Department of Economics
University of Southampton
Southampton SO17 1BJ
UK

Discussion Papers in Economics and Econometrics

HOW LIKELIHOOD AND IDENTIFICATION WENT BAYESIAN

John Aldrich

No. 0111

This paper is available on our website
http://www/soton.ac.uk/~econweb/dp/dp01.html

How Likelihood and Identification went Bayesian

John Aldrich
Department of Economics
University of Southampton
Southampton
SO17 1BJ
UK

Fax (0)(+44) 23 80593858
E-mail: john.aldrich@soton.ac.uk

Abstract
This paper considers how the concepts of likelihood and identification became part of Bayesian theory. This makes a nice study in the development of concepts in statistical theory. Likelihood slipped in easily but there was a protracted debate about how identification should be treated. Initially there was no agreement on whether identification involved the prior, the likelihood or the posterior.

October 2001

1 Introduction

Some of the concepts and terms associated with twentieth-century Bayesian theory were new, like "Bayesian" itself, a few were from the remote past, like "prior" and "posterior", but most were from non-Bayesian theory. This paper examines how two of these, *likelihood* and *identification*, passed into Bayesian theory. S*ufficiency* will also be involved and the story of the three makes a nice case-study in the development of concepts in statistical theory.

On present views identification and likelihood are related but their origins were quite separate. "Likelihood" came from R. A. Fisher's theory of estimation and some of his theory came too–propositions Harold Jeffreys (1939) thought could be established more easily in Bayesian theory. Later Bayesians were not so fixed on Fisher but likelihood stayed without much debate. "Identification" crystallised in the 1940s when the econometrician Tjalling C. Koopmans discussed a type of inference *distinct* from statistical inference. Identification then made a double passage, into statistical inference and into Bayesian econometrics. However there was no agreement on which propositions about identification were the important ones or even whether identification involved the prior, the likelihood or the posterior. "Whether Bayesian theory requires a different definition of identification from the classical one" was still deemed "unresolved" by Hsiao (p. 272) in 1983. A resolution seems to have been reached, though this disturbed past is still reflected in recent contributions to Bayesian econometrics.

The writers involved with likelihood and identification were seldom as explicit as Kolmogorov and Fomin (1970, p. 96) who commend one concept as "natural" or "fruitful" and explain the survival of another through "historical inertia", but they had to make similar judgements. Those judgements link decisions about definitions and similar minutiae to the bigger themes in statistical inference.

The account follows likelihood from its anti-Bayesian beginnings in Fisher's theory into

Jeffreys's theory and on to the appearance of the likelihood principle around 1960, since when there has been relative quiet. I mention some proto-identification theory but the identification story begins in the 1940s, with Bayesian interest peaking in the 1970s. That development too may have run its course.

2 The Fisher programme

In the early 1920s R. A. Fisher (1890-1962) developed a new approach to statistical theory based on the idea that the object in calculating a statistic is to extract as much relevant information from the data as possible; see Aldrich (1997) for a detailed account and Box (1978) for biographical data. Likelihood played an important part in the new theory and, reviewing the theory after a decade or so, Fisher (1935, p. 41) judged that he had amply demonstrated "the adequacy of the concept of likelihood for inductive reasoning". The claim and the concept require attention because they were carried into Bayesian theory. Incidentally Fisher rewrote the language of statistical theory on a scale unmatched by any twentieth century Bayesian.

2.1 Fisher's likelihood

Fisher formulated the notion of likelihood because he wanted to emphasise the distinctiveness of his approach. Soper, Young, Cave, Lee and Pearson (1917) interpreted Fisher's phrase "most likely value" as the value obtained from maximising the posterior distribution obtained from a uniform prior and they criticised his choice of prior. Fisher however insisted he was *not* using a prior–indeed he (1921, p. 24) rejected the whole set-up:

> Bayes (1763) attempted to find, by observing a sample, the actual probability that

> the population value lay in any given range. ... Such a problem is indeterminate without knowing the statistical mechanism under which different values of [the parameter] come into existence; it cannot be solved from the data supplied by a sample, or any number of samples, of the population.

However *if* the statistical mechanism *were* known Fisher (1925, p. 10) agreed that Bayes' solution could be applied.

Fisher defined likelihood in "On the Mathematical Foundations of Theoretical Statistics" (1922, p. 310):

> The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.

The proportionality derived from the presence of the differential element when writing the "chance of an observation falling in the range dx" (p. 323) but as Fisher was only ever interested in the part of the function involving the unknown parameter–he usually took logs and differentiated–he could be more radical. Thus in 1932 he (pp. 258-9) ignored "the numerical coefficient which is independent of x [the parameter]" and wrote the likelihood for the binomial as $x^a(1-x)^b$ or "some arbitrary multiple of it".

Fisher distinguished likelihood not only from the (usually) illegitimate posterior probability but from a legitimate frequency notion of probability. He (1921, p. 4) complained that "two radically distinct concepts have been confused under the name of 'probability'..." The roles of (legitimate) probability and likelihood were set out in the *Statistical Methods for Research Workers* (1925, pp. 9-11):

> The deduction of inferences respecting samples, from assumptions respecting the

> populations from which they are drawn, shows us the position in Statistics of the **Theory of Probability**. ...
>
> This is not to say that we cannot draw, from knowledge of a sample, inferences respecting the population from which the sample was drawn, but that the mathematical concept of probability is inadequate to express our mental confidence or diffidence in making such inferences and that the mathematical quantity [**Likelihood**] which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability.

Hitherto the task of expressing "mental confidence" had been assigned to the invalid posterior probability. On the final point in the passage Fisher long complained that "there is still a tendency to treat likelihood as though it were a sort of probability."

While likelihood was fixed by the definition of 1922, its status was ambiguous. The 1925 passage and the 1921 paper (which introduced likelihood) treat likelihood as a primitive measure of belief, as it was in Barnard (1949), Edwards (1972) and other later works. Yet the "Foundations" and Fisher's main writings into the 1930s did *not* stress the primitiveness of likelihood but rather propositions about the efficiency of likelihood-based methods in extracting information. The case for the "adequacy" of the concept of likelihood was not a matter of definition; it rested on two argument Fisher (1935, p. 53) summarised as:

> First, that the particular method of estimation, arrived at by choosing those values of the parameters the likelihood of which is greatest, is found to elicit not less information than any other method which can be adopted. Secondly, the residual information supplied by the sample, which is not included in a mere statement of

> the parametric values which maximize the likelihood, can be obtained from other characteristics of the likelihood function ...

The "information" here is a repeated sampling notion–the "Fisher information measure".

2.2 Information and proto-identification

Information underpinned Fisher's case for likelihood but information ideas are also relevant to the identification story and, although they did not come into play until much later, it is convenient to describe them here.

Behind the information measure was a primitive qualitative notion of a statistic containing *no* information about a parameter. The idea that a statistic contains *no* information about a parameter when its distribution does not depend on the parameter arrived with sufficiency. Sufficiency is described in the glossary of the "Foundations" (p. 310) thus

> A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the sample provides any additional information as to the value of the parameter to be estimated.

In the body of the paper Fisher (p. 316) is more formal about what the sufficiency of a statistic, θ_1, involves

> In mathematical language ... if θ be the parameter to be estimated, θ_1 a statistic which contains the whole of the information as to the value of θ, which the sample supplies and θ_2 any other statistic, then the surface of distribution of pairs of values of θ_1 and θ_2 , for a given value of θ, is such that for a given value of θ_1, the distribution of θ_2 does not involve θ. In other words, when θ_1 is known, knowledge of the value of θ_2 throws no further light on upon the value of θ.

This is hardly an elementary application of uninformativeness as variation-freeness yet it is clear that *this* is the idea that is being applied. In the notation of §6.2 below, suppose the probability density of Y, $f(y, \alpha)$, is known to belong to a family of densities with $\alpha \in A$. There is *no* "information in the sample as to the value of the parameter to be estimated" when

$$f(y, \alpha^1) = f(y, \alpha^2) \text{ for all } y$$

for *any* pair of parameter values α^1 and α^2 in A. Variation-freeness also figured later in Fisher's notion of ancillarity and the conditional inference theory based upon it.

Fisher discussed the elementary case in 1935 when he (p. 47) tested the information measure against "our pre-mathematical common sense" requirements of such a measure. The measure is zero when it should be: "when the probabilities of the different kinds of observation which can be made are all independent of a particular parameter, the observations will supply no information about the parameter."

