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 I : 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.

To begin with, let us agree to the use of the notation

$$\text{Inf } (\mathcal{E}, x)$$

only as a pseudo-mathematical short hand for the ungainly expression : 'the whole of the relevant information about ω contained in the data $(\mathcal{E}, x)$'. At this point an objection may well be raised to the following effect : The concept of information in the data $(\mathcal{E}, x)$ 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, $\mathcal{E}$ and x, let us hasten to point out that at the moment we are concerned with variations in $\mathcal{E}$ and x only and so we are holding fixed the other two elements of II and q. That $\text{Inf}(\mathcal{E}, x)$ 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 $\omega+1, \omega+2, \ldots \omega+100$ where ω is an unknown number. Let $\mathcal{E}_n$ 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 $\mathcal{E}_2$ and $\mathcal{E}_{25}$ then, other things being equal, we shall no doubt prefer $\mathcal{E}_{25}$ to $\mathcal{E}_2$. Consider now the hypothetical situation where $\mathcal{E}_2$ 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, $\text{Inf}(\mathcal{E}_2, x)$ tells us categorically that $\omega = 15$ or 16. Now, contrast the above with another hypothetical situation where $\mathcal{E}_{25}$ 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 $\text{Inf}(\mathcal{E}_{25}, x')$ 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 $\mathcal{E}_{25}$ is 'more informative' than $\mathcal{E}_2$, it is equally incontrovertable that the particular sample (17, 115) from experiment $\mathcal{E}_2$ will tell us a great deal more about the parameter than will the sample (17, 20, ..., 52) from $\mathcal{E}_{25}$. To be more specific, with

$$\text{Inf } \{\mathcal{E}_n, (x_1, x_2, \ldots, x_n)\}$$

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 smaller 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) $p = p(x|\omega)$, the probability function, is a map $p : \mathscr{X} \times \Omega \to [0, 1]$ that satisfies the identity

$$\sum_{x \in \mathscr{X}} p(x|\omega) \equiv 1 \qquad \text{for all} \quad \omega \in \Omega.$$

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

$$\sum_{\omega \in \Omega} p(x|\omega) > 0 \qquad \text{for all} \quad x \in \mathscr{X}.$$

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 $\mathscr{X}_t = \{x | T(x) = t\}$ of $\mathscr{X}$ and the family $\{\mathscr{X}_t | t \in \mathscr{T}\}$ 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

$$p_T(t|\omega) = \sum_{x \in \mathscr{X}_t} p(x|\omega).$$

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

$$\{\mathcal{E}_t^T : t \in \mathcal{T}\}$$

of conditional experiments, with mixture probabilities

$$\pi_t = p_T(t \mid \omega), \quad t \in \mathcal{T}$$

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 $x_1 \in \mathcal{X}_1$ and $\omega \in \Omega$. 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 *transition 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 $x_1 \in \mathcal{X}_1$) 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 $\omega \in \Omega$ and $x_2 \in \mathcal{X}_2$.

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 $x_1 \in \mathcal{X}_1$, 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 $\mathcal{E}$ 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 $\mathcal{E}$.' It follows then that once an experiment $\mathcal{E}_1$ has been carried out resulting in the outcome y, it is not possible to add to the information $\mathrm{Inf}(\mathcal{E}_1, y)$ so obtained by carrying out a further 'post-randomization' exercise—that is, by performing a secondary experiment $\mathcal{E}_{(y)}$ whose randomness structure may depend on the outcome y of $\mathcal{E}_1$ but is completely known to the experimenter. Let us formally rewrite the above in the form

$$\mathrm{Inf}(\mathcal{E}_1, y) = \mathrm{Inf}\{(\mathcal{E}_1 \to \mathcal{E}_{(y)}),\ (y, z)\}$$

where $(\mathcal{E}_1 \to \mathcal{E}_{(y)})$ stands for the composite experiment '$\mathcal{E}_1$ followed by $\mathcal{E}_{(y)}$' and y, z are the outcomes of $\mathcal{E}_1$ and $\mathcal{E}_{(y)}$ respectively.

