta$ is idealized as an observable random variable (or vector), X, associated with a sample space $\mathcal{X}$ and a family $\{p_\theta : \theta \in \Theta\}$ of probability functions (distributions) on $\mathcal{X}$ indexed by θ. We avoid all measurability difficulties by restricting ourselves only to discrete sample spaces. Given two spaces $\mathcal{X}$ and $\mathcal{Y}$, a transition function τ, from $\mathcal{X}$ to $\mathcal{Y}$, is a family

$$\tau = \{\tau_x : x \in \mathcal{X}\}$$

of probability functions, τ_x, on indexed by $x \in \mathcal{X}$. Thus, the family of Hypergeometric probability functions $\{h(N, n, x) : x = 0, 1, \ldots, N\}$ is a transition function from $\{0, 1, \ldots, N\}$ to $\{0, 1, \ldots, n\}$.

Let X and Y be two experiments with models $(\mathcal{X}, \{p_\theta : \theta \in \Theta\})$ and $(\mathcal{Y}, \{q_\theta : \theta \in \Theta\})$ respectively.

Definition (Blackwell): The experiment X is *sufficient for* (at least as informative as) the experiment Y and write $X > Y$ if there exists a transition function $\tau = \{\tau_x : x \in \mathcal{X}\}$ from $\mathcal{X}$ to $\mathcal{Y}$ such that

$$q_\theta(y) = \sum_x \tau_x(y) p_\theta(y) \qquad \ldots \quad (2.1)$$

for all $y \in \mathcal{Y}$ and $\theta \in \Theta$.

A transition function τ satisfying (2.1) is called here a Blackwell transition function. It is easy to check that the relation $>$ defines a partial order on the family of experiments related to θ.

If $T = T(X)$ is a sufficient statistic in the classical sense of Fisher (i.e., the conditional distribution of X given $\{T = t\}$ does not involve θ), then it follows at once that T is sufficient for X in the sense of Blackwell $(T > X)$. Of course, X is sufficient for T in either sense.

The intuitive content of the relation $X > Y$ is as follows:

If we perform the experiment X, note its outcome x, and finally carry out a postrandomization exercise that chooses a point $y \in \mathcal{Y}$ in accordance with the probability function τ_x, then the experiment Y^* of such a choice of y is in a sense indistinguishable from the experiment Y in that both are endowed with the same model $(\mathcal{Y}, \{q_\theta : \theta \in \Theta\})$. Any decision rule related to θ that is based on the experiment Y can therefore be perfectly matched (in terms of their average performance characteristics) by a randomized rule based on X.

For two fixed integers N and n ($n \leqslant N$), consider now the simple case where $X \sim b(N, \theta)$ and $Y \sim n(n, \theta)$, $0 < \theta < 1$. To prove that $X > Y$ we consider an experiment $W = (W_1, \ldots, W_N)$ where its components W_i, $i = 1, \ldots, N$, are i.i.d Bernoulli variables with parameter θ. Note that since $X^* = W_1 + \ldots + W_N$ is sufficient for W in the classical sense, $X^* > W$. On the other hand, for $Y^* = W_1 + \ldots + W_n$, $W > Y^*$. Therefore, $X^* > Y^*$. Since X and X^* (Y and Y^*) are indistinguishable in their models, $X > Y$.

To conclude our version of Skibinsky's characterization we note that

$$Y^* \mid X^* = x \sim h(N, n, x).$$

Then a Blackwell transition function $\{\tau_x\}$ for our problem is the family of Hypergeometric probability functions. That is,

$$P_R\{Y = y \mid \theta\} = \sum_x \tau_x(y) P_T\{X = x \mid \theta\} \qquad \ldots \quad (2.2)$$

for every $y \in \mathcal{Y}$ and every $\theta \in (0, 1)$ where $\tau_x(\cdot)$ is the Hypergeometric probability function with parameter (N, n, x). Finally, the uniqueness of $\{\tau_x\}$ as a Blackwell transition function follows from the fact that the family $\{b(N, \theta) : 0 < \theta < 1\}$ of probability distributions is complete. If $\{\tau'_x\}$ is another transition function satisfying (2.2), then

$$\sum_x [\tau_x(y) - \tau'_x(y)] P_T\{X = x \mid \theta\} = 0$$

for every $y \in \mathcal{Y}$ and therefore $\tau_x(y) = \tau'_x(y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