The identification concerns in Koopmans's writings of the 1940s–§6.1 below–were quite different and Fisher came closest to anticipating them, not in any theory, but in an example. In the *Statistical Methods* Fisher (1925 p. 24) considers the task of estimating θ in a restricted multinomial distribution with cell probabilities

$$\left(\frac{1}{4}(2+\theta), \quad \frac{1}{4}(1-\theta), \quad \frac{1}{4}(1-\theta), \quad \frac{1}{4}\theta\right).$$

θ, however, is not a fundamental parameter in Fisher's genetical example, it is the product of p and p', the "recombination fractions" for males and females. In the first edition Fisher did not explain why he does not estimate these parameters but subsequently he (1928 p. 241)

wrote:

> Since these [multinomial] probabilities involve only the quantity pp', it is only this and not the separate values of p and p' that the data can provide an estimate. ... If p and p' were equal, then $\sqrt{\theta}$ would give the recombination fractions in both sexes, and if these are unequal it will always give their geometric mean. The data before us, however, throw direct light only on the value of θ.

The information calculations that follow concern θ, not p and p' which drop out of sight.

This example and the issues it illustrates can be connected with the formal information ideas–see §6.2 below–but Fisher had no occasion to do so. There was another information context in which he did not go into details. Fisher was constantly making 'no no-information' assumptions. Thus, in taking the reciprocal of the information measure in his maximum likelihood theory, he was assuming that the measure is non-zero. *Any* careful description of the circumstances in which a technique will work involves some kind of identification condition and there were already such descriptions in the literature.

Fisher had been taught least squares after the fashion of Brunt (1917). The model involved, put in modern notation, is

$$y \sim N(X\beta, \sigma^2 I)$$

with X a matrix of known constants. The section (p. 79) on the "independence of the normal equations" considers how the normal equations may *not* yield a "determinate solution". The treatment can be traced back to Gauss's first publication on least squares (1809) where there is a version of the full rank condition for "determinacy"; see Aldrich (1999) for details. Another investigation of determinability appears in Pearson's (1894) paper on estimating a mixture of

normals. Pearson first establishes that a non-degenerate mixture is uniquely determined–"a curve which breaks up into two normal components can break up in one way, and one way only." (p. 74).

To return to likelihood, Fisher's case for "the adequacy of the concept of likelihood" rested on information. I have not detailed the case for Jeffreys (1935, p. 70) thought the argument "would be made much easier by an explicit use of probability" and likelihood first when Bayesian went Jeffreys made the case without information technicalities.

3 The Jeffreys programme

The earliest Bayesian work of the physicist Harold Jeffreys (1891-1989) was oriented more to logic and the philosophy of science than to statistical analysis but in the 1930s his seismological research took a statistical turn and he started writing about statistics and paying attention to Fisher. The *Theory of Probability* (1939) developed a science of statistics, of comparable scope to Fisher's, founded on the *theory of probability*–Jeffreys's name for the theory of inductive inference based on the principle of inverse probability or Bayes' theorem; see Cook (1991) for biographical information and Lindley (1986) for an account of Jeffreys's book.

The *Probability* recast Fisher's methods in a consistent form and showed how some of his larger objectives could be realised more easily. Jeffreys (1939, p. 323) admired Fisher's ability to grasp the essentials of a solution by "some brilliant piece of common sense" but lamented the lack of system. Jeffreys was more influenced by Fisher than by writers from the Bayesian past. He mentions earlier Bayesians, including Laplace, Edgeworth and Pearson for their views *about* Bayesian theory but hardly for their work *in* Bayesian theory. Pearson, for instance, is the author of the *Grammar of Science, not* the cooperator of 1917 or the 1907 investigator

of "the influence of past experience on future expectations". It has taken modern historians, like Dale (1991) and Hald (1998), to recover the pre-Jeffreys Bayesian past.

3.1 Jeffreys's likelihood

Jeffreys was the first to use the now standard terminology of Bayes' formula in its application to inference; see David (2001). Wrinch and Jeffreys (1921, p. 387) contracted the widely-used "probability a priori" and "probability a posteriori" to "prior probability" and "posterior probability", less, Jeffreys (1939, p. 29) explained, for linguistic streamlining than because "a priori" had other philosophical meanings and was open to misunderstanding. The adoption of "likelihood" was a more significant step and Jeffreys only took it after he had steeped himself in Fisher's work. Jeffreys chose to put likelihood into the *Probability*; he could have chosen another word or left the component without a name. He did not explain his decision beyond saying that Fisher's term was "convenient".

Jeffreys (1939, p. 29) wrote the "principle of inverse probability, first given by Bayes in 1763" as

$$P(q_r \,|\, pH) \propto P(q_r \,|\, H) P(p \,|\, q_r H)$$

where the H on which all the probabilities are conditional is "previous knowledge". He interpreted the terms as follows:

> If p is a description of a set of observations and the q_r a set of hypotheses, the factor $P(q_r \,|\, H)$ may be called the *prior probability,* $P(q_r \,|\, pH)$ the *posterior probability* and $P(p \,|\, q_r H)$ the *likelihood*, a convenient term introduced by Professor R. A. Fisher, though in his usage it is sometimes multiplied by a constant factor. It is the

probability of the observations given the original information and the hypothesis under discussion.

Jeffreys draws attention to and then *deviates* from Fisher's usage; he does *not* multiply by a constant. The form in which Jeffreys writes the principle is that which follows most directly from the definition of conditional probability. Presumably Jeffreys saw no point in the Fisher version. Indeed he may have thought that Fisher's separation of likelihood and probability was unhelpful. Fisher's point about likelihood not obeying the laws of probability could be made in Jeffreys's notation as

$$P(p \,|\, q_r \text{ or } q_r, H) \neq P(p \,|\, q_r H) + P(p \,|\, q_s H).$$

Jeffreys's likelihood is clearly *not* the same as Fisher's likelihood. The likelihoods do not refer to the same things. One is, or derives from, a family of conditional distributions where the conditioning variable is a parameter while the other comes from a family of unconditional distributions indexed by the parameter. This difference reflects the fundamental difference in the way parameters are conceived in the two theories. However the difference seemed to cause nobody any trouble because it is clear how the translations should be made. Although parameters are random variables in one paradigm only, the notions of random variable and conditional distribution are common to the paradigms. Fisher's term would have been inconvenient and there would have been something to argue about if Bayesian theory used both the indexing notion and the conditioning notion. On a more operational level Fisher's maximum likelihood and Jeffreys's maximum likelihood are the same functions of the data. Perhaps Jeffreys's likelihood is as near to Fisher's as it could be *given* the different status of parameters in the two theories. This is a big *given* and Bayesians could have introduced a new term.

Part of the *disanalogy* between Fisher's likelihood and Jeffreys's likelihood was Fisher's claim that likelihood is "appropriate for measuring our order of preference among different possible populations",something Jeffreys might say of the posterior distribution. However there was a *positive* analogy in that some of Fisher's information claims could be sustained by Jeffreys's likelihood. Thus when he came to interpret the magic formula

$$\text{Posterior Probability} \propto \text{Prior Probability} \times \text{Likelihood}$$

Jeffreys (1939, p. 46) wrote

> where by the likelihood we understand the probability that the observations should have occurred, given the hypothesis and the previous knowledge. The prior probability has nothing to do with the observations immediately under discussion, though it may depend on previous observations. Consequently the whole of the information that is relevant to the posterior probabilities of different hypotheses is summed up in the values they [the observations] give to the likelihood.

The last sentence led Berger and Wolpert (1984, p. 23) and Lindley (1986, p. 36) to suggest that Jeffreys understood the likelihood principle. However the likelihood principle involves a contrast between the probability distribution of the observations and *anything* that works in the formula. Jeffreys, however, was using "likelihood" for the probability distribution and was contrasting it with the prior distribution. The likelihood principle only emerged in Bayesian statistics around 1960–see §5.1 below. Jeffreys missed this corollary of the principle of inverse probability although admittedly it would have fitted in well with other parts of his system e.g. his (pp. 315-6) well-known criticism of the use of tail areas in significance testing: "*a hypothesis that may be true may be rejected because it has not predicted observable results that*

have not occurred."

Unlike Fisher, Jeffreys based no technical development on "information" or "the whole of the information". He did not need to for, however they are defined, they must operate through the likelihood function as the only way the observations affect the posterior is through the likelihood. It was easy for Jeffreys to recast some of Fisher's core theory. We have seen how he established the first part of the first of the "main results" of the "present system" (p. 351)

> a proof independent of limiting processes that the whole information contained in the observations with respect to the hypotheses under test is contained in the likelihood, and that where sufficient statistics exist other functions of the observations are irrelevant.