Now let $T : \mathcal{X} \to \mathcal{T}$ be a sufficient statistic for $\mathcal{E} = (\mathcal{X}, \Omega, p)$ and let $\mathcal{E}_T$ and $\{\mathcal{E}_t^T : t \epsilon \mathcal{T}\}$ be respectively the marginal experiment and the family of conditional experiments as defined in Section 2. Now, we may look upon a performance of $\mathcal{E}$ and the observation of the outcome x as 'a performance of the marginal experiment $\mathcal{E}_T$, observation of its outcome $t = T(x)$, followed by a post-randomization exercise $\mathcal{E}_t^T$ of identifying the exact location of x on the surface $\mathcal{X}_t = \{x' \mid T(x') = t\}$'. Since T is sufficient, the conditional experiment $\mathcal{E}_t^T$ is statistically trivial for every $t \epsilon \mathcal{T}$. 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 $\mathcal{S}$ (*The sufficiency principle*) : If, in the context of an experiment $\mathcal{E}$, the statistic T is sufficient then, for all $x \epsilon \mathcal{X}$ and $t = T(x)$,

$$\mathrm{Inf}(\mathcal{E}, x) = \mathrm{Inf}(\mathcal{E}_T, t).$$

If T is sufficient and $\mathcal{X}_t$ a particular T-surface, then from $\mathcal{S}$ it follows that $\mathrm{Inf}(\mathcal{E}, x)$ is the same for all $x \epsilon \mathcal{X}_t$. In the literature we often find the sufficiency principle stated in the following alternative (and perhaps a trifle less severe) form :

Principle $\mathcal{S}'$ (*Alternative version of* $\mathcal{S}$) : $\mathrm{Inf}(\mathcal{E}, x') = \mathrm{Inf}(\mathcal{E}, x'')$ 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 $\mathcal{S}'$:

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 $\mathscr{S}$. 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 $\mathscr{S}$, $\mathscr{S}'$, $\mathscr{S}$, $\mathscr{S}'$ and $\mathscr{L}'$ in the following:

Theorem 1: $\mathscr{S} \Longrightarrow \mathscr{S}'$, $\mathscr{S} \Longrightarrow \mathscr{S}' \Longleftrightarrow \mathscr{L}' \Longrightarrow \mathscr{S}'$.

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 $T : \mathscr{X} \to \mathscr{T}$ 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 $\mathcal{L}$ (*The likelihood principle*): If the data $(\mathcal{E}_1, x_1)$ and $(\mathcal{E}_2, x_2)$ generate equivalent likelihood functions on Ω, then $\mathrm{Inf}(\mathcal{E}_1, x_1) = \mathrm{Inf}(\mathcal{E}_2, x_2)$.

Before going into the far-reaching implications of $\mathcal{L}$, let us briefly examine the logical relationships in which $\mathcal{L}$ stands vis a vis the principles stated earlier. That $\mathcal{L} \Longrightarrow \mathcal{S}$ follows at once from the definition of similar experiments. From the definition of a sufficient statistic it follows that the likelihood functions $L(\omega \mid \mathcal{E}, x)$ and $L(\omega \mid \mathcal{E}_T, t)$ are equivalent, whenever T is sufficient and $t = T(x)$. So $\mathcal{L} \Longrightarrow \mathcal{S}$. Likewise, when T is ancillary, the likelihood functions generated by the data $(\mathcal{E}, x)$ and $(\mathcal{E}_t^T, x)$ are equivalent, whenever $t = T(x)$. Therefore, $\mathcal{L} \Longrightarrow \mathcal{C}$. The following theorem asserts that the two weak principles $\mathcal{S}$ and $\mathcal{C}$ are together equivalent to $\mathcal{L}$.

Theorem 2: $(\mathcal{S}$ *and* $\mathcal{C}) \Longrightarrow \mathcal{L}$.

Proof: Suppose the data $(\mathcal{E}_1, x_1)$ and $(\mathcal{E}_2, x_2)$ generate equivalent likelihood functions, that is, there exists $c > 0$ such that

$$L(\omega \mid \mathcal{E}_1,\ x_1) = cL(\omega \mid \mathcal{E}_2, x_2) \qquad \ldots \quad (*)$$

for all $\omega \epsilon \Omega$. Using $\mathcal{S}$ and $\mathcal{C}$ we have to prove the equality $\mathrm{Inf}(\mathcal{E}_1, x_1) = \mathrm{Inf}(\mathcal{E}_2, x_2)$. To this end let us contemplate the mixture experiment $\mathcal{E}$ of $\mathcal{E}_1$ and $\mathcal{E}_2$ 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 $\mathcal{E}$. In view of $(*)$ and our choice of the mixture probabilities, it is clear that the data $(\mathcal{E}, (1, x_1))$ and $(\mathcal{E}, (2, x_2))$ generate *identical* likelihood functions, and so from $\mathcal{S}$ it follows that

$$\mathrm{Inf}(\mathcal{E}, (1, x_1)) = \mathrm{Inf}(\mathcal{E}, (2, x_2)).$$

Now, applying $\mathcal{C}$ to each side of the above equality we arrive at the desired equality.

