On Resolving the Controversy in Statistical Inference

By G. N. Wilkinson

Rothamsted Experimental Station, Queen's University, Ontario and Bell Laboratories, Murray Hill, N.J., U.S.A.†

[Read before the Royal Statistical Society, at a meeting organized by the Research Section, on Wednesday, February 9th, 1977, Professor S. D. Silvey in the Chair]

Summary

The controversy concerning the fundamental principles of statistics still remains unresolved. It is suggested that one key to resolving the conflict lies in recognizing that inferential probability derived from observational data is inherently *noncoherent*, in the sense that their inferential implications cannot be represented by a single probability distribution on the parameter space (except in the Objective Bayesian case). More precisely, for a parameter space R^1, the class of all functions of the parameter comprise equivalence classes of invertibly related functions, and to each such class a logically distinct inferential probability distribution pertains. (There is an additional cross-coherence requirement for simultaneous inference.) The noncoherence of these distributions flows from the nonequivalence of the relevant components of the data for each.

Noncoherence is mathematically inherent in confidence and fiducial theory, and provides a basis for reconciling the Fisherian and Neyman–Pearsonian viewpoints. A unified theory of confidence-based inferential probability is presented, and the fundamental incompatibility of this with Subjective Bayesian theory is discussed.

Keywords: ANCILLARY; BAYESIAN THEORY; CONDITIONING PRINCIPLE; CONFIDENCE THEORY; FIDUCIAL THEORY; FISHER; INFERENTIAL PROBABILITY; LIKELIHOOD PRINCIPLE; NEYMAN–PEARSON; NONCOHERENCE; PARADOXES; PIVOTAL; PRINCIPLES OF INFERENCE; RELEVANCE PRINCIPLE; STATISTICAL INFERENCE; SUFFICIENT.

1. Introduction

A few preliminaries first. The discussion of statistical inference is restricted here to the parametric context of inference about one or more parameters of an otherwise known sampling distribution for the observational data. (The word "known" here signifies a precisely or approximately known property of the real world, in contradistinction to "specified" or "given" which can refer more widely to theoretical or mathematical postulates.) The omission of significance tests and nonparametric or robust inference is for reasons of space rather than of principle. Decision theory is also omitted from consideration. Its objectives differ formally from those of statistical inference and there are mathematical conflicts between the relevant principles of each. See Barnard (1949), Fisher (1956a), Cox (1958), Tukey (1960), Blyth (1970) and Barnett (1973).

1.1. *The Nature of the Controversy*

The controversy concerning the fundamental principles of statistical inference is still unresolved. Barnett (1973) gives a comparative discussion. See also Hacking (1965), Plackett (1966), Fraser (1972, 1974) and Edwards (1976). The principal disagreements are between Bayesian (or non-Bayesian likelihood) and frequentist theories of inference; and among the latter, of course, the long-standing disagreement between the Fisherian and Neyman–Pearsonian points of view. It must be emphasized that the substance of the

† Present address: Mathematics Research Centre, Madison, Wisconsin (from September, 1977).

controversy, and of this paper, concerns inference from finite (and relatively small) amounts of data. All theories have some form of asymptotic validity, in that actual numerical differences disappear asymptotically as the amounts of data become large; only the formal differences of expression remain.

I think that the continuing controversy has had, overall, a debilitating effect on Statistics, particularly if one considers the forward impetus that a unified theory of inference would produce. Much of the published literature would become irrelevant, teaching would be enhanced and simplified, and even if much of the practice of Statistics continued as before, there would be changes in the research on and published justifications for practical methods.

Contrary to Barnett's (1973) pessimistic view, I believe that a unified theory of inference *does* exist. The main thrust of this paper will be to demonstrate that a reconciliation of the frequentist viewpoints is possible, and hence to throw into a new and sharper focus the fundamental incompatibility of the frequentist and Bayesian viewpoints. One or other must be untenable as formulated, and the conflict can only be resolved by considering the necessary empirical properties that a theory of inference must encompass.

1.2. *Empirical Principles of Statistical Inference*

In considering why the controversy has continued for so long, one notices first a peculiar difficulty in invalidating a possibly erroneous theory of uncertain inference. Scientific theories about Nature ultimately stand or fall on the basis of clear observational disproof (or otherwise) of predicted properties of Nature. In the case of uncertain inference, however, the very uncertainty of uncertain predictions renders the question of their proof or disproof almost meaningless. Invalidation of an inference theory therefore depends on the discovery of extreme examples from which the derived statements of uncertainty are quite clearly incorrect from an intuitive or empirical point of view. Pathological freak examples are insufficient for this purpose. An element of continuity between the extreme and more practical cases is needed, to argue properly that the detected form of incorrectness must be present to some degree in nearly all such cases.

Thus Cox (1958) gives an example which throws into sharp relief the inferential conflict between Fisher's (1956a) Conditioning principle and the Neyman–Pearsonian principle of optimizing power or its equivalent in Confidence theory; and from which it is clear that the latter principles are in some respects inferentially unsound. This led Cox to suggest that a conditional confidence theory is needed. Certainly the Conditioning principle enjoys wide empirical acceptance, and I know of no sound counter-examples against it.

Secondly, one may question whether mathematical reasoning is being used correctly in statistical inference about the real world. Fisher (1956a) has discussed the peculiar features of uncertain inference, and his views may be summarized as:

Relevance Principle

Valid uncertain inference requires that all relevant information be properly utilized in, and all irrelevant or spurious information be excluded from, the reasoning of it.

This may be considered the fundamental principle of uncertain inference. Something like it has been understood for centuries as a fundamental tenet in Law (or at least in British-based Law). Though its general import is clear, application of it requires definition of the terms *relevant* and *properly*. Resolution of this question of definition is not simply a matter of logic but ultimately one of empirical verification. Statistical inference in this respect is as fundamentally empirical as the sciences in which it is applied. In the unified theory to be outlined, Fisher's Conditioning principle provides the operational means of separating relevant and irrelevant information.

Rigorous application of the Relevance principle does exclude some approaches to inference as fundamentally invalid, for instance Bayesian theories invoking artificial priors (invariant, conjugate etc.). Nor is it right to argue that introducing such artificial information is valid because the resulting conclusions are almost independent of it. One would have to argue that the same conclusions could be reached without introducing the false information, and this may not follow in general. One might also argue that parametric inference itself is fundamentally unsound, because it cannot be known with certainty that the sampling distribution of the data belongs to the family assumed. But here at least one can argue that such a family may be sufficiently representative of the undoubtedly larger class of possible distributions to validate the inference.

Thirdly, one may ask, is there some fundamental property of statistical inference that has been overlooked? I believe there is, and this is really the crux of the paper. I shall argue that the following principle describes an inherent and empirically recognizable property of finite observational data:

Noncoherence Principle

The inferential implications of observational data alone (with sampling distribution known in parametric form) are noncoherent, in that they cannot be represented by a single inferential probability distribution on the parameter space.

A detailed explanation of noncoherence will be given in Section 3, but I draw attention here to the radical implication that inferential probability does not conform in general to the classical Kolmogoroff axioms of mathematical probability.

1.3. *A Confidence-based Theory of Inferential Probability*

There are two forms of inferential probability that have an objective frequency interpretation. The first, Objective Bayesian probability, is a special form rarely applicable in practice since the additional knowledge of a prior frequency distribution is needed. The other is the fiducial form discovered by Fisher (1930). This has (in simpler cases) the familiar confidence interpretation which Neyman (1934) subsequently exploited in his theory of confidence intervals.

Bayesian probability is coherent, but fiducial probability is not. This Fisher apparently did not perceive. The noncoherence, however, is deducible from confidence theory. Recognition of it leads to a resolution of all the known difficulties with fiducial theory, and hence to a mathematically consistent calculus of confidence-based inferential probability that will be outlined in Section 4. The fundamental conflict between this and Subjective Bayesian theory will be examined in Section 6. Structural distributions (Fraser, 1968) are a special form of fiducial distribution (Section 4.9). (*Added in proof*: Fraser disagrees, see Discussion and Reply.)

2. NOTATION

Actual observations will be distinguished from mathematical variables with a subscript a as in x_a. An actual but unknown parameter will be similarly distinguished, e.g. θ_a, but representation of the uncertainty about the value of θ_a in the form of a probability distribution will require that θ_a be *formally* regarded as a random variable in inferential statements. The subscript a will sometimes be omitted when no ambiguity arises. To avoid confusion the term *probability* will be used only in its inferential sense (Section 3). Sampling distributions will be described with the term *frequency*. The same distinction will apply in the notations used: p, P for probability and f, F for frequency. The notation P(*proposition*) will denote the probability that the proposition is true. However, the particular form $P(\theta < \theta_1)$ may be condensed to the usual form $P^{(\theta)}(\theta_1)$ for a distribution function or to $P(\theta)$ when no mathematical variable θ_1 need be specified. Similarly for frequency, F.

The only sample and parameter spaces considered are the real line (R^1) or, more generally R^p; with their associated Borel algebras. Distributions will be represented by a differential notation, for example, $dP(\theta_a)$, $dF(x)$; or more fully $dP(\theta_a;\, x_a)$, $dF(x;\, \theta)$ to indicate functional dependence on x_a and θ respectively.

Functions (transformations) are restricted to be essentially continuous (e.c.), with at most a countable number of discontinuities, so that Borel structure is preserved under transformation. Invertibility will be indicated by the operator symbol $\rightleftharpoons$ as in $g(\theta) \rightleftharpoons \theta$ or in $g(\theta_1, \theta_2) \rightleftharpoons \theta_1$, where the invertibility is with respect to θ_1 given θ_2. Monotone relations are specified with the operators $\uparrow$, $\downarrow$, $\uparrow\downarrow$ (meaning $\uparrow$ or $\downarrow$). Compound symbols $x_2 | x_1$ or $\theta_2 | \theta_1$ are used with the meaning "x_2 given x_1", etc. When applicable, the phrase "for all (values of other variables)" will usually be omitted and should be taken as understood in the absence of other qualifications.

3. Inferential Probability and Noncoherence

I take it as fundamental that inferential probability is a measure of belief attaching to propositions about the real world, like "it will rain here tomorrow" or "$2.1 < \mu < 5.3$", which are quite definitely true or false, but whose truth value is currently unknown. The probability is an estimate of the truth value. This use of the term *probability* is logically distinct from its other use as a synonym for theoretical frequency, in describing factual properties of random variables. The distinction is so important here that I shall not use the word *probability* at all in the latter sense, referring instead to *frequency*. See Shafer (1976).

3.1. *Basic Definitions*

Inferential probability may be determined objectively in two ways, Bayesian and Fiducial, which we now consider. The common feature of both is that the particular proposition under consideration is identified as one of a relevant conceptual class of similar propositions in which the relative frequency of true propositions is known. The required probability is then equated (in magnitude only, not logically) to that relative frequency.

We are concerned here only with nontrivial inference from observational data, and consider the simplest case of an observation x_a, the observed value of a random variable x, with distribution function $F(x;\, \theta)$ known in terms of a mathematical variable θ representing an unknown parameter which actually is θ_a. The sample and parameter spaces are the real line.

Since any nontrivial expression of uncertainty will be functionally dependent on the observed value x_a, we are led to consider propositions about θ_a of the form $\theta_a < \theta_p(x_a)$, for which the assigned probability p will be functionally independent of x_a:

(i) *The Objective Bayesian case.* If it is known additionally that θ_a is itself the unknown but realized value of a physcial random variable θ with known frequency distribution function $F(\theta)$, then the pair (θ, x) has a known bivariate frequency distribution from which the conditional frequency distribution function $F(\theta | x_a)$ is immediately deducible. The inferential probability of $\theta_a < \theta_p(x_a)$ is then defined as

$$P(\theta_a < \theta_p(x_a)) = F^{(\theta|x_a)}(\theta < \theta_p(x_a)) = p, \tag{3.1}$$

with $\theta_p(x_a)$ identified as the pth percentile value of θ given x_a.

(ii) *The Fiducial case.* If nothing is known about θ_a other than inherent in x_a and its sampling distribution, no Bayesian conditional frequency distribution exists. However, if there is a well-defined *simple ordering relation* between x and θ, in that for every p, $0 < p < 1$, the equation $F(x;\, \theta) = p$ defines a monotone relation between x and θ, then an inferential probability distribution for θ_a is determined by the definition

$$P(\theta_a < \theta_p(x_a)) = F^{(x|\theta)}(\theta < \theta_p(x)) = p, \tag{3.2}$$

where now $\theta_p^{-1}(\theta)$ is identified as the pth percentile value of x given θ if $F(x;\, \theta) \uparrow \theta$, or of $-x$ if $F(x;\, \theta) \downarrow \theta$. Note there is an additional requirement that the ordering relation be *irreducible* (see Section 4.1), and also that fiducial distributions may be incomplete, with some probability unassigned (see Section 3.4).