Fisher's proof of this first part had required the notion of Fisher information which Jeffreys thought could only be justified by a large sample argument.

Jeffreys also had his own method for the second sufficiency part of the result. He replaced Fisher's original definition of sufficiency (§2.2 above) with one based on the factorisation criterion. (Fisher had used the criterion but not to define sufficiency.) Jeffreys (pp. 89-90) writes:

> Whenever the likelihood, apart from factors independent of the unknown parameters to be estimated, can be expressed as a function of the unknown parameters, the number of observations, and a number of functions of the observations equal to the number of unknown parameters, those functions of the observations are called sufficient statistics.

Thus he (p. 90) could say that "the whole of the information with respect to the [parameters] that is contained in the observations is summmarized in the [sufficient statistics]".

In the third edition of the *Theory* (1961, pp. 165-6) Jeffreys shifted the emphasis in his account of sufficiency to the way the posterior distribution depends on the data and showed how the Fisher 1922 definition could be "proved" as a theorem. Raiffa and Schaifer (1961, p. 32) adopted a similar posterior-based definition: the posterior depends on the data through the sufficient statistic. They (p. 34) establish the "complete equivalence" of the Bayesian and classical (factorisation) definitions of sufficiency; the criterion is that the two concepts have the same extension, i.e. the same statistics from the same distributions are sufficient according to the two definitions. Raiffa and Schaifer prefer their definition because it "leads naturally" (p. xi) to the concept of marginal, or partial, sufficiency which concerns the way the data acts on the marginal posterior. This notion (p. 35) is relative to the prior distribution that permits the nuisance parameter to be integrated out.

Jeffreys installed likelihood in Bayesian theory. His definition corrected Fisher's but the claims Fisher made about likelihood and information could be made–and proved more easily–for their Jeffreys counterparts. So there was continuity of purpose as well as continuity of reference. When Fisher's concerns became less pressing there was more of a cost in maintaining likelihood. Schlaifer (1959, p. 338) warned his readers: "Again we emphasise that the new term is introduced purely for convenience: *a likelihood is a probability in the same sense as any other probability."* Against the "convenience" had to be set the fact that this was an unnatural term with the wrong associations. Convenience can be another name for historical inertia–it was not for Jeffreys in 1939.

4 Classical interlude: Neyman and Wald

The Bayesians who came after Jeffreys were not enthralled to Fisher; Raiffa and Schlaifer's (1961, p. ix) global acknowledgment is confined to "to Neyman, Pearson, Jeffreys, Von Neumann, Wald, Blackwell, Girshick and Savage." Fisher's information/likelihood theory had been replaced as the norm by the "classical" decision theory of Jerzy Neyman (1894-1981) and Abraham Wald (1902-50). For biographical information on Neyman and Wald see Reid (1982) and Wolfowitz (1952).

Neyman's "Outline of a Theory of Statistical Estimation based on the Classical Theory of Probability" (1937) confronted Jeffreys by distinguishing (p. 341) Jeffreys's theory from the "classical theory". It did not confront Fisher, though an essential point of the new theory was that it was *not* based on likelihood or information. Neyman (1935, p. 74) had pressed Fisher on whether likelihood is outside probability and whether it is possible to construct a theory of mathematical statistics based solely on the (classical!) theory of probability. Neyman (p. 75) preferred to found estimation on the "frequency of errors in judgement". This provides a "sufficiently simple and unquestionable principle in statistical work" while the "amount of information is too complicated and remote to serve as a principle". Likelihood and maximum likelihood had figured in Neyman's inference theory since the Neyman-Pearson 1928 paper on likelihood ratio tests. Neyman and Pearson (pp. 184-7) used the same definitions as Fisher's but their likelihood, likelihood ratios and maximum likelihood were not enchanted but merely useful, their usefulness registering in results on error frequencies.

Wald (1939) gave a unified treatment of testing and estimation based on the notions of loss ("weight") functions, risk functions and admissibility. He (p. 302) repeated Neyman's objections to a priori probability distributions but used the concept in a technical capacity: "it proves to be useful in deducing certain theorems and in the calculation of the best systems

of regions of acceptance". In his *Statistical Decision Functions* Wald (1950) extended this line of research emphasising the minimax criterion and including (p. v) a "treatment of the design of experimentation as a part of the general decision problem."

5 Bayesian decision theory

There were many publications that contributed to the modern Bayesian revival but I concentrate on two of the most influential: *The Foundations of Statistics* (1954) by L. J. Savage (1917-1971) and *Applied Statistical Decision Theory* (1961) by Howard Raiffa (b 1924) and Robert Schlaifer (d 1994); for biographical information on Savage, see Lindley (1980). The new Bayesian line was very different from Jeffreys's. The foundations combined "personalism" with the "behavioralism" (decision-orientation) of Neyman and Wald. Savage (p. 276) considered the *Theory of Probability* "an ingenious and vigorous defense of a necessary view, similar to, but more sophisticated than Laplace's". Raiffa and Schlaifer refer to some of Jeffreys's higher-level contributions, though not to his recasting of Fisher.

Unlike Fisher, these authors were not profoundly dissatisfied with their elders. Savage (1954, p. 4) judged the methods of the "British-American School" as "on the whole consistent" with the theory of probability he was proposing–indeed he tried to develop a subjectivist interpretation of Wald's minimax theory. Raiffa and Schlaifer (1961, p. vii) stated:

> the so-called "Bayesian" principles underlying the methods of analysis presented in this book are in no sense in conflict with the principles underlying the traditional decision theory of Neyman and Pearson.

The term "Bayesian", incidentally, had only come into use around 1950–see David (2001).

While Raiffa and Schlaifer played down the conflict between classical and Bayesian approaches, important differences were emerging. These differences affected methods as well as foundations. Lindley (1980, p. 7) recalls Savage's (1962a, p. 307) admission "I came to take ... Bayesian statistics ... seriously only through recognition of the likelihood principle.".

5.1 The likelihood principle: terminal and preposterior analysis

"Likelihood" on its own hardly figures in Savage's 1954 book. He (p. 140) writes "The concept of likelihood ratio, sometimes simply called likelihood, is now one of the most pervasive concepts of statistical theory". Likelihood ratios are discussed at length and there is some discussion of maximum likelihood. Raiffa and Schlaifer discuss likelihood *per se* and use Jeffreys's definition, thus writing for the discrete case (p. 29): "If the conditional measure has a mass function, we shall denote by $l(z \mid \theta)$ [likelihood] the probability given θ that e [the experiment] results in z".

Raiffa and Schlaifer introduced the refinement of the "likelihood kernel". The kernel was useful in extracting distributional information from Bayes' formula and helpful in setting up the conjugate prior theory. It is presented (p. 30) as follows:

> if ρ and κ are functions on Z [the observable] such that for all z and θ
>
> $$l(z \mid \theta) = \kappa(z \mid \theta)\rho(z)$$
>
> i.e. if the ratio $\kappa(z \mid \theta)/l(z \mid \theta)$ is a constant as regards θ, we shall say that $\kappa(z \mid \theta)$ is a (not "the") *kernel* of the likelihood ...

While most modern Bayesians define likelihood as the conditional measure, some, e.g. Box and Tiao (1973, pp. 10-1), slide between this and the likelihood kernel in the Fisher manner.

Savage previewed the likelihood principle when he (1960, p. 544) wrote

> it is becoming increasingly accepted that, once an experiment has been done, any analysis or other reaction to the experiment ought to depend on the likelihood-ratio function and on it alone, without any further regard to how the experiment was actually planned or performed.

He had first heard this argument–applied to sequential sampling–from Barnard in 1952, as he (1962, p. 76) later recalled. Savage (1962, p. 17) set out the argument more fully, replacing the "likelihood-ratio function" by the likelihood:

> According to Bayes's theorem $Pr(x \mid \lambda)$, considered as function of λ constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and precisely, if y is the datum of some other experiment, and if it happens that $Pr(y \mid \lambda)$ and $Pr(x \mid \lambda)$ are proportional functions of λ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the value of λ .... I, and others, call this principle the likelihood principle. The function $Pr(x \mid \lambda)$–rather this function together with all others that results from it by multiplication by a positive constant–is called the likelihood function.