Since $\mathcal{S} \Longrightarrow \mathcal{S}'$, we immediately arrive at the following corollary which was proved earlier by A. Birnbaum (1962).

Corollary: $(\mathcal{S}$ *and* $\mathcal{C}) \Longrightarrow \mathcal{L}$.

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 $d = (\mathcal{E}, x)$ 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 $d = (\mathscr{E}, x)$, where $\mathscr{E} = (\mathscr{X}, \Omega, p)$ is a typical statistical experiment concerning ω and x a typical outcome that may arise when $\mathscr{E}$ is performed. But what data are attainable, in other words, what triples $(\mathscr{X}, \Omega, p)$ 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 $(\mathscr{X}, \Omega, p)$ 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 $(\mathscr{X}, \mathscr{A})$. 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 $\mathscr{E}$ for which $\mathscr{X} = \{1, 2, 3, \ldots\}$ and $p(x\,\omega) = \omega(1-\omega)^{x-1}$, $x \in \mathscr{X}$. It is not difficult to see how we can plan a (marginal) experiment $\mathscr{E}_1$ for which $\mathscr{X}_1 = \{0, 1\}$ and $p_1(0\,|\,\omega) = 1/(2-\omega)$. But can we plan an experiment $\mathscr{E}_2$ for which $\mathscr{X}_2 = \{0, 1\}$ and $p_2(0\,|\,\omega) = \sqrt{\omega}$ 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

$$L(\omega) = \text{Prob}\,(A\,|\,\omega)$$

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 $(\mathscr{X}, \Omega, p)$, 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 $\mathscr{E}_1$ and $\mathscr{E}_2$ 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 $t \in \mathcal{T}$, 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 $\mathcal{E}_t^T \in \mathfrak{E}$. [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 $\mathcal{E} \in \mathfrak{E}$, let us finally turn our attention to the range of Inf. If we accept the likelihood principle, i.e., if we agree that

$$\text{Inf}(\mathcal{E}_1, x_1) = \text{Inf}(\mathcal{E}_2, x_2)$$

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 $p(x|\theta)$ as a function of θ, the Fisher Information is defined as

$$I(\theta) = E_\theta\left\{\frac{\partial}{\partial\theta}\log p(X|\theta)\right\}^2$$

$$= -E_\theta\left\{\frac{\partial^2}{\partial\theta^2}\log p(X|\theta)\right\}$$

where X is regarded as a random variable ranging over $\mathcal{X}$. 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 $\Lambda(\theta) \doteq \log L(\theta)$—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

$$E_\theta J(\theta) = E_\theta \left\{ -\frac{\partial^2}{\partial \theta^2} \log L(\theta | X) \right\}$$

$$= \sum_x \left\{ \frac{\partial}{\partial \theta} \log p(x|\theta) \right\}^2 p(x|\theta) \qquad \ldots \quad (**)$$

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

$$q^*(\omega | \mathcal{E}, x) = L(\omega | \mathcal{E}, x) q(\omega) / \Sigma L(\omega | \mathcal{E}, x) q(\omega)$$
$$\sim L(\omega | \mathcal{E}, x) q(\omega).$$

In view of the above, a Bayesian should not have any qualms about identifying $\mathrm{Inf}(\mathcal{E}, x)$ with the likelihood function $L(\omega | \mathcal{E}, x)$.

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

$$\mathfrak{D} = (\mathcal{X}, \Omega, p, \mathcal{A}, W),$$

(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 $T : \mathcal{X} \to \mathcal{I}$ 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 $\mathscr{E}$) that subsumes H_0 as the hypothesis $\omega = \omega_0$. Consider now the case where we recognize two points x and x' in $\mathscr{X}$ 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 $\mathscr{S}'$, 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 $(\mathscr{E}, x)$ and $(\mathscr{E}, x')$ was made relative to a particular model $(\mathscr{X}, \Omega, p)$ 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 $\mathscr{S}'$ in the present instance, but is only doubting the adequacy or appropriateness of the particular statistical model $(\mathscr{X}, \Omega, p)$.

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 $\mathscr{E}$, the same data $(\mathscr{E}, x)$ 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 $\mathcal{L}$? 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 $\mathcal{L}$ assert that is not sufficiently firm or wide ? We must recognize this basic fact that $\mathcal{L}$ does not assert anything that can be measured in terms of its operating characteristics. It appears that in this instance Barnard is confusing principle $\mathcal{L}$ 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 $\mathcal{L}$, let us quote once again from Barnard and Sprott (1971, p. 176) :

"$\mathcal{L}$ applies to problems for which the model consists of a sample space $\mathcal{X}$, a parameter space Ω and a family of probability functions $p : \mathcal{X} \times \Omega \to R^+ \ldots$. For two such problems $(\mathcal{X}, \Omega, p)$ and $(\mathcal{X}', \Omega, p')$, principle $\mathcal{L}$ asserts that if $x \epsilon \mathcal{X}$ and $x' \epsilon \mathcal{X}'$ and $p(x \mid \omega)/p'(x' \mid \omega)$ is independent of ω, then the inference from x must be the same as the inference from x'. We may distinguish three forms of $\mathcal{L}$:

1. *Strongly restricted* $\mathcal{L}$: Principle $\mathcal{L}$ applicable only if $(\mathcal{X}, \Omega, p) = (\mathcal{X}', \Omega, p')$. This is equivalent to the sufficiency principle.

2. *Weakly restricted* $\mathcal{L}$: Principle $\mathcal{L}$ applicable (a) whenever $(\mathcal{X}, \Omega, p) = (\mathcal{X}', \Omega, p')$ and (b) when $(\mathcal{X}, \Omega, p) \neq (\mathcal{X}', \Omega, p')$ but there are no structural features of $(\mathcal{X}, \Omega, p)$ (such as group structures) which have inferential relevance and which are not present in $(\mathcal{X}', \Omega, p')$.

3. *Unrestricted* $\mathcal{L}$: Principle $\mathcal{L}$ applicable to all situations which can be modelled as above.

4. *Totally unrestricted* $\mathcal{L}$: 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 likelihood 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 $\mathcal{L}$ implies that the likelihood ratio $\rho = L(\omega_1)/L(\omega_0)$ generated by the data $(\mathcal{E}, x)$ 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 $\mathcal{L}$ 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 $\mathcal{L}$.

(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 $(\mathcal{E}, x)$ 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 $L : J \to [0, 1]$ defined as

$$(*) \qquad L(\theta) = \theta^2 / \sum_{\theta \in J} \theta^2,$$

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

$$(\bullet\bullet) \qquad \bar{l}(\theta)d\theta = 3\theta^2 d\theta$$

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 $\bar{L}_2 : J \to (0, 1]$ 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 $\theta \to \theta^2 = \phi$ 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 $\theta \to \phi$ 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 Likelihood 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

$$L(\theta \mid x) = \begin{cases} .02 & \text{for } \theta = x \\ .001 & \text{for } \theta = x a_i^{-1}, \quad i = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

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 .08. 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 $J = (x/(10.1), x/(0.9))$ is .08. 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, $\mathcal{E}$, 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

$$(\bullet) \qquad L(\theta) = L(\theta, \phi \mid x) = \begin{cases} .08 & \text{when} \quad \theta = \quad x \\ .02 & \text{when} \quad \theta = -x \end{cases}$$

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 .08.

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

$$(\bullet\bullet) \qquad L(\theta, \phi \mid x, y) = \begin{cases} .001 & \text{when} \quad \theta = x, \quad \phi \epsilon J \\ .02 & \text{when} \quad \theta = -x, \ \phi = y \\ 0 & \text{otherwise} \end{cases}$$

and then eliminating ϕ from $(\bullet\bullet)$ as per method (f) of Section 7. The eliminated likelihood function is

$$(\bullet\bullet\bullet) \qquad L_e(\theta) = \sup_{\phi} L(\theta, \phi \mid x, y) = \begin{cases} .001 & \theta = x \\ .02 & \theta = -x \end{cases}$$

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) does 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 0 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 ($\mathcal{L}$) it then follows that at the time of analysing the information contained in the data $(\mathcal{E}, x)$, we need not concern ourselves about the exact nature of the experiment $\mathcal{E}$—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 $\mathcal{L}$ 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 $\mathscr{S}$ the following set of three classical (Fisherian) methods of statistical inference.

The $\mathscr{S}$ 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)

$$\text{(i)} \qquad L(\theta) \sim \exp\{-n(\theta - x_{(n)})^2/2\}$$

where $x_{(n)} = (x_1 + x_2 + \ldots + x_n)/n$. Under the above circumstances, let $\mathscr{S}$ consist of the trilogy of statistical methods:

$\mathscr{S}(a)$: If $|x_{(n)} - \theta_0| > 3/\sqrt{n}$, then reject the null-hypothesis $H_0: \theta = \theta_0$ and declare that the data is highly significant.

$\mathscr{S}(b)$: The statement $\theta \in (x_{(n)} - 3/\sqrt{n},\ x_{(n)} + 3/\sqrt{n})$ may be made with a great deal (well over 99%) of 'self-assurance' or 'confidence'.

$\mathscr{S}(c)$: 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 $x_{(n)}$ satisfies the inequality $|x_{(n)}| > 3/\sqrt{n}$.

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 $E(N \mid 0)$ 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)

$$\text{(ii)} \qquad L(\theta) = (\sqrt{2\pi})^{-N} \exp\{-\Sigma(x_i - \theta)^2/2\}$$

$$\sim \exp\{-N(\theta - x_{(N)})^2/2\}.$$

*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 hypothesis $H_0 : \theta = 0$.

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

$$\pi(\theta) = \Pr(\text{Test ends with rejection of } H_0 | \theta)$$

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 $|\bar{x}_{(n)}| > 3/\sqrt{n}$ 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 $L(\theta|x)$ 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 $\theta \leq 0$ 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

$$\text{(iii)} \qquad \Pr(H_\delta|x) = \int_{-\delta}^{\delta} \frac{\sqrt{N}}{\sqrt{2\pi}} \exp\{-N(\theta - x_{(N)})^2/2\} d\theta$$

$$= \Pr\{-\delta\sqrt{N} - \sqrt{N}x_{(N)} < Z < \delta\sqrt{N} - \sqrt{N}x_{(N)}\}$$

where Z is a $N(0, 1)$ variable.

The stopping rule $\mathcal{T}_M$ is such that with a fair sized N the sample mean $x_{(N)}$ is either roughly equal to $\pm 3/\sqrt{N}$ or is some number in between. Let us consider the situation when $x_{(N)}$ is just above $3/\sqrt{N}$ and ignore the overshoot. Formula (iii) now becomes

$$\text{(iv)} \qquad \Pr(H_\delta|x) = \Pr(-\delta\sqrt{N} - 3 < Z < \delta\sqrt{N} - 3)$$

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 $\mathcal{S}(\alpha)$ 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 $|x_{(n)}| > 3/\sqrt{n}$, 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 $|\sqrt{n}\, x_{(n)}|$ 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 $\sqrt{n}\, x_{(n)} = 3$, 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 $\mathcal{H}$ 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 $\mathcal{H}$ 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 $\int_0^\infty f(y)dy$ divergo. 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

$$\text{(iv)} \qquad \Pr(Y > 10\theta|\theta) = \int_{100}^{b\theta} p(y|\theta)dy$$

$$= \int_{10}^{b} f(y)dy = 0.99$$

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 $\mathcal{L}$. 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 $\mathcal{L}$ 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 $Y \sim N(\theta, \sigma)$, 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 M^{r}_{ijk} and the LR methods of Section 7.

The other slippery route that will generate the 3σ-intervals (vi) and (vii) from $\mathcal{L}$ 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 ritual 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 $\mathcal{LB}$: 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 $\mathcal{LB}$ is a corollary to $\mathcal{L}$. 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 $\mathcal{LB}$ (and not $\mathcal{L}$) 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

$$\text{(a)} \qquad \Pr(A \mid Y = y) > 0.99 \quad \text{for all } y \epsilon (0, \infty),$$

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

$$q(\theta|y) = \begin{cases} a\theta^{-1} \exp\left\{-50\left(\frac{\theta}{y} - 1\right)^2\right\}, & b^{-1}y \leqslant \theta < \infty \\ 0 & 0 < \theta < b^{-1}y \end{cases}$$

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 $\mathcal{JS}$, 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_S 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. Acc. 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 $i = 1, 2, ..., N$ 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 $\mathcal{F}(a)$ and $\mathcal{F}(b)$ 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 $\omega \varepsilon A$ and $\omega \varepsilon B$ 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 scientist.

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 choice 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_A 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_A 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, which 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 $H \in \Phi$, 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.