3. Further characterizations

Consider now an urn with N balls of k ($k \leqslant N$) different colors. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ be the vector of frequency counts for the colors; that is, x_i ($i = 1, 2, \ldots, k$) is the number of balls with the i-th color. If a simple random sample of n balls is drawn from the urn, then the sample vector of frequency counts has the multivariate Hypergeometric distribution with parameters, $N, n,$ and $\boldsymbol{x}$. This distribution is denoted by $H(N, n, \boldsymbol{x})$ and its support by

$$Z_n^k = \{(x_1, \ldots, x_k) : x_i \in Z, \ \Sigma x_i = N\}$$

where $Z = \{0, 1, \ldots\}$.

Writing $M(N, \boldsymbol{\theta})$ for the Multinomial distribution over Z_N^k, we state the natural extension of Skibinsky's characterization. Here the parameter space is the simplex

$$\mathcal{S} = \{(p_1, \ldots, p_k) : p_i \geqslant 0, \Sigma p_i = 1\}.$$

Proposition 1 : Let $X \sim M(N, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathcal{S}$, and let $\{\tau_{\boldsymbol{x}} : \boldsymbol{x} \in Z_N^k\}$ be a family of probability distributions on Z_n^k where $n \leqslant N$. Consider a random vector $Y \mid X = \boldsymbol{x} \sim \tau_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in Z_N^k$. Then $Y \sim M(n, \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathcal{S}$ if and only if $\tau_{\boldsymbol{x}}$ is $H(N, n, \boldsymbol{x})$ for all $\boldsymbol{x} \in Z_N^k$.

The proof follows the same steps of the univariate case discussed in Section 2. Here, we consider the experiment $W = (W_1, \ldots, W_N)$ where the components are i.i.d. with the common distribution being $M(1, \boldsymbol{\theta})$.

We write $X \sim \text{Poi}(\theta)$, $\theta > 0$, to indicate that X has Poisson distribution with parameter $\theta \in (0, \infty)$. Consider an additional experiment Y such that for a known number $r \in (0, 1)$, $Y \sim \text{Poi}(r\theta)$, $\theta > 0$. To prove that $X > Y$ we consider an experiment $W = (W_1, W_2)$ where its components W_1 and W_2 are independent with distributions $\text{Poi}(r\theta)$ and $\text{Poi}((1-r)\theta)$ respectively. Since $W > W_1$, $X^* = W_1 + W_2$ is sufficient for W in the classical sense, and X and X^* (Y and W_1) are indistinguishable in their models, it follows that $X > Y$.

Since $W_1 \mid X^* = x \sim b(x, r)$ for all $x \in Z$, a Blackwell transition function $\{\tau_x\}$ is the family of Binomial probability functions. That is,

$$\frac{e^{-\theta r}(\theta r)^y}{y!} = \sum_x \tau_x(y) \frac{e^{-\theta}\theta^x}{x!}$$

for every $y \in Z$ and all $\theta \in (0, \infty)$ where $\tau_x(\cdot)$ now is the Binomial probability function with parameter (x, r). The uniqueness of this family $\{\tau_x\}$ of Binomials as a Blackwell transition function follows from the completeness of the family $\{\text{Poi}(\theta) : \theta \in (0, \infty)\}$ of probability distributions on Z.