Raiffa and Schlaifer do not state the likelihood principle but their detailed comparison of the Bayesian and classical treatments of optional stopping (pp. 36-42) became a standard illustration. They (p. 42) summarised:

> the likelihoods for all noninformative stopping processes have a common kernel, and therefore all lead to the same posterior distribution. ...[O]n the other hand

> we shall also be concerned with the problems of *experimental design* as they look before any sample has actually been taken, and then we *shall* want to ask what can be expected to happen if we predetermine r [the number of successes] rather than n [the number of trials] and so forth.

Raiffa and Schlaifer [p. x] distinguish between "terminal" (post-experimental) analysis and "preposterior" (pre-experimental) analysis, between choice of an act after an experiment has been performed, and choice of the experiment to be performed. In pre-experimental analysis the probability distribution of the observable is considered but post-experimental analysis is conditional on the observed outcome and based on the likelihood function, not on the density. Thus something of Fisher's distinction (§2.1) between probability and likelihood–and of elaborations such as that by Barnard (1949)–reappears in the Raiffa and Schlaifer scheme.

Bayesians have varied in their attitudes to the likelihood principle. Against the enthusiasm of Savage and Lindley, of the econometricians Leamer (1978) and Poirier (1988), is the caution of Gelman, Carlin, Stern and Rubin (1995, pp. 9, 190) for whom the principle is of limited interest and its suggestion that analysis be done "without further regard to how the experiment was actually planned or performed" potentially misleading. However, despite disagreements about the significance of the likelihood principle and differences in the use of likelihood and likelihood kernel, likelihood's passage into Bayesian theory was untroubled. This was not the case with identification.

6 The troubles with identification

There are two obvious ways of transferring a concept into a new theory so that some continuity is maintained. Either take its meaning as fixed and consider how the concept functions in the

new theory or find something in the new theory with the same or similar function. When Jeffreys appropriated likelihood he took its meaning as fixed–more or less–and found that his likelihood had a similar function to Fisher's in that some of the theorems in which it figured were the same. The situation with identification was more complicated for both ways of managing the transfer were tried–either filling the blank in "from the Bayesian viewpoint, classical identification theory is really concerned with —" (from Drèze (1972, p. 27)) or filling the blank in "an appropriate Bayesian definition of identification is —" (from Rothenberg (1973, p. 158)). Both projects were pursued, but identification carried so much baggage that there was more than one way of filling both blanks.

It may be worth noting some points about the language of the identification community. "Identified" and "identifiable"–and "identification" and "identifiability"–circulated as synonyms, linguistic anomalies it would take too long to unravel. "Identification problem" signified both a topic for discussion and something requiring a solution. If there were a problem, i.e. the parameter were not identifiable, or under-identified, the solutions were to lower one's sights to an "identified function" of the parameter or to get more information.

6.1 Koopmans's identifiability

The topic of identification–with its concepts of identifiability and observational equivalence–was formalised in the 1940s by T. C. Koopmans (1910-85), first as applied to the simultaneous equations model of econometrics and then to scientific inference in general. In econometrics identification became a big topic, warranting a book in 1966 (F. M. Fisher's *Identification Problem in Econometrics)* and a chapter in the1983 *Handbook of Econometrics* (by C. Hsiao). The story of identification in econometrics is a complex one and various facets have been discussed by Qin (1989), Morgan (1990, chapter 6), Aldrich (1994), Rayner and Aldrich (1997)

and others.

Koopmans's most general ideas appear in "The Identification of Structural Characteristics" by Koopmans and Reiersøl (1950). The authors were particularly interested in econometrics and factor analysis but the genetic example from Fisher's *Statistical Methods* (§2.2 above) can illustrate their viewpoint. The "structural characteristics" are the elements of (p, p'), while the distribution of the observations depends on the multinomial probabilities. Given "exact knowledge" of those probabilities, can the value of (p, p') be deduced? Clearly not, for infinitely many pairs of values generate the same value of the product pp' and hence the same probability distribution for the observables. In Koopmans's language, the pair (p, p') is not *identifiable* since for every pair there is another (in fact infinitely many) *observationally equivalent* pair(s). Koopmans and Reiersøl (p. 170) state that "the study of identifiability proceeds from a hypothetical exact knowledge of the probability distribution of observed variables rather than from a finite sample of observations." Their "identification problem" was outside the scheme of statistical analysis set down in R. A. Fisher's "Foundations" in which the statistician specifies a population and makes inferences from the sample to the parameters of the population.

Writing of the economic context, Koopmans (1949, Abstract p. 125) separates the process of inference from sample to theoretical structure into a step from sample to population and one from population to structure.

> Statistical inference, from observations to economic behavior parameters, can be made in two steps: inference from the observations to the parameters of the assumed joint distribution of the observations, and inference from that distribution to the parameters of the structural equations describing economic behavior. The latter problem of inference [is] described by the term "identification problem".

Step one is the domain of statistical inference in the narrow sense, step two the domain of

economic theory or the structural theory of whichever substantive field the observations come from. Identifiability turns on the invertibility of the mapping–e.g. $pp' = \theta$ or $B^{-1}\Gamma = \Pi$ (see below)–from the theory parameters to the parameters that can be estimated from the data and the typical identification theorem involved a condition on the rank of a matrix. Koopmans was interested in the total process of inference from sample to structure and failure at the first stage wrecked the total process. For Koopmans this involved a data deficiency of some kind–the most familiar being an X matrix of deficient rank in regression–*not* a failure of identification.

Koopmans's identification concerned a layer of inference beyond statistical inference and belonged in the structural theory. One can imagine a Bayesian and the classical Koopmans agreeing that identification is a property of the structural specification and is the same whether considered classically or from the Bayesian approach. However I have not found any Bayesian taking this line for there was a powerful deflecting influence in the idea that identification is related prior information. The relationship was built into the way the simultaneous equations model was usually presented.

Following Koopmans, Rubin and Leipnik (1950), each equation of the simultaneous equations model has an economic "identity" as a demand, a supply or some other theoretically meaningful equation. There is a vector z_t of conditioning (exogenous) variables and a vector y_t of conditioned (endogenous) variables. The system of "structural" equations is organised into two parts: an unrestricted structural model with parameters B, Γ, Σ and a set of "a priori restrictions" expressed through the function Ψ:

$$\begin{aligned} By_t &= \Gamma z_t + \varepsilon_t : \varepsilon_t \backsim IN(0, \Sigma) : t = 1, ..., T \\ \Psi(B, \Gamma, \Sigma) &= 0. \end{aligned}$$

The commonest form of restriction was the exclusion restriction–a particular element of B or Γ is zero signifying that a particular variable does not appear in a particular equation.

The implied "joint distribution of the observations" or "population" is given by the reduced form equations

$$\begin{aligned} y_t &= \Pi z_t + v_t : v_t \backsim IN(0, \Omega), \text{ with } \Pi = B^{-1}\Gamma \text{ and } \Omega = B^{-1}\Sigma B^{-1'} \\ \textit{or } \ y_t &\backsim IN(\Pi z_t, \Omega). \end{aligned}$$

The structure is identified when only one (B, Γ, Σ) *satisfying the restrictions* can be obtained from an attainable Π and Ω.

If there are *more than one* (B, Γ, Σ), Koopmans, Rubin and Leipnik (p. 73) call them observationally equivalent; they represent "mathematically equivalent ways of writing down" the distribution. Econometricians came to speak of identification being "achieved" or the parameters "identified" by the imposition of "restrictions". In this spirit the identification of the parameters in the Fisher genetic model would be achieved if, say, the restriction $p = p'$ were imposed.

Econometricians generally took the hierarchical formulation of the structural model for granted, though as Malinvaud (1966, p. 545) noted,

> the distinction between the general statistical hypothesis and a priori restrictions is purely conventional. Its only object is to facilitate the discussion of identification problems, and it is based on no necessary logical principle.

Koopmans's mathematical technique for establishing identification results exploited this distinction, but it does not usually match the economics of the situation for the unrestricted model has no obvious economic interpretation. One could just write down the 'restricted'

model as Haavelmo (1944, p. 99) originally did. Bayarri, DeGroot and Kadane (1988) have drawn attention to the arbitrariness of the division into prior and likelihood in a hierarchical model; the hierarchy was an important part of the Koopmans identification legacy.