Let me emphasize again the similarity of the two definitions above, the principal difference being in conceptual class of propositions relevant to each case:

Proposition	*Bayesian class*	*Fiducial class*
$\theta_a < \theta_p(x_a)$	$\{\theta < \theta_p(x_a)\}$ (θ variable)	$\{\theta < \theta_p(x)\}$ (θ, x variable)

(3.3)

Fiducial probability has a *confidence interpretation.* The 2-dimensional class of propositions $\{\theta < \theta_p(x)\}$ comprises 1-dimensional subclasses $\{(\theta < \theta_p(x));\ \theta\ \textit{given}\}$ for every one of which the associated *confidence frequency* (frequency of true propositions) is p. Fiducial probability may also be interpreted geometrically as *projected frequency* (see Section 3.3).

An important operational concept in fiducial theory is that of a *pivotal variate* $u(x, \theta)$ monotone in both x and θ and with known frequency distribution $dF(u)$ independent of θ. (I have substituted "variate" for "quantity" in Fisher's terminology to emphasize that there is an associated frequency distribution.) The distribution function $F(x;\ \theta)$, if monotone in θ, is itself a pivotal variate, but simpler transforms of it often exist. For instance, if $x \sim N(\theta, 1)$, $x-\theta$ is a pivotal variate. With $u_a = u(x_a, \theta_a)$ the proposition $\theta_a < \theta_p(x_a)$ is logically equivalent to the proposition $u_a < u_p$ if $F(x;\ \theta) \uparrow \theta$, of which the subject is the *unknown* but realized value u_a of the variate u.

For the Normal example above, if z denotes a Standard Normal variate, the projection of frequency as a measure of inferential probability in (3.2) may be represented symbolically by the implicational expression

$$x = \theta + z \overset{\text{inf}}{\Rightarrow} \theta_a = x_a - z. \tag{3.4}$$

The qualifier *inf* is inserted to emphasize that the implication is an *inference,* not a *deduction*; $\theta_a = x_a - z$ is not a statement of fact, but formally confers on the unknown θ_a the inferential status of random variable.

In conformity with Fisher's terminology, I shall also refer to a simple ordering relation as a *pivotal relation,* and the component monotone relations as *percentile relations.*

3.2. *The Noncoherence of Fiducial Probability*

The *definitional equivalence* in (3.1) and (3.2) of *inferential probability* and *frequency* determines an *isomorphism* between them that defines the relevant calculus of probability. More precisely the isomorphism is between corresponding classes of probability and frequency distributions, respectively, generated by a class of transformations g on the relevant spaces, such that definitional equivalence is *preserved.*

In the Objective Bayesian case it is clear that definitional equivalence is preserved under any 1–1 or many–1 transformation g, with $\theta_a \rightarrow g(\theta_a)$, $\theta \rightarrow g(\theta)$ respectively. Hence the calculus of probability and of frequency is the same in this case. We shall say that Objective Bayesian probability measure is *coherent* for all e.c. functions $g(\theta)$.

However, in the fiducial case, definitional equivalence is preserved only under *invertible* transformations g, with $\theta_a \rightleftharpoons g(\theta_a)$, $\theta \rightleftharpoons g(\theta)$ and $x \rightleftharpoons g(x)$, since it is easily shown that noninvertible transformations as above alter the associated confidence frequencies. Thus a fiducial probability measure $dP(\theta)$ is coherent only for invertible functions $g(\theta) \rightleftharpoons \theta$. This, of course, has radical implications for the calculus of fiducial probability.

Invertible transformations form a group which defines, as cosets in the class of e.c. transformations, equivalence classes of *invertibly related* transformations, and to each equivalence class a logically distinct probability distribution pertains. Fiducial probability distributions (if they exist) for noninvertibly related functions of θ are *noncoherent,* and each needs to be deduced directly from the relevant components of the observational data and the pivotal relations which they imply. Noncoherence flows essentially from the mathematical nonequivalence of the relevant ordering relations. There is an additional cross-coherence requirement for a simultaneous probability distribution, described in Section 4.

Empirical evidence for noncoherence as an inherent property of (finite) observational data will be considered in Section 3.5, and other logical implications of it discussed in Section 3.6. See also Sections 4.6 and 5.

3.3. *Geometrical Interpretation of Fiducial Probability*

Mathematically the definition (3.2) may be considered a definition of probability measure by *set-intersection* (A. T. James, personal communication, 1962) or, as I shall term it here, *frequency projection*, in contradistinction to *frequency conditioning* as in (3.1). A geometrical representation is given in Fig. 1. All the known data are indicated in the figure. The known family of frequency distributions $dF(x;\ \theta)$ assigns a frequency measure on all (horizontal) x-lines. This has been indicated by drawing the corresponding x-intervals (say with frequency measure 5 per cent) defined by two of the monotone percentile relations between x and θ. The observed value x_a is represented by the (vertical) θ-line $x = x_a$, on which an inferential probability measure is to be determined.

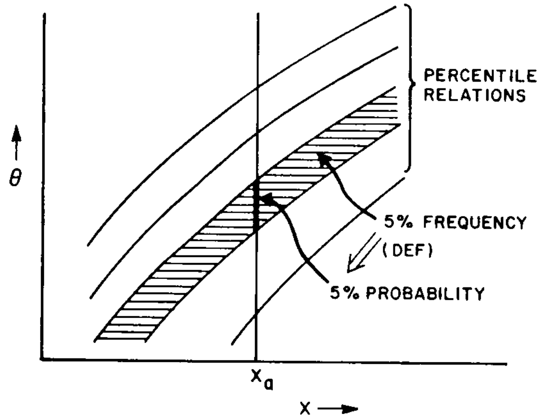

FIG. 1. Frequency measures in the (x, θ) plane.

No bivariate frequency distribution exists on the (x, θ) plane, and hence no conditional frequency distribution on any θ-line. However, the union of the corresponding x-intervals shown generates a 2-dimensional strip which in a well-defined sense constitutes 5 per cent of the (x, θ) plane. In this way a frequency measure can be established for the special σ-algebra of 2-dimensional sets generated from percentile-defined strips.

A projected frequency measure can now be assigned to the θ-line $x = x_a$ by assigning the frequency associated with each percentile strip to its interval-intersection with $x = x_a$. The projected distribution for θ_a is invariant with respect to prior monotone transformation of the sample space $\{x\}$, and more generally with respect also to any nonmonotone but invertible e.c. transformation of $\{x\}$,† though in the latter case the discontinuities induce a reordering of frequency elements in the projection process, since the simple ordering relation between x and θ is not preserved.‡

† *Added in proof*: provided that the corresponding transforms of the monotone percentile relations are used for projection. I am grateful to Pedersen (see Discussion) for detecting the fault in wording here.

‡ Note that the unique ordering of points on the real line is essential to the definition given of a pivotal relation. On the circle, for instance, there are no fixed special points analogous to ±∞, and the existence of a uniquely relevant pivotal relation depends on the particular form of sampling distribution assumed. Invariance under rotation of either the entire distributional form or else of symmetry about a diameter appears to be necessary.

3.4. *Incomplete Fiducial Distributions*

A pivotal relation between x and θ is termed *complete* if all the monotone percentile relations span the whole space $\{x\}$; otherwise it is *incomplete.* An incomplete pivotal relation produces an incomplete inferential probability distribution, with some probability unassigned.

An incomplete distribution for θ usually arises when there is some specified restriction on the range of possible values for θ, say to a semi-infinite or finite interval, as illustrated in Fig. 2, in which upper and lower amounts of frequency P_U, P_L remain unassigned by the projection process, and are considered as unassigned probability.

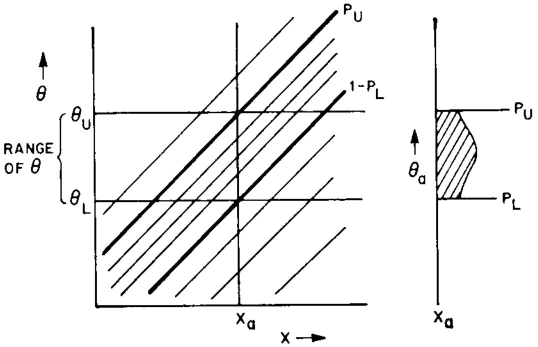

FIG. 2. An incomplete fiducial distribution produced by a range restriction. Unassigned probability is shown as condensations P_L, P_U at $\theta = \theta_L$, θ_U respectively.

It is formally convenient to represent the unassigned probability as finite probability condensations superimposed on (but not logically identified with) the endpoints θ_U, θ_L of the range of θ. This representation is mathematically consistent with upper and lower truncation of the distribution function for θ, and also with the fact that, in considering the confidence frequency interpretation, truncation (to the limits θ_U, θ_L) of conceptual confidence intervals extending beyond the specified range does not affect the confidence frequency.

From the inferential viewpoint, however, the unassigned probability corresponds to unassigned belief and the resulting inferential distribution is more aptly regarded as partly indeterminate. The incomplete distribution assigns an objective lower bound (with confidence interpretation) to belief in any propositional θ-interval, and an upper bound is obtained by adding the amounts P_L, P_U to the lower bound. That the data themselves provide no additional information on how to distribute the unassigned belief is intuitively clear. To distribute it in proportion to that already distributed (that is by conditioning the assigned probability to add to one) would be an unwarranted assumption or an exaggeration of the evidence. For consider a limiting case where the specified range of θ is so small, relative to the standard deviation of x, that nearly all the probability is unassigned. Considering how uninformative is an observation x_a in such a case, I suggest that the almost completely indeterminate distribution is a far more plausible expression of inference than the precisely determined, complete and almost uniform inferential distribution produced by conditioning as above. The latter is tantamount to assuming Bayes' Axiom.

An interpretation of unassigned probability sometimes applicable (though not in the case described above) ties in with significance tests of the mathematical specification of distribution. An unassigned probability is often the complement of the significance probability produced by such a test, and thus a statistical measure of the degree of incompatibility (significant or otherwise) between the data and the mathematical specification. An example is given in Section 3.5.

Another possible interpretation is when there is *a priori* a finite (nonzero) chance, under the relevant scientific theory, that the unknown parameter lies exactly at a point of condensation. The associated probability may then be regarded as an estimated upper bound (*a posteriori*) for belief in such a possibility.

3.5. *Empirical Evidence for Noncoherence as an Intrinsic Property of Observational Data*

Stein (1959), referring to remarks made by Fisher, was led to produce an extreme example in which a derived fiducial distribution was clearly absurd if judged by its confidence properties. This example, reinterpreted in the present context, provides the clearest empirical evidence for noncoherence of which I am aware, and we consider it now. Other illustrations of noncoherence will occur elsewhere in the paper.

Taking some knowledge of fiducial theory for granted, we consider n independent random variables $x_i \sim N(\mu_i, 1)$, with observed values x_{ia}, $i = 1, 2, \ldots, n$, and let $\mathbf{x}$, $\mathbf{x}_a$, $\boldsymbol{\mu}$ denote the corresponding vectors in R^n. We assume there is no prior knowledge about any μ_i. The confidence-based fiducial distribution for any μ_i is formally specified by $\mu_i \sim N(x_{ia}, 1)$, and similarly for any specified contrast $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mu}$ by $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mu} = \mathrm{N}(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{x}_a, \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\lambda})$. The totality of such distributions is formally specified by assigning $\boldsymbol{\mu}$ a Spherical Normal distribution,

$$\boldsymbol{\mu} \sim \mathrm{N}(\mathbf{x}_a, \mathbf{I}_n). \tag{3.5}$$

Now consider inference about the length $|\boldsymbol{\mu}|$ of $\boldsymbol{\mu}$, or equivalently the invertibly related function $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} = |\boldsymbol{\mu}|^2$. Note immediately that real interest in such a function logically contradicts our previous statement of no prior information about any μ_i, for clearly the origin $\mathbf{0}$ is now designated *a priori* to be a special point in the space $\{\boldsymbol{\mu}\}$.

In view of the mathematical noncoherence of fiducial probability deduced above we would not expect the marginal distribution of $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$ derived from (3.5) and specified by

$$\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} \sim \chi^2(n, \mathbf{x}_a^{\mathrm{T}}\mathbf{x}_a) \tag{3.6}$$

to be inferentially valid, since the confidence properties of (3.5) would be destroyed by the marginal transformation. This is precisely what Stein demonstrated.