The above result in its extended form may be summarized as :

Proposition 2 : Let $X \sim \text{Poi}(\theta)$, $\theta > 0$, and let $\{\tau_x : x \in Z\}$ be a family of probability distributions on the set $Z^k = \{(y_1, \ldots, y_k) : y_i \in Z\}$. Consider a random vector $Y = (Y_1, \ldots, Y_k)$ such that $Y \mid X = x \sim \tau_x$ for all $x \in Z$. For a known vector $\boldsymbol{r} = (r_1, \ldots, r_k) \in \mathcal{S}$, the components, Y_i $(i = 1, \ldots, k)$, of Y are mutually independent and $Y_i \sim \text{Poi}(r_i\theta)$, $\theta > 0$ if and only if τ_x is $M(x, \boldsymbol{r})$ for all $x \in Z$.

We end this section with a parallel characterization of the Dirichlet-Multinomial distribution. In Basu and Pereira (1980) we studied in details this distribution and indicated its use in statistics. We define the Dirichlet-Multinomial $DM(N; \alpha_1, \ldots, \alpha_k)$ on $\mathcal{Z}_N^k$ as the mixture of the Multinomial family $\{M(N, \boldsymbol{p}); \boldsymbol{p} \in \mathcal{S}\}$ with $\boldsymbol{p}$ distributed as Dirichlet (on $\mathcal{S}$) with parameter $(\alpha_1, \alpha_2, \ldots, \alpha_k)$. Its probability function is given by

$$f(z_1, \ldots, z_k) = \frac{N!\,\Gamma(\alpha)}{\Gamma(\alpha+N)} \prod_{i=1}^{k} \frac{\Gamma(\alpha_i+z_i)}{z_i!\,\Gamma(\alpha_i)}$$

for all $(z_1, \ldots, z_k) \in \mathcal{Z}_n^k$ where $\alpha = \sum_1^k \alpha_i$. When $k = 2$, in place of $(Z_1, Z_2, \sim DM(N, \alpha_1, \alpha_2)$, we write $Z_1 \sim Bb(N; \alpha_1, \alpha_2)$ to indicate that Z_1 is distributed as Beta-Binomial with parameter $(N; \alpha_1, \alpha_2)$.

Consider a sequence of Bernoulli trials with probability of success $\theta \in (0, 1)$. If $X+\alpha$ is the number of trials needed to obtain a fixed number α of success, then X is said to be a Negative Binomial experiment with parameter $(\alpha; \theta)$ and we write $X \sim nb(\alpha; \theta)$, $0 < \theta < 1$. Its probability function is

$$g(x|\theta) = \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!}\,\theta^{\alpha}(1-\theta)^{x} \qquad \ldots \quad (3.1)$$

for every $x \in \mathcal{Z}$ and all $\theta \in (0, 1)$. Note that

$$\sum_{x=0}^{\infty} \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!}\,(1-\theta)^{x} = \theta^{-\alpha} \text{ for every } \alpha \in (0, \infty) \text{ and all } \theta \in (0, 1).$$

Then the following results hold not only for $\alpha \in Z$ but in general for any $\alpha \in (0, \infty)$. In this case, we still write $X \sim nb(\alpha; \theta)$, $0 < \theta < 1$, to indicate that the family of probability functions associated with the experiment X is (3.1). It is easy to check that this family is complete.

For $\alpha \geqslant \alpha_1 > 0$, let X and Y be two experiments such that $X \sim nb(\alpha; \theta)$ and $Y \sim nb(\alpha_1; \theta)$, $0 < \theta < 1$. To prove that $X > Y$, consider the experiment $W = (W_1, W_2)$ where now W_1 and W_2 are independent with distributions $nb(\alpha_1; \theta)$ and $nb(\alpha-\alpha_1; \theta)$ respectively. Following our previous chain of arguments, one can easily check that (i) $W_1+W_2 \sim X$, (ii) $W_1+W_2 > W > W_1$, (iii) $X > Y$, (iv) $W_1 | W_1+W_2 = x \sim Bb(x; \alpha_1, \alpha-\alpha_1)$, and (v) the family $\{Bb(x; \alpha_1, \alpha-\alpha_1) : x \in Z\}$ of probability functions is the unique Blackwell transition function, and thus arrive at a Skibinsky type characterization of the Beta-Binomial.