Koopmans put the distinction, between what can be discovered from data under favourable–even ideal–circumstances and what cannot, to the service of a view that economic theory is indispensable. Thus Koopmans, Rubin and Leipnik (p. 64) insist that

> Statistical observation will in favourable circumstances permit [the investigator] to estimate ... the characteristics of the probability distribution of the variables. Under no circumstances whatever will passive statistical observation permit him to distinguish between mathematically equivalent ways of writing down that distribution.... The only way in which he can hope to identify ... equations ... is with the help of a priori specifications of the form of each structural equation.

"Favourable circumstances"–enough variety in z to deliver Π and Ω–will not necessarily deliver (B, Γ, Σ).

6.2 Identification goes statistical

In the 50s and 60s the terminology of identification came to be applied to *statistical* inference, mainly in the context of regularity conditions for statistical inference procedures–see Aldrich (1994, introduction). T. J. Rothenberg's "Identification in parametric models" (1971) consolidated this movement with a body of theorems. In the 1977 reprint of the *Identification Problem* F. M. Fisher (p. v) distinguished the treatment of identification in a "general statistical setting" with his own which treated identification as a "branch of economics". The contrast is genuine although the Koopmans-Fisher notion had always been amphibious for,

while it lodged in economic theory, it was always associated with statistical inference actual or possible–Koopmans and Reiersøl published in the *Annals of Mathematical Statistics* after all.

The root idea of the statistical notion was R. A. Fisher's uninformativeness as variation freeness (§2.2) and the first of Rothenberg's theorems treats the non-vanishing of the Fisher information measure as a condition for local identifiability. Rothenberg (p. 577) remarks, "lack of identification is simply the lack of sufficient information to distinguish between alternative structures [parameter values]".

Rothenberg (p. 578) gives a pair of definitions for the parametric case. Here Y represents the n-dimensional outcome of some random experiment with density function f and A represents the m-dimensional parameter space:

> DEFINITION 1: Two parameter points (structures) α^1 and α^2 are said to be observationally equivalent if $f(y, \alpha^1) = f(y, \alpha^2)$ for all y in R^n.
>
> DEFINITION 2: A parameter point α^0 in A is said to be *identifiable* if there is no other α in A which is observationally equivalent.

These descend from definitions in Koopmans and Reiersøl (p. 169) but these had not reflected Koopmans's concerns with inference from population to structure and the role of prior information in making such inference possible. Rothenberg's α may be a structural parameter or a reduced form parameter. In terms of Koopmans's two steps, the "statistical identification problem" can be interpreted as encompassing both steps or just the first. Rothenberg (p. 577) opts for the encompassing possibility, describing the "identification problem" as concerned with the "possibility of drawing inferences from observed samples to an underlying theoretical structure."

7 Bayesian treatments of identification

The "solution" of the identification problem would be seen in the posterior distribution and so there was a very direct way of extending identification into Bayesian theory. Morales (1971, p. 20) deemed "identified" a model in which the posterior density is not flat. Rothenberg (1973, p. 158) also considered the possibility:

> From the Bayesian point of view, the posterior density function summarises all the prior and sample information that is available. If the posterior distribution for the structural parameter α is highly concentrated around its mean, then we are in an excellent position to distinguish between different values of α. ... Thus one is tempted to define identification in terms of the degree of concentration of the posterior density.

Rothenberg drew back because it becomes possible for a parameter to be identifiable "even though the data were completely irrelevant". So he used "estimable" for this concentration notion, leaving "unanswered the question of an appropriate Bayesian definition of identification".

Koopmans had established a body of interrelated propositions about identification in econometrics: lack of identification matters because inference about structural parameters is prevented or impeded; identification is about prior information making possible–or not–the transition from the reduced form to the structural form; theorems about when identification is achieved usually involve the rank of certain matrices. The Morales-Rothenberg proposal built on the first of these propositions but by picking different ones authors could extend identification into Bayesian theory in any number of ways–all would satisfy one criterion of naturalness, that when 'contracted' they correspond to classical identification. The next four

sections consider different Bayesian responses to identification. For accounts of the origins of Bayesian econometrics see Drèze and Richard (1983) and Qin (1996).

7.1 Broadening the concept of identification

It would be too much to say that Bayesian econometrics was solely a response to the identification problem but the problem was seen as one where the new approach could make a difference–see Drèze (1972, p. 8). Lindley's (1971, p. 46n) remark, "it might be noted that underidentifiability causes no real difficulty in the Bayesian approach" became a talisman for Bayesian econometrics. Raiffa and Schlaifer had shown how to treat the singular $X'X$ regression case when there is extra-sample information and the new Bayesian theory–unlike Jeffreys's–opened up the possibility of using uncertain prior knowledge to solve or alleviate the identification problem. From this new perspective Koopmans had entertained either certain prior knowledge (about the restrictions) or no prior knowledge (about the unrestricted coefficients). (There is a considerable literature considering the interaction of identification and impropriety of priors–e.g. Erickson (1989) and Gelfand and Sahu (1999)–but it would take too long to explore this.)

Jacques Drèze (b 1929) and Arnold Zellner (b 1927), the founding fathers of Bayesian econometrics, devised formal constructions extending the Koopmans notion of identification. They both emphasised how Bayesian analysis permits the introduction of uncertain prior information into the simultaneous equations model. Zellner (1971, p. 254) argues

> Usually in the sampling theory framework exact restrictions are imposed on the parameters of a model to achieve identification ... Since prior information in the Bayesian framework need not necessarily take the form of exact restrictions ... there is a need to broaden the concept of identification to allow for the more

general kind of prior information used in Bayesian analysis.

Drèze and Zellner worked on the simultaneous equations model but–following the example of Leamer (1978)–their approaches can be illustrated in the simpler regression framework. Suppose we take as the *basic scheme* the unrestricted structural form and corresponding reduced form the following regression specification

$$\begin{aligned} y &\sim N(X(\alpha_1 + \alpha_2), \sigma^2 I) \\ y &\sim N(X\pi, \sigma^2 I) : \pi = \alpha_1 + \alpha_2. \end{aligned}$$

To simplify matters further suppose that σ^2 is known. Koopmans assumes that there is no obstacle to knowing π. However, as neither α_1 or α_2 can be determined from knowledge of π, neither is identifiable. Consider imposing the a priori restriction $\alpha_2 = 0$. Now identification of α_1 is achieved because π determines α_1. This is an identifying restriction.

Drèze reproduces in Bayesian terms the two step inference procedure described by Koopmans. The first step is inference to the reduced form and the second is inference from the reduced form to the structural form. Drèze shows how this decomposition is valid in that the data is informative only about the reduced form parameters; the structural parameters and the data are independent given the reduced form parameters.

$$f(\alpha \,|\pi, y) = f(\alpha \,|\pi)$$

With the restriction, $\alpha_2 = 0$, a single α_1 is associated with any given π. Without it, infinitely many $\alpha_1's$ are associated with any given π. The identifying restriction could be interpreted as a prior probability distribution for α_2 concentrated on 0. A way of weakening this restriction

would be to take an uninformative prior for α_1 and choose an informative prior for α_2, say:

$$\alpha_2 \sim N(0, M).$$

To take advantage of this extension, Drèze introduces the notion of "identification in probability." Define the parameter α_2 as "identified in probability" when the conditional prior probability of α_2 given π is proper. In the present case

$$\alpha_1 | \pi \sim N(\pi, M)$$

Identification is the limiting case of identification in probability when the probability distribution is concentrated at a single point.

The "more general kind of prior information" envisaged by Zellner may take the form of an informative prior on α_2, e.g.

$$\alpha_2 \sim N(0, M)$$

then we can integrate α_2 out of the likelihood and obtain

$$y \sim N(X\alpha_1, \sigma^2(I + XMX'))$$

In this new distribution no distinct values of α_1 are observationally equivalent. Here is a "broadened" concept of observational equivalence on which to base a "broadened" concept of identification. These rest on a broadened notion of likelihood–a marginal likelihood reflecting both the prior for α_2 and the original likelihood.

These broadened concepts involve both the likelihood and prior–as they are meant to!.An immediate consequence is that identification in probability, say, may be lost if the prior is changed. Of course this characteristic was inherited from the ancestor, the Koopmans hierarchical specification of unrestricted model plus a priori restrictions. However the project of broadening the scope of identification seems to have been suddenly abandoned. Drèze (1975, p. 167) had been presenting identification in probability in lectures.

> Reactions to this presentation have led me to recognise that it was misleading to use the word "identification" in defining a property of the prior density for the parameters of unidentified models. I agree with Kadane's view "that identification is a property of the likelihood function and is the same whether considered classically or from the Bayesian approach."