A correct, confidence-based fiducial distribution for $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$ must derive from a fully relevant pivotal relation between some component function of $\mathbf{x}_a$ and $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$. The relevant component is clearly $\mathbf{x}_a^{\mathrm{T}}\mathbf{x}_a$, which has a noncentral χ^2 distribution,

$$\mathbf{x}^{\mathrm{T}}\mathbf{x} \sim \chi^2(n, \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}), \tag{3.7}$$

depending only on the relevant parameter $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$. The other component of observational information, namely the angular orientation of $\mathbf{x}_a$, is clearly irrelevant to inference about $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$ because of the spherical symmetry of the problem (see Section 4.5). Fiducial inversion of (3.7) leads to an inferential distribution for $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$ with the requisite confidence property. The inferential distribution function is

$$\begin{aligned} P(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}) &= 1 - F^{(\mathbf{x}^{\mathrm{T}}\mathbf{x})}(\mathbf{x}_a^{\mathrm{T}}\mathbf{x}_a;\, \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}) \\ &= F\{\chi^2(n, \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}) > \mathbf{x}_a^{\mathrm{T}}\mathbf{x}_a\}. \end{aligned} \tag{3.8}$$

Note this distribution is incomplete, with a condensation at zero of unassigned probability, $p_0 = F(\chi_n^2 > \mathbf{x}_a^{\mathrm{T}}\mathbf{x}_a)$. A high value of p_0 would indicate significant evidence that the observed point $\mathbf{x}_a$ is too close to $\mathbf{0}$ to be statistically compatible with the assumed covariance matrix $\mathbf{I}_n$ of $\mathbf{x}$ or else with the Normal form of distribution.

Stein showed that in the limit, with n increasing, the actual confidence frequencies associated with the (incorrect) fiducial distribution (3.6) tend to zero. The following calculations (for which I am indebted to W. N. Venables) show how extreme the difference between (3.6) and (3.8) can be, even for moderate n. With $n = 50$ and $\mathbf{x}_a^{\mathrm{T}}\mathbf{x}_a = 100$, the central 95 per cent fiducial intervals produced are (to nearest integral values)

Incorrect, from (3.6): $109 < \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} < 196$,

Correct, from (3.8): $21 < \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu} < 89$, (3.9) the latter interval consistent with the fact that $E(\mathbf{x}^{\mathrm{T}}\mathbf{x}) = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}+50$, which suggests a central value $100-50 = 50$ for $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$.

Since the spherical distribution (3.5) produces entirely plausible inferential distributions for any μ_i or specified contrast $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\mu}$, and at the same time produces an intuitively absurd distribution for $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}$, the real force of Stein's example is to demonstrate empirically that noncoherence is no mere mathematical artifact but an intrinsic property inherent in the observation themselves, and flowing from the nonequivalence of the relevant information for each kind of inference. For the fundamental ingredient of a confidence-based inferential distribution is a simple ordering relation between a relevant statistic and a parameter, and even a single observational variable implies a multiplicity of inferentially nonequivalent relations.

3.6. *More about Noncoherence*

Consider again the set of all e.c. functions $g(\theta)$, $\theta\in R^1$. Each equivalence class C of invertibly related functions g corresponds to a unique σ-algebra on $\{\theta\}$ which is the inverse image of the σ-algebra on the range $\{g(\theta)\}$ to which the full Borel algebra on $\{\theta\}$ is mapped by the transformation $\theta\rightarrow g(\theta)$. Invertibly related functions determine the same inverse σ-algebra, which is either the full Borel algebra on $\{\theta\}$, corresponding to invertible functions, or else a subalgebra of it, for instance, the subalgebra generated from intervals symmetric about zero in the case of θ^2 (and functions invertibly related to it). Thus the restricted coherence of fiducial probability implies that a logically distinct probability measure pertains to each such σ-algebra.

Noncoherence goes deeper than this, however. As indicated in Stein's example we must distinguish between different parametric or topological representations of the parameter space on *a priori* grounds. Consider the simple case $x\sim N(\theta,1)$. If the relevant scientific context indicates that any value $\theta\in R^1$ is possible for θ_a, none being especially indicated, then logically our inferential interest is confined to invertible functions of θ. If, however, the scientific context indicates 0 as a special value (corresponding to a special form of the theory) then the logically appropriate representation of the parameter space $\{\theta\}$ is as the product space $\{|\theta|\}\otimes\{\mathrm{sign}(\theta)\}$, and inferential interest is directed to the components $|\theta|$ and $\mathrm{sign}(\theta)$. Though the details are omitted here, the latter representation leads to a fiducial distribution for θ, expressed in the form $|\theta|\,\mathrm{sign}(\theta)$, which is noncoherent with the distribution $\theta\sim N(x_a, 1)$, exhibiting relative to the latter a shrinkage towards the value 0 and a condensation of probability at 0 equal to $P(\chi_1^2 > x_a^2)$. It thus appears that noncoherence provides the logical rationale for inference based on shrinkage estimators as in ridge-regression, without invoking arbitrary optimization principles. Noncoherence also accounts for the *inadmissibility* (Stein, 1956) of the mean of a multivariate Normal sample with respect to a quadratic loss function, since the decision-theoretic formulation directs attention to this one function on the parameter (and sample) space.

Logicians may note that the probability of a particular proposition depends (in general) not only on that proposition but on the inferentially relevant Boolean algebra of propositions over which belief is to be distributed. Thus the proposition $\theta>0$ might be regarded as embedded in the algebra of propositions generated from $\{(\theta>c);\ c\in R^1\}$ or in the algebra comprising only $\theta>0$, $\theta\leqslant 0$ and their union; depending on whether θ or $\mathrm{sign}(\theta)$ is the variable of inferential interest. Likewise a proposition of the generic form $\theta^2<c$ (or $-c<\theta<c$) would be referred not to $dP(\theta)$ but to $dP(\theta^2)$, the latter being appropriate to the subalgebra generated from symmetric intervals.

4. CONFIDENCE-BASED THEORY OF INFERENTIAL PROBABILITY

Because of space restrictions only a brief outline of the theory will be given here, with detailed proofs omitted. Some topics discussed in an earlier and longer version of the paper will be briefly mentioned in Section 4.9.

The first point to note is that a uniquely relevant simple ordering relation exists only in a 2-dimensional statistic-parameter product space. Application of the frequency projection concept to higher-dimensional product spaces therefore depends on a factorization being found of the simultaneous frequency distribution of the observational variables, to identify 2-dimensional product subspaces in which relevant ordering relations exist. In other words, the joint information must be completely separated into relevant and irrelevant components for each kind of inference, and for this Fisher's Conditioning principle provides the necessary operational tool.

For the present we continue with the case of one observation, one parameter as in Section 3. The basic definition (3.2) of fiducial probability for this case will also apply to more general situations with appropriate notational identification.

4.1. *Irreducible Pivotal Relations*

The existence of a pivotal relation between x and θ may not ensure that it is fully relevant for inference about θ, though this will almost always be so. One mathematical possibility to be excluded is that there exists an interval I for which the conditional distribution $dF(x \mid x \in I)$ is functionally independent of θ. For then, given the information $x_a \in I$, the further knowledge of the actual value x_a is clearly irrelevant and ought therefore to be excluded in defining a fully relevant pivotal relation, according to the Relevance principle. If such an interval I did exist, the pivotal relation could be reduced to fully relevant form by contracting the sample space $\{x\}$, replacing I by a single representative point x_I. This would induce vertical steps in the percentile relations at x_I, so that, if $x_a \in I$, a partly indeterminate probability distribution for θ would result, each vertical step defining an upper and lower bound for the corresponding percentile point θ_p. Probability distributions for θ at other values of x_a would be unaffected.

A second possibility to be excluded is the existence of an ancillary statistic $g(x)$, defining a more relevant subset $\{x;\, g(x) = g(x_a)\}$ of $\{x\}$ on which inference should be conditioned. (Cf. Buehler, 1959.) Robinson (1975) gives an important though unusual example of this.

A pivotal relation will be termed irreducible if neither of the above possibilities exists. The first possibility is excluded by reduction to minimal sufficiency, and the second if x is boundedly complete.† (We need not consider subsets of measure zero here.)

4.2. *Extension of Fiducial Probability to Functions* $\phi(\theta, x)$

This extension and the related expectation theory (Section 4.3) is crucial to fiducial theory for two or more parameters (Section 4.5).

As noted in Section 3.1, if a pivotal variate $u(x, \theta)$ exists with known frequency distribution independent of θ, then, considering the case $u(x, \theta) \uparrow \theta$, the proposition $\theta_a < \theta_p(x_a)$ can be re-expressed as $u_a < u_p$, in which the subject of the proposition, $u_a = u(x_a, \theta_a)$, is the realized (but unknown) value of the random variable u. Thus it can be seen that classical probability theory, which is directly applicable to propositions whose subject is the value of a physical random variable, constitutes a sub-theory of the more general theory of noncoherent probability outlined in this paper.

In summary, since the equations $\theta = \theta_p(x)$ and $u(x, \theta) = u_p$ are functionally equivalent for given p, $0 < p < 1$, both defining the same percentile relation between x and θ, we shall say that θ is *pivotally equivalent* to $u(x, \theta)$ or, more briefly, $\theta \overset{\mathcal{P}}{\sim} u(x, \theta)$, and hence that $dP(\theta_a)$ is pivotally equivalent (equivalent under projection) to $dF(u)$, that is, $dP(\theta_a) \overset{\mathcal{P}}{\sim} dF(u)$. We also have $\phi(\theta) \overset{\mathcal{P}}{\sim} u(x, \theta)$, $dP(\phi_a) \overset{\mathcal{P}}{\sim} dF(u)$ for any invertible function $\phi(\theta)$ of θ.

In the special case of the function $u(x, \theta)$ there is a *direct* equivalence, $dP(u_a) \sim dF(u)$, in the sense that $P^{(u_a)}(u) \equiv F(u)$, and likewise $dP(h(u_a)) \sim dF(h(u))$ for any e.c. function h. More generally there is a pivotal equivalence,

$$dP(\phi(x_a, \theta_a)) \overset{\mathcal{P}}{\sim} dF(h(u)) \tag{4.1}$$

† This weaker condition replaces the conjectured monotone likelihood ratio condition in the read version.

for any function $\phi(x, \theta)$ such that

$$\phi(x, \theta) = \phi(x, \theta_p(x)) \Leftrightarrow h(u) = h(u_p), \quad 0 < p < 1, \tag{4.2}$$

for some e.c. function h, that is, both equations in (4.2) define the same transformed percentile relation between x and θ. Expressing this in words, a proposition about the value ϕ_a of such a function is equivalent to a proposition about $h(u_a)$ and hence about u_a, so that $dF(u)$ determines the inferential probability distribution for ϕ_a. Examples of pivotally equivalent functions for the case $x \sim N(\theta, 1)$, $u = \theta - x$ are (in terms of u rather than x) $\theta + u^2$, $u^2 e^\theta$, $u^2 - \theta^2$; θu is not.

Since all such functions ϕ are pivotally equivalent to each other (including θ itself) we shall find it convenient to describe any such ϕ with the notation $\phi(x, \theta) \overset{P}{\sim} \theta$, omitting explicit reference to a pivotal variate u except when necessary.

Referring to Fig. 1 it can be seen that the foregoing extension of fiducial probability corresponds geometrically to projection of frequency not on the family of vertical lines $x = x_a$, but onto some family of loci which, jointly with the loci $u = u_p$, determines a coordinate system in the (x, θ) plane. The change of coordinate system for projection induces the change of variable $\phi = \phi(x, \theta)$ to which the projected distributions relate. Noting that frequency projection is unaffected by invertible transformation of the percentile relations (Section 3.3), and that relations $u = u_p$ may be mapped onto themselves with any e.c. transformation $u \rightarrow h(u)$, the most general form of transformation appears to be

$$(\theta, u) \rightarrow (g(\theta, u), h(u)), \tag{4.3}$$

where $g(\theta, u) \rightleftharpoons \theta$ for all u.

4.3. *Fiducial Expectations*

In view of the multiplicity of inferential distributions that derive from the given data, it is logically essential that the inferential expectation $E(\phi_a)$ of a function $\phi(x_a, \theta_a)$ be determined with respect to its own inferential distribution, that is,

$$E(\phi_a) = \int \phi_a \, dP(\phi_a). \tag{4.4}$$

Re-expressed in the form of a restriction on the applicability of $dP(\theta_a)$,

$$E(\phi_a) = \int \phi_a \, dP(\theta_a) \quad \text{only if} \quad \phi(x, \theta) \overset{P}{\sim} \theta. \tag{4.5}$$

In the special case $\phi = \phi(\theta)$, the condition reduces to $\phi(\theta) \rightleftharpoons \theta$.