The following is a summary of an extended version of the above results.

Proposition 3 : Let $X \sim nb(\alpha; \theta)$, $0 < \theta < 1$, and let $\{\tau_x : x \in Z\}$ be a family of probability distributions on the set Z^k. Consider a random vector $Y = (Y_1, \ldots, Y_k)$ such that $Y \mid X = x \sim \tau_x$ for all $x \in Z$. For a fixed vector $(\alpha_1, \ldots, \alpha_k)$ where $0 < \alpha_i < \infty$, $i = 1, 2, \ldots, k$, and $\alpha = \sum_1^k \alpha_i$, the components Y_i $(i = 1, \ldots, k)$ of Y are mutually independent with $Y_i \sim nb(\alpha_i, \theta)$, $0 < \theta < 1$, if and only if τ_x is $DM(x; \alpha_1, \ldots, \alpha_k)$ for all $x \in Z$.

References

Basu, D. and Pereira, C. A. B. (1982) : On the Bayesian analysis of categorical data : The problem of nonresponse. *Journal of Statistical Planning and Inference*, **6**, 345-362.

Blackwell, D. and Girshick, M. A. (1954) : *Theory of Games and Statistical Experiments*, John Wiley, N. Y.

Lehmann, E. L. (1959) : *Testing Statistical Hypothesis*, John Wiley, N. Y.

Skibinsky, M. (1970) : A characterization of hypergeometric distributions. *JASA*, **65**, 926-929.

Paper received : March, 1981.

13

Learning Statistics from Counter Examples: Ancillary Statistics

D. Basu [1]

Abstract

Bayesian objection to the analysis of data in frequency theory terms is amplified through several counter examples in which an ancillary statistic exists and there is a temptation to choose a reference set after looking at the data. It is argued that Fisher insisted on conditioning by an ancillary statistic, because conditioning the data $\mathbf{x}$ by an ancillary Y does not change the likelihood. In this sense Fisher discovered the supremacy of the likelihood function.

Key words and Phrases: Ancillary statistics; Conditional frequentist inference; Information; Liklihood principle; Reference set; Sufficiency principle.

1. INTRODUCTION

This paper is especially addressed to the statisticians who have not yet fully grasped the Bayesian objection to the analysis of data in repeated sampling terms. Let $\mathbf{x}$ be the sample, $f(\mathbf{x}|\theta)$ the model and θ the parameter. A statistic $Y = Y(\mathbf{x})$ is *ancillary* if the sampling distribution of Y, given θ, is θ-free (is the same for all values of θ). A statistic $T = T(\mathbf{x})$ is *sufficient* if the distribution of the sample $\mathbf{x}$, given T and θ, is θ-free. An ancillary statistic Y by itself contains no information about the parameter, whereas a sufficient statistic T is fully informative in a sense. R.A. Fisher's attempt to make sense of the notion of *information in the data* led him to these two important concepts in Statistics.

Let $L(\theta) = f(\mathbf{x}|\theta)$ be the *likelihood function* determined by the sample $\mathbf{x}$ and let $\hat{\theta}$ be the *maximum likelihood* (ML) estimate of θ. If $\hat{\theta}$ is a sufficient statistic then, according to Fisher, there would be no loss of information if the performance characteristics of $\hat{\theta}$ as an estimate of θ is sought to be evaluated in terms of the sampling distribution of $\hat{\theta}$. We shall repudiate this in the end with an example.

---

[1] Indian Statistical Institute and Florida State University

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9_34,

If the ML estimate $\hat{\theta}$ is not a sufficient statistic then Fisher sought to recover the information lost in the sampling distribution of $\hat{\theta}$ with the help of an *ancillary complement* Y to the estimator $\hat{\theta}$. The ancillary statistic Y has to complement $\hat{\theta}$ in the sense that the pair $(\hat{\theta}, Y)$ is jointly sufficient. The Fisher Information $I_{\hat{\theta},Y}(\theta)$ in the sufficient statistic $(\hat{\theta}, Y)$ is then the same as the full information

$$I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} \log L(\theta)]$$

in the sample $\mathbf{x}$. (Note that $I(\theta)$ does not relate to the particular sample $\mathbf{x}$ but is obtained by averaging the quantity $-\frac{\partial^2}{\partial\theta^2} \log L(\theta)$ over the sample space.) The Fisher Information in the statistic $\hat{\theta}$ is less than the full information $I(\theta)$. The cornerstone of the Fisher argument lies in the identity

$$I(\theta) = I_{\hat{\theta},Y}(\theta) = E[I_{\hat{\theta}}(\theta|Y)],$$

where $I_{\hat{\theta}}(\theta|Y)$ is the conditional information in the statistic $\hat{\theta}$, given Y, and the expectation on the right hand side is with respect to the ancillary statistic Y. Thus, the conditional information in $\hat{\theta}$, given Y, depends on Y and can be, for a particular value of the statistic Y, much less or much greater than the full information $I(\theta)$. The *conditionality argument* of R.A. Fisher rests on the proposition that the performance characteristics of the estimator $\hat{\theta}$ ought to be evaluated conditionally, holding the ancillary statistic Y fixed at its observed value y. As Fisher argued, the event $Y = y$, even though uninformative by itself, has a lot of latent information about θ in the sense that it helps us discern how good or bad the estimate $\hat{\theta}$ is in the present instance. The set $S(y) = \{\mathbf{x} : Y(\mathbf{x}) = y\}$ defines what Fisher called the *reference set*. Sir Ronald was trying to cut down the *sample space* S to size. We illustrate the conditionality argument with several examples.

2. EXAMPLES

Example 1: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be iid observations on a random variable that is uniformly distributed over the interval $[\theta, 2\theta]$, where $\theta > 0$ is the unknown scale parameter. With

$$m = \min x_i \text{ and } M = \max x_i,$$

the likelihood function $L(\theta)$ equals $1/\theta^n$ over the interval $[M/2, m]$ and zero outside the interval. The ML estimator $\hat{\theta} = M/2$ is not sufficient, the minimal sufficient statistic being the pair (m, M). Since the two statistics m and M are stochastically independent in an asymptomatic sense, it is clear that there will be a substantial loss of information if we marginalize the data to the ML estimator $M/2$. Comparing the mean squared error (MSE) of $M/2$ with that of m as estimators of θ, we find that the former is exactly four times better than the latter. Consider, therefore, the estimator

$T = (2M+m)/5$ which is the weighted average of $M/2$ and m with weights 4 and 1 respectively. Both $M/2$ and T are equivariant estimators of the scale parameter θ, and so their MSE's are constant multiples of θ^2. It works out that the ratio of the two MSE's tends to $25/12$ as the sample size n tends to infinity. The ML estimator $\hat{\theta}$ can hardly be called an efficient estimate of θ in the usual sense of the term. Over thirty-six years ago, when I came upon this counterexample, it was pointed out to me by C.R. Rao that the ML estimator $\hat{\theta}$ ought to be judged conditionally after holding fixed its ancillary complement $Y = M/m$ at its observed value. That Y is an ancillary statistic follows from the facts that Y is scale invariant and that θ is a scale parameter. As we noted before, the likelihood mass is spread over the interval $[M/2, m]$ pinpointing the parameter θ within that interval. The statistic $Y = M/m$ varies over the range $[1, 2]$ and is indeed a measure of how good the sample is – the nearer Y is to 2 the better the sample is. While evaluating the ML estimate $\hat{\theta}$ we ought to take note of the observed value y of the statistic Y. That is, instead of referring $\hat{\theta}$ to the full sample space S, we ought to refer it to the *reference set* $S(y)$.

In terms of the full sample space S the ML estimator $M/2$ is not sufficient. But when it is conditioned by Y it suddenly becomes fully informative (sufficient, that is). Note that the other two estimators m and T also become fully informative when they are referred to the set $S(y)$. Indeed, the three statistics $M/2$, m and T become functionally related when conditioned by Y.

This example beautifully illustrates what Fisher meant by *recovery of ancillary information*. The next example illustrates how a weak pivotal quantity can be strengthened by proper conditioning with an ancillary statistic.

Example 2: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be n iid observations on a random variable with pdf $f(x-\theta)$, where f is known but θ (the location parameter) is unknown. Consider the statistic x_1 and its ancillary complement $D = (x_2 - x_1, x_3, -x_1, \ldots, x_n - x_1)$. The statistic x_1 by itself carries very little information about θ, but it becomes fully informative (sufficient) when conditioned by D. The conditional pdf of x_1, given D, has θ embedded in it as a location parameter. Fisher derived the *fiducial distribution* of the parameter θ by inverting the pivotal quantity $x_1 - \theta$ after conditioning it by the ancillary statistic D.

The previous example raises many questions. Some sample questions and answers are listed below.

Question: What is the status of the ancillary statistic D? Is it the *maximum ancillary* in the sense that every other ancillary statistic is a function of D?

Answer: No. D is never the maximum ancillary. However, in some situations D will be a *maximal ancillary* in the sense that no larger (with respect to the partial order of functional relationship) ancillary statistic exists. A

multiplicity of maximal ancillaries is a fact of life in this situation.

Question: Is the fiducial distribution of θ in Example 2 critically dependent on the choice of the pivotal quantity $x_1 - \theta$?

Answer: No. Another pivotal quantity like, say, $\mathbf{x} - \theta$, when conditioned by D, will result in the same fiducial distribution of θ. This is because $\bar{x} = x_1 + (\bar{x} - x_1)$ and $\bar{x} - x_1$ is a function of D.

Question: Can we interpret the fiducial distribution of θ probabilistically?

Answer: It was pointed out by Harold Jeffreys that the fiducial distribution of the location parameter (as derived by Fisher) coincides with the posterior distribution of θ corresponding to the uniform prior (over the entire real line) for the parameter.

In the presence of multiple ancillaries, the choice of the proper reference set is a problem. The dilemma is best exemplified by the following example.

Example 3: Let $(x_1, y_1), i = 1, 2, \ldots, n$, be n iid observations on (X, Y) whose joint distribution is Bivariate Normal with zero means, unit variances and covariance θ, which is the parameter of interest. In this case both $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are ancillary statistics. Note that the pair $(\mathbf{x}, \mathbf{y})$ is the entire data and therefore is sufficient. Holding the ancillary $\mathbf{x}$ as fixed and regarding $\mathbf{y}$ as the variable, we may want to estimate θ by $\sum x_i y_i / \sum x_i^2$ and then regard the estimate as unbiased with variance $(1 - \theta^2)/\sum x_1^2$. But how about holding $\mathbf{y}$ fixed and reporting that $\sum x_i y_i / \sum y_i^2$ is an unbiased estimate with variance $(1 - \theta^2)/\sum y_1^2$? It is tempting to opt for the ancillary with the larger sum of squares. But would it not be a statistical heresy to choose the reference set after looking at the data?!

3. COX ON ANCILLARIES

D.R. Cox (1971) suggested a way to deal with the problem of multiple ancillaries. Looking back at the Fisher identity $I(\theta) = EI(\theta|Y)$, Cox argued that the basic role of the conditioning ancillary Y is to discriminate between samples with varying degrees of information. So in the presence of multiple ancillaries we should choose that Y for which $I(\theta|Y)$ is most variable in Y. So opt for the Y for which Var $I(\theta|Y)$ is maximum. One snag in the Cox argument is that Var $I(\theta|Y)$ is a function of θ and so there may not exist a Y that maximizes the function uniformly in θ. Also note that in our Example 3 the Cox method fails because, in view of the perfect symmetry between $\mathbf{x}$ and $\mathbf{y}$, Var $I(\theta|\mathbf{x})$ = Var $I(\theta|\mathbf{y})$.

But the real snag in the Cox argument is the meaninglessness of the notion of Fisher Information as a measure of the evidential meaning of the particular data at hand. Fisher's preoccupation with the elusive notion of information in the data led him to the likelihood function which he recognized as the carrier of all the information in the data. The likelihood was then partially summarized in the two statistics $\hat{\theta}$, the ML estimate, and $Z(\hat{\theta})$, the second derivative of $-\log L(\theta)$ at $\theta = \hat{\theta}$. Note that $Z(\hat{\theta})$ is the reciprocal of the radius of curvature of the log likelihood at its mode, the

larger the value of $Z(\hat{\theta})$ the sharper is the fall of the likelihood function as θ moves away from $\hat{\theta}$. We have to stretch our minds a little to regard $Z(\hat{\theta})$ as a rough measure of the concentration of the likelihood mass around $\hat{\theta}$. The greater the concentration the more informative is the likelihood. The Fisher Information $I(\theta)$ is obtained from $Z(\hat{\theta})$ by first replacing $\hat{\theta}$ by θ and then taking the average value of $Z(\theta)$ over the whole sample space S. But how can we regard $I(\theta)$ as information in the data?

Why did Fisher insist that the conditioning statistic Y has to be ancillary? Because, conditoning the data $\mathbf{x}$ by an ancillary Y does not change the likelihood. Fisher discovered the supremacy of the likelihood but got carried away by his amazing craftmanship with sample space mathematics.

4. E.L. LEHMANN ON ANCILLARIES

Eric Lehmann (1981) finally recognized the conditionality argument. And now he has to cope with the disturbing presence of ancillary statistics. Invoking the *Sufficiency Principle*, Eric would reduce the data $\mathbf{x}$ to the minimal sufficient statistic $T = T(\mathbf{x})$. Since T is sufficient, all reasonable inference procedures ought to depend on $\mathbf{x}$ only through $T(\mathbf{x})$. This data reduction sweeps away much of the ancillary dust under the rug. But, as in Example 1, some functions of the minimal sufficient statistic T may still be recognized as ancillary statistics. Eric has yet to come out openly on the question of how to deal with such persistent ancillaries.

From what Eric writes in his 1981 article, it seems that he feels quite comfortable with statistical models for which the minimal sufficient statistic T is complete. In such cases no nontrivial function of T can be ancillary. Furthermore, thanks to the so called Basu Theorem, every ancillary statistic Y is stochastically independent (conditionally on θ) of T. Therefore, no T- based decision procedure can be altered by conditioning with an ancillary Y. So who needs to think of the conditionality argument when we have a complete sufficient statistic? Remember, Fisher looked for an ancillary complement to the ML estimate $\hat{\theta}$ only when the statistic $\hat{\theta}$ was not sufficient. So in the most favorable set up where $\hat{\theta}$ is a complete sufficient statistic, can anyone object if we evaluate the estimate $\hat{\theta}$ in terms of the sampling distribution of the estimator? We give an example to prove both Fisher and Lehmann wrong on this question.

Example 4: Consider a sequence of Bernoulli trials with parameter p that results in a finite sequence $w = SFFS \ldots FS$ of successes S and failures F. Let $X(w)$ and $Y(w)$ denote, respectively, the number of S's and the number of F's in the sample sequence w. We picture w as a sample path, the locus of a point that begins its journey at the original and travels through the lattice points of the positive quadrant, moving one step to the right for each S and one step up for each F. The lattice point with coordinates $X(w)$ and $Y(w)$ is where the sample path w ends. Our example relates to a particular

sampling (stopping) rule **R**. Writing (X, Y) for the location of the moving point, the rule is described as:

Rule **R**: Continue sampling as long as (I) $Y < 2X + 1$, (II) $Y > X - 2$, and (III) $X + Y < 100$. Alternatively, the rule may be defined as: Stop sampling as soon as the sample path hits one of the three boundary lines (i) $y = 2x + 1$, (ii) $y = x - 2$, and (iii) $x + y = 100$.

As always, the likelihood does not recognize the stopping rule and comes out as

$$L(p) = f(w|p) = p^{X(w)}q^{Y(w)}$$

where $q = 1 - p$. The pair $X(w), Y(w)$ constitute the minimal sufficient statistic. The ML estimate is $\hat{p} = X/(X + Y)$. The range of the sufficient statistic (X, Y) consists of the boundary points

$(0,1)$,	$(1,3)$,	...,	$(33,67)$	on line	(i),
$(34,66)$,	$(35,65)$,	...,	$(50,50)$	on line	(iii), *and*
$(51,49)$,	$(50,48)$,	...,	$(2,0)$	on line	(ii).

The ML estimator $\hat{p} = X/(X + Y)$ monotonically increases from zero to unity as (X, Y) moves through the above set of boundary points. Hence $\hat{p}$ itself is minimal sufficient. Let us assert here without proof that $\hat{p}$ is a complete sufficient statistic in this case and that no nontrivial ancillary statistic exists.

Sir Ronald is no longer with us. So let me address the following questions to my good friend Eric Lehmann who is a living legend among us for his unparalleled erudition in Statistical Mathematics. The questions relate to Example 4.

Question: What should be our criterion for the choice of an estimate of p?

(The unbiasedness criterion is sort of vacuous in this case. There is only one unbiased estimator, which is zero or unity depending on whether the first trial results in an F or an S.)

Question: If ML is the chosen criterion, then how should we evaluate the estimate $\hat{p} = X/(X + Y)$? Does it make sense to evaluate $\hat{p}$ in terms of some average performance characteristics?

Question: Are all sample paths w equally informative?

(Even though there are no ancillary statistics in this case, we can still detect major qualitative differences between different sample paths. For instance, short sample paths like F or SS have very little to say about the parameter, whereas long paths that end on line (iii) are clearly much more informative.)

Question: Why do we need to decipher what the sample w has to say about the parameter p in terms of a sample space? Does the sample F

obtained following the rule **R** say anything different from the statement: A single Bernoulli trial has resulted in a failure?

Question: Do sample space ideas like bias, variance, risk function, etc., make any sense in this case?

Question: Why not act like a Bayesian and analyze the particular likelihood function generated by the data? Isn't is quite clear in this case that all that the data has to say about the parameter is summarized in the likelihood?

REFERENCES

Basu, D. (1988), *Statistical Information and Likelihood: A Collection of Critical Essays* by D. Basu, ed. J.K. Ghosh, Springer Verlag, New York.

Cox, D. R. (1971), The Choice Between Alternative Ancillary Statistics. *Jour. Royal Statist. Soc.* **(B)33**, 251-255.

Lehmann, E.L. (1981), An Interpretation of Completeness and Basu's Theorem. *Jour. Amer. Stat. Assoc.*, **76**, 335-340.

ERRATUM

Erratum to: Selected Works of Debabrata Basu

Anirban DasGupta

Department of Statistics and Mathematics
Purdue University
150 N. University Street
West Lafayette, IN 47907, USA
dasgupta@stat.purdue.edu

A. DasGupta (ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5825-9, © Springer Science+Business Media, LLC 2011

DOI 10.1007/10.1007/978-1-4419-5825-9_35

Table of Contents

The original table of contents we received when initially producing this title did not include some crucial permission informations. The updated TOC is as follows:

The online version of the original book can be found at
http://dx.doi.org/10.1007/978-1-4419-5825-9