Drèze's Damascene conversion was significant for the Louvain school of Bayesian econometrics followed him.

7.2 The role of identification in Bayesian theory

The "view" from J. B. Kadane's "The Role of Identification in Bayesian Theory" (1975, p. 175) that

> identification is a property of the likelihood function and is the same whether considered classically or from the Bayesian approach.

is the best remembered part of the paper that brought the statistical notion of identification into the Bayesian debate.

The "likelihood" is evidently the Jeffreys likelihood, not the likelihood kernel. The force of the remark is *first* that identification is *not* a property of the prior and *second* that the

definition of identification–Rothenberg's in §6.2–is meaningful in the Bayesian framework. The remark might also be taken, as by Hsiao (1983, p. 272), as a declaration that a Bayesian definition is unnecessary.

Kadane demonstrates the role of identification through theorems of relevance to Bayesians. These expand a remark he (p. 180n) attributes to Savage: "identification is properly a part of the study of the design of experiments". Kadane (pp. 184-189) applies the preposterior decision machinery of Raiffa and Schaifer, considering an experiment "valuable" if it can yield information leading to a better decision, i.e. if the minimum expected loss is reduced by performing the experiment (see §7.4 below for examples). His Theorem 4 is a "characterization of identified functions in terms of valuable experiments".

Zellner's (1971) comprehensive book on Bayesian econometric treats *one* preposterior topic–a multi-period control problem in which control also gives an opening for learning about the structure of the economy (Sections 11.5-6)–and econometricians might have concluded from Kadane's theorems that identification was none of their business as they rarely design experiments–see §7.4 below. However they produced variants of Kadane's results. Leamer (1978, p. 192) and Hsiao (p. 273) depart from the decision formulation and consider an experiment "informative" about θ if it *might change your opinions* about θ. A common feature of this work and most that has followed is the emphasis on the *change* from the prior to the posterior not, as in Morales (see §7.1), an emphasis on the posterior and its properties.

The various theorems differ in scope but agree in showing that the situations–the experiments–characterised by identification and informativeness are the same. When Kadane and Leamer describe their theorems they bring to the surface an issue running through this entire story: what is sameness? Kadane (p. 181) argues that "the dependence on the prior indicates that the concepts of identification and informativeness are different" while Leamer (p. 193) claims that

"the words identifiable ... and publicly informative ... are interchangeable", which presumably means that the corresponding concepts are the same. In §3.1 we met various definitions of sufficiency and Raiffa and Schaifer's (p. 34) demonstration of the "complete equivalence" of the Bayesian and classical definitions. The differences involved here seem philosophical rather than statistical. Different inference theories employ different terms and the terms of the characterisations may not make sense in other theories although the objects characterised are the same. (However the example of Wald (1939) in which a prior distribution has only a formal significance indicates that we must take care in interpreting "making sense".) The case of Jeffreys's co-option of likelihood suggests that one consideration underlying the judgement of similar enough to be considered the same is that enough significant propositions remain true when the change is made. There has been a great deal of discussion amongst historians and philosophers of science of the way concepts change–or keep–their meaning as the theory in which they are embedded changes. See Sankey (1994) provides a useful review. The literature is suggestive but it does not generate exploitable conclusions.

Returning to Kadane's "view", there is a second argument in his paper which provided backing of a different kind.

7.3 The duality between parameter and data

Kadane (p. 178) expounds the "analogy of the theory of identified functions to the theory of sufficient statistics". This idea has been developed in conjunction with the notion of a duality between parameter and data in Bayesian theory. This notion was discussed by several authors including Florens and Mouchart (1977), Picci (1977) and Dawid (1979); Dawid also refers to earlier literature. The most detailed and abstract treatment of identification on these lines is by Florens, Mouchart and Rolin (1990, §§4.5-6) but I will take as texts the more elementary

discussions in Gourieroux and Montfort (1989/95) and Dawid (1979).

On the Bayesian view parameters as well as observations are random variables. Gourieroux and Montfort (1989/95, pp. 102 and -8) present a pair of definitions–of sufficiency of a statistic "in the Bayesian sense" and of sufficiency of a parameter "in the Bayesian sense":

- A statistic S is sufficient in the Bayesian sense if θ and Y are conditionally independent given $S(Y)$, i.e. $\theta \perp Y \mid S(Y)$.
- A function $g(\theta)$ of the parameter is said to be sufficient in the Bayesian sense if $Y \perp \theta \mid g(\theta)$.

Next they define a function of the parameter as minimal sufficient if it is sufficient and a function of any other sufficient function. They (p. 108) then state a "property"–not a definition–relating identification (Rothenberg style) to minimal sufficiency: "θ is identified if and only if θ is minimal sufficient in the Bayesian sense."

Gourieroux and Montfort do *not* refer to "identification in the Bayesian sense" but their definitions and properties could be re-choreographed to produce such a notion. In his account of conditional independence in statistical theory Dawid (1979, p. 4) considers the distribution of X as determined by a pair of parameters (Θ, Φ). In the case when $X \perp \Phi \mid \Theta$, the parameter Φ is "redundant once Θ is known". In such a situation the "full parameter (Θ, Φ) is said not to be identified". Thus the full parameter is not identified when it contains redundant parameters. In the basic scheme of §7.1 the parameter vector (α, π) contains redundant parameters. However the redundancy idea is better applied reparametrising from α to (π, δ) where δ $(= \alpha_1 - \alpha_2)$ is the redundant part of α.

Using properties of conditional independence Dawid glides through definitions and results, thus:

> If Θ is a sufficient parameter, so that $X \perp \Phi \mid \Theta$, and the parameters have a

prior distribution , then $\Phi \perp X \mid \Theta$, so that $p(\phi \mid x, \theta) = p(\phi \mid \theta)$. We see that the conditional distribution for the redundant part Φ of the parameter, given the sufficient parameter Θ, is the same in the posterior distribution as in the prior: once we have learned about Θ from the data, we can learn nothing about Φ, over and above what we knew already.

The demonstration of the duality between parameters and data and the parallel between identified functions and sufficient statistics may have given identification a new naturalness in Bayesian theory. Of course Bayesians may not be equally impressed by this insight. For Florens, Mouchart and Rolin the theory of reduction is a major part of the subject while Lindley (1979, p. 16) is unimpressed by the "trick" of sufficiency, classifying the difficulties for Bayesian theory created by its absence as "numerical not inferential".

7.4 Informativeness of observations

Neath and Samaniego (1997, p. 226) emphasise their distance from Kadane's concern with learning from an *experiment* when they present some *outcome* level analysis pertaining to nonidentifiable models. However they do not investigate the possibility of *not* learning from outcomes. Drèze (1972, p. 7) had mentioned the possibility in his outline of the inferential use of Bayes' formula

> if $P(x \mid B_i)$ [the likelihood of the observation] is the same for all i, then $P(B_i \mid x)$ is equal to $P(B_i)$, and the observation is noninformative; when this property holds for all x, the B_i's are called "observationally equivalent".

However, he did not dwell on noninformative observations nor relate them to the identification discussion later in his paper.

The noninformative observation can provide the basis for observational equivalence but it has no role in unconditional classical theory. Yet it is a prima facie interesting concept to anyone who accepts the likelihood principle and wants to do terminal analysis (§5.1). Noninformativeness is to outcomes what Fisher's statistic supplying no information about a parameter is to experiments (see §2.2).

In the paradigm cases of the identification problem in the regression and simultaneous equations models *all* the points in the sample space are noninformative if any is–the same is true for Neath and Samaniego's binomial model with parameter $p = p_1 + p_2$ and Fisher's genetic model. Yet it is clearly possible for an identified experiment to generate *some* noninformative outcomes. I have two biassed coins: for one the probability of heads is 0.8; for the other 0.2. I have forgotten which is which and contemplate experimenting to help me decide. One experiment would be to toss the–arbitrarily determined–'first' coin twice, record the total number of heads, Y, and base the decision on that number. Another would be to toss the first coin and then the second and record the total number of heads and base the decision on that number. The parameter space comprises the two possible identities of the'first' coin. Using notation based on that of §6.2, with $f(0, 0.8)$ as the probability that $Y = 0$ given that the first coin is the 0.8 coin, the densities for the first experiment are

$$\begin{aligned} f(0, 0.8) &= 0.04 : f(1, 0.8) = 0.32 : f(2, 0.8) = 0.64 \\ f(0, 0.2) &= 0.64 : f(1, 0.2) = 0.32 : f(2, 0.2) = 0.04 \end{aligned}$$

and for the second

$$\begin{aligned} f(0, 0.8) &= 0.16 : f(1, 0.8) = 0.68 : f(2, 0.8) = 0.16 \\ f(0, 0.2) &= 0.16 : f(1, 0.2) = 0.68 : f(2, 0.2) = 0.16. \end{aligned}$$

The first experiment is valuable in Kadane's sense (assuming a reasonable loss function and nondogmatic prior) and can help me decide which coin is which but the second experiment is worthless. The first coin is identifiable from the first experiment but not from the second. Consider now the observation $Y = 1$ which is noninformative in both experiments. According to the likelihood principle, it does not matter whether this observation has come from the informative, valuable experiment or from the uninformative, valueless one. The division of experiments into identified ones and unidentified ones is not essential for terminal analysis; the essential division is into informative and noninformative observations.