4.4. *Fiducial Theory for Several Observations, One Parameter*

Given a vector of n observations $\mathbf{x}_a$ relating to one parameter θ_a the necessary and sufficient condition for a fiducial distribution $dP(\theta_a)$ to exist (with the direct confidence interpretation implied by (3.2)) is that there exist an invertible transformation of the observational variable $\mathbf{x}$ to a scalar $y \in R^1$ and a vector $\mathbf{z} \in R^{n-1}$ for which the joint sampling distribution factorizes in the form

$$dF(y, \mathbf{z};\ \theta) = dF(y \mid \mathbf{z};\ \theta)\, dF(\mathbf{z}) \tag{4.6}$$

with the properties

(i) $dF(\mathbf{z})$ is independent of θ;†

(ii) $dF(y \mid \mathbf{z})$ defines an irreducible pivotal relation between $y \mid \mathbf{z}$ and θ.

† It is also necessary that z be a *proper* ancillary: Define an ancillary statistic z to be (i), ***fully relevant*** if a function of the minimal sufficient statistic for θ, (ii), *canonically irrelevant* if independent of the minimal sufficient statistic, (iii), *proper* if a function only of fully relevant and/or canonically irrelevant ancillaries, (iv), *improper* otherwise. A counter-example of Dawid (see Discussion and Reply) shows that conditioning with respect to improper ancillaries causes a loss of relevant information.

The fiducial distribution $dP(\theta_a)$ is then determined by frequency projection from the pivotal relation between $y|\mathbf{z}$ and θ.

This important extension of fiducial theory is implicit in Fisher (1934), and is discussed and illustrated in Fisher (1956a). I shall term a factorization of the form (4.6) with the properties above a *fully relevant* factorization. Fisher termed $\mathbf{z}$ an *ancillary* (vector) statistic and $y|\mathbf{z}$ may be termed a *conditionally sufficient* statistic. Note that $\mathbf{z}$ in (4.6) is uniquely determined, modulo invertible transformation,† but that the *informative* statistic y is arbitrary, subject to it being pivotally related to θ. Nevertheless $y|\mathbf{z}$ leads to a uniquely determined distribution for θ; a remarkable property of Fisher's solution. As he expressed it, conditioning with respect to the ancillary statistic recovers the information about θ not inherent in the marginal distribution of y. For instance if $x_i \sim N(\theta, 1)$, $i = 1, 2$, the conditional distribution $dF(x_1|x_1-x_2)$ is equivalent to that of the sufficient statistic $\bar{x}$, differing from $dF(\bar{x})$ only by a displacement $(x_1-x_2)/2$.

As noted earlier the inferential distribution $dP(\theta_a)$ is coherent only for invertible functions $\phi(\theta)$ or more generally for pivotally equivalent functions $\phi(x, \theta) \overset{P}{\sim} \theta$. Inference about non-invertible functions $\phi(\theta)$ and other related functions involves mapping to a different representation of the parameter space,

$$\theta \rightleftharpoons (\phi_1(\theta), \phi_2(\theta), \ldots), \tag{4.7}$$

and thus formally belongs in the domain of simultaneous inference (see below).

When no factorization (4.6) exists, no distribution $dP(\theta_a)$ exists with a direct confidence interpretation. However, the theory is applicable to a much wider class of cases, in that *derived conditional* distributions for θ_a may exist, see Section 4.8.

4.5. *Fiducial Theory for Several Parameters*

As in Section 4.4. the existence of a simultaneous inferential distribution depends on there being a fully relevant factorization of the sampling distribution of the data. To simplify notation I shall restrict attention to a vector $\boldsymbol{\theta} \in R^2$ comprising two parameters θ_1 and θ_2. The first requirement of a fully relevant factorization is that there exist a factorization of the form (4.6)‡ with $\mathbf{y} \in R^2$ and $\mathbf{z} \in R^{n-2}$. This separates out the inferentially irrelevant component $dF(\mathbf{z})$. Next we consider the necessary factorization of $dF(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta})$, namely (with conditioning with respect to the ancillary statistic $\mathbf{z}$ taken for granted),

$$dF(\mathbf{y}; \boldsymbol{\theta}) = dF(y_1; \theta_1) . dF(y_2|y_1; \theta_1, \theta_2). \tag{4.8}$$

This factorization, if fully relevant§, will uniquely determine a simultaneous inferential distribution $dP(\theta_{1a}, \theta_{2a})$ in the factorized form

$$dP(\theta_{1a}, \theta_{2a}) = dP(\theta_{1a}) . dP(\theta_{2a}|\theta_{1a}). \tag{4.9}$$

Sufficient conditions for full relevance are as follows:

(i) $dF(y_1)$ defines an irreducible pivotal relation between y_1 and θ_1. This determines $dP(\theta_{1a})$ as a function of y_{1a}.

(ii) $dF(y_2|y_1)$ defines a *complete* irreducible pivotal relation between $y_2|y_1$ and $\theta_2|\theta_1$. This determines $dP(\theta_{2a}|\theta_{1a})$ as a function of θ_{1a}, y_{1a} and y_{2a}.

† *Added in proof*: and modulo trivial measure-preserving transformations. (See Discussion (Pedersen) and Reply.)

‡ Inference is similarly possible from an alternative form of factorization $dF(y_1|y_2; \theta_1) . dF(y_2; \theta_1, \theta_2)$. (See Discussion (Sprott) and Reply.)

§ It is important that the parameters be range-independent, since otherwise the factorization cannot represent a complete separation of information. (See Discussion (Dawid) and Reply.)

(iii) The distribution function $P(\theta_{2a}\,|\,\theta_{1a})$, considered as a function of θ_{1a}, y_{1a}, must be pivotally related to θ_{1a}, that is,

$$P(\theta_2\,|\,\theta_1)\overset{P}{\sim}\theta_1. \tag{4.10}$$

The first two conditions are interlocking, each ensuring that the pivotal relation of the other is *fully relevant* for inference about the corresponding parameter (θ_1 or $\theta_2\,|\,\theta_1$). In particular, marginal inference about θ_1 from y_1 alone is justified only if the information about θ_1 in $dF(y_2\,|\,y_1)$ is *completely aliased* with that about θ_2 and thus unusable unless θ_2 is known. Condition (ii) ensures this is so. See Sprott (1975) for somewhat similar conditions.

The additional requirement of *completeness* in (ii) may not always be necessary, but there are cases where an incomplete distribution of belief in $dP(\theta_2\,|\,\theta_1)$ does indicate recoverable information about θ_1, which would invalidate inference about θ_1 from y_1 alone. With conditions (i) and (ii) as above, y_1 is termed marginally sufficient for θ_1 and $y_2\,|\,y_1$ conditionally sufficient for $\theta_2\,|\,\theta_1$. Likewise y_1 is termed an ancillary statistic for inference about θ_2, *even if* $dF(y_2\,|\,y_1)$ depends only on θ_1 and not on y_1, since it supplies the necessary information about θ_1.

Condition (iii), which may be termed the *cross-coherence* condition, is extremely important since it eliminates, in conjunction with other coherence considerations, all of the alleged paradoxes in fiducial theory (Creasy (1954), Mauldon (1955), Tukey (1957), Brillinger (1962), Bennett and Cornish (1963), Dempster (1963a, b), Geisser and Cornfield (1963)). For the simultaneous distribution (4.9) can be considered logically as specifying a *θ_1-distribution of distributions* for $\theta_2\,|\,\theta_1$ from which an *expected* fiducial distribution function for θ_2 may be derived as

$$\bar{P}(\theta_2) = E_{\theta_1}\{P(\theta_2\ \ \theta_1)\} = \int P(\theta_2\,|\,\theta_1)\,dP(\theta_1). \tag{4.11}$$

Thus (4.10) is a *logical* requirement flowing from inferential expectation theory (Section 4.3).

Note that the pivotal equivalence of $P(\theta_2\,|\,\theta_1)$ to θ_1 does not imply pivotal equivalence of $P\{(a<\theta_2<b)\,|\,\theta_1\}$ to θ_1 for all a, b. In the case of $dP(\mu\,|\,\sigma)$ derived from a Normal sample (see below), $P\{(a<\mu<b)\,|\,\sigma\}$ is pivotally equivalent to σ only if the interval (a, b) includes the central point $c = \bar{x}_a$, at which $P\{(\mu<c)\,|\,\sigma\}$ is independent of σ. This is the point where the relation of $P(\mu\,|\,\sigma)$ to σ changes from monotone increasing to monotone decreasing.

The coherence properties and confidence interpretations of (4.9) and (4.11) are discussed in Section 4.7. Extension to 3 or more parameters is straightforward.

4.6. *Examples*

(*i*) *Location and scale.* Given a set of n independent and identically distributed (i.i.d.) observations x_{ia}, $i = 1, 2, \ldots, n$ with sampling density function $\sigma^{-1} f\{(x-\mu)/\sigma\}$, f known, the simultaneous fiducial distribution for μ and σ is, with a constant of integration c,

$$dP(\mu, \sigma) = c\prod_{i=1}^{n}\,[\sigma^{-1} f\{(x_{ia}-\mu)/\sigma\}]\,d\mu\,d\sigma/\sigma. \tag{4.12}$$

See Fisher (1948, 1956a) for derivation.

(*ii*) *Normal sample.* A special case of the above, in which the sample mean $\bar{x}$ and variance s^2 are jointly sufficient for μ and σ^2. If z denotes a standard normal variate the relevant pivotal relations are

$$(n-1)\,s^2 = \sigma^2\chi^2_{n-1}, \quad \bar{x} = \mu+(\sigma/\surd n)\,z, \tag{4.13}$$

fiducial inversion of which gives the inferential relations

$$\sigma^2_a = (n-1)\,s^2_a\,\chi^{-2}_{n-1}, \quad \mu_a = \bar{x}_a-(\sigma_a/\surd n)\,z. \tag{4.14}$$

These specify $dP(\mu_a, \sigma_a)$ in the factorized form $dP(\sigma_a).dP(\mu_a|\sigma_a)$. Note that $P(\mu|\sigma)$ is monotone in σ and independent of s, so that $P(\mu|\sigma)\overset{p}{\sim}\sigma$, as required. Substituting for σ_a in the expression for μ_a in (4.14) determines the *expected* fiducial distribution $d\bar{P}(\mu_a)$, for μ_a, specified by the relation

$$\mu_a = \bar{x}_a - (s_a/\sqrt{n})\, t_{n-1}. \tag{4.15}$$

This derivation was given by Fisher (1939) and Yates (1939). Fisher's first derivation (1935) of (4.15) was directly from the pivotal variate $t_{n-1} = (\bar{x}-\mu)\sqrt{n}/s$ which defines, in the present context, a marginal pivotal relation independent of σ between $\bar{x}/s$ and the function μ/s. An alternative factorization of $dF(\bar{x}, s)$ determines a marginal pivotal relation between the variables $r = \bar{x}/s$ and $\rho = \mu/\sigma$, specified by a noncentral t-distribution,

$$r = t(n-1, \rho). \tag{4.16}$$

Inversion of this determines a distribution for ρ, $dP(\rho)$, which can also be derived from, and is thus coherent with, $dP(\mu, \sigma)$. See Section 4.7.

(*iii*) *Behrens–Fisher problem.* Two independent samples (n_1, n_2) from two normal populations with distinct unknown variances. It is notationally convenient to let σ_1^2, σ_2^2 denote the sampling variances of $\bar{x}_1, \bar{x}_2$ respectively, and s_1^2, s_2^2 the usual χ^2 sample estimates of them. Clearly there is a simultaneous distribution $dP(\mu_1, \mu_2, \sigma_1, \sigma_2)$, represented by two pairs of inferential relations of the form (4.14), which may be used to derive inferential distributions for pivotally equivalent functions on the parameter space (see Section 4.7). Fisher (1935) derived a fiducial distribution for $\mu_1-\mu_2$ directly from two relations of the form (4.15), obtaining the inferential relation (reversing one sign since t has a symmetric distribution)

$$(\mu_{1a}-\mu_{2a}) = (\bar{x}_{1a}-\bar{x}_{2a}) + s_{1a}\, t_1 + s_{2a}\, t_2. \tag{4.17}$$

This specifies an *expected* distribution for $\mu_1-\mu_2$ with distribution function

$$\bar{P}(\mu_1-\mu_2) = \int\!\!\int P(\mu_1-\mu_2|\sigma_1, \sigma_2)\, dP(\sigma_1)\, dP(\sigma_2). \tag{4.18}$$

Dividing (4.17) by $s_a = \sqrt{(s_{1a}^2+s_{2a}^2)}$, one obtains the Behrens–Fisher statistic (Behrens, 1929; Fisher, 1935),

$$\{(\mu_{1a}-\mu_{2a})-(\bar{x}_{1a}-\bar{x}_{2a})\}/s_a = (s_{1a}/s_a)\, t_1 + (s_{2a}/s_a)\, t_2, \tag{4.19}$$

from which it is clear that s_{1a}/s_{2a} is an important ancillary statistic. The confidence property of $d\bar{P}(\mu_1-\mu_2)$ is discussed in Section 4.7.

Added in proof: As mentioned at the meeting, I have subsequently discovered that the Behrens–Fisher distribution does not satisfy the cross-coherence requirement. In the factorized distribution $dP(\mu_1-\mu_2|\sigma).dP(\sigma|\eta).dP(\eta)$, where $\sigma^2 = \sigma_1^2+\sigma_2^2$ and $\eta = (\sigma_1/\sigma_2)^2$, $dP(\sigma^2|\eta)$ is not cross-coherent with $dP(\eta)$, depending on η only through the function $|\rho|$, where $\rho = \ln(\sigma_1/\sigma_2)$, when $n_1 = n_2$. Clearly in this case only the component $|r|$ of s_1/s_2, where $r = \ln(s_1/s_2)$, is relevant to inference about σ and $\mu_1-\mu_2$, the sign of r being irrelevant. In a forthcoming paper Professor A. T. James and I will describe a modified Behrens–Fisher distribution derived from a simultaneous distribution for $\mu_1-\mu_2$, σ and $|\rho|$, for the case $n_1 = n_2$, and will consider also the case $n_1 \neq n_2$. The behaviour of the modified solution is similar to that described at the end of Section 4.7.

(*iv*) *Ratio of means* (*the Creasy–Fieller paradox*). Given $x_i \sim N(\mu_i, 1)$, $i = 1, 2$, let ψ, ω denote the angular and radial coordinates of the point (μ_1, μ_2), with ψ measured relative to the radial line through (x_1, x_2), and let $S^2 = x_1^2+x_2^2$. James *et al.* (1974), see also Sprott (1963), give the simultaneous fiducial distribution $dP(\psi, \omega)$ in the factorized form $dP(\omega;\, S_a).dP(\psi|\omega;\, \omega S_a)$, where $dP(\psi|\omega)$ has density function (Fisher, 1956a, Ch. V)

$$p(\psi|\omega) = \frac{\exp(\omega S_a \cos\psi)}{2\pi\, I_0(\omega S_a)}, \quad -\pi < \psi \leqslant \pi, \tag{4.20}$$

and $dP(\omega)$ is determined by fiducial inversion of the noncentral χ^2 relation $S^2 = \chi^2(2, \omega^2)$, giving the distribution function

$$P(\omega^2) = F\{\chi^2(2, \omega^2) > S_a^2\} \tag{4.21}$$

(*Added in proof*: This replaces an incorrect equation in the read version.)

Note that the dispersion of $dP(\psi \mid \omega)$ on the circle is monotone in ωS_a. Hence it can be shown that $dP(\psi \mid \omega)$ is pivotally equivalent to ω, as required.

Averaging $dP(\psi \mid \omega)$ with respect to $dP(\omega)$ gives an expected distribution $d\bar{P}(\psi)$ for ψ, and hence, by transformation, for the ratio μ_1/μ_2. This is the correct distribution according to the present theory, and gives more conservative fiducial limits than obtained by either Fieller (1954) or Creasy (1954). Creasy's derivation is invalid because $dP(\omega)$ (and hence $d\bar{P}(\mu_1/\mu_2)$) is not coherent with $dP(\mu_1, \mu_2)$; Fieller's, through failing to condition on the relevant ancillary statistic S. Also, the Fieller "pivotal" statistic is not pivotal in the sense defined in the present theory. Sprott (1963) gives another distribution for μ_1/μ_2 which is also invalid from coherence considerations.

(*v*) *Combining information on means and variances.* A. T. James and I, in a forthcoming paper which extends results of Fisher (1961a, b), give exact fiducial theory for combining information arising from experiments of the incomplete-block type. See also Section 4.8.

(*vi*) *Multivariate Normal.* This will be considered in Section 5.

4.7. *Confidence and Coherence Properties*

Returning to the context of Section 4.5 it is clear that $dP(\theta_1, \theta_2)$ has a *piece-wise confidence* interpretation, associated with its component factors $dP(\theta_1)$, $dP(\theta_2 \mid \theta_1)$. The marginal distribution $dP(\theta_1)$ has the *direct confidence* interpretation of (3.2), and so has $dP(\theta_2 \mid \theta_1)$, except that now the confidence frequencies are functions, $p(\theta_1)$, of θ_1.

The *expected* distribution $d\bar{P}(\theta_2)$ has the *expected confidence* interpretation associated with its definition, the expected confidence frequencies $\bar{p}$ being defined by

$$\bar{p} = \int p(\theta_1)\, dP(\theta_1). \tag{4.22}$$

Thus in the case of the Normal sample, Section 4.6(ii), the conditional confidence frequencies $p(\sigma)$ for $dP(\mu \mid \sigma)$ are determined from the Standard Normal distribution in (4.14), and the expected confidence frequencies $\bar{p}$ for $d\bar{P}(\mu)$ derive from the t distribution in (4.15).

The inferentially essential preservation of these confidence properties under transformation, and of cross-coherence also, defines in general the class of distributions $dP(\phi_1, \phi_2)$ coherent with $dP(\theta_1, \theta_2)$ to be those where

$$\left.\begin{array}{ll} \text{(i)} & \phi_1 = \phi_1(y_1, \theta_1) \overset{\mathcal{P}}{\sim} \theta_1, \\ \text{(ii)} & \phi_2 = \phi_2(y_2, \theta_2, y_1, \theta_1) \overset{\mathcal{P}}{\sim} \theta_2 \mid \theta_1, \\ \text{(iii)} & P(\phi_2 \mid \phi_1) \overset{\mathcal{P}}{\sim} \phi_1. \end{array}\right\} \tag{4.23}$$

However, the class of coherent distributions can be larger in special cases, which we now consider.

In general, a fully relevant factorization (4.8), if it exists, will be unique in the sense that the only functions on the sample space $\{(y_1, y_2)\}$ with distribution depending on only one parameter are of the form $g(x_1) \rightleftharpoons x_1$. Thus only θ_1 and functions pivotally equivalent to it have a distribution with direct confidence interpretation. In special cases, however, a multiplicity of such factorizations may exist, and the class of coherent distributions is then correspondingly enlarged. These special cases are characterized by the existence of a transformation $(\theta_1, \theta_2) \rightarrow (\phi_1, \phi_2)$ such that

$$dP(\phi_1, \phi_2) = dP(\phi_1) . dP(\phi_2), \tag{4.24}$$

that is, such that ϕ_1 and ϕ_2 are statistically independent fiducially. An equivalent, dual condition is the existence of statistically independent statistics y_1, y_2, that is, such that (conditioning on relevant ancillaries as before)

$$dF(y_1, y_2;\ \theta_1, \theta_2) = dF(y_1;\ \theta_1).dF(y_2;\ \theta_1, \theta_2). \tag{4.25}$$

In the case of the Normal sample the statistics $\bar{x}$ and s are statistically independent, and by transformation of $\bar{x}$ to $r = \bar{x}/s$, μ to $\rho = \mu/\sigma$ (see Section 4.6(ii)), it can also be seen that ρ and σ are fiducially independent.

In the special case where θ_1 and θ_2 are fiducially independent, which corresponds to the special form of factorization

$$dF(y_1, y_2;\ \theta_1, \theta_2) = dF(y_1;\ \theta_1).dF(y_2 | y_1;\ \theta_2), \tag{4.26}$$

it is immediately evident that $dP(\theta_1, \theta_2)$ has a direct, *simultaneous confidence* interpretation, in that the conceptual class of simultaneous propositions

$$\{\theta_1 < \theta_{1p_1}(y_1) \quad \text{and} \quad \theta_2 < \theta_{2p_2}(y_2 | y_1)\} \tag{4.27}$$

contains a known fraction $p = p_1 p_2$ of true propositions with p derivable directly from $dF(y_1, y_2)$. In the general case, however, θ_{2p_2} will also be a function of θ_1, which precludes p from being expressible as a simultaneous repeated-sampling frequency.

Likewise the expected fiducial distribution $d\bar{P}(\theta_2)$ will not have a direct confidence interpretation unless θ_2 is statistically independent of θ_1 fiducially. The Behrens–Fisher distribution $d\bar{P}(\mu_1 - \mu_2)$ in Section 4.6(iii) is an example.

Let us invoke now Fisher's (1956a) concept of a *relevant reference set* in which inferential probabilities are verifiable as freqencies. The relevant set will be some subspace of the sample-parameter product space and is represented here by a condensed notation indicating both the component variables and their assigned frequency distribution, if any. Summarizing the theory so far we have, assuming all specified factorizations are fully relevant,

Reference Sets

(i) Objective Bayesian inference for one parameter:

$$\{dF(\theta | x_a)\} \Rightarrow dP(\theta_a;\ x_a).$$

(ii) Fiducial inference for one parameter:

$$\{\theta, dF(y | \mathbf{z}_a;\ \theta)\} \Rightarrow dP(\theta_a;\ y_a, \mathbf{z}_a).$$

(iii) Simultaneous fiducial inference for two parameters:

$$\{\theta_1, dF(y_1;\ \theta_1)\} \Rightarrow dP(\theta_{1a}), \tag{4.28}$$

$$\{dP(\theta_1;\ y_{1a})\} \otimes \{\theta_2, dF(y_2 | y_{1a};\ \theta_1, \theta_2)\} \Rightarrow dP(\theta_{1a}, \theta_{2a}).$$

(iv) Special case when $dF(y_2 | y_1)$ is independent of θ_1:

$$\{\theta_1, dF(y_1;\ \theta_1\} \Rightarrow dP(\theta_{1a}),$$

$$\{\theta_2, dF(y_2 | y_{1a};\ \theta_2\} \Rightarrow dP(\theta_{2a}).$$

We are now at the crucial point of divergence between the Confidence and Fiducial theories. In generalizing to two or more parameters, Confidence theory formally assumes that the simultaneous confidence property is a necessary perquisite of a confidence region (and of course this might very well be so in decision-theoretic applications such as quality control). However, it is not a logical requirement in the theory of inferential probability presented here. The only logically necessary requirement for forming a simultaneous inferential distribution from confidence-based component distributions $dP(\theta_1)$, $dP(\theta_2 | \theta_1)$ is the cross-coherence condition $P(\theta_2 | \theta_1) \overset{p}{\sim} \theta_1$, which implies in general no more than the piece-wise confidence property. Furthermore, this leads to an expected confidence interpretation for $d\bar{P}(\theta_2)$, see (4.22), which is easily understood, given that the direct confidence frequencies associated with $dP(\theta_2 | \theta_1)$ depend on the unknown θ_1, so that some averaging of them with respect to possible values of θ_1 is intuitively appropriate.

A second logical point. Under the present theory no simultaneous fiducial distribution exists with a simultaneous confidence interpretation numerically different from its piece-wise confidence interpretation. Indeed, if a simultaneous distribution with conflicting confidence properties did exist, it would be in logical conflict not only with Fisher's Conditioning principle and with the fundamental Relevance principle, but with a third, *Diffusion* principle which I shall describe shortly.

Consider the implications of the Conditioning and Relevance principles first. Except in the special case where $dP(\theta_2 | \theta_1)$ is independent of both y_1 and θ_1, y_{1a} is a relevant ancillary statistic, on which $d\bar{P}(\theta_2)$, or $dP(\theta_2)$ if θ_2 is fiducially independent of θ_1, functionally depends. Fixing y_{1a} in defining the relevant conceptual reference set then logically dictates that the sampling variation of y_1 be transferred fiducially to the unknown parameter θ_1, as in (4.28) (iii), since this is the only way of preserving, in the reference set, the known datum that y_1 and θ_1 have a random, pivotal relation. A reference set in which both y_1 and θ_1 were fixed would violate the Relevance principle. Indeed, the conditional distribution $dP(\theta_2 | \theta_1)$, if considered in isolation, does violate the Relevance principle in this way (if dependent on both θ_1 and y_1), and is inferentially relevant only as a component of the simultaneous distribution $dP(\theta_1, \theta_2)$, and hence for determining $d\bar{P}(\theta_2)$.

Thus it is clear that the simultaneous confidence interpretation, even when it exists, is not inferentially relevant, since simultaneous repeated-sampling frequencies derive from a reference set in which y_1 is not fixed. The logical difficulty is apparent in (4.28) (iv), since the two 2-dimensional reference sets there cannot be symbolically combined into a 4-dimensional product space unless the subscript a is removed from the term $y_2 | y_{1a}$. To illustrate with the case of the Normal sample, it is the expected confidence property of $d\bar{P}(\mu_a;\, \bar{x}_a, s_a)$ in a reference set with s_a fixed and σ varying fiducially that is inferentially relevant. The numerically equivalent, direct confidence property of $dP(\mu_a/s_a)$ derived from the pivotal relation between $\bar{x}/s$ and μ/s is not relevant in interpreting $d\bar{P}(\mu_a)$, since s is not fixed in the related reference set.

Finally, not only is $d\bar{P}(\theta_2)$ the only inferential distribution for θ_2 logically consistent with the component distributions $dP(\theta_1)$ and $dP(\theta_2 | \theta_1)$ and utilizing fully all the relevant information; it also exhibits a property which in my view is empirically essential. If θ_1 were known, then $dP(\theta_2 | \theta_1)$ would be the correct inferential distribution. If we now eliminate "θ_1 known" as a datum, the theory dictates that the dependence of $dP(\theta_2 | \theta_1)$ on this false datum be eliminated by averaging with respect to the distribution $dP(\theta_1)$, thus forming $d\bar{P}(\theta_2)$. Now, in the precise mathematical sense of diffusion theory, $d\bar{P}(\theta_2)$ is a diffused form of the unknown $dP(\theta_2 | \theta_{1a})$, the diffusion representing an increase in uncertainty about θ_2 (just like the increase in entropy that physical diffusion causes). It is this property that I regard as empirically essential, and formulate therefore as a principle of inference:

Diffusion Principle

Lack of relevant ancillary information increases uncertainty, by diffusing the distribution of belief.

We now need just one convincing example to show that insistence on a repeated-sampling property for inferential distributions would be in logical conflict with the Diffusion principle. This is provided by the Behrens–Fisher problem, particularly in the case of equal sample sizes n, which we now assume. If it were known that $\sigma_1 = \sigma_2$, in the notation of Section 4.6(iii), then the Behrens–Fisher statistic would have inferentially a t-distribution,

$$d_a = \frac{(\mu_1-\mu_2)-(\bar{x}_{1a}-\bar{x}_{2a})}{\sqrt{(s_{1a}^2+s_{2a}^2)}} \overset{\text{inf}}{=} t_{2n-2}. \qquad (4.29)$$

This also has a direct repeated-sampling interpretation, corresponding to removing the subscripts a in (4.29). If, however, nothing is known about σ_1 and σ_2, the Diffusion principle dictates that d_a be assigned an inferential distribution more diffuse than that of t_{2n-2}, which implies that the positive tabular value of d for any central p per cent fiducial interval should be greater than that of t_{2n-2}. In fact, unless n is very small, the tabular values of the Behrens–Fisher distribution for d_a depend only slightly on the ancillary statistic s_{1a}/s_{2a} and are approximately equal to those of t_{n-1}, rather than t_{2n-2}. This loss of $n-1$ degrees of freedom is highly plausible as a measure of the increase in uncertainty when σ_1/σ_2 is unknown. At the same time it is clear that since fiducial intervals based on t_{2n-2} have a direct repeated-sampling interpretation when $\sigma_1 = \sigma_2$, the wider intervals dictated by the Diffusion principle cannot possibly have such an interpretation.

Much of the foregoing reasoning is of course inherent in Fisher's discussion of the Behrens–Fisher solution (1937, 1939), following criticisms of it by Bartlett (1936) and Welch (1937), and in Yates (1939). Further arguments by Fisher (1941, 1945, 1955, 1956a, b) against equating "strength of evidence" with repeated-sampling frequencies were not widely accepted. See comments by Neyman (1956), Bartlett (1956), Welch (1956) and Yates (1964).

It is perhaps of interest to note that if *a priori* knowledge indicated a finite (but unknown) chance that $\sigma_1/\sigma_2 = 1$ the fiducial distribution for σ_1/σ_2 derived from the pivotal relation with s_1/s_2 would no longer be inferentially relevant, but rather a modified distribution for σ_1/σ_2 expressed as $\exp(|\rho|.\text{sign}(\rho))$, where $\rho = \ln(\sigma_1/\sigma_2)$, and derived from the simultaneous distribution $dP(|\rho|, \text{sign}(\rho))$ (*cf.* Section 3.6). As the observed ratio s_{1a}/s_{2a} mathematically approaches unity the modified distribution for σ_1/σ_2 will exhibit an increasing degree of shrinkage towards $\sigma_1/\sigma_2 = 1$, with an increasing condensation there of inferential probability providing an *a posteriori* upper-bound measure of belief in $\sigma_1/\sigma_2 = 1$. Although I have not worked out the details, it is clear that for the correspondingly modified $d\bar{P}(\mu_1 - \mu_2)$, the appropriate tabular values of the d-statistic would decrease towards those of t_{2n-2} as $s_{1a}/s_{2a} \to 1$, somewhat like the tabular values of the Welch test (Welch, 1947; Aspin, 1949), though the latter have the defect of being less than those of t_{2n-2} over a central range of s_{1a}/s_{2a}. (*Added in proof*: See also comments on a modified solution in Section 4.6(iii).)

4.8. *Derived Conditional Distributions*

The distribution $dP(\theta_1, \theta_2) = dP(\theta_1)\,dP(\theta_2 | \theta_1)$ also implies a family of conditional distributions, when θ_1 and θ_2 are not independent,

$$dP(\theta_1 | \theta_2) = \frac{p(\theta_2 | \theta_1))}{\bar{p}(\theta_2)}\, dP(\theta_1), \tag{4.30}$$

where $p(\theta_2 | \theta_1)$, $\bar{p}(\theta_2)$ are the density functions of $dP(\theta_2 | \theta_1)$, $d\bar{P}(\theta_2)$ respectively. These distributions do not have the direct confidence property of the distributions $dP(\theta_2 | \theta_1)$, hence the term *derived.* Thus the order of the arguments θ_1, θ_2 in $dP(\theta_1, \theta_2)$ is inferentially significant.

The conditional distribution (4.30) is formally a Bayesian posterior distribution for θ_1, with prior $dP(\theta_1)$ and fiducial likelihood $p(\theta_2 | \theta_1)$ for θ_1, though the latter is not generally proportional to the sample likelihood $f(y_{2a} | y_{1a};\, \theta_1, \theta_2)$.

Derived conditional distributions are inferentially inapplicable unless the value of the conditioning parameter is known. However, they have an important application in extending the theory to cases where no fully relevant factorization exists (Fisher, 1961a; Fraser, 1961). Consider the simplest case of two independent observations x_1, x_2 which each have a pivotal relation to a single unknown parameter θ. If no conditionally sufficient statistic can be derived from the combined observations, one could provisionally suppose that the observations related to different parameters θ_1 and θ_2 and hence obtain the simultaneous distribution $dP(\theta_1, \theta_2)$. Imposing the constraint $\theta_1 = \theta_2$ would then identify a derived conditional distribution for θ. There is an apparent element of indeterminacy here in the choice of conditioning variable, $\theta_1 - \theta_2$ or θ_1/θ_2, etc., and the question of uniqueness needs further study. The main fiducial requirement, clearly, is that there exist a transformation to new variables $(\theta_1, \theta_2) \to (\theta, \phi)$, such that $\phi(\theta_1, \theta_2) = 0$ is equivalent to the constraint $\theta_1 = \theta_2$, and for which a valid simultaneous distribution $dP(\theta)\, dP(\phi \mid \theta)$ exists, corresponding to an alternative, fully relevant factorization of $dF(x_1, x_2)$. In some cases a requirement of this kind leads to a unique solution, and I conjecture this is true in general.

Consider, for example, inference about the ratio $\eta = (\sigma_1/\sigma_2)^2$ from three independent χ^2 statistics $S_i = n_i s_i^2 = \sigma_i^2 \chi_{n_i}^2$, $i = 1, 2, 3$, with the constraint $\sigma_2^3 = \sigma_1^2 + \sigma_2^2$. Let $\xi = \sigma_3^2/(\sigma_1^2 + \sigma_2^2)$, $y = s_1^2/s_2^2$ and $z(\eta) = s_3^2/s^2(\eta)$, where $(n_1 + n_2)\, s^2(\eta) = (S_1/\eta + S_2)(1 + \eta)$. The following statistically independent relations involving F-variables,

$$y = \eta F_{n_1, n_2}, \quad z(\eta) = \xi F_{n_3, n_1 + n_2}, \tag{4.31}$$

determine a simultaneous distribution $dP(\eta, \xi)$ and hence a derived conditional distribution $dP(\eta \mid \xi = 1)$ for η. (*Added in proof*: This solution needs modifying in the way described for the Behrens–Fisher problem, Section 4.6(iii).)

4.9. *Miscellaneous Topics*

Lack of space prohibits any but a brief mention of the following topics, some of which will be developed elsewhere.

Discrete observations

The extension of the theory to discrete sample spaces is relatively straightforward, but the resulting inferential distributions are partly indeterminate. Percentile relations become step-functions which define, in the frequency projection process, only upper and lower bounds for each inferential percentile point. In special cases when the parameter space is also discrete, a discrete pivotal variate may exist, e.g. with $x \sim N(\theta, 1)$, $\text{sign}(\theta)/\text{sign}(x)$, with distribution dependent on $|x|$ and $|\theta|$.

Uniqueness

There appear to be no nonuniqueness problems in the theory described here, except possibly in the application of derived conditional distributions (Section 4.8). Nonunique ancillary statistics, discussed by Basu (1964), Barnard and Sprott (1971) and Cox (1971), appear to arise only when a fully relevant factorization does not exist.

Structural distributions

Fraser's (1968) theory of structural distributions is an important sub-theory of the theory described here. Structural distributions are a special form of fiducial distribution which pertains when the relevant pivotal relations are not merely monotone but have a group invariant structure. Structural distributions *formally* resemble Bayesian posterior distributions, but the theory makes it clear that the invariant differential measure which multiplies a likelihood does not represent prior knowledge but characterizes the sampling variation of the likelihood. (Fraser disagrees with the above description of structural theory, see Discussion and Reply.)

Approximate theory

The nonexistence of a fully relevant factorization implies that the empirical information about the parameter of interest cannot be completely untangled, there being some partial confounding of relevant with irrelevant information, or between the relevant information for two or more parameters. An approximate fiducial theory is then required to optimize as far as possible the extraction of information relevant to each kind of inference. Fraser's Local Analysis (1968) produces approximate fiducial distributions with excellent asymptotic properties. See also Fraser (1964a, b, c) and Sprott (1973).

5. MULTIVARIATE NORMAL INFERENCE

Many of the difficulties and apparent paradoxes encountered in inference about the parameters of a multivariate normal population are explainable in terms of the noncoherence of the relevant inferential probability distributions, and provide further empirical evidence for noncoherence as an intrinsic property of inference. However, I emphasize that only a glimpse of the full story is provided below.

Consider the following, statistically independent random variables relating to a sample from an m-dimensional multivariate normal distribution,

$$\mathbf{x} = \mathrm{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \mathbf{S} = \mathrm{W}_m(\boldsymbol{\Sigma}, n), \tag{5.1}$$

where $\mathbf{S}$ is a Wishart estimator of $\boldsymbol{\Sigma}$ with n degrees of freedom. We suppose that $\boldsymbol{\Sigma}$ is non-singular; that $n \geqslant m$, so that $\mathbf{S}$ is non singular with probability 1; and that $\mathbf{x}_a$ and $\mathbf{S}_a$ ($\mathbf{S}_a$ non-singular) are the jointly sufficient statistics for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ derived from the sample.

5.1. *Inference about the Mean*

Two mathematically noncoherent distributions for $\boldsymbol{\mu}$ may be derived, one of which may be characterized as relevant for *coordinate-dependent* inference about $\boldsymbol{\mu}$, the second for a *coordinate-free* form of inference, in a way to be explained. The first was derived by Cornish (1961), generalizing a special case of Fisher (1954); the second is a marginal distribution (Bennett and Cornish, 1963) of the simultaneous fiducial distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ derived by Segal (1938). I shall refer to them respectively as the Cornish–Fisher and Segal distributions. (Composite forms of these distributions are also possible.)

Both distributions have a similar form involving the empirical distance metric

$$\delta_a^2 = (\boldsymbol{\mu} - \mathbf{x}_a)^{\mathrm{T}} \mathbf{S}_a^{-1} (\boldsymbol{\mu} - \mathbf{x}_a), \tag{5.2}$$

namely

$$dP(\boldsymbol{\mu}) = c_q(1 + \delta_a^2/n)^{(n+m-q)/2}\, d\boldsymbol{\mu}, \tag{5.3}$$

with $q = 0$ (Cornish–Fisher) or $q = m-1$ (Segal) and c_q the appropriate constant of integration.

For any linear function $\mu = \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{\mu}$, and with $x = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}$, $s^2 = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{S} \boldsymbol{\lambda}$, the marginal distribution $dP(\mu)$ derived from (5.3) is specified by the inferential relation

$$\mu = x_a + s_a t_{n-q} \tag{5.4}$$

and this has two mutually noncoherent forms, with $q = 0$ or $m-1$ as above. Direct fiducial derivation shows that (5.4) with $q = 0$ is applicable when $\boldsymbol{\lambda}$ comprises the direction-cosines of an *a priori* specified direction, whereas (5.4) with $q = m-1$ is applicable when $\boldsymbol{\lambda}$ specifies the *estimated* direction of maximum linear discrimination of $\mathbf{x}$ relative to a specified point $\boldsymbol{\mu}_0$, that is, $\boldsymbol{\lambda}$ is such that $(\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}_a - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{\mu}_0)^2 / \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{S}_a \boldsymbol{\lambda}$ is maximized and is therefore $\boldsymbol{\lambda} = \mathbf{S}_a^{-1}(\mathbf{x}_a - \boldsymbol{\mu}_0)$. Each $\boldsymbol{\mu}_0$ uniquely determines a $\boldsymbol{\lambda}$ and vice versa. The relevant distribution theory (Kshirsagar, 1972) shows that the normalized sample discriminant function $(x-\mu)/s$, where $x = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{x}$ and $s^2 = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{S} \boldsymbol{\lambda}$, is distributed as t_{n-m+1}, whence (5.4) with $q = m-1$.

The reason for the noncoherence is now intuitively clear. Noncoherent inferential probability theory is making a proper distinction between the direction $\boldsymbol{\lambda}$ being specified *a priori* or being selected for an optimal property on the basis of the observed data; and makes due allowance for the selection effect by decreasing the relevant degrees of freedom by $m-1$.

The reason for describing the inferences as coordinate-dependent or coordinate-free is also clear. The discriminant function $x-\mu$ is invariant under affine transformation of the coordinate system.

The empirical distance metric δ_a^2 in (5.2) is similarly invariant, and has an inferential distribution $dP(\delta_a^2)$, specified by the relation

$$\delta_a^2 = \frac{mn}{n-m+1} F(m, n-m+1). \tag{5.5}$$

This distribution may be derived directly from Hotelling's frequency distribution for δ^2, and is coherent only with the Segal distribution for $\boldsymbol{\mu}$, as one would expect.

5.2. *Inference about the Dispersion Matrix*

Both the noncoherent forms (5.3) of $dP(\boldsymbol{\mu})$ are expected fiducial distributions, averaged with respect to corresponding, noncoherent fiducial distributions for $\boldsymbol{\Sigma}$ which may be formally expressed by the relation

$$\boldsymbol{\Sigma}^{-1} = \mathbf{W}_m\{\mathbf{S}_a^{-1}, n+(m-1)-q\} \tag{5.6}$$

with $q = 0$ (Cornish–Fisher) or $q = m-1$ (Segal). Segal used a simultaneous pivotal variate $\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{S}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ in his derivation. The Cornish–Fisher form follows from the mathematical symmetry of the two cases, both leading to the similar forms (5.3) for $dP(\boldsymbol{\mu})$.

Clearly the noncoherent forms of (5.6) must be applicable respectively to coordinate-dependent and coordinate-free forms of inference, as before. With respect to any *a priori* specified linear coordinate system, it can be shown that the Cornish–Fisher form correctly specifies the fiducial distributions for all marginal variances, with relations of the form

$$\sigma^2 = s_a^2\,\chi_n^{-2}, \tag{5.7}$$

which are otherwise directly deducible; and likewise for all conditional variances, the χ^2 degrees of freedom on (5.7) being then reduced by the number of conditioning variables. The Segal form, however, is invalid for this purpose, producing a loss of $(m-1)$ degrees of freedom throughout, but I conjecture that it is applicable for coordinate-free inference relating to the parameters of size, shape and orientation of the dispersion ellipsoid.

The Cornish–Fisher form of (5.6) does not give correct fiducial distributions for simple correlations ρ (Geisser and Cornfield, 1963), even though it leads to a correct averaged distribution $dP(\boldsymbol{\mu})$ for coordinate-dependent inferences. I hope to give a further explanation of this elsewhere, but consideration of the bivariate case suggests that the averaging process depends essentially only on ρ^2 rather than ρ. It appears in fact that no valid simultaneous distribution $dP(\rho, \sigma_1, \sigma_2)$ exists for these particular parameters. Fisher's (1956a) distribution for them does not satisfy the cross-coherence requirement, and gives incorrect marginal distributions for σ_1 and σ_2 (Bennett and Cornish, 1963; Dempster, 1963a, b; Geisser and Cornfield, 1963).

6. CONFLICT WITH BAYESIAN THEORY

A logically consistent theory of inferential probability has been outlined. It includes Objective Bayesian inference and classical probability (in the inferential sense) as special cases, but is more generally a theory of confidence-based fiducial probability. It is free from the earlier difficulties and paradoxes encountered in the development of fiducial theory, and provides, I believe, a mathematical reconciliation of confidence and fiducial viewpoints. The radical noncoherence property of confidence-based probability leads to highly plausible expressions of inference in my view, that could not be derived without it.

Subjective Bayesian theory also appears to be a logically consistent theory of inference, if improper priors are excluded (see Dawid *et al.*, 1973), but clearly it is fundamentally incompatible with the inferential theory in this paper. If one theory is right the other must be wrong, and this is ultimately a matter of empirical resolution, perhaps requiring many more convincing examples than presently available. I personally think that Bayesian theories of inference are wrong (except in the asymptotic sense for large samples), since they fail to give expression to the intrinsic, noncoherent implications of finite data, and some further arguments are given below to support this view. However, I shall also make some more constructive remarks concerning Bayesian methods.

6.1. *The Likelihood Principle*

Which of the competing theories is right or wrong depends crucially on whether the Likelihood principle is empirically valid in general and likewise for non-Bayesian Likelihood inference. Returning to the simple context of Section 3.1, the objective information for inference about θ comprises two components, $\{x_a, dF(x;\ \theta)\}$. In (6.1) this information is expressed in a logically equivalent form in terms of the realized likelihood function $L_a(\theta)$ and its sampling distribution,

$$\{x_a, dF(x;\ \theta)\} \Leftrightarrow \{L_a(\theta), dF(L(\theta', x);\ \theta)\}. \tag{6.1}$$

(The second symbol θ' is needed so that the functional dependence on θ can be indicated.) Then the Likelihood principle states that only the actual likelihood $L_a(\theta)$ is relevant for inference about θ_a.

Now this is unequivocally true only in the Objective Bayesian case where θ_a is known to be the outcome of a physical random process, with frequency distribution $dF(\theta)$, for then it is a fact of frequency theory that $dF(\theta | x_a) \propto dF(\theta).L_a(\theta)$, i.e. Bayes' theorem. Except for this case, why is the second, logically distinct component of information $dF\{L(\theta', x);\ \theta\}$ to be always judged irrelevant? In general it does depend on θ and in the equivalent form $dF(x;\ \theta)$ determines the pivotal relation between x and θ shown in Fig. 1, in which the clustering of the percentile relations gives information on how close any observed value will be to the corresponding parameter value. Only in the special case where the specified family of sampling distributions is invariant under a group of transformations is $dF\{L(\theta', x);\ \theta\}$ logically equivalent, *in conjunction with the invariance information*, to $L_a(\theta)$. For instance, if we know that θ is a pure location parameter, so that $L_a(\theta) \equiv L(x_a - \theta)$, then given only $L_a(\theta)$ we can infer also the pivotal relation between x and θ and hence that the fiducial distribution $dP(\theta)$ is proportional to $L_a(\theta)\, d\theta$. The additional invariance information is represented, in $dP(\theta)$, by the multiplying element $d\theta$, which is the invariant differential measure with respect to translations. We may also infer other distributions noncoherent with $dP(\theta)$ given $L_a(\theta)$, such as $dP(\theta^2)$, but their density functions will not be proportional to a sample likelihood as in $dP(\theta)$.

In general it can be argued, therefore, that the additional sampling information is relevant and necessary for a proper interpretation of a sample likelihood, and that any theory which always suppresses this information violates the fundamental principle of uncertain inference. A Bayesian counter-argument to this would be that there is a sound axiomatic basis for Subjective Bayesian theory. I return to this in Section 6.3.

Birnbaum (1962) gave a purported proof that the Conditioning and Sufficiency principles jointly imply the Likelihood principle. The present theory of inferential probability contradicts this, since sufficiency and conditioning are fully utilized to give inferential distributions that do not satisfy the Likelihood principle, except in the special case described above. The notion of irreducibility (Section 4.1) is important here. Durbin (1970) showed that Birnbaum's proof fails if the Sufficiency principle is first applied to eliminate irrelevant information, so that conditioning is restricted to relevant variables only. G. A. Barnard (personal communication) has also indicated that Birnbaum's crucial Lemma 1 of the Sufficiency principle in invalid—(*added in proof*) whenever there is relevant ordering information in the sample and parameter spaces. (See Discussion (Barnard) and Reply.)

6.2. *Inferential Completeness*

An inferential statement (or set of statements) about a parameter θ could be termed *complete* if it suffices as a representation of the currently available information for all future applications. Here we shall consider a more restricted notion of completeness, namely that the inferential statement, if complete, should suffice for combining with further observational information about θ that may come to hand, to form an appropriately modified statement of inference.

A self-evident corollary of this is that a complete statement of inference will be symmetrically (invertibly) related to the relevant empirical information for it, since otherwise it could be argued that relevant information has been lost in the mathematical transformation to inferential form.

A remarkable difference can now be seen between an Objective Bayesian posterior and a fiducial distribution. Given that θ itself is a random variable with known frequency distribution $dF(\theta)$, Bayes' theorem is applicable for combining further observational information about the unknown realized value θ_a, and requires only the actual likelihood function $L_a(\theta_a)$ determined by the observations. The sampling variation of $L(\theta)$ is irrelevant. The posterior distribution $dP_a(\theta_a)$ is thus complete in the above sense, and can be described as the *known* distribution of belief regarding θ_a, given the observations. A fiducial distribution $dP_a(\theta_a)$, however, is inferentially incomplete. To preserve a symmetrical relation with the relevant data for it, $\{x_a, dF(x\colon \theta)\}$, it needs to be specified additionally as a function $dP(\theta_a; x_a)$ of x_a. In equivalent terms not only $dP_a(\theta_a)$ but also its sampling variation is relevant. In a statistical sense therefore, a fiducial distribution is more logically described as an *estimated* distribution of belief.

In general, Bayes' theorem cannot be applied to combine additional observational information with a prior fiducial distribution. This was proved by Lindley (1958). The basic reason is that the additional information *invalidates* the fiducial prior and a new fiducial posterior distribution must be deduced afresh from a fully relevant pivotal relation derived from the combined observations. The additional sampling information about a fiducial distribution is needed for this combination of information.

6.3. *Subjective Priors*

Savage (1954) proposed a set of axioms on the basis of which it is possible to deduce a personal, prior distribution of belief in any circumstances. However, Savage's formulation involves the use of an auxiliary random mechanism in determining a prior. This, in my view, violates the fundamental Relevance principle; outcomes of such a random guessing process are of questionable relevance in inference. Quite apart from this point it is clear that such a prior is itself a random variable, with random variation determined partly and often mostly by the guessing mechanism; and thus has a logical status similar to that of a fiducial distribution, as a (subjectively) *estimated* distribution of belief. Since Bayes' theorem is not applicable in general to fiducial priors, I see no reason for believing it to be applicable to personal priors either.

6.4. *Noncoherence*

A Bayesian counter-argument to the above is the view advocated by Lindley that one must behave *coherently* (in a decision-theoretic sense), and that coherent behaviour dictates the use of Bayes' theorem (Ramsey, 1931). However, in view of the other recognizable conflicts between statistical inference and decision theory, the validity of this argument is questionable. Furthermore, since there are empirically recognizable noncoherent implications in observational data, I suspect that decision-theoretic coherence as a necessity may be an artifact of the decision-theoretic formulation. Consider, for instance, the noncoherent implications of Stein's (1959) example in a quality control context, with two classes of customers specifying conflicting quality criteria. Where is the necessity for coherence here?

I also find it difficult to see why a prior distribution, formulated on the vaguest of prior information, should be considered capable of obliterating the empirical noncoherence of observational data, especially when the noncoherence is as extreme as that in the Stein example. Indeed, a symmetry argument may be invoked here: If reasoning from observational data alone leads only to noncoherent probability, what additional property does prior information possess that would lead instead to coherent prior probability?

A fundamental property of "ignorance" is its preservation under transformation—ignorance about θ implies ignorance about any function of θ, and one would expect the same to hold approximately for "near-ignorance". Such a property is compatible with the view that belief is intrinsically noncoherent, but not with the Bayesian view that belief is representable by a single probability distribution. I believe, indeed, that the Noncoherence Principle is as fundamental in Statistical Inference as the Heisenberg Uncertainty Principle is in Quantum theory—the parallel is unusually apt.

6.5. *The Principle of Precise Measurement*

I do not think this principle is valid as usually stated, for justifying the use of otherwise dubious diffuse priors; for any prior, however diffuse, has the remarkably strong effect, when used in conjunction with Bayes' theorem, of suppressing all information other than expressed in the sample likelihood function, and thus, in particular, the noncoherent implications of the observational data. The main justification for Bayesian methods is in terms of their large-sample asymptotic properties.

6.6. *A Modified Bayesian Theory*

In spite of the fundamental objections voiced above, Bayesian methods clearly have an important practical role, and I think that a noncoherent form of Bayesian theory could be developed as a logically consistent adjunct to the confidence-based inferential theory described here. Some common ground already exists with respect to the subtheory of structural distributions and of derived conditional distributions (Section 4.8). The latter suggest that *fiducial* likelihood functions may have an important role in a modified Bayesian calculus.

7. Acknowledgements

The research described here was carried out at the three institutions named in the title and at the University of Adelaide. The basic concepts were evolved at Rothamsted. A resolution of uniqueness problems in simultaneous inference was developed at Queen's University, and the final mathematical formulation was completed at Bell Laboratories. My personal indebtedness is first to the late R. A. Fisher. Companionship with him in Adelaide during his last years has had a profound influence on my understanding of Inference. I am also indebted to A. T. James, whose understanding of fiducial theory has contributed greatly to my own. Lastly, I am grateful to G. A. Barnard, C. R. Blyth, J. A. Nelder and W. N. Venables for comments which influenced the final formulation; and to R. Gnanadesikan, J. Kruskal and C. Mallows for valuable comments on presentation.

References

Aspin, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. (With an Appendix by B. L. Welch). *Biometrika*, **36**, 290–296.

Barnard, G. A. (1949). Statistical inference (with discussion). *J. R. Statist. Soc.* B, **11**, 116–149.

Barnard, G. A. and Sprott, D. A. (1971). A note on Basu's examples of anomalous ancillary statistics. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds), pp. 163–170. Toronto: Holt, Rinehart and Winston.

Barnett, V. (1973). *Comparative Statistical Inference*. New York: Wiley.

Bartlett, M. S. (1936). The information available in small samples. *Proc. Camb. Phil. Soc.*, **32**, 560–566.

—— (1956). Comment on Sir Ronald Fisher's paper: "On a test of significance in *Pearson's Biometrika Tables* (No. 11)". *J. R. Statist. Soc.* B, **18**, 295–296.

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

Behrens, W.-V. (1927). Ein Beitrag zur Fehlerberechung bei wenigen Beobachtungen. *Landw. Jb.* **68**, 807–837.

Bennett, G. W. and Cornish, E. A. (1963). A comparison of the simultaneous fiducial distributions derived from the multivariate normal distribution. *Bull. Int. Statist. Inst.*, **40**, 902–919.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Ass.*, **57**, 269–306.

Blyth, C. R. (1970). On the inference and decision models of statistics (with Discussion). *Ann. Math. Statist.*, **41**, 1034–1058.

Brillinger, D. R. (1962). Examples bearing on the definition of fiducial probability with a bibliography. *Ann. Math. Statist.*, **33**, 1349–1355.

Buehler, R. J. (1959). Some validity criteria for statistical inference. *Ann. Math. Statist.*, **30**, 845–863.

Cornish, E. A. (1961). Simultaneous fiducial distributions of location parameters. *C.S.I.R.O. Aust. Div. Math. Statist. Tech. Paper*, **8**.

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

—— (1971). The choice between alternative ancillary statistics. *J. R. Statist. Soc.* B, **33**, 251–255.

Creasy, M. A. (1954). Limits for the ratio of means. *J. R. Statist. Soc.* B, **16**, 186–192.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with Discussion). *J. R. Statist. Soc.* B, **35**, 189–233.

Dempster, A. P. (1963a). Further examples of inconsistencies in the fiducial argument. *Ann. Math. Statist.*, **34**, 884–891.

—— (1963b). On a paradox concerning inference about a covariance matrix. *Ann. Math. Statist.*, **34**, 1414–1418.

Durbin, J. (1970). On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood. *J. Amer. Statist. Ass.*, **65**, 395–401.

Edwards, A. W. F. (1976). Fiducial probability. *Statistician*, **25**, 15–35.

Fieller, E. C. (1954). Some problems in interval estimation. *J. R. Statist. Soc.* B, **16**, 175–185.

Fisher, R. A. (1930). Inverse probability. *Proc. Camb. Phil. Soc.*, **26**, 528–535.

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

—— (1935). The fiducial argument in statistical inference. *Ann. Eugen.*, **6**, 391–398.

—— (1936). Uncertain inference. *Proc. Amer. Acad. Arts Science*, **71**, 245–258.

—— (1937). On a point raised by M. S. Bartlett on fiducial probability. *Ann. Eugen.*, **7**, 370–375.

—— (1939). The comparison of samples with possibly unequal variances. *Ann. Eugen.*, **9**, 174–180.

—— (1941). The asymptotic approach to Behren's integral, with further tables for the *d* test of significance. *Ann. Eugen.*, **11**, 141–172.

—— (1945). The logical inversion of the notion of the random variable. *Sankhyā*, **7**, 129–132.

—— (1948). Conclusions fiduciaires. *Ann. Inst. Henri Poincaré*, **10**, 191–213.

—— (1954). Contribution to discussion of papers by Creasy and Fieller. *J. R. Statist. Soc.* B, **16**, 212–213.

—— (1955). Statistical methods and scientific induction. *J. R. Statist. Soc.* B, **17**, 69–78.

—— (1956a). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd (3rd ed.: Hafner).

—— (1956b). On a test of significance in Pearson's *Biometrika Tables* (No. 11). *J. R. Statist. Soc.* B, **18**, 56–60.

—— (1961a). Sampling the reference set. *Sankhyā*, **23**, 3–8.

—— (1961b). The weighted mean of two normal samples with unknown variance ratio. *Sankhyā*, **23**, 103–114.

—— (1971). Collected papers of R. A. Fisher, Vol. 1–5 (J. H. Bennett, ed.). Adelaide, South Australia: The University of Adelaide.

Fraser, D. A. S. (1961). The fiducial method and invariance. *Biometrika*, **48**, 261–280.

—— (1964a). On local unbiased estimation. *J. R. Statist. Soc.* B, **26**, 46–51.

—— (1964b). Local conditional sufficiency. *J. R. Statist. Soc.* B, **26**, 52–62.

—— (1964c). On local inference and information. *J. R. Statist. Soc.* B, **26**, 253–260.

—— (1968). *The Structure of Inference.* New York: Wiley.

—— (1972). Bayes, likelihood or structural? *Ann. Math. Statist.*, **43**, 777–790.

—— (1974). Comparison of inference philosophies. In *Information, Inference and Decision* (X. Menges, ed.), pp. 77–98. Dordrecht: Holland/Boston: Reidel.

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. R. Statist. Soc.* B, **25**, 368–376.

Hacking, I. (1965). *Logic of Statistical Inference.* Cambridge University Press.

James, A. T., Wilkinson, G. N. and Venables, W. N. (1974). Interval estimates for a ratio of means. *Sankhyā*, A, **36**, 177–183.

Kshirsagar, A. M. (1972). *Multivariate Analysis.* New York: Marcel Dekker.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *J. R. Statist. Soc.* B, **20**, 102–107.

Mauldon, J. G. (1955). Pivotal quantities for Wishart's and related distributions and a paradox in fiducial theory. *J. R. Statist. Soc.* B, **17**, 79–85.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. R. Statist. Soc.*, **97**, 558–625.

—— (1956). Note on an article by Sir Ronald Fisher. *J. R. Statist. Soc.* B, **18**, 288–294.

Plackett, R. L. (1966). Current trends in statistical inference. *J. R. Statist. Soc.* A, **129**, 249–267.

Ramsey, F. P. (1931). *The Foundations of Mathematics and Other Essays.* London: Kegan, Paul, Trench, Trubner & Co.

Robinson, G. K. (1975). Some counter examples to the theory of confidence intervals. *Biometrika*, **62**, 155–161.

SAVAGE, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

SEGAL, I. E. (1938). Fiducial distribution of several parameters with application to a normal system. *Proc. Camb. Phil. Soc.*, **34**, 41–47.

SHAFER, GLENN (1976). *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

SPROTT, D. A. (1963). Fiducial distributions associated with independent normal variates. *Sankhyā*, A, **25**, 403–406.

—— (1965). Transformations and sufficiency. *J. R. Statist. Soc.* B, **27**, 479–485.

—— (1973). Normal likelihoods and their relation to large sample theory of estimation. *Biometrika*, **60**, 457, 465.

—— (1975). Marginal and conditional sufficiency. *Biometrika*, **62**, 599–605.

STEIN, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp.*, **1**, 197–206.

—— (1959). An example of wide discrepancy between confidence intervals and fiducial intervals. *Ann. Math. Statist.*, **30**, 877–881.

TUKEY, J. W. (1957). Some examples with fiducial relevance. *Ann. Math. Statist.*, **28**, 687–695.

—— (1960). Conclusions vs. decisions. *Technometrics*, **2**, 423–433.

WELCH, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, **29**, 350–362.

—— (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, **34**, 28–35.

—— (1956). Note on some criticisms made by Sir Ronald Fisher. *J. R. Statist. Soc.* B, **18**, 297–302.

YATES, F. (1939). An apparent inconsistency arising from tests of significance based on fiducial distributions of unknown parameters. *Proc. Camb. Phil. Soc.*, **35**, 579–591.

—— (1964). Fiducial probability, recognizable subsets and Behrens' test. *Biometrics*, **20**, 343–360.

DISCUSSION OF MR WILKINSON'S PAPER

Dr A. W. F. EDWARDS (Cambridge University): My pleasure at being asked to propose the vote of thanks to Mr Wilkinson is exceeded only by my disappointment at being unable to do so in person; the administrative machinery of the University of Cambridge is relentless, and not even a signal contribution to our understanding of fiducial probability causes it to release one of its prisoners on parole.

I first encountered fiducial probability a few weeks before graduating from Fisher's department in 1957, having bought his recently published *Statistical Methods and Scientific Inference*. When I took it to him for his autograph he willingly signed it "Good luck—Ronald A. Fisher", and never was that phrase in greater need. Sixteen years later I had the temerity to write inside the cover the date and "I have now reached the end of this book".

Until today's paper, I have found most published contributions on fiducial probability disappointing. Notable exceptions are those of Sir Harold Jeffreys, who from an early stage was quite clear that with respect to location and scale parameters his invariant prior distributions and Fisher's fiducial postulate were different ways of saying the same thing; those of D. A. S. Fraser, whose early papers on fiducial probability were most illuminating, but who then fell into the black hole of structural probability and was inclined to forget the route by which he had travelled; and those of Ian Hacking, who took us back to the work of Jeffreys and Fisher at a time when it was more fashionable to move in the opposite direction.

Barnard, Dempster, Sprott and Yates all kept worrying away at the argument, but for the most part the published comment ranged from outright hostility, through papers in which respected statisticians made it abundantly clear to the informed reader that they had totally failed to grasp the essential point, to the comments of one or two courageous names who publicly admitted their confusion.

Mr Wilkinson has put us in his debt by taking fiducial probability seriously. Anyone who does that is forced into one of two positions. Either he limits the field of application of fiducial probability to those models which do not generate paradoxes, which means essentially location-and-scale parameter models, though as Fraser has shown one can eliminate paradoxes which arise through a multiplicity of pivotal quantities by specifying the pivot as part of the model: hence structural probability. Or he takes the bull by the horns and argues that fiducial probabilities are not ordinary probabilities, and that some new principle has to be admitted. Hacking wheedled the Principle of Irrelevance out of the writings of Fisher and Jeffreys, and the development of that line led to the restricted viewpoint described in my paper "Fiducial probability" to which