The special uniform distribution discussed by Barndorff-Nielsen and Cox (1994, pp. 34-5) provides another example where an informative outcome may fail to be realised. Consider a random sample of size 2 from a uniform distribution on $(\theta - 1, \theta + 1)$ with $\theta \in \{\theta^1, \theta^2\}$. Let $Y_{(1)}$ and $Y_{(2)}$ be the order statistics and $\overline{Y}$ the mean and R the range $Y_{(2)} - Y_{(1)}$. Given data $y_{(1)}$ and $y_{(2)}$, the likelihood is flat on the interval $(\overline{y} - 1 + \frac{1}{2}r,\ \overline{y} + 1 - \frac{1}{2}r)$ of the parameter space. If θ^1 and θ^2 are both in this interval the observation is noninformative.

Poirier (1998) has written about noninformative observations under the rubric "uninformative data". He envisages a nonidentified model where part of the parameter ψ is identified and part λ is unidentified. He (p. 485) defines

> The data y are marginally uninformative for λ iff $f(\lambda \mid y) = f(\lambda)$. The data y are conditionally uninformative iff $f(\lambda \mid \psi, y) = f(\lambda \mid \psi)$

One could imagine analogous concepts at the experiment level.

Poirier's (pp. 485-6) results are on the lines of the sentences *following* Lindley's (1971, p. 46n) celebrated remark (his θ_1 corresponds to λ and the "remaining parameters" to ψ):

> In passing it might be noted that underidentifiability causes no real difficulty in the Bayesian approach. If the likelihood does not involve a particular parameter, θ_1 say, when written in the natural form, then the conditional distribution of θ_1, given the remaining parameters, will be the same before and after the data. This will not typically be true of the marginal distribution of θ_1 because of the changes in assessment of the other parameters caused by the data, though if θ_1 is independent of them, it will be. For example, unidentifiable (or *unestimable*) parameters in linear least squares theory are like θ_1 and do not appear in the likelihood.

Bauwens, Lubrano and Richard (1999, p. 42) consider the same set-up with identified and unidentified sub-parameters and consider whether the marginal and conditional priors are revised by the sample. These results have a family resemblance to those of Dawid discussed in §7.3 as Gelfand and Sahu (1999, p. 248) point out. However only Poirier seems to be working at the outcome level.

The tags, "identification ... is the same whether considered classically or from the Bayesian approach" and "underidentifiability causes no real difficulty in the Bayesian approach", sit together uneasily even if they are not precisely contradictory. Poirier (p. 483) reducers the tension by showing that in a nonidentified model there is "no Bayesian free lunch ... there exist quantities about which the data are [marginally] uninformative." For Morales, say (§7.1), the price of the "free lunch" was obtaining the prior information. Value-added accounting

focusses on the properties of the transformation of prior to posterior, not on the properties of the posterior.

8 Does Bayesian theory require a different definition?

By way of summary, consider the generalised Hsiao question from the Introduction, "whether Bayesian theory requires a different definition of — from the classical one" and the received answers–received *today*, that is, for these are matters of judgement not of necessity. Of course the fact that the question can be posed reflects the continuing subordinate status of Bayesian theory.

For likelihood, the received answer is tacitly *no*, for the differences are not considered material: allowing for the altered status of parameters the definitions are the same (see §3.1 and §5.1). For sufficiency there is no received answer: the answer will depend on how "required" and "different" are interpreted. Although the original Fisher definition has not been used by Bayesian writers, it could be used without creating scandal; it would have to be linked to the scene of its application by a chain of equivalences: equivalences in the sense of having the same extension. It is not a natural definition for it points to features of the extension which are irrelevant for Bayesian analysis. Jeffreys used the factorisation definition which requires no–or at least a shorter–chain but he and Raiffa and Schlaiffer saw merit in a more explicitly Bayesian definition; this may have the same extension as the classical definition but it employs terms with no place in classical theory. The nature of the "requirement" is to facilitate the development of other concepts. (see §3.1).

In Bayesian econometrics today the identification question is routinely asked and answered *no* (see §§7.2-3). Thus Piorier (1998) and Bauwens, Lubrano and Richard (1999) endorse

the Kadane declaration while at the same time recalling the econometric past. That past contained radical alternatives to classical identification–so radical that it is misleading to speak of different definitions of the same thing. In the general statistics literature there is not the same burden of history, only the divergence seen with sufficiency. Neath and Samaniego (1997) give the classical definition and get on with things. Gelfand and Sahu (1999) give a Bayesian definition of identification and get on with things; their definition is based on Dawid (1979) which relates to the classical definition the way that Raiffa and Schlaifer's definition of sufficiency relates to the classical definition. (see §7.3).

Identification seems to be secure in Bayesian econometrics and the project of broadening the concept of identification has lapsed but in a sense identification has bifurcated with a second branch based on the informativeness of observations–see §7.4. Koopmans's original notion of identifiability in relation to "hypothetical exact knowledge" appears to be of no current interest.

References

Aldrich, J. (1994) Haavelmo's Identification Theory, *Econometric Theory*, **10**, 198-219.

_____ (1997) R. A. Fisher and the Making of Maximum Likelihood 1912-22, *Statistical Science*, **12**, 162-176.

_____ (1999) Determinacy in the Linear Model: Gauss to Bose and Koopmans, *International Statistical Review*, **67**, 211-219.

Barnard, G. A. (1949) Statistical Inference (with discussion) *Journal of the Royal Statistical Society B*, **11**, 115-149.

Barndorff-Nielsen, O. and D. R. Cox (1994) *Inference and Asymptotics*, London: Chapman and Hall.

Bauwens, L., M. Lubrano and J.-F. Richard (1999) *Bayesian Inference in Dynamic Econometric Models*, Oxford: Oxford University Press.

Bayarri, M. J., M. H. DeGroot and J. B. Kadane (1988) What is the Likelihood Function, in S. Gupta (ed) *Statistical Decision Theory and Related Topics, IV, volume 1*, New York: Springer. (1993).

Berger, J. O. and R. L. Wolpert (1984) *The Likelihood Principle*. Volume 6 in the Institute of Mathematical Statistics Lecture Notes Series. Hayward, CA: IMS.

Birnbaum, A. (1962) On the Foundations of Statistical Inference, *Journal of the American Statistical Association*, **57**, 269-306.

Box, G. E. P. and G. C. Tiao (1973) *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Box, J. F. (1978) *R. A. Fisher: The Life of a Scientist*, New York, Wiley.

Brunt, D. (1917) *The Combination of Observations*. Cambridge University Press, Cambridge.

Cook, A. (1991) Sir Harold Jeffreys, *Biographical Memoirs of Fellows of the Royal Society*, **37**, 303-331.

Dale, A. I. (1991) *A History of Inverse Probability from Thomas Bayes to Karl Pearson*, New York: Springer-Verlag.

David, H. A. (2001) First (?) Occurrence of Common Terms in Statistics and Probability, Appendix B and pp. 219-228 of H. A.David & A. W. F. Edwards (ed) *Annotated Readings in the History of Statistics*, Springer New York.

Dawid, A. P. (1979) Conditional Independence in Statistical Theory, (with discussion) *Journal of the Royal Statistical Society B,* **41***, 1-31.*

Drèze, J. (1972) Econometrics and Decision Theory, *Econometrica*, **40**, 1-17.

_____ (1975) Bayesian Theory of Identification in Simultaneous Equations Models. In S.

E. Fienberg and A. Zellner (eds) *Studies in Bayesian Econometrics and Statistics*, Chapter 5.1 and pp. 159-174. Amsterdam: North-Holland.

_____ and J.-F. Richard (1983) Bayesian Analysis of the Simultaneous Equations Model. In Z. Griliches and M. Intriligator (eds) *Handbook of Econometrics, Vol. I*, Chapter 9 and pp. 517-598. Amsterdam: North-Holland.

Edwards, A. W. F. (1972) *Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference*, Cambridge: Cambridge University Press.

Erickson, T. (1989) Proper Posteriors from Improper Priors for an Unidentified Errors-in-Variables Model, *Econometrica*, **57**, 1299-1316.

Fisher, F. M. (1966) *The Identification Problem in Econometrics*. Huntingdon, New York: Krieger, 1977. (Reprint with new preface of book originally published in 1966 by McGraw-Hill.)

Fisher, R. A. (1921) On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample, *Metron*, **1**, 3-32.

_____ (1922) On the Mathematical Foundations of Theoretical Statistics, *Philosophical Transactions of the Royal Society, A*, **222**, 309-368.

_____ (1925) *Statistical Methods for Research Workers*, second edition in 1928, Edinburgh: Oliver & Boyd.

_____ (1932) Inverse Probability and the Use of Likelihood, *Proceedings of the Cambridge Philosophical Society*, **28**, 257-261.

_____ (1935) The Logic of Inductive Inference, (with discussion) *Journal of the Royal Statistical Society, **98**, 39-82.*

Florens, J.-P. and M. Mouchart (1977) Reduction of Bayesian Experiments, CORE DP No. 7737.

Florens, J.-P, M. Mouchart and J-M. Rolin (1990) *Elements of Bayesian Statistics*, New York: Dekker.

Gauss, C. F. (1809) *Theory of the Motion of Heavenly Bodies Moving around the Sun in Conic Sections*, English translation by C. H. Davis of Latin publication, reprinted 1963, Dover, New York.

Gelfand, A. E. and S. K. Sahu (1999) Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models, *Journal of the American Statistical Association,* **94**, 247–253.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (1995) *Bayesian Data Analysis*, London, Chapman & Hall.

Gourieroux, C. and A. Montfort (1989/95) *Statistique et Modèles Économétriques*, vol. 1, Paris, Economica. References are to English translation by Q. Vuong (1995) as *Statistics and Econometric Models*, Cambridge: University Press.

Haavelmo, T. (1944) The Probability Approach in Econometrics. Supplement to *Econometrica*, **12**, i-viii, 1-118.

Hald, A. (1998) *A History of Probability and Statistics 1750-1930*, New York: Wiley.

Hsiao, C. (1983) Identification. In Z. Griliches and M. Intriligator (eds) *Handbook of Econometrics, Vol. I,* Chapter 4 and pp. 224-283. Amsterdam: North-Holland.

Jeffreys, H. (1935) Contribution to the discussion of Fisher (1935), *Journal of the Royal Statistical Society,* **98**, 70-72.

_____ (1939) *Theory of Probability*, with a second edition in 1948 and a third in 1961, Oxford: University Press.

Kadane, J. B. (1975) The Role of Identification in Bayesian Theory. In S. E. Fienberg and A. Zellner (eds) *Studies in Bayesian Econometrics and Statistics*, Chapter 5.2 and pp. 175-191. Amsterdam: North-Holland.

Kolmogorov, A. N. and S. V. Fomin (1970) *Introductory Real Analysis*, New York: Dover.

Koopmans, T. C. (1949) Identification Problems in Economic Model Construction. *Econometrica*, **17**, 125-144.

_____ and O. Reiersøl (1950) The Identification of Structural Characteristics, *Annals of Mathematical Statistics*, **21**, 165-181.

_____ , H. Rubin and R. B. Leipnik (1950) Measuring the Equation Systems of Dynamic Economics. In T. C. Koopmans (ed) *Statistical Inference in Dynamic Economic Models*, Chapter 2 and pp. 52-237. New York: John Wiley.

Leamer, E. E. (1978) *Specification Searches*, New York: John Wiley.

Lindley, D. V. (1971) *Bayesian Statistics: A Review*, Philadelphia: SIAM.

_____ (1979) Discussion of Dawid (1979), *Journal of the Royal Statistical Society B, **41**, 15-16.*

_____ (1980) L. J. Savage–His Work in Probability and Statistics, *Annals of Statistics*, **8**, 1-24.

_____ (1986) On Re-reading Jeffreys, pp. 35-46 of I. S. Francis et al (eds) *Pacific Statistical Congress*, New York: Elsevier.

Malinvaud, E. (1966) *Statistical Methods of Econometrics*, Amsterdam: North-Holland.

Morales, J. A. (1971) *Bayesian Full Information Structural Analysis*, Berlin: Springer.

Neath, A. A. and F. J. Samaniego (1997) On the Efficacy of Bayesian Inference for Nonidentifiable Models, *American Statistician*, **51**, 225-232.

Neyman, J. (1935) Contribution to the discussion of Fisher (1935), *Journal of the Royal Statistical Society, **98**, 73-76.*

_____ (1937) Outline of a Theory of Statistical Estimation based on the Classical Theory of Probability, *Philosophical Transactions of the Royal Society*, **236**, 333-380.

Neyman, J. and E. S. Pearson (1928) On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference, *Biometrika*, **20**, 175-240 and 263-294.

Pearson, K. (1892) *The Grammar of Science*, London: Walter Scott.

_____ (1894) Contributions to the Mathematical Theory of Evolution, *Philosophical Transactions of the Royal Society A*, **185**, 71-110.

_____ (1907) On the Influence of Past Experience on Future Expectation, *Philosophical Magazine*, **13**, 365-378.

Picci, G. (1977) Some Connections between the Theory of Sufficient Statistics and the Identifiability Problem, *SIAM Journal of Applied Mathematics*, **33**, 383-398.

Poirier, D. (1988) Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics (with discussion), *Journal of Economic Perspectives*, **2**, 120-170.

_____ (1998) Revising Beliefs in Nonidentified Models, *Econometric Theory*, **14**, 483-509

Qin, D. (1989) Normalisation of Identification Theory, *Oxford Economic Papers*, **41**, 73-93.

_____ (1996) Bayesian Econometrics: the First Twenty Years, *Econometric Theory*, **12**, 500-516.

Raiffa, H. A. and R. Schlaifer (1961) *Applied Statistical Decision Theory*, Boston: Graduate School of Business Administration, Harvard University.

Rayner, J. and J. Aldrich (1997) Koopmans after Johansen: Identifiability in the Simultaneous Equations Model, paper given at the 1997 Australasian meeting the Econometric Society.

Reid, C. (1982) *Neyman–from Life*, New York: Springer-Verlag.

Rothenberg, T. J. (1971) Identification in Parametric Models, *Econometrica*, **39**, 577-91.

_____ (1973) *Efficient Estimation with A Priori Information*, New Haven: Yale University Press.

Sankey, H. (1994) *The Incommensurability Thesis*, Aldershot, Hants: Avebury.

Savage, L. J. (1954/72) *The Foundations of Statistics*. References are to the expanded second edition published by Dover (New York). This preserved the pagination of the first edition published by Wiley (New York).

_____ (1960) Recent Tendencies in the Foundations of Statistics, pp. 550-554 of *Proceedings of the 8th International Congress of Mathematicians*, Cambridge: Cambridge University Press.

_____ (1962) Subjective Probability and Statistical Practice, in L. J. Savage et al, *The Foundations of Statistical Inference*, London: Methuen.

_____ (1962a) Discussion of Birnbaum (1962), *Journal of the American Statistical Association*, **57**, 307-308.

Schlaifer, R. (1959) *Probability and Statistics for Business Decisions*, New York: McGraw-Hill.

Soper, H. E., A. W. Young, B. M. Cave, A. Lee and K. Pearson (1917) On the Distribution of the Correlation Coefficient in Small Samples. Appendix II to the Papers of 'Student' and R. A. Fisher. A Cooperative Study, *Biometrika*, **11**, 328-413.

Wald, A. (1939) Contributions to the Theory of Statistical Estimation and Testing Hypotheses, *Annals of Mathematical Statistics*, **10**, 299-326.

_____ (1950) *Statistical Decision Functions*, New York: Wiley.

Wolfowitz, J. (1952) Abraham Wald, 1902-1950, *Annals of Mathematical Statistics*, **23**, 1-13.

Wrinch, D. and H. Jeffreys (1921) On Certain Fundamental Principles of Scientific Inquiry, *Philosophical Magazine*, **42**, 369-390.

Zellner, A. (1971) *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley.