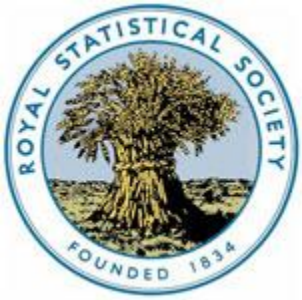

Conditional Independence in Statistical Theory
Author(s): A. P. Dawid
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 41, No. 1 (1979), pp. 1-31
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2984718
Accessed: 23-08-2017 20:21 UTC

J. R. Statist. Soc. B (1979),
41, *No.* 1, *pp.* 1–31

Conditional Independence in Statistical Theory

By A. P. DAWID†

University College London

[Read before the ROYAL STATISTICAL SOCIETY at a meeting organized by the RESEARCH SECTION on Wednesday, October 18th, 1978, the Chairman Professor J. F. C. KINGMAN in the Chair]

SUMMARY

Some simple heuristic properties of conditional independence are shown to form a conceptual framework for much of the theory of statistical inference. This framework is illustrated by an examination of the rôle of conditional independence in several diverse areas of the field of statistics. Topics covered include sufficiency and ancillarity, parameter identification, causal inference, prediction sufficiency, data selection mechanisms, invariant statistical models and a subjectivist approach to model-building.

Keywords: INDEPENDENCE; CONDITIONAL INDEPENDENCE; MARKOV CHAINS; SUFFICIENCY; ANCILLARITY; IDENTIFICATION; SIMPSON'S PARADOX; INVARIANCE; BAYESIAN INFERENCE; PREDICTION; ADEQUACY; TRANSITIVITY; TOTAL SUFFICIENCY; DATA SELECTION; OPTIONAL STOPPING

1. INTRODUCTION

INDEPENDENCE and conditional independence are familiar concepts of probability theory, where they form the basis of several areas of study, such as limit theorems and Markov chains. It is the aim of this paper to show that independence and (more particularly) conditional independence are equally fundamental in the theory of statistical inference.

It transpires that many of the important concepts of statistics (sufficiency, ancillarity, etc.) can be regarded as expressions of conditional independence, and that many results and theorems concerning these concepts are just applications of some simple general properties of conditional independence. By taking conditional independence as basic, and expressing other properties in terms of it, we achieve a unification of many areas of statistics which appear, at first sight, to be entirely unrelated. This unification in its turn promotes cross-fertilization: when two different areas of study are found to be isomorphic, known results in one area immediately become available for use in the other. Thus I would claim that, rather than just being another useful tool in the statistician's kitbag, conditional independence offers a new language for the expression of statistical concepts and a framework for their study.

The above claim is a big one, and the reader must determine for himself whether he feels it is substantiated by this paper. My plan has been to take several areas of statistics, express them in terms of conditional independence and examine the consequences of this point of view. Most (though not all) of the results are well known, but have not before been considered from this angle. (Notable exceptions are Hall *et al.*, 1965; Pitman and Speed, 1973; and to some extent Lauritzen, 1974; who have made fundamental use of conditional independence in their work.) Some of the areas studied may seem entirely unconnected with each other: this goes to illustrate the power and scope of the underlying structure.

One theme which runs through much of the material here is the simple way in which Bayesian results can be "read off" as consequences of sampling theory properties, as soon as these have been expressed in terms of conditional independence. The theory is particularly useful for a Bayesian, as in it the distinction between data and parameters is largely irrelevant. Nevertheless, the theory also has important applications outside Bayesian inference, and must

† Present address: Department of Mathematics, The City University, Northampton Square, London EC1V OHB.

be regarded as philosophically neutral in the arguments between the various competitive viewpoints on inference.

No attempt has been made in this paper to give rigorous statements or proofs. A general treatment of conditional independence, in terms broad enough to cover all the applications considered here, seems to require some specialized mathematics: details may be found in Dawid (1980). However, the principal notions are easy to express and work with heuristically. It is my hope that others will find these heuristics as fruitful in their own investigations as I have done.

Sections 2, 3 and 4 below introduce and develop some basic general properties o conditional independence. The next five sections can be read largely independently of one another. Section 5 treats the use of covariates in causal inference; Section 6 considers problems of statistical prediction; Section 7 deals with processes which bias the observation of data; Section 8 provides a unifying treatment of various results on invariance; and Section 9 presents the rudiments of a subjectivist theory of model-building.

2. INDEPENDENCE

2.1. *Definitions*

Let X and Y be random variables. We denote by $p(x,y)$, $p(x)$ and $p(x|y)$, respectively, the joint density of (X, Y), the marginal density of X, and the conditional density of X given $Y=y$; and so on.

We write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent. The usual mathematical expression of this property, in terms of densities, is

(Ia) $$p(x,y) = p(x)\,p(y).$$

However, while this factorization forms the basis of further development of the probability theory (for example extensions to several variables), it is itself a derivative of a more basic, though mathematically equivalent, formulation. Intuitively, $X \perp\!\!\!\perp Y$ if any information received about Y does not alter uncertainty about X; that is

(IIa) $$p(x|y) = p(x).$$

There are other, seemingly weaker but in fact equivalent, formulations. Thus instead of (Ia) we may require

(Ib) $$p(x,y) = a(x)\,b(y),$$

expressing a factorization of $p(x,y)$, but not insisting that the factors be the marginal densities.

In place of (IIa) we may similarly substitute

(IIb) $$p(x|y) = a(x),$$

expressing the fact that the distributions of X given Y do not depend on the value at which the conditioning variable is fixed, but not insisting that the common conditional distribution should be the marginal distribution of X.

It is a matter of trivial mathematics to verify the equivalence of the above four formulations of the property $X \perp\!\!\!\perp Y$. Nevertheless, this equivalence is, to be at any rate, somewhat surprising. The intuitive notion captured by (II)(a or b) appears non-symmetrical, treating X and Y on quite different footings. However, (I) shows that the property really is symmetrical, after all. Thus, if we take (II)(a or b) as the definition of $X \perp\!\!\!\perp Y$, the following result is by no means vacuous.

Theorem 2.1. If $X \perp\!\!\!\perp Y$, then $Y \perp\!\!\!\perp X$.

An example of the use of Theorem 2.1 arises in retrospective studies. Let X denote exposure ($X=1$) or non-exposure ($X=0$) to some possibly carcinogenic agent, and let Y indicate

development ($Y = 1$) or non-development ($Y = 0$) of a certain kind of cancer. It may be important to investigate whether or not $Y \perp\!\!\!\perp X$ in the sense of (IIb): that is, whether or not the incidence of cancer is the same in those who were, and in those who were not, exposed to the agent. It is common to collect two groups of subjects, one in which the subjects developed the cancer ($Y = 1$) and a control group of those who did not ($Y = 0$), and investigate whether or not the proportion exposed to the agent is the same in both groups. Thus the data throw light directly on whether or not $X \perp\!\!\!\perp Y$, and hence, by means of Theorem 2.1, indirectly on whether or not $Y \perp\!\!\!\perp X$.

2.2. *Introduction of Parameters*

One advantage of expressing $X \perp\!\!\!\perp Y$ in terms of (IIb) is that the concept it expresses extends to cases where the variables do not have a joint probability distribution. For example, suppose we have a statistical model in which the distributions of the data X are determined by the value of a parameter Θ, and let S be a function of X. Then, given only this structure, it makes sense to talk about the conditional distribution of S, given $\Theta = \theta$, but not so of the joint distribution of S and Θ, nor of the marginal distribution of S. We can still write $S \perp\!\!\!\perp \Theta$, if we interpret this in the sense of (IIb), although the other characterizations are now meaningless. With this understanding, $S \perp\!\!\!\perp \Theta$ expresses the fact that the sampling distribution of S is the same for all values of Θ; that is, in the usual statistical terminology, S is an *ancillary statistic*.

We cannot apply Theorem 2.1 in this case, since it depends on the existence of a joint distribution. However, if we add further structure, in the form of a (prior) distribution for Θ, such a joint distribution is created, and it therefore follows that $\Theta \perp\!\!\!\perp S$, for any ancillary statistic S. That is, by (IIa), $p(\theta|s) = p(\theta)$, or, the distribution of Θ posterior to observing S is unchanged from that in the prior distribution. This is a well-known result in Bayesian statistics: the above proof is perhaps marginally more trivial than any other.

Even without taking a Bayesian viewpoint, one could interpret $\Theta \perp\!\!\!\perp S$ as saying that S provides no information about Θ, thus turning Theorem 2.1 into a *principle of inference*: no information about Θ can be extracted from observing an ancillary statistic. This is very close to being a statement of the *conditionality principle* (Birnbaum, 1962).

3. CONDITIONAL INDEPENDENCE

3.1. *Definitions*

We now introduce a further variable Z, and write $X \perp\!\!\!\perp Y|Z$ to denote that X and Y are independent in their joint distribution given $Z = z$, for any value of z. Once again, this property has several equivalent expressions in terms of density functions:

(1a) $p(x, y|z) = p(x|z)\,p(y|z)$; (2a) $p(x|y, z) = p(x|z)$;

(1b) $p(x, y|z) = a(x, z)\,b(y, z)$; (2b) $p(x|y, z) = a(x, z)$.

A caution is called for here concerning the use (for example in pseudo-Bayesian theory) of *improper distributions* for random variables. It is shown in Dawid *et al.* (1973) that, in such circumstances, it is possible for (2b) to hold and, at the same time, for (2a) to fail; this is referred to as the *marginalization paradox*. However, such behaviour cannot occur when only proper distributions are admitted, which we henceforth assume. The fact that (2b) implies (2a) gives a useful trick for finding conditional distributions: see, for example, Dempster (1969, Theorem 13.4.2) and Dawid (1977b, Theorem 3).

To my mind, the intuitive (and non-symmetrical) content of $X \perp\!\!\!\perp Y|Z$ is best captured by (2b), which says that the conditional distribution of X, given Y and Z, is in fact completely determined by the value of Z alone, Y being superfluous once Z is given.

In parallel with Theorem 2.1, and equally non-vacuous for the interpretation (2)(a or b), we have

Theorem 3.1. If $X \perp\!\!\!\perp Y | Z$, then $Y \perp\!\!\!\perp X | Z$.

As a possible application of this result, consider the problem of operating a "fair" procedure for the selection of minority group members for university admission (Cole, 1973; Bickel *et al.*, 1977). One solution is to require that the probability of such selection should depend only on the academic promise of the candidates, and not on race, sex and so on. Let X denote selection ($X = 1$) or rejection ($X = 0$), let Y denote (say) sex, and let Z be a test-score regarded as a good assessment of academic promise. It is intended that $X \perp\!\!\!\perp Y | Z$. This could be monitored by checking whether or not $Y \perp\!\!\!\perp X | Z$: that is to say by taking two test-groups, one of successful and another of unsuccessful candidates, and looking to see whether the proportions of those getting any particular Z-score who are male are the same in both groups. (Further aspects of this problem are discussed in Section 5 below.)

Once again, it is not necessary for the concept of conditional independence that X, Y and Z have a joint distribution. Any of the expressions of $X \perp\!\!\!\perp Y | \Theta$ makes sense, where X and Y are random variables with a joint distribution governed by the value of the parameter Θ.

3.2. *Sufficiency*

Likewise, if Θ governs the distribution of X, and T is a function of X, the property $X \perp\!\!\!\perp \Theta | T$ is meaningful in the sense of (2b), and expresses the fact that the conditional sampling distributions of X given T are the same for all values of the parameter: that is, T is a *sufficient statistic*. Application of Theorem 3.1 to the case where Θ is given a prior distribution produces another important result in Bayesian inference: $\Theta \perp\!\!\!\perp X | T$ says that $p(\theta | x) = p(\theta | t)$, that is, the posterior distribution based on the full data is the same as that based only on a sufficient statistic.

For a non-Bayesian, $\Theta \perp\!\!\!\perp X | T$ encapsulates the *sufficiency principle*: no information about Θ is contained in X over and above that contained in T.

3.3. *Identification*

Now let the distribution of X be determined by a pair of parameters (Θ, Φ). Then we can interpret $X \perp\!\!\!\perp \Phi | \Theta$ in the sense of (2b). This says that the distributions of X are in fact completely determined by the value of Θ, Φ being redundant once Θ is known. A familiar example is given by the estimable functions in the normal linear model of less than full rank: $\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. Two different values of $\boldsymbol{\beta}$ can yield the same value of $\boldsymbol{\Gamma}\boldsymbol{\beta}$, and hence the same distribution for $\mathbf{X}$. We can take $\Theta = (\boldsymbol{\Gamma}\boldsymbol{\beta}, \sigma^2)$, and Φ any set of parameters which, together with Θ, determine $\boldsymbol{\beta}$: for example $\Phi = (\mathbf{I} - \boldsymbol{\Gamma}^{-}\boldsymbol{\Gamma})\boldsymbol{\beta}$, where $\boldsymbol{\Gamma}^{-}$ is a fixed g-inverse of $\boldsymbol{\Gamma}$; or we could even take Φ to be the full parameter $(\boldsymbol{\beta}, \sigma^2)$. Other important examples occur in econometric simultaneous equation models (Drèze, 1974) and the linear structural relationship (Reiersøl, 1950) under normality.

In the above situation, the full parameter (Θ, Φ) is said not to be identified. Barankin (1961) terms Θ a *sufficient parameter*. The concepts of sufficient statistic and sufficient parameter are, in fact, very closely related: both have similar expressions in terms of conditional independence (Picci, 1977). This relationship is just one aspect of a fuller duality that exists between data and parameter in a statistical model; see, for example, Kudō (1967) and Florens and Mouchart (1977).

If Θ is a sufficient parameter, so that $X \perp\!\!\!\perp \Phi | \Theta$, and the parameters have a prior distribution, then $\Phi \perp\!\!\!\perp X | \Theta$, so that $p(\phi | x, \theta) = p(\phi | \theta)$. 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 more about Φ, over and above what we knew already. This result, although easy to prove and long a part of the folk-lore of Bayesian statistics, is hard to trace in the literature. Drèze (1974) gives it in the particular context of simultaneous equation models. Theorem 5 of Kadane (1974) comes very close to this result.

4. Elementary Properties

We have so far needed only the symmetry property of conditional independence. Some further simple general results are the following.

Lemma 4.1. $X \perp\!\!\!\perp Y \mid Z$ if and only if $(X,Z) \perp\!\!\!\perp (Y,Z) \mid Z$.

Lemma 4.2. If $X \perp\!\!\!\perp Y \mid Z$, and U is a function of X, then
(i) $U \perp\!\!\!\perp Y \mid Z$ and (ii) $X \perp\!\!\!\perp Y \mid (Z, U)$.

Lemma 4.3. If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y,Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.

Important special cases of Lemmas 4.2 and 4.3 arise when the conditioning on Z is absent. Note that the converse of Lemma 4.3 holds, by Lemma 4.2 and the symmetry property.

We can define the joint independence of n random variables $X_1, \ldots, X_n$, denoted $\otimes_{i=1}^n X_i$, inductively as follows:

$$\mathop{\perp\!\!\!\perp}_{i=1}^{2} X_i \Leftrightarrow X_1 \perp\!\!\!\perp X_2, \quad \text{and} \quad \mathop{\perp\!\!\!\perp}_{i=1}^{n} X_i \Leftrightarrow \left(\mathop{\perp\!\!\!\perp}_{i=1}^{n-1} X_i \text{ and } (X_1, \ldots, X_{n-1}) \perp\!\!\!\perp X_n \right) \quad (n > 2).$$

It is an easy exercise to show, using the properties already established, that this property does not depend on the ordering of the X's.

Many derivations in probability theory and statistics may be conveniently stated and performed using the notation and general results of conditional independence. As an example, suppose (X_t) is a discrete-time Markov chain. Denote by $Y_t = (X_t, X_{t+1}, \ldots)$ the *future* at time t, and by $Z_t = (\ldots, X_{t-1}, X_t)$ the *past* at time t. (It is convenient, but not necessary, to consider the present state X_t as part of both the past and the future). Then the Markov property is expressible in terms of conditional independence: for all t, $Y_t \perp\!\!\!\perp Z_t \mid X_t$.

In particular, $(X_t, X_{t+1}) \perp\!\!\!\perp Z_{t-1} \mid X_{t-1}$, by Lemma 4.2 (i), and so, by Lemma 4.2 again,

$$X_t \perp\!\!\!\perp Z_{t-1} \mid (X_{t-1}, X_{t+1}). \tag{4.1}$$

Also, $(Z_{t-1}, X_t) \perp\!\!\!\perp Y_{t+1} \mid X_{t+1}$, so that

$$X_t \perp\!\!\!\perp Y_{t+1} \mid (X_{t-1}, X_{t+1}, Z_{t-1}). \tag{4.2}$$

Combining (4.1) and (4.2) by means of Lemma 4.3 yields

$$X_t \perp\!\!\!\perp (Z_{t-1}, Y_{t+1}) \mid (X_{t-1}, X_{t+1}). \tag{4.3}$$

Thus we have demonstrated the *nearest-neighbour property* of a Markov chain: given the states at all times other than t, the conditional distribution of X_t is in fact determined by the states at times $(t-1)$ and $(t+1)$ only.

Numerous further examples of the application of these simple properties of conditional independence will be found throughout the rest of this paper.

5. Causal Inference and Simpson's Paradox

Simpson's paradox (Simpson, 1951) has been much discussed (see, for example, Blyth, 1972; Lindley and Novick, 1979) but continues to trap the unwary. Conditional independence proves to be helpful in resolving the paradox and in setting up meaningful null hypotheses for testing causal effects.

Consider a population of experimental units. Several treatments, labelled by X, are available for application to any unit, and interest revolves around their differential effects on a response variable Y. Also available is information on a set of covariates, denoted by Z. In its simplest form, Simpson's paradox arises because it is possible to have $Y \perp\!\!\!\perp X \mid Z$ without $Y \perp\!\!\!\perp X$, and conversely. For example, with X taking values T (treatment) or $\overline{T}$

(control), Y taking values R (recovery) or $\bar{R}$ (no recovery), and Z indexing sex, the data of Table 1 seem to exhibit a beneficial effect of treatment for either sex, which disappears when the table is collapsed over sex; the opposite effect occurs with Table 2.

TABLE 1

	Male		*Female*	
	R	$\bar{R}$	R	$\bar{R}$
T	100	50	100	150
$\bar{T}$	160	120	20	60

TABLE 2

	Male		*Female*	
	R	$\bar{R}$	R	$\bar{R}$
T	20	40	100	50
$\bar{T}$	20	40	20	10

It is not difficult to fabricate examples involving several covariates, where the treatment effect appears to come and go as each new covariate is taken into account. Similar behaviour can occur in the linear model. What is an appropriate inference in such a case?

Let us label the individual units of the population by a variable I. We can consider the family of distributions for the response Y on unit I when treatment X is applied, as I and X vary. This is largely a conceptual entity, since only one treatment can in fact be applied to any unit.

The hypothesis of no treatment effect, at the level of individual units, can be written as

$$Y \perp\!\!\!\perp X \mid I. \tag{5.1}$$

Without further input, however, (5.1) is a conceptual, untestable assertion. Rubin (1978) considers a subjectivist Bayesian approach to inference about (5.1), in a framework which differs slightly from ours. We shall take a more standard view.

Suppose that we have a set of covariates Z (a function of I).

Definition. We say that Z is a *sufficient* set of covariates if

$$Y \perp\!\!\!\perp I \mid (X, Z). \tag{5.2}$$

If (5.2) holds, then any further information about I is irrelevant to the response distribution, for any treatment, when Z is known.

If Z is sufficient, then it is meaningful to discuss the distribution of Y given (X, Z). In particular, a meaningful statement is

$$Y \perp\!\!\!\perp X \mid Z. \tag{5.3}$$

Moreover (5.3) is, in principle, testable by standard statistical methods (although the practical difficulties increase with the dimensionality of Z).

Now if (5.2) and (5.3) hold, then, by Lemma 4.3, $Y \perp\!\!\!\perp (I, X) \mid Z$, whence in particular $Y \perp\!\!\!\perp X \mid (I, Z)$, which is (5.1). Conversely, if (5.2) and (5.1) hold then, within the sub-population of units having a fixed value of Z, the distribution of Y given (I, X) depends

neither on I, for known X, nor on X, for known I. Since any combination of I and X is conceptually possible, we must have $Y \perp\!\!\!\perp (I, X) \mid Z$, so that (5.3) holds.

We therefore find that, so long as Z is a *sufficient* set of covariates, (5.1) and (5.3) are logically equivalent, and it is valid to base inference about the existence of differential treatment effects on a test of (5.3). However, the equivalence breaks down for insufficient covariates. In particular, one must not ignore covariates altogether unless the population is homogeneous (so that $Y \perp\!\!\!\perp I \mid X$).

How can one be certain that (5.2) holds? If it does, and W is another set of covariates, then

$$Y \perp\!\!\!\perp W \mid (X, Z) \tag{5.4}$$

and (5.4) may be testable. If (5.4) is rejected, then so is (5.2). However, (5.4) does not imply (5.2), so that its acceptance does not automatically justify that of (5.2). At some stage it is necessary to make a subjective judgement that a set of covariates Z^* is sufficient. Usually Z^* will be of high dimension, and it is desirable, in the interests of efficiency and parsimony, to find a sufficient subset of Z^*. It is easy to show that Z forms such a reduction if and only if (5.4) holds, with $Z^* = (Z, W)$.

In the situation of Table 1, we might be willing to suppose that sex is a sufficient covariate. The significant differences in recovery rates for the two sexes, on either treatment, indicate that this covariate cannot be discarded, so that the appropriate inference is that the treatment is beneficial. By the same argument, the treatment would be judged valueless for Table 2. The assumption that sex is sufficient could be examined by introducing more covariates, but at some stage this process must stop, leaving a weak link at the very beginning of the chain of inference, which can only be reinforced by the statistician's informed judgement that the sufficiency condition is met. (This may be regarded as a judgement of partial exchangeability: see Lindley and Novick, 1979.)

Throughout the above analysis, we have supposed the covariates to be completely determined by I. This assumption is essential to the argument. For example, if it fails, then from (5.2) and (5.3) we can still deduce $Y \perp\!\!\!\perp X \mid (I, Z)$, but this is no longer equivalent to (5.1) and may not be interesting. Lindley and Novick give a thorough discussion of these points. Note that the inference (5.1) *will* still be possible in this case if also $Z \perp\!\!\!\perp X \mid I$: that is, if there is no effect of treatment on the covariates.

Another variant of the above analysis arises when it is not possible to assign treatments to units arbitrarily. For example, Bickel *et al.* (1977) examine whether there is any effect ascribable to sex (X) on the probability of acceptance (Y) onto a graduate programme. It is not even conceptually possible to vary X for a given individual application I, so that (5.1) becomes meaningless and some other interpretation of "no sex effect" is necessary. We can still define a sufficient set of covariates by (5.2), expressing the fact that the response distribution is the same for all individuals having the same value for (X, Z). Then it is easy to see that (5.3) is equivalent to

$$Y \perp\!\!\!\perp I \mid Z. \tag{5.5}$$

This says that Z (without X) is sufficient information on an individual to predict Y. (If (5.5) holds, then so does (5.2) automatically.) If this is regarded as an adequate interpretation of "no sex effect", this may again be tested by means of (5.3), always assuming Z is sufficient.

The adequacy of (5.5) must depend on the context. In particular, it is difficult to formalize the notion that X does not enter into Z (what if Z includes information on whether the candidate was of the same sex as the interviewer?). Or it might be that Z included height, and that women tend to be smaller than men. Should use of height in deciding on admission be regarded as a form of sexual discrimination?

6. Prediction Sufficiency

6.1. *Adequacy*

Let X represent the data from an experiment, and let Y denote the outcome of a further experiment, about which prediction is required. We suppose (X, Y) to have a joint sampling distribution depending on a parameter Θ.

Let T be a function of X. Then T is *adequate* (Skibinsky, 1967) for this prediction problem if (i) T is sufficient for Θ based on X, and (ii) in the sampling distributions, the distribution of Y given X depends on T alone. That is to say, (i) $X \perp\!\!\!\perp \Theta \,|\, T$, and (ii) $Y \perp\!\!\!\perp X \,|\, (T, \Theta)$. An equivalent single condition is, of course, $X \perp\!\!\!\perp (Y, \Theta) \,|\, T$. If Θ has an arbitrary prior distribution, then (X, Y, Θ) has an induced joint distribution still satisfying $X \perp\!\!\!\perp (Y, \Theta) \,|\, T$, so that $Y \perp\!\!\!\perp X \,|\, T$. That is, in the unconditional joint distribution of (X, Y), T is all that need be retained of X for the purpose of predicting Y.

Consider again the adequacy condition $X \perp\!\!\!\perp (Y, \Theta) \,|\, T$. We may construct the family of distributions of X given Y and Θ, and this property says that T is a sufficient statistic for this family. A rigorous statement and proof of this result may be found in the paper by Skibinsky. As he notes, an immediate consequence of this is that, under regularity conditions, there exists a *minimal* sufficient statistic for this extended family of distributions, that is to say, a *minimal adequate* statistic T^*, which is all that need be retained of X for prediction of Y.

6.2. *Decision Problems*

If, marginally for any prior distribution on Θ, $Y \perp\!\!\!\perp X \,|\, T$ (as holds if, for instance, T is adequate), then all information about X, other than T, may be lost, without destroying the ability to predict Y. Thus, given any decision problem based on data X, with loss determined by Y only, one might expect the decision rules based on T to form an essentially complete class. This is proved by Torgersen (1977). Conversely, if this essential completeness holds for all such problems, then $Y \perp\!\!\!\perp X \,|\, T$ for any prior distribution.

We have already seen that adequacy of T $(X \perp\!\!\!\perp (Y, \Theta) \,|\, T)$ is sufficient to ensure that $Y \perp\!\!\!\perp X \,|\, T$ holds for any prior, but it is not necessary. For example, another sufficient condition is $Y \perp\!\!\!\perp (X, \Theta) \,|\, T$, expressing the conditional independence of X and Y given T in the sampling distributions $(Y \perp\!\!\!\perp X \,|\, (\Theta, T))$ together with the condition that the distribution of Y given T is completely known $(Y \perp\!\!\!\perp \Theta \,|\, T)$. We could also have a mixture of these conditions: $X \perp\!\!\!\perp (Y, \Theta) \,|\, T$ for some values of T, and $Y \perp\!\!\!\perp (X, \Theta) \,|\, T$ for the remainder. Torgersen shows that this mixture is the most general situation under which $Y \perp\!\!\!\perp X \,|\, T$ for any prior.

However, if the loss function is allowed to depend on Θ as well as on Y, then the set of decision rules based on T forms an essentially complete class, for every such decision problem, if and only if T is adequate for Y based on X (so that T is "sufficient" for the "parameter" (Θ, Y): Takeuchi and Akahira, 1975).

6.3. *Transitivity*

Closely related to adequacy are the concepts of *transitivity* (Bahadur, 1954; Hall *et al.*, 1965) and *total sufficiency* (Lauritzen, 1974) in sequential experiments. Let $X = (X_1, X_2, \ldots)$ have a joint distribution depending on Θ, and for each n let T_n be a function of $X_{(n)} = (X_1, X_2, \ldots, X_n)$. The sequence (T_n) is *transitive* if $T_{n+1} \perp\!\!\!\perp X_{(n)} \,|\, (T_n, \Theta)$. Of particular interest are transitive sequences that are also sufficient, that is, T_n is sufficient for Θ based on $X_{(n)}$. Then $X_{(n)} \perp\!\!\!\perp (T_{n+1}, \Theta) \,|\, T_n$; that is, T_n is adequate for T_{n+1} based on $X_{(n)}$. Bahadur shows that, for any sequential decision problem with loss determined by Θ, attention may be restricted to rules based on a sufficient and transitive sequence.

Hall *et al.* consider a slightly more general notion of transitivity, in which ξ_n, the data available at stage n, need not necessarily be the whole of $X_{(n)}$. The same definition applies: (T_n) is transitive for (ξ_n) if T_n is a function of ξ_n, and $T_{n+1} \perp\!\!\!\perp \xi_n \,|\, (T_n, \Theta)$. They derive a number of results about transitivity using conditional independence. An example is their Theorem 4.2, which may be stated as follows.

Theorem. Suppose (T_n) is transitive for (ξ_n), and let (W_n), (η_n) be further sequences such that η_n is a function of ξ_n, and W_n is simultaneously a function of T_n and (separately) of η_n. Suppose further that $\eta_n \perp\!\!\!\perp T_n \,|\, W_n$ (in the sampling distributions). Then (W_n) is transitive for (η_n).

Proof. We have (all the time given Θ): $T_{n+1} \perp\!\!\!\perp \xi_n \,|\, T_n$ by transitivity, whence, since W_{n+1} is a function of T_{n+1} and η_n a function of ξ_n, $W_{n+1} \perp\!\!\!\perp \eta_n \,|\, T_n$. Hence $W_{n+1} \perp\!\!\!\perp \eta_n \,|\, (T_n, W_n)$, as W_n is a function of η_n. We also have $T_n \perp\!\!\!\perp \eta_n \,|\, W_n$, which together with the previous result yields $(T_n, W_{n+1}) \perp\!\!\!\perp \eta_n \,|\, W_n$, so that $W_{n+1} \perp\!\!\!\perp \eta_n \,|\, W_n$, and (W_n) is transitive for (η_n).

6.4. *Total Sufficiency*

Lauritzen (1974) calls a function S_n of $X_{(n)}$ *totally sufficient* if S_n is adequate for $(X_{n+1}, X_{n+2}, \ldots)$ based on $X_{(n)}$; a sequence (S_n) is totally sufficient if each S_n is. A sequence may be sufficient and transitive without being totally sufficient, and vice versa. Suppose, however, that (S_n) is totally sufficient, and each S_{n+1} can be expressed as a function of S_n and X_{n+1}. Then, in the sampling distributions, $S_{n+1} \perp\!\!\!\perp X_{(n)} \,|\, (X_{n+1}, S_n)$ (the distributions are degenerate!), while also $X_{n+1} \perp\!\!\!\perp X_{(n)} \,|\, S_n$. Thus $(S_{n+1}, X_{n+1}) \perp\!\!\!\perp X_{(n)} \,|\, S_n$, whence $S_{n+1} \perp\!\!\!\perp X_{(n)} \,|\, S_n$, so that (S_n) is (sufficient and) transitive.

In particular, suppose S_n is *minimal* totally sufficient, for each n. We have

$$X_{(n)} \perp\!\!\!\perp (X_{n+1}, X_{n+2}, \ldots, \Theta) \,|\, S_n,$$

whence

$$X_{(n)} \perp\!\!\!\perp (X_{n+2}, X_{n+3}, \ldots, \Theta) \,|\, (S_n, X_{n+1}),$$

and so

$$X_{(n+1)} \perp\!\!\!\perp (X_{n+2}, X_{n+3}, \ldots, \Theta) \,|\, (S_n, X_{n+1}).$$

So (S_n, X_{n+1}) is totally sufficient based on $X_{(n+1)}$, and hence, by minimality, S_{n+1} is a function of S_n and X_{n+1}. The previous paragraph therefore applies: (S_n) is sufficient and transitive.

(More generally, it is easy to show that, if (X, Y, Z) have a joint distribution, if S is adequate for (Y, Z) based on X and T adequate for (X, Z) based on Y, then (S, T) is adequate for Z based on (X, Y).)

We further have, for a minimal totally sufficient sequence (S_n), $X_{n+1} \perp\!\!\!\perp X_{(n)} \,|\, (S_n, \Theta)$, whence $(S_n, X_{n+1}) \perp\!\!\!\perp X_{(n)} \,|\, (S_n, \Theta)$. Since S_{n+1} is a function of (S_n, X_{n+1}), and $S_{(n)} = (S_1, S_2, \ldots, S_n)$ a function of $X_{(n)}$, we get $S_{n+1} \perp\!\!\!\perp S_{(n)} \,|\, (S_n, \Theta)$. That is, in the sampling distributions, the sequence (S_n) is a *Markov chain.* It follows that $S_{(n)} \perp\!\!\!\perp (S_{n+1}, S_{n+2}, \ldots) \,|\, (S_n, \Theta)$, while we know $S_{(n)} \perp\!\!\!\perp \Theta \,|\, S_n$ by sufficiency. Hence $S_{(n)} \perp\!\!\!\perp (S_{n+1}, S_{n+2}, \ldots, \Theta) \,|\, S_n$ (and thus (S_n) is totally sufficient with respect to itself). We deduce that $S_{n-1} \perp\!\!\!\perp \Theta \,|\, S_n$, so that the *backward transition distributions* of the Markov chain (S_n) are completely known, independently of the parameter. Also, for any prior distribution on Θ, $S_{(n)} \perp\!\!\!\perp (S_{n+1}, S_{n+2}, \ldots) \,|\, S_n$, so that (S_n) remains a Markov chain in its marginal distribution. It is clear that the same backward transition distributions apply to this marginal chain.

The above theory has been used by Lauritzen (1974) in an approach to model-building. Starting with a given backward transition structure for a Markov chain (S_n), he investigates the class of all possible distributions for (S_n) consistent with this. This class is convex, and the extremal elements are taken as constituting the appropriate family of sampling distributions.

7. DATA SELECTION PROCESSES

7.1. *Introduction*

There are many situations in which data arise from some random process, but observation of these data is biased by selection effects, so that not every outcome is equally likely to be observed. The simplest case is *truncation*, where a generated outcome X is observed if and only if X falls in some pre-assigned set A. The distribution after truncation will differ from the untruncated distribution, but the *conditional* distributions, given that $X \in A$, will agree.

More generally, let Y be a function of X, and suppose the probability that the generated outcome X is selected for observation is determined by Y alone. We can express this in terms of conditional independence by defining an indicator variable Q, with $Q = 1$ if the outcome is selected, $Q = 0$ otherwise. Then we are supposing $Q \perp\!\!\!\perp X \mid Y$. It then follows trivially that $X \perp\!\!\!\perp Q \mid Y$, which may be interpreted, in terms of (2a) of Section 3, as saying that the distribution of X *given* Y is the same for both the selected and unselected data. In particular, estimation of this distribution from selected data may proceed without the need for any corrective adjustment.

7.2. *Diagnosis*

An application of the above arises in a statistical approach to medical diagnosis (Dawid, 1976). With each individual in a population is associated a pair (S, D), where D signifies his disease, and S the full set of symptoms, signs, etc. on which diagnosis is to be based. The selection variable Q represents admission to the centre in which the data are collected. This selection is supposed to be governed entirely by the *presenting symptoms* S_0, a subset of S, together with further non-medical personal information G. Thus $Q \perp\!\!\!\perp (S, D) \mid (S_0, G)$.

The non-medical nature of G may be formalized as $G \perp\!\!\!\perp D \mid S$, expressing the fact that, for inference about D from (S, G), S is sufficient.

We therefore have $Q \perp\!\!\!\perp D \mid (S, S_0, G)$ (where S_0 is redundant), and $G \perp\!\!\!\perp D \mid S$, which together give $(Q, G) \perp\!\!\!\perp D \mid S$, so that $D \perp\!\!\!\perp Q \mid S$. It follows that the distributions of D given S are unaffected by selection and, in the absence of any further biases, may be estimated directly. This contrasts with a widespread emphasis on estimation of the distributions of S given D: distributions which *are* subject to distortion by selection bias.

Another application of the general result arises in regression analysis, in which the regression of a response variable on the explanatory variables is the same, no matter whether these explanatory variables arise at random, or are selected in any way, so long as such a selection does not involve the response.

7.3. *Sufficiency and Adequacy*

Let the distribution of X be governed by a parameter Θ, and suppose $Q \perp\!\!\!\perp \Theta \mid X$, so that selection only depends on the data. If T is sufficient before selection then $X \perp\!\!\!\perp \Theta \mid T$. Since $Q \perp\!\!\!\perp \Theta \mid (X, T)$, we get $(Q, X) \perp\!\!\!\perp \Theta \mid T$, and so $X \perp\!\!\!\perp \Theta \mid (Q, T)$. That is, T remains sufficient in the selected distributions (Tukey, 1949).

For a Bayesian, $Q \perp\!\!\!\perp \Theta \mid X$ immediately yields $\Theta \perp\!\!\!\perp Q \mid X$, so the selection may be ignored when forming posterior distributions based on the full data X.

Now suppose (X, Y) have a joint distribution depending on Θ, and that T is adequate for predicting Y from X: $X \perp\!\!\!\perp (\Theta, Y) \mid T$. Let selection be governed purely by the value of X: $Q \perp\!\!\!\perp (\Theta, Y) \mid X$. We derive $(Q, X) \perp\!\!\!\perp (\Theta, Y) \mid T$ (remember T is a function of X), so that $X \perp\!\!\!\perp (\Theta, Y) \mid (Q, T)$, and T remains adequate after selection (Padmanabhan, 1972). For a Bayesian, $Y \perp\!\!\!\perp (X, Q) \mid T$: both the fact of selection and the value of X are irrelevant to prediction when T is known.

In contrast to the above results, Padmanabhan shows that a statistic ancillary before selection is not generally ancillary afterwards.

7.4. *Sequential Sampling*

Suppose the components of $X = (X_1, X_2, \ldots)$ are observed sequentially, until the process terminates by the operation of a deterministic *stopping-time*. Thus if N denotes the number of observations made, the event A_n, that $N = n$, is completely determined by the value of $X_{(n)} = (X_1, X_2, \ldots, X_n)$. Such stopping times are of importance both in practice (since they often operate) and in theory (for example, optional stopping of Markov processes or martingales).

Several extensions of this notion to the non-deterministic case have been suggested (Bahadur, 1954; Kemperman, 1961; Siegmund, 1967). Pitman and Speed (1973) show, using simple properties of conditional independence, that these various definitions are all equivalent to (or special cases of) the following

Definition. N is a *randomized stopping time* for X if, for all n, $A_n \perp\!\!\!\perp X \,|\, X_{(n)}$.

They go on to show that N can be considered as a deterministic stopping time on a suitable sequence of σ-fields and that the Markov or martingale properties carry over when X is regarded as adapted to this sequence. Thus the theory of randomized stopping times may be studied by means of the well-known established theory of deterministic stopping times.

Suppose now that the distribution of X depends on a parameter Θ. The stopping-time property is extended to: $A_n \perp\!\!\!\perp (X, \Theta) \,|\, X_{(n)}$, and the observed data may be represented by $(A_n, X_{(n)})$. Let T_n be sufficient for Θ based on $X_{(n)}$. Then $X_{(n)} \perp\!\!\!\perp \Theta \,|\, T_n$, and since $A_n \perp\!\!\!\perp \Theta \,|\, (X_{(n)}, T_n)$ it easily follows that $(A_n, X_{(n)}) \perp\!\!\!\perp \Theta \,|\, T_n$, so that T_n is a sufficient summary of the observed data in the stopped experiment. Likewise, the properties of transitivity and total sufficiency are preserved by such optional stopping.

For a Bayesian, $\Theta \perp\!\!\!\perp A_n \,|\, X_{(n)}$. Thus the posterior distribution based on the data of the stopped experiment is the same as that based on the experiment which always produces $X_{(n)}$. (This is a reflection of the fact that a stopping-time does not alter likelihood-ratios.)

Similarly, for a Bayesian, $(X_{n+1}, X_{n+2}, \ldots) \perp\!\!\!\perp A_n \,|\, X_{(n)}$, so that prediction is unaffected by optional stopping.

The above theory can be applied in non-sequential situations in which the observed set of data may be a selected subset of the full data X, for example missing-data problems. More details may be found in Rubin (1976) and Dawid and Dickey (1977).

8. Conditional Independence and Invariance

8.1. *Invariance in Parametric Models*

It is useful to have some theory which gives us conditions under which we shall have conditional independence. Consider first a statistical model in which the distributions of X are governed by Θ, and suppose these distributions are *equivariant* under the action of transformation groups G on X and $\bar{G} = \{\bar{g}: g \in G\}$ on Θ: that is, for each $g \in G$, the distribution of gX, given $\Theta = \theta$, is the same as that of X given $\Theta = \bar{g}\theta$. It is then easy to show (see, for example, p. 220 of Lehmann, 1959) that, if Z is a maximal invariant function of X under G, and Ψ a maximal invariant function of Θ under $\bar{G}$, then the distribution of Z depends on Ψ alone.

The above result may be expressed in terms of conditional independence:

$$Z \perp\!\!\!\perp \Theta \,|\, \Psi. \tag{8.1}$$

8.2. *Sufficiency and Invariance*

Now consider a sufficient statistic S in the above model, and suppose that G acts on S in the sense that, if $S(x_1) = S(x_2)$, then $S(gx_1) = S(gx_2)$. (This will be true if S is minimal sufficient: Fraser, 1966.) For $g \in G$, we get g^* acting on S defined by: $g^*(S(x)) = S(g(x))$.

The distributions of X given S do not depend on Θ, and may themselves be considered as a parametric family of distributions, with S now playing the rôle of the parameter. Moreover, this family is easily seen to be equivariant under the action of G on X and $G^* = \{g^*: g \in G\}$ on S. Let U be a maximal invariant function of S under G^*. Then the previous result may be directly applied, to yield

$$Z \perp\!\!\!\perp S \,|\, U. \tag{8.2}$$

(This is made precise in Hall *et al.*, 1965.) Since S is sufficient, $Z \perp\!\!\!\perp \Theta \,|\, S$, whence $Z \perp\!\!\!\perp \Theta \,|\, (S, U)$ (U is a function of S). Taking this together with (8.2), Lemma 4.3 yields $Z \perp\!\!\!\perp (S, \Theta) \,|\, U$; that is to say, U is *adequate* for S based on Z. In particular, $Z \perp\!\!\!\perp \Theta \,|\, U$, so that U is sufficient for Θ based on Z. Hall *et al.* call U *invariantly sufficient.*

Example 8.1 (Fraser, 1966). Let $X=(X_1,\ldots,X_n)$ be a random sample from $N(\Theta,\sigma^2)$ (σ^2 known) and let G be the location group with typical member

$$(x_1,x_2,\ldots,x_n)\xmapsto{g}(x_1+a,x_2+a,\ldots,x_n+a).$$

The parameter transforms as $\theta\xmapsto{\bar{g}}\theta+a$; and the minimal sufficient statistic is $S=\bar{X}$, transforming by $s\xrightarrow{g^*}s+a$. We have $Z=(X_1-\bar{X},\ldots,X_n-\bar{X})$, and U is trivial. We deduce that $\bar{X}$ and $(X_1-\bar{X},\ldots,X_n-\bar{X})$ are independent given Θ.

A similar argument, using the scale group, shows that if $X=(X_1,\ldots,X_n)$, with $X_i\sim\Gamma(\lambda_i,\beta)$ independently, then $S=\Sigma X_i$ is independent of $(X_1/S,\ldots,X_n/S)$.

Example 8.2. For a case where U is non-trivial, let $X=(\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n)$ be a random sample from the Fisher–von Mises distribution on the circle. Thus $\mathbf{X}_i$ is a unit vector in the plane, whose density (with respect to Lebesgue measure on the circumference of the unit circle) is given by $c(\|\boldsymbol{\theta}\|)\exp(\boldsymbol{\theta}'\mathbf{x}_i)$. Here the parameter-value $\boldsymbol{\theta}$ is an arbitrary vector. A minimal sufficient statistic is $\mathbf{S}=\Sigma\mathbf{X}_i$, summing over i from 1 to n.

We take, as a typical element g of G, a rotation of each $\mathbf{X}_i$ through the same angle α. Then $\bar{g}$ operates on $\boldsymbol{\Theta}$, and g^* on $\mathbf{S}$, by means of the same rotation. We may take Z to consist of the *configuration* of the $\mathbf{X}_i$'s, that is, the set of angular distances between them, while $\Psi=\|\boldsymbol{\Theta}\|$ and $U=\|\mathbf{S}\|$. We therefore find: (i) the distribution of the configuration Z depends only on $\|\boldsymbol{\Theta}\|$, and (ii) given $\|\mathbf{S}\|$, the configuration Z is independent of (the orientation of) $\mathbf{S}$.

Example 8.3. As an extension of the above, suppose that the sample size n is chosen by a random process, with known probabilities, before the experiment is performed. If $\Psi=\|\boldsymbol{\Theta}\|$ is regarded as fixed, then $\mathbf{S}$ is still minimal sufficient (in particular, the value of n is *not* required for inference). The only change is that n must be appended to Z. We thus find, in particular, that the distribution of (the orientation of) $\mathbf{S}=\Sigma\mathbf{X}_i$, given n and $\|\mathbf{S}\|$ (and the parameters), is the same for all n.

Now when $\Psi=\|\boldsymbol{\Theta}\|$ is known, $U=\|\mathbf{S}\|$ is ancillary. This suggests that inference about $\boldsymbol{\Theta}$ might then be based on the distribution of $\mathbf{S}$ given $\|\mathbf{S}\|$. Apart from any theoretical arguments for this, the above result shows that this distribution is not affected by the randomness of n. In contrast, the full distribution of $\mathbf{S}$ does depend on the distribution of n. Although the information about n is discarded in the reduction to the minimal sufficient statistic $\mathbf{S}$, many statisticians would like to condition on its value. The conditioning on $\|\mathbf{S}\|$ is an alternative which accomplishes this end.

For a general discussion of problems of inference raised by examples such as this, see Dawid (1975).

8.3. *Bayesian Invariance*

Consider again the equivariant sampling model of Section 8.1, and suppose that the parameter Θ has a prior distribution that is *invariant* under $\bar{G}$. That is, for $\bar{g}\in\bar{G}$, $\bar{g}\Theta$ has the *same* distribution as Θ. It follows easily (Dawid, 1977a: Appendix) that the family of posterior distributions for Θ given X is equivariant under the action $\bar{G}$ on Θ and G on X (we are now considering Θ as a random variable with distributions governed by the "parameter" X). By the same argument that yielded (8.1), we have

$$\Psi\perp\!\!\!\perp X\,|\,Z. \tag{8.3}$$

It follows that the posterior distribution of Ψ depends on the data X through the value of Z alone.

An important special case arises when $\bar{G}$ consists of the identity transformation alone. In this case, any of the sampling distributions of X given Θ is unchanged by any transformation $g\in G$ operating on X. Then $\Psi=\Theta$, and *any* prior distribution for Θ is invariant under $\bar{G}$. From (8.3), $\Theta\perp\!\!\!\perp X\,|\,Z$ for any prior, so that the posterior distribution depends on the

data through Z alone, which is therefore Bayesian sufficient for Θ. Under regularity conditions, this is equivalent to ordinary sufficiency:

$$X \perp\!\!\!\perp \Theta \mid Z. \tag{8.4}$$

We have therefore derived the following result: if every member of the family of sampling distributions is invariant under the group G, then the maximal invariant under G is sufficient. (For a rigorous statement and proof, see Farrell, 1962).

Example 8.4 (Basu, 1970). Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ be a random sample from $N(\mu, \sigma^2)$, with both parameters unknown. A typical element of G takes $\mathbf{X}$ into $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{A}$ is an orthogonal matrix such that $\mathbf{A1} = \mathbf{1}$. Then $\mathbf{Y}$ has the same distribution as $\mathbf{X}$ for every (μ, σ^2). A maximal invariant under G is $(\Sigma X_i, \Sigma X_i^2)$, which is therefore sufficient.

9. Intersubjective Models

From the point of view of de Finetti's theory of subjective probability (1975), there is no need for parametric (or non-parametric!) statistical models. Let Y represent the whole class of *observables* with which a subjectivist is concerned in a particular application. (This will include variables about which prediction may be required, as well as X, the data to be collected.) Then such an individual is supposed to have a joint subjective distribution over Y, which is completely known to him. After observing X, he has a new distribution, derived by simple conditioning, which is now relevant for prediction.

Consider now a collection of such individuals $\Lambda = \{\lambda\}$, all concerned with the same variables Y. Each λ will have his own subjective distribution Q_λ for Y. The family $\mathcal{Q} = \{Q_\lambda : \lambda \in \Lambda\}$ looks like a statistical model for Y, with λ as the parameter, although its interpretation is really very different. Nevertheless, we can apply standard statistical concepts to this family. For example, suppose a variable Θ (a function of Y) is *sufficient* for $\mathcal{Q}$. Then all the subjectivists *agree* about the distribution of Y given Θ. This motivates the following definition, in which "*I*" denotes "intersubjective":

Definition 9.1. If Θ is sufficient for $\mathcal{Q}$, we call Θ an *I-parameter*; the (common) distributions of Y given Θ constitute an *I-model.*

There is, of course, some arbitrariness about what is included in Y, but once this is settled we can introduce the concept of a *minimal I*-parameter, corresponding to a minimal sufficient statistic for $\mathcal{Q}$.

Example 9.1. Let $Y = (Y_1, Y_2, \ldots, Y_N)$, where each Y_i is the outcome of a toss of the same coin, coded 0 or 1. Suppose every individual agrees that the Y_i's are *exchangeable*, so that each distribution Q_λ is unchanged when the components of Y undergo an arbitrary permutation. It follows easily that $T_N = \Sigma_{i=1}^N Y_i$ is sufficient for $\mathcal{Q}$, and can be taken as the *I*-parameter. All individuals agree on the *I*-model: given T_N, all possible sequences of Y's are equally likely; although they may disagree about the marginal distribution for T_N.

Example 9.2. To continue from the previous example: if all individuals agree that (conceptually at least) the coin may be tossed infinitely often, while exchangeability is retained, then they might try to model the larger world, consisting of $Y^* = (Y_1, Y_2, \ldots, ad\ inf.)$. By de Finetti's celebrated theorem (1937), each Q_λ may be represented as a mixture of Bernoulli sequences for Y^*, with probability ξ, where ξ has some distribution π_λ over $[0, 1]$. In fact more is true (Kendall, 1967). Define $\Theta = \lim_{n\to\infty} n^{-1} \Sigma Y_i$, (summing over i from 1 to n) if this exists, $\Theta = 0$ (say) otherwise. The limit exists with probability one for each Q_λ, and we may identify Θ with ξ. For each Q_λ, the distribution of Y^* given Θ may be taken as Bernoulli trials with probability Θ. Thus Θ is an *I*-parameter for Y^*, and the *I*-model is just that of Bernoulli trials. We have therefore *derived* the traditional model which forms the starting point of most investigations of Y^*, as an agreed intersubjective model for all those individuals who are willing to accept exchangeability.

Once we have the concept of an *I*-parameter, we can introduce appropriate versions of other statistical concepts. For example, a statistic T (a function of X) is *I-sufficient* if $X \perp\!\!\!\perp \Theta \mid T$.

(Note that the distributions of X given Θ, and hence of X given Θ and T, are well defined in the sense that they are the same for every Q_λ.) Similarly, S is *I-ancillary* if $S \perp\!\!\!\perp \Theta$. We can go on to introduce *I-adequacy*, and all the other statistical concepts which may be expressed in terms of conditional independence. The classical results will continue to hold.

It is important not to confuse the technical and intuitive contents of such concepts. For example, an I-parameter is technically sufficient for $\mathcal{Q}$; an I-sufficient statistic is not. It is, however easy to show that an I-sufficient statistic is also technically sufficient based on X alone, and so can serve as an I-parameter for X.

The technical concept of adequacy was used in Dawid (1980) in an investigation of the general structure of conditional independence. Again, this use must be distinguished from the intuitive concept, now re-expressed as I-adequacy.

In the new framework, no real distinction is made between data and parameters: all are random variables. Many statistical definitions and properties become simple expressions of conditional independence. In this way the duality between data and parameters, touched upon in Section 3.2, is seen to be inevitable. For example, an I-parameter Θ is *identified* if and only if it is technically *minimal* sufficient for $\mathcal{Q}$.

The invariance (under permutations) of the distribution of Y or Y^* in Examples 9.1 and 9.2 is the key property which generates the relevant I-models. Such an argument, based on the theory of Section 8, is applicable in many other circumstances. It is hoped to investigate this approach to model-building (which has close ties with that of Lauritzen, 1974) in a future paper.

10. Conclusion

Although it may not appear so, I have resisted the temptation to say everything I can think of about conditional independence in this paper. Apart from containing a more mathematical development of the simple properties considered here, another paper (Dawid, 1980) investigates further problems, such as the implications of the pairs of properties: (i) $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$; or (ii) $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Y$. Application of these results in a paper of mine as yet unpublished is shown to lead to corrected forms for a number of fallacious arguments which do not, on the face of it, appear to have anything in common—until their structure is expressed in terms of conditional independence.

Another paper, by Dawid and Dickey, not published yet, expresses various concepts of "partial sufficiency in the presence of nuisance parameters" in terms of conditional independence, and uses the general results to obtain connections between them.

The present paper has, I hope, given some support to my claim that ideas of conditional independence pervade statistical theory, and that it might be illuminating to focus their light on all aspects and into all corners of the subject. I have no doubt that further applications and examples of conditional independence could be found in profusion, and I look forward to learning of such discovery at the hands of others.

References

Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.*, **25**, 423–462.

Barankin, E. W. (1961). Sufficient parameters: solution of the minimal dimensionality problem. *Ann. Inst. Statist. Math.*, **12**, 91–118.

Basu, D. (1970). On sufficiency and invariance. In *Essays in Probability and Statistics* (R. C. Bose, I. M. Chakravati, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith, eds), pp. 61–84. Chapel Hill: University of North Carolina Press.

Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1977). Sex bias in graduate admissions: data from Berkeley. In *Statistics and Public Policy* (W. B. Fairley and F. Mosteller, eds), pp. 113–130. Reading, Massachusetts: Addison-Wesley.

Birnbaum, A. (1962). On the foundations of statistical inference (with Discussion). *J. Amer. Statist. Ass.*, **57**, 269–326.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Ass.*, **67**, 364–366.

Cole, N. S. (1973). Bias in selection. *J. Educ. Meas.*, **10**, 237–255.

DAWID, A. P. (1975). On the concepts of sufficiency and ancillarity in the presence of nuisance parameters. *J. R. Statist. Soc.* B, **37**, 248–258.
—— (1976). Properties of diagnostic data distributions. *Biometrics*, **32**, 647–658.
—— (1977a). Invariant distributions and analysis of variance models. *Biometrika*, **64**, 291–297.
—— (1977b). Spherical matrix distributions and a multivariate model. *J. R. Statist. Soc.* B, **39**, 254–261.
—— (1980). Conditional independence for statistical operations. *Ann. Statist.* (to appear).
DAWID, A. P. and DICKEY, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Amer. Statist. Ass.*, **72**, 845–850.
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.
DE FINETTI, B. (1937). Foresight: its logical laws, its subjective sources (in French). English translation in *Studies in Subjective Probability* (1964), (H. E. Kyburg and H. E. Smokler, eds), pp. 93–158. New York: Wiley.
—— (1975). *Theory of Probability* (English translation). London: Wiley.
DEMPSTER, A. P. (1961). *Elements of Continuous Multivariate Analysis.* Reading, Massachusetts: Addison-Wesley.
DRÈZE, J. H. (1974). Bayesian theory of identification in simultaneous equations models. In *Studies in Bayesian Econometrics and Statistics* (S. Fienberg and A. Zellner, eds), Amsterdam: North Holland.
FARRELL, R. H. (1962). Representation of invariant measures. *Illinois J. Math.*, **6**, 447–467.
FLORENS, J.-P. and MOUCHART, M. (1977). Reduction of Bayesian experiments. CORE Discussion Paper 7737.
FRASER, D. A. S. (1966). On sufficiency and conditional sufficiency. *Sankhyā* A, **28**, 145–150.
HALL, W. J., WIJSMAN, R. A. and GHOSH, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Ann. Math. Statist.*, **36**, 575–614.
KADANE, J. B. (1974). The role of identification in Bayesian theory. In *Studies in Bayesian Econometrics and Statistics* (S. E. Fienberg and A. Zellner, eds). Amsterdam: North-Holland.
KEMPERMAN, J. H. B. (1961). *The Passage Problem for a Stationary Markov Chain.* Chicago: University of Chicago Press.
KENDALL, D. G. (1967). On finite and infinite sequences of exchangeable events. *Studia Sci. Math. Hungar.*, **2**, 319–327.
KUDŌ, H. (1967). On partial prior information and the property of parametric sufficiency. *Proc. Fifth Berk. Symp.* I, 251–265.
LAURITZEN, S. L. (1974). Sufficiency, prediction and extreme models. *Scand. J. Statist.*, **1**, 128–134.
LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* New York: Wiley.
LINDLEY, D. V. and NOVICK, M. R. (1979). The role of exchangeability in inference. *Ann. Statist.* (to appear),
PADMANABHAN, A. R. (1972). Local efficiency, efficiency, adequacy, ancillarity and truncation. *Sankhyā* A. **34**, 145–152.
PICCI, G. (1977). Some connections between the theory of sufficient statistics and the identifiability problem. *SIAM J. Appl. Math.*, **33**, 383–398.
PITMAN, J. W. and SPEED, T. P. (1973). A note on random times. *Stoch. Procs and their Applns*, **1**, 369–374.
REIERSØL, O. (1950). Identifiability of a linear relationship between variables which are subject to error. *Econometrica*, **18**, 375–389.
RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.
—— (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.
SIEGMUND, D. (1967). Some problems in the theory of optimal stopping. *Ann. Math. Statist.*, **38**, 1627–1640.
SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *J. R. Statist. Soc.* B, **13**, 238–241.
SKIBINSKY, M. (1967). Adequate subfields and sufficiency. *Ann. Math. Statist.*, **38**, 155–161.
TAKEUCHI, K. and AKAHIRA, M. (1975). Characterizations of prediction sufficiency (adequacy) in terms of risk functions. *Ann. Statist.*, **3**, 1018–1024.
TORGERSEN, E. N. (1977). Prediction sufficiency when the loss function does not depend on the unknown parameter. *Ann. Statist.*, **5**, 155–163.
TUKEY, J. W. (1949). Sufficiency, truncation and selection. *Ann. Math. Statist.*, **20**, 309–311.

DISCUSSION OF PROFESSOR DAWID'S PAPER

Professor D. V. LINDLEY: To unite several different ideas through a single concept is a valuable notion and Dawid's use of independence is a notable example of this. But is independence the most satisfactory concept to use, for there are several difficulties with it? Firstly, it rarely exists: it is hard to find things that are truly independent. I had an example of this recently when I misread an Ordnance Survey map because the spot heights were in metres but the contour lines were at 50-foot intervals. Presumably we have gone metric because our screws won't work overseas, so

my country walks are dependent on some features of the engineering industry. Secondly, independence is too closely related to hypothesis-testing for my taste. This is brought out in the discussion of Simpson's paradox, which, to my mind, is concerned with the behaviour of measures of association and not with the narrower point of whether or not that measure is zero. A third point is that independence is not just a relation between pairs of random variables but between members of a set. The basic idea is $\perp\!\!\!\perp(X_1, X_2, \ldots, X_n)$, not $X_1 \perp\!\!\!\perp X_2$, and the definition of the former in terms of the latter is cumbersome. A fourth point concerns the difficulty of the concept of independence itself. It is such an elusive concept: for example, in pairwise independence not implying independence. But perhaps these four objections are not real objections because the ideas that independence is meant to unite are themselves difficult. So let me turn to these ideas, like sufficiency and ancillarity.

Are these ideas important in the Bayesian view? I suggest much less so than in the sampling-theory view. In the coherent approach one has data x and parameters θ of interest, and λ nuisance. One needs to evaluate $p(\theta \mid x)$ and the only tools available for this are the rules of the probability calculus. If $t_1(x)$ is sufficient for (θ, λ), then this will come out in the analysis: or we may get $t_2(x)$ marginally sufficient for θ for the specific marginal $p(\lambda \mid \theta)$ that is being used—a difficult concept outside the Bayesian view. In other words, the calculus provides the answer and no tricks of sufficiency are required. At best they may be helpful. In the Bayesian view it is nearly as easy to investigate samples from a t-distribution as it is a normal one, despite the fact that a three-dimensional sufficient statistic $(n, \bar{x}, s^2)$ exists for the latter but not the former. Such difficulties as do arise are numerical, not inferential.

In the paper it is suggested that $X \perp\!\!\!\perp \theta$ can be useful when θ is not a random variable but a fixed quantity. This can be dangerous when allied to other ideas. Suppose θ is 0 or 1, and X is 0, 1 or 2; the distribution of X having $p(X = 0) = p(X = 1) = a$, for some $0 \leqslant a \leqslant \frac{1}{2}$, for both θ-values. Then $g(X, \theta) = 0$ if $X = \theta$, and 1 otherwise, is a pivot, so that $X = 0$ gives $\theta = 0$ a probability of a, despite X and θ being independent; and this for any a. It may legitimately be said that this is a criticism of pivots and not independence. Godambe has given more complicated, practical examples.

Finally, some remarks about Simpson's paradox. It is usual to employ randomization in experimental design, which, in the notation of the paper, presumably means $X \perp\!\!\!\perp I \mid Z$. The strength of this is that if it is not true that $X \perp\!\!\!\perp Y \mid Z$, then it cannot happen that $X \perp\!\!\!\perp Y \mid I$: that is, the observed dependence is real. This result can be strengthened by using a measure of association rather than dependence. Novick and I in our study of the paradox felt that we did not gain a proper appreciation of it until we had changed the random variables into uncertain quantities: until we had thought about the effect of a treatment on a particular person, You. Random variables, with their essential, allied notion of a population, did not seem adequate fully to explain the paradox. Consider two cases:

X = treatment, Y = recovery, Z = sex, and
X = variety, Y = yield, Z = height,

with identical data in the two cases exhibiting the paradox, having positive association for all Z, and negative unconditionally. Then in the first case, for a given person, where Y is uncertain, one would increase X (that is, the positive association is the real one) whereas in the other one would choose the variety with least X (the negative, unconditional association is the real one). The same applies with the weaker concept of independence and I feel that $X \perp\!\!\!\perp Y$ for random variables that makes no mention of the population is less satisfactory than $X_1 \perp\!\!\!\perp Y_1$, where X_1 is the treatment for You, and Y_1 is Your recovery.

These comments do not detract from a remarkable paper. I am enormously impressed by Dawid's scholarship. He seems to have read everything and knitted it into his thesis. That alone demonstrates the power of his approach. If a Bayesian thinks it good, then a sampling-theorist, to whom independence is more important, must find it superb. Perhaps he could even use it to find out which ancillary to condition on. It is a real pleasure to propose the vote of thanks.

Mr A. D. McLaren (University of Glasgow): We have had a very instructive paper tonight. The author has combined erudition with a disarming simplicity to show us that much of what appears in the journals is yet another example of conditional independence. I can certainly give empirical confirmation, if any is needed, that conditional independence crops up regularly in

practical statistical work. For instance it is now routine (Akhmeteli *et al.*, 1977) to calculate odds that a woman with a family history of a sex-linked disease such as haemophilia is in fact a (symptomless) carrier of the disease. The calculation is based both on clinical tests on the woman, and possibly her female relatives, and on analysis of her family tree including any descendants. This purely genetical analysis can be complicated (Elston and Stewart, 1971; Lange and Elston, 1975) if the pedigree is a long one, but it is much simplified by the fact that conditional on the woman's own genotype her descendants are independent of kin related to her through her parents. Indeed the conditional independence of the development of the offspring and any characters acquired by the parents, given their genotypes, comes close to defining the anti-Lamarckian view of heredity.

Another practical example is suggested by the paper's mention of missing data (Section 7.4). The key formula here is usually, in Bayesian notation, $\pi(\theta \mid A) = \Sigma_M \pi(\theta \mid A, M)\, \pi(M \mid A)$, where θ are the parameters of interest and A, M represent the available and missing parts of the data respectively. The characteristic inferential, rather than computational, difficulty is posed by the terms $\pi(M \mid A)$. Suppose, for example as in McLaren (1967), that graduates of a training course take a test (scores denoted S) at the end of the course and later only some of them take another (scores denoted R) to see what knowledge they have retained. Thus $A = (\mathbf{S}_1, \mathbf{R}_1, \mathbf{S}_2)$ and $M = (\mathbf{R}_2)$. Now $\pi(M \mid A) = \pi(\mathbf{R}_2 \mid \mathbf{S}_2, \mathbf{S}_1, \mathbf{R}_1) = \pi_1(\mathbf{R}_2 \mid \mathbf{S}_2)$, say. Can the function π_1 be supplied from a straightforward interpretation of the scatter diagram of $\mathbf{R}_1$ against $\mathbf{S}_1$, perhaps using standard prediction formulas? The answer is yes, provided the probability that a graduate goes on to take the second test depends only on his score in the first test. This is merely an illustration of Section 7.2, one of the simplest but most important in the whole paper.

Would the author like to clarify the relationship of Section 7.3, particularly the sentence "For a Bayesian, . . . the selection may be ignored . . .", to Section 7.2?

Section 5 of the paper is extremely interesting. After equation (5.4) we read: "At some stage it is necessary to make a subjective judgement that a set of covariates Z^* is sufficient." As a Bayesian I must agree with this statement; but practitioners will console themselves with the thought that this subjective judgement can be made less painful if the usual random allocation $X \perp\!\!\!\perp I \mid Z$ of treatments to units is made. Note, however, Jeffreys' (1961, p. 243) caution against trying to randomize out a covariate with a large effect.

Turning to the theoretical aspects of tonight's paper, I found Theorem 2.1 both amusing and thought-provoking. When two variables are associated it may be in a distinctly unsymmetrical manner but, as the theorem states, if stochastic independence does strike it does so symmetrically. Would it be interesting to consider weaker versions of (conditional) independence in the hope of finding an unsymmetrical one? Suppose a vector random variable Y is said to be *weakly independent* of another, X, if $E(\mathbf{Y} \mid \mathbf{X})$ does not depend on X. This definition is equivalent to demanding that each component of Y is uncorrelated with any function of X. It does not imply that X is weakly independent of Y: consider a bivariate example in which the joint density of X and Y is uniform over an isosceles triangle with its odd side parallel to the Y-axis.

Another area of interest is the theory of partial sufficiency as developed by Kagan and Shalaevskii (see Linnik, 1975, p. 4). Here one is concerned with statistics T such that $E(f \mid T, \theta)$ is free of θ for statistics f belonging to a specified linear space.

As indication that, even armed with the notion of conditional independence, one must still take care, consider a statistic U known to be independent of a statistic T sufficient in the ordinary sense. The consideration that $p(U \mid t, \theta)$ is free of θ (sufficiency) and free of t (independence) might suggest that U must be ancillary, but this is not so as the following artificial example (Basu, 1958) shows. Let X_1 and X_2 be independent from a uniform distribution over $[2\mu, 2\mu+1]$ where μ is an unknown integer. X_1 is sufficient, in fact determining μ precisely; however, the independent statistic X_2, far from being ancillary, is itself sufficient. The explanation of this "paradox" is in fact extremely simple, as a moment's reflection will show.

My own favourite example of conditional independence is as follows; no doubt Professor Dawid knows it. Two statisticians start rival schools of inference. Colleagues are approached in turn and each joins one school or the other. Of the first n approached ($n = 1, 2, \ldots$) V_n join the first school. If $V_n = v$, the probability that the next statistician will join the first school is $(v+1)/(n+2)$. What happens to V_n/n as $n \to \infty$? The author virtually gave the answer to-day in Example 9.2. The process $\{V_1, V_2, V_3, \ldots\}$ is a Markov chain with non-stationary transition probabilities: $(v+1)/(n+2)$, the complementary probability, and zeros. It is therefore isomorphic to a process $\{R_1, R_2, R_3, \ldots\}$ where R_n is the cumulative number of successes in a Bernoulli process whose success probability is

itself unknown and is assigned a prior distribution uniform over [0, 1]. This may be verified by the routine Bayesian calculation $\Pr(R_{n+1} = r+1 \mid R_n = r) = \ldots = (r+1)/(n+2)$. Hence the limiting behaviour as $n \to \infty$ of V_n/n is the same as that of R_n/n, namely almost sure convergence to a limiting random variable whose distribution is uniform over [0, 1].

I am sure that I have said enough to demonstrate that, conditional on the discussant, contributions to these meetings may be independent of the paper. It only remains to second the vote of thanks to our speaker for sharing his splendid insights with us tonight. We who are about to dine salute you.

The vote of thanks was passed by acclamation.

Professor J. B. KADANE (Carnegie–Mellon University): Professor Dawid has given us a very enlightening synthesis of concepts in his paper. It turns out that many of the key concepts in statistical theory, such as ancillarity, sufficiency and identification, are special cases, and that results thought to require individual proofs are special cases of very simple facts concerning conditional independence. Thus an important unification is reported in this paper.

Another aspect of statistics that Professor Dawid's paper illuminates is a duality between the sample space and the parameter space. The relation among several concepts, such as sufficiency of a statistic and sufficiency of a parameter, are made much clearer using the very general approach suggested by this paper. Although he declares his paper to be philosopically neutral, several of these dualities are expressible only in terms of a (proper) subjectivist view of probability theory.

This subjectivist view is strained in Professor Dawid's treatment of Simpson's paradox. How are we to be sure, when considering some set of covariates Z^* which we would like to believe is sufficient, that tomorrow some brilliant scientist will not find an extremely good predictor of treatment success and failure, one which was excluded from Z^*? If we hold dogmatically to the sufficiency of Z^*, then, foolishly, we are unable to learn from the work of this hypothetical brilliant scientist. Thus the flat declaration that I consider Z^* to be a sufficient set of covariates (with probability one) is more than I am willing to say of a set of covariates in any applied problem I have encountered. I am, however, willing to specify a set of covariates and do an analysis as if I believed them to be a sufficient set. In deciding to do such as an "as if" analysis, I take into account my current beliefs about the phenomenon I am trying to predict, the sensitivity of the conclusions I expect to draw to this assumption, and how serious that sensitivity is to my goal (that is, my utility function enters importantly). A special case of this problem is the question of choosing regressions in econometric models, for which see Kadane and Dickey (1979), where our views are spelled out at greater length. But suffice it to say here that having a formal definition of the sufficiency of a set of covariates Z^* does not resolve Simpson's paradox for a subjective Bayesian, in my view. Thus I agree with Professor Dawid about a "weak link" at the very beginning of the chain of analysis, but doubt that much reinforcement can come from a statistician's informed judgement that the condition is met. We make the assumption not because we believe it to be true, but rather because we find it useful.

Professor J. M. DICKEY (University College of Wales, Aberystwyth): It has seemed like full-time activity lately just to keep informed on Professor Dawid's work. The present paper is an important product of years of his considering an apparent diversity of problems and theories which are now unified through the view of conditional independence.

I find the following heuristic way of thinking useful in considering the mathematical content of conditional independence. The relation $X \perp\!\!\!\perp Y \mid Z$ requires that ordinary independence holds for each of a family of joint distributions for X and Y. This family is formed from the conditional distributions in an overall joint distribution for X, Y, Z. However, the latter requirement can be generalized to situations where Y and/or Z are not random variables, or are not jointly distributed with X (Dawid, 1979).

The "elementary properties" of conditional independence are easily remembered in the following forms:

Lemma 4.1 seems empty of content when stated in terms of sigma fields.

Lemmas 4.2 and 4.3 have fully equivalent forms stated without the conditioning on Z. They hold identically as theorems without Z if and only if they hold as theorems for classes of distributions defined by the values of a variable Z.

Lemma 4.2. If $X \perp\!\!\!\perp Y$ and $U \equiv (X)$, then (i) $U \perp\!\!\!\perp Y$, (ii) $X \perp\!\!\!\perp Y \mid U$.

Lemma 4.3 has a similarly simplified statement. In the case that X and Y are joint random variables, the converse of Lemma 4.2 is Lemma 4.3 under the identification

Lemma 4.2		Lemma 4.3
Y	$\leftrightarrow$	X
X		(Y, W)
U		Y

I was surprised to read that a statistic T sufficient for Θ based on X remains sufficient for selected data $X \mid Q$ (Section 7.3). However, note that this holds under the explicit assumption that selection depends only on the data which form the basis for the report, $Q \perp\!\!\!\perp \Theta \mid X$. So T and Q are both based on the same X. This excludes the common situations where Q selects a subset or other distortion, $Z_Q = Z_Q(X)$, and then one wonders what to make of the data (Z_Q, Q). The sufficiency or not of a statistic based on Z_Q will depend on the form of Q. For example, the number of successes in a selected Bernoulli subsequence is not, in general, a sufficient function of the subsequence. Dawid and Dickey (1977) treat such questions of distortion of the face-value inference by report-selection processes. "Unknown" selection processes are particularly important in practice.

Mr S. M. Rizvi (National Coal Board): In the fourth line of Section 2.1 Professor Dawid defines independence: "We write $X \perp\!\!\!\perp Y$ to denote that X and Y are independent." This prompts the question: independent of what? This question can be interpreted in three different ways:
(1) X and Y are independent of some mysterious element Z;
(2) X and Y are independent of each other; and
(3) X is independent of Y, without saying that Y is independent of X.
His formulations enshrined in (IIa) and (IIb) are the consequences of the third interpretation; Theorem 2.1 is a consequence of the second. The word "are" in his definition of $\perp\!\!\!\perp$ supports the second interpretation. The third interpretation gains substance from his assertion that $X \perp\!\!\!\perp Y$ if any information received about Y does not alter uncertainties about X. There is a fundamental conflict between the second and third interpretation in that the second interpretation imparts a bidirectional attribute to the relational operator $\perp\!\!\!\perp$ by definition; whereas the third interpretation assigns to it a unidirectional character. The latter is logically robust as any relational operator R must initially denote only a unidirectional relationship between a given pair X, Y; thus leaving the possibility of a symmetrical, non-symmetrical, asymmetrical and anti-symmetrical character for R to be settled on a logico-empirical basis.

In his application of Theorem 2.1 to the investigation of cancer development from some carcinogenic agent, I suspect Professor Dawid is in an inferential trap. In real-life problems of this nature when one relates X to Y, one really is testing the hypothesis:
(a) whether X and Y have an association (a symmetrical relationship);
(b) whether X and Y have a causal link (an asymmetrical relationship); and
(c) whether X depends on Y (a non-symmetrical relationship).
Supposing his formulation of Theorem 2.1 was correct, and supposing his null hypothesis of independence were to fail, thus implying either association or causality or dependence, then Professor Dawid would not be able to reach logically valid conclusions without framing the following theorems:
Theorem.
(a) If X[ass] Y, then Y[ass] X; where [ass] denotes "has association with".
(b) If X[d] Y, then Y[d] X; where [d] denotes "is dependent on".
(c) If X[cau] Y, then Y(cau) X; where [cau] denotes "causes", and (cau) denotes "is caused by".
It is useful to consider Professor Dawid's Table 1 in terms of indices as follows:

TABLE 1A (indexed)

	Male		*Female*	
	R	$\bar{R}$	R	$\bar{R}$
T	67	33	40	60
$\bar{T}$	57	43	25	75

when consolidated over sex Table 1A looks as follows:

TABLE 1A (unisexed)

	R	$\bar{R}$
T	107	93
$\bar{T}$	82	118

One would not fail to make the same inference about the efficacy of the treatment from the consolidated table as one would from the fragmented table. Perhaps Simpson's paradox has more faces to it than is appreciated!

Mr E. F. HARDING (Cambridge University): This is indeed a superb paper, on which it is difficult to comment when it seems to settle so many questions. I would like to make just two points simply.

The first concerns Section 2.2 which, apart from the Bayesian paragraph, says only that S, whose distribution does not vary with θ, provides no information about Θ; which may be taken as a principle of inference. The section appears to say more, in "turning Theorem 2.1 [unprovable; indeed meaningless until $\Theta \perp\!\!\!\perp S$ is defined] into a principle of inference". First, here $S \perp\!\!\!\perp \Theta$ denotes only $p(s \mid \theta) = a(s)$; second, "Theorem 2.1" here only defines that $\Theta \perp\!\!\!\perp S$ denotes the same as $S \perp\!\!\!\perp \Theta$; thirdly, interpreting $\Theta \perp\!\!\!\perp S$ as "S provides no information about Θ" is the same, therefore, as interpreting $S \perp\!\!\!\perp \Theta$ likewise, which is the above "principle". I am commenting on presentation, not content: using different notations $S \perp\!\!\!\perp \Theta$ and $\Theta \perp\!\!\!\perp S$ for the same thing, with evoked overtones of the stochastic independence they denote in their different meaning in another context, appears (without really doing so) to beg big questions. The same point occurs in Section 3.2, implicitly perhaps also in Sections 3.3 and 8.2.

In passing, the above "principle" needs interpreting, as shown for instance by the existence of ancillary functions S of a minimal sufficient $T = (R, S)$ say (example: $s = x_{\max} - x_{\min}$ for the uniform distributions on $(\theta, 1+\theta)$)—when can we maintain that $\inf(., s)$ is independent of s? Under various interpretations of inf (hypothesis-testing, estimation, confidence intervals variously derived, Bayesian methods, etc.) this is sometimes so, sometimes not. There is an important distinction between such ancillaries and ancillaries S^* of which no function is any function of T, which can be hidden by attaching inferential interpretations to the notation $S \perp\!\!\!\perp \Theta$ (or, worse, $\Theta \perp\!\!\!\perp S$!).

My second point concerns Section 5. I do not consider the Simpson phenomenon paradoxical, but *fallacy* can arise if data are inappropriately pooled. Suppose Table 1 gives the results of randomly allocating T to $150M$ out of 430 and to $250F$ out of 330: then, for instance, we can report the data allowing estimation of expected frequencies were T to be allocated to a random 400 individuals out of 760 (sexless) by pooling as the weighted (430 : 330) sum, namely 220/180//160/200 (to nearest integers). If T is beneficial for both M and F it must remain so in the collapsed table. To use the simple sum for this purpose is an arithmetic fallacy like using $\frac{1}{2}(\bar{x}+\bar{y})$ as the combined mean of two unequal samples. On the other hand, if getting T depends on access to a treatment centre, then the row-margins, regional incidence of the disease, recovery rate and sex may covary all together. Here no causal inference about T is possible, but knowing the T-category of an individual enables prediction of his recovery, for which the appropriate collapsed (sexless) table is the simple sum.

This is implicit in Professor Dawid's very illuminating analysis of covariates. It is more nearly explicit in Professor Lindley's reference to randomization in the experimental situation. Above are quite explicit details of just two possible situations and I think it is worth while looking at it in this way.

Professor G. A. BARNARD (Essex): Professor Dawid, as we have come to expect, presents us with a set of wide ranging and original ideas in this paper, and we must be grateful to him for the resulting stimulus to rethink basic ideas. And I for one would grant his claim to the extent that he has shown that the formal similarities between many of the ideas involved in statistical inference are well worth noting, if only from the point of view of providing an introduction to deeper understanding. He is also to be congratulated on the skill with which he walks the tightrope, as we all

must, nowadays, between the smooth and easy, but perhaps not very fertile territory of simplistic frequentism, and what Professor Bartlett has referred to as the Bayesian bog.

Professor Dawid seems to suggest that it is possible to take a notion of conditional independence as fundamental, to provide it with a set of formal axioms and on this basis to develop a reasonably complete theory of statistical inference. Of course he does not claim to have done this in the present paper. I would look forward very much indeed to seeing such an attempt made, even though I feel it would be likely to come up against some very severe snags. For the present it seems to me that the various ideas which can be formalized by Professor Dawid's notation have no more than a formal similarity—useful no doubt, but not to be regarded as fundamental. Many years ago I attempted to draw a distinction between "statistical independence" and "absolute independence". The first refers to the property expressed by Professor Dawid's I and II, A and B, but the second relates to something more fundamental which might be described as "absence of causal link between". From a formal point of view a difference between these concepts can be seen to consist in the fact that with three variables, U, V, W, we may have pairwise statistical independence but not total independence as a triplet: but if in fact there are no causal links between U and V or between V and W nor between U and W then there are no causal links between the three events as a group. From a more semantic point of view the difference lies in the fact that statistical independence is a property of populations, while absolute independence is not.

One's feeling that Professor Dawid may have overlooked this distinction is perhaps strongest in relation to his Section 5. Simpson's paradox seems to me to involve little more than our naive tendency to think that the sum of two fractions can be obtained by adding numerators and adding denominators. I do not think that it could ever be for a statistician, as such, to make an "informed judgement" about whether all possible causal factors have been examined. The most he can do on the basis of observing samples from a population is to make statements about that population, leaving open the question whether in another population another set of relationships might hold.

Dr P. J. Bickel (University of California): Professor Dawid has given us an interesting overview of the role of conditional independence in aiding our understanding of the structure of statistical models. A particularly fruitful field of application is log-linear models where fundamental work has already been done by Goodman (1971) and Haberman (1974). For instance, one can essentially restate one of Goodman's (1971) results as follows. Hierarchical models for complete m-way contingency tables with all main effects present yield closed form (in a suitable sense) maximum likelihood estimates of the cell probabilities if and only if the cell probabilities vary subject to the following restrictions (and only these): There is a partition of the m variables of the table into disjoint subsets $A_1, \ldots, A_k$ of (possibly) unequal size with the following properties.

For each A_j, $j \geq 2$, $\exists$ a set of variables B_j such that:

(i) $B_j \subset \bigcup_{i<j} A_i$, ($B_j = \emptyset$ is possible)

(ii) If $A_{i_1} \cap B_j \neq \emptyset$, $A_{i_2} \cap B_j \neq \emptyset$, and $i_1 < i_2 < j$, then $A_{i_1} \cap B_j \subset B_{i_2}$.

(iii) $A_j \perp\!\!\!\perp \bar{B}_j \mid B_j$, where $\bar{B}_j = [\bigcup_{i<j} A_i] - B_j$.

This remark follows from Theorems 3.4.1 and 3.4.2 in Bishop *et al.* (1975), for instance. The form that the estimates must take is easy to see from this formulation.

Restriction (ii) says that if variables included in A_{i_1} and A_{i_2} are needed for predicting A_j then all such variables included in A_{i_1} are also needed for predicting A_{i_2}. This restriction is equivalent to the hierarchical property of the models. By dropping it one gets all models which are simply interpretable and, I believe, all models which have closed form maximum likelihood estimates of the cell probabilities.

Turning briefly to Professor Dawid's remarks concerning Bickel *et al.* (1977) I always felt that the main inadequacy in our study was the insufficiency of Z, in our case the department applied to. The difficulties in determining whether X enters Z are discussed in my correspondence with Kruskal in Bickel *et al* .(1977). In any case, there is no interview for admission to our graduate school.

Professor M. H. DeGroot (Carnegie–Mellon University): I congratulate Phil Dawid on a nicely written paper which clearly shows the relevance of the concept of conditional independence to a wide variety of areas of statistics. One area that he does not mention in which this concept is

useful is the selection of a regression model. Suppose that a choice must be made between two regression models M_1 and M_2. Under M_1, the regression function of the dependent variable Y is specified as a function of some vector $\mathbf{X}$ of independent variables, and under M_2 the regression function of Y is specified as a function of a vector $\mathbf{W}$ of other independent variables. In order to be sure that the models M_1 and M_2 are mutually exclusive, it is appropriate to impose the additional conditions that $Y \perp\!\!\!\perp \mathbf{W} \mid \mathbf{X}$ under M_1 and that $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{W}$ under M_2. This approach is followed in Davis and DeGroot (1978). Other areas in which notions of conditional independence can be helpful are the study of qualitative probability, where Theorem 3.1 need not hold, and the study of sufficient experiments, rather than sufficient statistics in a given experiment.

The concept of conditional independence can also be embedded in a more general setting. Given some joint distribution of X, Y and Z, we might ask whether X and Y are conditionally independent given a particular partition (or subfield) Π of the sample space $\mathscr{Z}$ of Z. We might denote this relation $X \perp\!\!\!\perp Y \mid \Pi$. Let Π_0 denote the trivial partition containing just the two sets $\mathscr{Z}$ and ϕ, and let Π^0 denote the total partition containing each individual point of $\mathscr{Z}$. At one extreme there is the possibility that $X \perp\!\!\!\perp Y \mid \Pi_0$, which means that $X \perp\!\!\!\perp Y$. At the other extreme there is the possibility that $X \perp\!\!\!\perp Y \mid \Pi^0$, which means that $X \perp\!\!\!\perp Y \mid Z$. Can we characterize problems in which there will exist a partition Π such that $X \perp\!\!\!\perp Y \mid \Pi$? Can we find coarsest and finest partitions for which this property holds? In problems of sufficient statistics, we are interested in coarsest partitions. In problems relating to Simpson's paradox, we are interested in finest partitions.

The adequacy of (5.4) is considered in the final paragraph of Section 5. I believe that the only possible formalization of "the notion that X does not enter into Z" is that $Z \perp\!\!\!\perp X$. But the critical issue is not whether $Z \perp\!\!\!\perp X$, but whether Z and a student's performance V in the program are independent. Thus, the use of an applicant's height in deciding on admission should be regarded as a form of sex discrimination if height is not relevant to one's ability to succeed in the program. Performance V is an important variable missing from the discussion in Section 5. The justification for having acceptance Y depend on Z is that V depends on Z. If $V \perp\!\!\!\perp Z$ but neither $Z \perp\!\!\!\perp X$ nor $Y \perp\!\!\!\perp Z$ is true, then there is sex discrimination (against one sex or the other).

Professor D. A. S. FRASER (University of Toronto): Conditional independence involves two of the most basic concepts in statistics—independence and conditional probability.

Independence provides the primary means by which we build larger statistical models from smaller initial components.

Conditional probability extends these means for building larger models, but also—of fundamental importance—it has to do with what probabilities are the proper or correct probabilities in any particular application.

The paper claims that conditional independence is more than a useful tool and that it offers a new language for expressing and studying statistical concepts. Independence as a concept arose asymmetric—the conditional of X is independent of Y—was symmetrized in its pure form by the more mathematically inclined, and now is re-emphasized in its original, more natural asymmetric form.

The central contribution of the paper is the notation for independence, both marginal and conditional. The symbol for orthogonality is useful in mathematics and statistics—so also the symbol for independence, $\perp\!\!\!\perp$, will be useful in statistics, and its arrival on the scene is overdue.

I am sceptical, however, that the notation can contribute to statistical inference as suggested in the paper. The concepts of sufficiency and ancillarity are already heavily overburdened with various theoretical results—and these totally dwarf the limited range of applications of the concepts. Rather, for statistical inference, attention is needed to the applications of probabilities, the determination of the *appropriate conditional probabilities*. This is an area of great need. We have progressed somewhat from the lack of clarity of some years ago when a prominent statistician asserted that the probability for an event that has occurred is zero or one and in context we may not know which, and yet in his introductory text asked for the probability of a finesse or a split given the observed two hands in a bridge distribution! There is much to be done in the area of inference and, in particular, in the determination of appropriate conditional probabilities. We will need more than the notation that expresses concepts we now have in hand.

Mr R. F. GALBRAITH (University College London): Professor Dawid has emphasised the concept of conditional independence in statistical inference. An explicit notation and theory for this

property can also be of value in some areas of probability theory and it is surprising that its use is not more widespread. One such area is that of probability models for processes in two and higher dimensions, where notation using probability densities can be cumbersome and even elementary probability manipulations hard to follow.

As one example, consider the binary process $\{X_{ij}\}$ on an $m \times n$ rectangular lattice discussed by Pickard (1977). This process has the properties that successive rows form a vector Markov chain, as do successive columns, and that these properties remain true when the process is restricted to an $s \times t$ sublattice. Theorem 5 of that paper asserts that the conditional distribution of X_{ij} given all other site variables depends only on the neighbours—those site variables that are horizontally, vertically or diagonally adjacent. This may be expressed as

$$X_{ij} \perp\!\!\!\perp L_{ij} \mid N_{ij}, \tag{1}$$

where L_{ij} denotes all variables other than X_{ij}, and N_{ij} denotes the neighbours of X_{ij}. A proof using probability densities is indicated by Pickard but the following is more easily understood (especially with the aid of a diagram).

The nearest neighbour property of Markov chains applied to the rows says

$$R_i \perp\!\!\!\perp S_i \mid (R_{i-1}, R_{i+1}),$$

where R_i denotes all variables in the ith row and S_i denotes all other variables. If we now denote by U_{ij} all variables in the ith row except X_{ij} we have, by Lemmas 4.2 and 4.1,

$$R_i \perp\!\!\!\perp (S_i, U_{ij}) \mid (R_{i-1}, R_{i+1}, U_{ij}).$$

In particular

$$X_{ij} \perp\!\!\!\perp L_{ij} \mid (N_{ij}, P_{ij}), \tag{2}$$

where the conditioning variables have been re-expressed in terms of N_{ij} (the neighbours of X_{ij}) and P_{ij} (everything in rows $i-1$, i and $i+1$ except for N_{ij} and X_{ij}). Now the same argument applied to the columns of the $3 \times n$ sublattice consisting of rows $i-1$, i and $i+1$ must give the corresponding statement

$$X_{ij} \perp\!\!\!\perp P_{ij} \mid N_{ij} \tag{3}$$

and, by Lemma 4.3, (3) and (2) together imply (1).

Instead of $X \perp\!\!\!\perp Y \mid Z$ some authors have used the notation $X \mid (Y, Z) \overset{D}{=} X \mid Z$. This latter seems inferior as it is less concise and obscures the symmetry between X and Y, a view that is confirmed when one tries to prove the converse of Lemma 4.3. On the other hand, it states the property differently (saying two things are the same, rather than saying one thing does not depend on another). The argument of the previous paragraph becomes

$$R_i \mid S_i \overset{D}{=} R_i \mid (R_{i-1}, R_{i+1}) \Rightarrow R_i \mid (S_i, U_{ij}) \overset{D}{=} R_i \mid (R_{i-1}, R_{i+1}, U_{ij}) \Rightarrow X_{ij} \mid L_{ij} \overset{D}{=} X_{ij} \mid (N_{ij}, P_{ij}) \overset{D}{=} X_{ij} \mid N_{ij},$$

which seems more obvious than the first version and there is apparently no need for Lemma 4.3.

Professor S. Geisser (University of Minnesota): It is remarkable how many concepts can be artfully woven together by the thread of conditional independence and its discovery no less so. However, where this unification leads us and what it achieves is obscure. In fact, in some respects it conceals almost as much as it reveals. Specifically, the notation

$$X \perp\!\!\!\perp Y \mid (Z, W) \tag{1}$$

cloaks a variety of factorizations—to name but a few in an obvious and more informative notation, though admittedly more cumbersome, that satisfy (1), we can have

$$(X, Y \mid Z, W) = (X \mid Z, W)(Y \mid Z, W)$$

or

$$(X, Y \mid Z, W) = (X \mid Z)(Y \mid W), \quad \text{etc.}$$

Thus the relationship $D \perp\!\!\!\perp Q \mid S$, in Section 7.2, derives from the easily obtained factorization $(D, Q \mid G, S) = (D \mid S)(Q \mid G)$, which displays the full anatomy of the situation. This particular case notwithstanding, such a factorization will often be considerably more enlightening than the less precise $D \perp\!\!\!\perp Q \mid (G, S)$, granted that the latter notation remains sufficient for the ostensible unification aims of the paper.

Professor V. P. GODAMBE (University of Waterloo): My comments would be restricted to Section 2.2 of Professor Dawid's very interesting paper. In Section 2.2 he introduces a *principle of statistical inference*: "no information about the parameter can be extracted from observing an ancillary statistic". In Professor Dawid's notation "if $s \perp\!\!\!\perp \theta$ then $\theta \perp\!\!\!\perp s$". But this principle is contradicted by the following obvious mode of inference. Let a finite population P consist of N individuals, *i.* $P = \{i: i = 1, ..., N\}$. With P is associated a parameter $\theta = (\theta_1, ..., \theta_i, ..., \theta_N)$ where each $\theta_i = 1$ or 0. It is *known* that $\Sigma\theta_i = N/2$ summing over i from 1 to N. But it is *not* known for which individuals i, $\theta_i = 1$ or 0. Hence the total number of possible values for θ are $\binom{N}{\frac{1}{2}N}$. Let s be a *random sample* of size n drawn (without replacement) from P; $s \subset P$. Now the very meaning of randomization implies the following *inference* about θ on the basis of s. "For a number k, suitably specified in terms of n, N, the values of θ for which $|\Sigma\theta_i/n - \frac{1}{2}| > k$ (summing over $i \in s$), are implausible, s being the random sample actually drawn." This I submit contradicts the above principle of statistical inference, namely "if $s \perp\!\!\!\perp \theta$ then $\theta \perp\!\!\!\perp s$", for the distribution of s is independent of θ $\left(p(s \mid \theta) = 1\Big/\binom{N}{n}, \forall\, \theta\right)$ but the inference contradicts the assertion $\theta \perp\!\!\!\perp s$. For instance, let $N = 10$, $n = 4$, $\theta^1 = (0000011111)$ and $\theta^2 = (0101010101)$. Then on the basis of $s = (1, 4, 5, 10)$ we prefer θ^2 and reject θ^1. Yet the distribution of s is identical on θ^1 and θ^2. Almost all statisticians with whom I discussed this agreed that they would accept the just-mentioned inference based on the random sample. Of course, there were some notable disagreements. I would be interested to read Professor Dawid's reaction. I am not quite clear about the relationship of the above principle of statistical inference with the conditionality principle as introduced by Birnbaum (1962). However, Birnbaum's sufficiency principle (and therefore his likelihood principle) clearly implies the above principle. But I plan to elaborate upon this in an independent publication.

I should emphasize that the mode of inference discussed in the preceding paragraph is free of any assumption concerning prior probabilities of possible θ's. This is clarified further by the above illustration with just two vectors θ^1 and θ^2.

Apart from the case discussed here, I believe there may possibly be many other cases where the principle of statistical inference "if $s \perp\!\!\!\perp \theta$ then $\theta \perp\!\!\!\perp s$" is contradicted. Possibly these cases could be unearthed by analysing situations where one feels reluctant to condition on a certain ancillary statistic. The underlying feeling is that the ancillary statistic (with its distribution) contains information about the unknown parameter.

The statistical inference of the type mentioned above in connection with simple random sampling $\left(p(s \mid \theta) = 1\Big/\binom{N}{n}, \forall\theta\right)$ can also be made for more complicated unequal probability sampling designs. But this again will be discussed in an independent publication.

Professor D. V. HINKLEY (University of Minnesota): Professor Dawid has nicely demonstrated the variety of his interests and has tied them together with a thin thread made of $\perp\!\!\!\perp$. Some claims are made, particularly at the beginning of the paper, but the present results do not suggest to me that we have a powerful new tool here. However, the later parts of the paper hold definite promise.

I was somewhat disappointed by Section 5, where the initial purpose dissolves after equation (5.4) with the necessary concession to subjective judgement. As always, the difficulty is that of giving a simple mathematical representation for the complex process of scientific inference. Although Professor Dawid is strictly interested in causal inference, rather than association, it is not entirely out of place to remark that analyses and tables for different sets of covariates will often be meaningful, particularly when covariates are approximately confounded. The question posed at the end of Section 5 is a subtle one, both for us and the U.S. Supreme Court!

The topic of prediction sufficiency has been of some interest to me since my dissatisfaction with the classical frequency treatment of predictive inference led me to examine definitions of predictive likelihood. Although my own work (Hinkley, 1979) was not closely related to the mathematical theory of the present paper, it did raise some interesting points. For example, where an adequate function of X may be *larger* than the usual minimal sufficient function, that part of Y which requires prediction *via* X may be *less* than the minimal sufficient function of Y.

One of Professor Dawid's aims is to evoke more explicitly the "duality between data and parameters", which I think is laudable. I believe that R. A. Fisher's appreciation of this duality,

and his corresponding understanding of the nature of probability models, were the key ingredients of fiducial theory. I have no doubt that Professor Dawid will use the language of conditional independence in his successful efforts to aid our understanding of statistical theory.

Professor H. KUDŌ (Kwansei Gakuin University, Japan): I welcome the opportunity to comment on this excellent paper by Professor Dawid, which unifies many concepts in statistical theory in terms of conditional independence. I agree completely with Professor Dawid's unification and am grateful to him for his presentation of such a beautiful theory.

Several concepts of relation among variables, which in this paper Professor Dawid rearranges in terms of conditional independence, have been treated as distinct subjects in the history of statistical research. Such a traditional way of treatment of the problems is mainly based on the fact that the relevant variables are characterized into several categories—especially, those of parameter and statistic. Why should we make such a strict distinction between parameter and statistic? The reason seems to me to have a very close connection with the notion of cause–effect relation *via* the old and traditional concept of probability. However, nowadays, our view of probability has changed greatly from the old concept of probability, which was strongly connected with causality.

An alternative and more natural classification of variables appearing in statistical research is the distinction between the "observable" and the "unobservable". The observable is a variable whose value becomes known after, and just after, the relevant observation has been accomplished, while the unobservable is one whose value remains unknown at that moment.

On the other hand, one might easily think that the variables in statistical problems are all random variables, irrespective of whether their probability distribution is (partly or completely) known or not. Here it should be noted that knowing the existence of the probability distribution of a variable does not imply getting any information about the value of the variable concerned. If it is known that the distribution μ of a random variable X belongs to a certain class Σ of distributions, then every statement about the value of X can be allowed to have an exceptional set (a set of measure zero w.r.t. every $\mu \in \Sigma$). Thus the problem of what kind of universality of the statements is attached to a category of variable (e.g. to a parameter, to a statistic or to a future variable in a prediction problem) should be solved to complete the theory in Professor Dawid's paper.

It is hoped that this paper by Professor Dawid may give rise to a new arrangement of statistical theory.

Professor K. V. MARDIA (University of Leeds): In Examples 8.2 and 8.3, it should be noted that if $C(\Psi)\exp(\boldsymbol{\theta}'\mathbf{x})$, $\Psi = \|\boldsymbol{\theta}\|$, is von Mises–Fisher distribution $M(\boldsymbol{\theta})$, then the distribution of $\mathbf{S} = \Sigma \mathbf{X}_i \mid U = \|\mathbf{S}\|$ is simply $M(U\boldsymbol{\theta})$. Hence the distribution of $\mathbf{S}$ given U and n is not only the same for all n but is of the same form as the parent distribution.

Suppose one wishes to test the hypothesis $\eta = \eta_0$ ($\eta = \theta/\Psi$) based on $M(U\boldsymbol{\theta})$ against $\eta \neq \eta_0$ where n is given. We can then obtain a confidence interval for η if Ψ is known. However, if Ψ is unknown, this method can be extended but any such extension will involve estimating the parameter Ψ. One "efficient" method is to estimate Ψ from $M(\boldsymbol{\theta})$ through the maximum likelihood method. An alternative approach to the above null hypothesis is to use a conditional test based on $U \mid C$ where $C = \Sigma \mathbf{X}_i' \boldsymbol{\eta}_0$ since the distribution of $U \mid C$ does not involve Ψ. How does this latter method fit in Professor Dawid's framework where there is a nuisance parameter? The second method also provides a confidence interval for η. Which of the two confidence zones is theoretically most satisfactory in the presence of nuisance parameter Ψ? For a detailed discussion of this problem, see Mardia (1975, section 4).

Professor M. MOUCHART (Catholic University of Louvain, Belgium): In the theorem of Section 6.3, one may drop the conditions that W_n is a function of η_n and that T_n is a function of ξ_n; i.e. this theorem is still valid when written as follows: "if (i) $\eta_n \subset \xi_n$, (ii) $W_n \subset T_n$, (iii) $T_{n+1} \perp\!\!\!\perp \xi_n \mid T_n$, (iv) $T_n \perp\!\!\!\perp \eta_n \mid W_n$ then (v) $W_{n+1} \perp\!\!\!\perp \eta_n \mid W_n$" where "$\subset$" among random variables is read as "is a function of". This is so because the first three conditions already imply $\eta_n \perp\!\!\!\perp T_{n+1} \mid T_n, W_n$. Note also that this theorem is true both in the sampling process (i.e. properties (iii), (iv) and (v) understood as conditionally on Θ) and in the predictive process (i.e. after integrating out the parameters).

These (fairly trivial) extensions call, in my opinion, for two more general comments. Firstly, conditional independence is a simple, yet rather powerful, concept, well known to statisticians,

worked out as early as the fifties by C. Stein in 1950 (in unpublished work) and "rediscovered independently by Burkholder in 1958, Hall in 1959 and Ghosh in 1960" as mentioned in the preface of Hall *et al.* (1965)—see also Martin *et al.* (1973) and the references in Professor Dawid's paper. I find it therefore somewhat surprising that this concept has not been used more systematically in the textbook literature and more efficiently in the specialized literature. These are good reasons for thanking Professor Dawid for his clear and lucid survey; it should help to popularize a method which deserves wider use.

A second comment is that, in my opinion, conditional independence is basically a probabilistic concept: $X \perp\!\!\!\perp Y \mid Z$ makes sense only if X, Y and Z are endowed with a well-defined probability distribution; in this case, the concept is symmetric in X and Y (in this respect, I found that Theorems 2.1 and 3.1 give more trouble than information by avoiding a clear (and symmetric) definition of independence in probability). I therefore tend to disagree with the last sentence of Section 3.1: to give meaning to $X \perp\!\!\!\perp Y \mid \Theta$ we need either to consider a well-defined (and proper) distribution on Θ or to consider a family of (possibly, all) distributions on Θ, the difference being essentially a matter of null-sets to be taken into account in condition (1) or (2) of the paper. Even if the independence property is stated "for every Θ" (instead of "for almost every Θ"), Θ has still to be considered as subject to (a family of) probability distributions and the statement is concerned with a (family of) induced probability(ies) on (X, Y, Θ). Failure to recognize this basic fact may induce "fallacious arguments", the marginalization paradox mentioned in Section 3.1 being just one example. Furthermore, this probabilistic framework gives precisely the power of conditional independence as a tool for statistical inference: it allows a systematic treatment of reductions on both the parameter and the sample space (see, for example, Florens and Mouchart, 1977) to eliminate nuisance parameters, to treat prediction problems (as in Section 6) or to handle partially observable processes. The full strength of this tool is still better appreciated when combining conditional independence with other properties, such as invariance (see, for example, Hall *et al.* 1965), measurable separability or strong identification (see, for example, Mouchart and Rolin, 1978). I think therefore that even in a sampling theory framework, it is preferable to stick to a full probabilistic approach to conditional independence (by considering properties defined on a family of probabilities induced by the set of all prior probabilities) instead of introducing a well-defined concept for Bayesian analysis and a "loose" one for a sampling theory analysis.

Dr I. NIMMO-SMITH (MRC Applied Psychology Unit): It is worth examining in more detail the "logical equivalence" established between (5.1) and (5.3), given (5.2); it turns out that the result depends on the universe of discourse in an asymmetric, but satisfactory, way. For let $A \subseteq X \times I$ denote a fixed "conceptual entity" of *applicable* treatment-to-unit allocations, and assume that (5.1), (5.2) and (5.2)+(5.3) have the meanings

$$p(y \mid x, i, z) = p(y \mid x', i, z), \tag{A1}$$

$$p(y \mid x, i, z) = p(y \mid x, i', z), \tag{A2}$$

and

$$p(y \mid x, i, z) = p(y \mid x', i', z) \tag{A3}$$

for all $x, x' \in X$, $i, i' \in I$ and $z \in Z$ where consistent and A-applicable. Then, regardless of the structure of A, (A2) and (A3) imply (A1), for $(x, i) \to (x', i)$ is a special case of $(x, i) \to (x', i')$. However (A1) and (A2) imply (A3) only if each valid substitution $(x, i, z) \to (x', i', z)$ may be made from a chain of ones like $(x, i, z) \to (x', i, z)$ and $(x, i, z) \to (x, i', z)$. This is equivalent to the connectedness of A within the strata of Z, or to the restriction that there should be no complete "conceptual" confounding, in any Z-stratum, between a subset of treatments and a subset of units. Now, for a fixed A, one may consider the effect of different sufficient covariates Z on the implication (A1) and (A2) $\to$ (A3), and one sees that the implication holds so long as Z contains the information of the "applicability" of treatments. Clearly one can use the above considerations to extend the approach of Dawid (1977a) to less complete and less balanced designs, the group being subsidiary to its orbits or (as here) derived equivalence classes.

A measurement technique which relies (perhaps optimistically) on conditional independence is described by Warner (1965) and Boruch (1972).

On another tack, there is a sense in which this paper is largely irrelevant to being, or understanding, a day-to-day, probability-revising person. The information he uses is rarely independent,

even conditionally, and yet has to be accommodated sequentially. In a recent paper, Navon (1978) has observed that treating this sequence of information as independent typically may lead to over-extreme revisions of probabilities. Turning this on its head, he suggests that the experimentally determined "conservatism" of humans (e.g. Edwards, 1968) may in part be explained by an inappropriate application of a natural facility that corrects (correctly!) for non-independence. This may indicate an important obstacle that confronts those who hold a normative brief for the Bayesian model of human inference processes. In no way does it diminish my thanks to Professor Dawid for his valuable and neat paper.

Professor M. R. Novick (University of Iowa): Professor Dawid has asserted that conditional independence is fundamentally important in statistical inference. I concur. I have for many years used his statement (Ia) as a definition of independence in my teaching, and my related work (with Lindley) on causal inference and Simpson's paradox goes back to 1975.

I shall suggest now, in response to Professor Dawid's call, that another major application of the concept of conditional independence ought to be identified. I note that Professor Dawid has referred to Birnbaum's classical paper on foundations of inference, but has not referred to his work on latent trait theory (Birnbaum, 1968) in which the concept of conditional independence finds central application. In fact, there is no reference in Professor Dawid's paper to any publication in *Psychometrika* or the psychometric literature generally, a literature in which conditional independence is used extensively.

Consider the joint distribution of $(X_1, X_2, \ldots, X_N)$ and suppose that for $n < N$ given $(\theta_1, \theta_2, \ldots, \theta_n)$, but not otherwise, the variables $X_1, \ldots X_2, \ldots, X_N$ are independent. The attainment of this local or conditional independence through the construction of the so-called latent variables θ is what psychometric latent trait theory is all about. (See, for example, Lord and Novick, 1968, Chapter 16; and Anderson, 1959; or almost any issue of *Psychometrika* for the past quarter century.) Classical test theory and factor analysis are parallel, though less satisfactory theories, in which conditional independence is replaced by uncorrelatedness. However, Novick (1965; see also Lord and Novick 1968, Chapters 2 and 24) has shown that these theories can be re-expressed in terms of conditional *linear* independence, which is a small but meaningful improvement. The work of Birnbaum (1968) and the earlier work of Lord (1952), however, correctly show that conditional independence is the key concept of structural modelling. Latent trait theory is more fundamental than a factor analytic decomposition.

To characterize this work in simple, if imprecise, language we can say that a structure for a system or set of variables has been provided if we can write structural parameters as functions of the observables and if, given the structural parameters, the observables are conditionally independent. Of course, nothing useful has happened unless n is very much less than N, and in practice we must disregard small eigenvalues and settle for near conditional independence. Similar applications have been made in biometrics and econometrics.

I am pleased to join Professor Dawid in looking forward to learning of further applications and examples of conditional independence from others.

Professor Donald B. Rubin (Educational Testing Service, Princeton): I want to congratulate Professor Dawid on a lucid discussion of conditional independence, a concept fundamental for building statistical models. I hope that Professor Dawid's presentation and examples are not read as implying that the primary practical objective of statistics is testing hypotheses, e.g. whether cancer is independent of an agent or whether selection to school is conditionally independent of sex given a test score. My experience suggests that the estimation of effects defined in terms of observable quantities is of far more interest than the testing of idealized mathematical models. In this context of trying to draw inferences about observable quantities, Professor Dawid's statement that to a Bayesian ". . . the distinction between data and parameters is largely irrelevant" is mathematically correct but statistically misleading unless the parameters are phenomenological in the sense of being essentially equivalent to functions of observables. I believe that Professor Dawid's discussion of causal inference suffers from being non-phenomenological.

The emphasis on the central role of a sufficient set of covariates (i.e. effectively all causally relevant covariates) and the conclusion that "At some stage it is necessary to make a subjective judgement that a set of covariates Z^* is sufficient" miss the major thrust of experimental design. We are never able to record a sufficient set of covariates, and if we needed to assume that a sufficient

set were recorded in order to draw valid causal inferences, then experiments would be of little value in the real world. In a randomized experiment, the validity of the randomization theory probability statement does not depend on an assumption that blocking has controlled a sufficient set of covariates; similarly, the validity of a Bayesian probability statement for causal effects does not require an assumption that a sufficient set of covariates is controlled by blocking or adjustment in data analysis. The statements are valid when data analysis methods are employed that are appropriate to the experimental design that was used, and the uncertainties in the probability statements reflect the potential impact of uncontrolled variables.

For example, it is relatively clear from the distributions of males and females in Professor Dawid's Tables 1 and 2 that neither table arose from a completely randomized experiment. It is likely that either (a) undisplayed covariates correlated with sex were used to assign treatments, or (b) sex was used as a blocking variable with differential probability of assignment to T and $\bar{T}$. In case (a), valid probability statements about the causal effects of the treatment must be conditional on the undisplayed covariates used for treatment assignment if they are recorded, or reflect a model for the non-ignorable assignment mechanism if they are not recorded. In case (b), valid causal probability statements must be conditional on sex, but there is no need to suppose that sex is a sufficient covariate. Of course, as we successfully condition on more causally relevant covariates (e.g. perhaps age), the resultant probability statements become more specific and thus more relevant to an individual unit with observed values of those covariates. But this does not imply that less conditional causal inferences are invalid or useless for deciding the typical benefits of the treatment for males and females.

I feel that a source of the problem with Professor Dawid's formulation of causal inference is his choice to let the outcome variable, Y, have a joint distribution with treatment, X, rather than be t-variate where t is the number of levels of X. Letting Y be t-variate as in Rubin (1978) allows us to define causal effects phenomenologically; that is, as comparisons among t observable quantities rather than as comparisons among t hypothetical conditional distributions of Y given X for fixed values of a sufficient set of covariates. With phenomenological definitions of causal effects, valid inferences for causal effects are predictions of unobserved observables conditional on recorded data, and thus generally change as more covariates are recorded, just as valid predictions change as more predictors are recorded. With distributional definitions, inferences for causal effects are apparently viewed as incorrect unless they arrive at the correct distribution, that is, unless they are conditional on a sufficient set of covariates, an unachievable goal in real world experiments.

Professor R. A. WIJSMAN (University of Illinois): Professor Dawid deserves to be commended highly for putting conditional independence (c.i.) on a pedestal and illuminating it from many different angles. The idea, introduced in Section 2.2, of manipulating Θ formally as a random variable in independence relations I found very useful. Lemmas 4.1–4.3, modestly labelled "simple general results", should not be misunderstood to be obvious. In fact, some of the statements, especially Lemma 4.3, I find sophisticated and not intuitive. To emphasize the power of these lemmas I shall present an application to a seemingly rather complicated problem which I met several years ago. When I first solved it I used *ad hoc* methods. In the future it should be easier to deal with this sort of problem now that Professor Dawid has so neatly singled out the crucial lemmas.

Here is the problem. Let U, V, W, X, Y, Z be statistics with a joint distribution depending on a parameter Θ, and $X = f(U, V)$. The following c.i. relations are given: (1) $U \perp\!\!\!\perp (V, Y, Z) \mid \Theta$, (2) $V \perp\!\!\!\perp Y \mid (W, Z, \Theta)$, (3) $V \perp\!\!\!\perp W \mid (Z, \Theta)$, (4) $(Y, Z) \perp\!\!\!\perp \Theta$. To show (5) $(X, Y, Z) \perp\!\!\!\perp \Theta \mid (X, Z)$, i.e. (X, Z) is a sufficient statistic for the family of distributions of (X, Y, Z). For the proof I shall need one more lemma to add to those in Section 4:

Lemma 4.4. $\perp\!\!\!\perp_i^n X_i \mid Z \Leftrightarrow [\perp\!\!\!\perp_i^{n-1} X_i \mid Z$ and $(X_1, \ldots, X_{n-1}) \perp\!\!\!\perp X_n \mid Z]$.

Without the conditioning this is stated as a definition in Section 4. Now the proof of (5) (the use of the various lemmas is indicated in parentheses). We have [(2) and (3)] $\Rightarrow$ (by Lemma 4.3) $V \perp\!\!\!\perp (W, Y) \mid (Z, \Theta) \Rightarrow$

$$V \perp\!\!\!\perp Y \mid (Z, \Theta), \tag{6}$$

(by Lemma 4.2(i)). Further, (1) $\Rightarrow U \perp\!\!\!\perp (V, Y) \mid (Z, \Theta)$ (by Lemma 4.2) $\Rightarrow (U, V, Y) \mid (Z, \Theta)$ (by Lemma 4.4 and (6)) $\Rightarrow (U, V) \perp\!\!\!\perp Y \mid (Z, \Theta)$ (by Lemma 4.4) $\Rightarrow$

$$X \perp\!\!\!\perp Y \mid (Z, \Theta) \tag{7}$$

(by Lemma 4.2(i)) since $X = f(U, V)$. Finally, (4) $\Rightarrow Y \perp\!\!\!\perp \Theta \mid Z$ (by Lemma 4.2) $\Rightarrow Y \perp\!\!\!\perp (X, \Theta) \mid Z$ (by Lemma 4.3 and (7)) $\Rightarrow Y \perp\!\!\!\perp \Theta \mid (X, Z)$ (by Lemma 4.2) $\Rightarrow$ (5) (by Lemma 4.1).

For quick and easy repeated use of Lemmas 4.1–4.3 it is convenient to express them in words. Lemma 4.1 says that the conditioning variable can be "added" to one or both of the c.i. variables. Lemma 4.2 says that the c.i. variables can be replaced by functions of them, and that such functions can be "added" to the conditioning variable. Lemma 4.3 is reminiscent of the multiplication law $P(AB) = P(A)\,P(B \mid A)$ for events. Indeed, in its simplest form Lemma 4.3 reads: if X is independent of Y and also c.i. of Z given Y, then X is independent of (Y, Z). This can be extended to any number of variables; e.g. $[X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z \mid Y$ and $X \perp\!\!\!\perp W \mid (Z, Y)] \Rightarrow X \perp\!\!\!\perp (Y, Z, W)$. This can then be conditioned further on another variable.

The AUTHOR replied later as follows.

Perhaps, as Professor Lindley and Mr Nimmo-Smith suggest, independence is a poor model, both for the physical world and for our subjective ideas of it; but *conditional* independence is surely vital for constructing such models. Suppose that we wish the conditional distributions for X given Z, in some mathematical model, to express in a meaningful way some property of the world. Then we must believe (or at least pretend) that these distributions are unaffected to any material extent by further factors Y that we have chosen to leave out of our description: that is, at least approximately, $X \perp\!\!\!\perp Y \mid Z$. Otherwise we should restructure our model, perhaps by adding further variables to Z (we can, of course, always achieve conditional independence by putting everything conceivable into Z). As Professor Fraser notes, the problem of determining *appropriate* conditional probabilities is paramount. It is the cornerstone of our understanding of Simpsons' paradox, and is of fundamental importance in many related areas, for example the treatment of regression by Davis and DeGroot. While new notation will not solve such problems by itself, its lubricating action may free the wheels of thought to run in the right tracks.

Professors Lindley and Godambe, and Mr Harding, object to my "proof" of the principle that no information is contained in an ancillary statistic (I note with interest that nobody raises similar objections to the sufficiency principle). Of course, this was not meant as a rigorous derivation, but rather to show how close to such a demonstration one can come, using only the ideas of (conditional) independence. Personally, it seems to me only natural that, if a statement such as "S carries no information about Θ" is rigorously demonstrable when Θ has an entirely arbitrary prior distribution, then it should continue to have force when no such prior is available. Nonetheless, Godambe's interesting example deserves a detailed response.

Let S, $\boldsymbol{\Theta}$ be the variables "sample" and "parameter-vector", having generic realized values s and $\boldsymbol{\theta}$, and let A be the event: "$|\Sigma\, \Theta_i/n - \frac{1}{2}| > k$", summing over $i \in S$. Godambe argues that $P(A \mid \boldsymbol{\Theta} = \boldsymbol{\theta})$ is small (say ε) for all $\boldsymbol{\theta}$, and that we should deduce that, even after observing $S = s$, it is much more plausible that A does not obtain rather than it does. This, he claims, suggests that we should favour a value $\boldsymbol{\theta}$ for which A fails over one for which A holds.

Now a Bayesian would agree with this to the extent that, for him, the marginal probability $P(A) = \varepsilon$, so that $P(A \mid S = s)$ must be small on the average; and in the particular case that he regards the components of $\boldsymbol{\Theta}$ as *a priori* exchangeable, he in fact finds $P(A \mid S = s) \equiv \varepsilon$, so that A is highly implausible in his posterior distribution. But this merely reflects the fact that, for given s, there are many *more* values for $\boldsymbol{\theta}$ for which A fails rather than holds, and does not imply that any *particular* such $\boldsymbol{\theta}$ is itself more favoured. Godambe's argument carries no more force than the belief that, in a fair lottery with 1,000 tickets, the number 500 is less likely to win than 267, because the latter is more "typical".

Mr McLaren raises many interesting points. In particular the idea of weaker, non-symmetrical forms of independence may well be worth pursuing, and may help to meet Mr Rizvi's objections. McLaren rightly points out the care needed in handling conditional independence arguments. Paradoxical arguments such as that of Basu are discussed in Dawid (1979).

McLaren asks me to relate Section 7.3 to 7.2. This is most nearly done by the identifications $\Theta \to D$, $X \to (S, G)$, but the underlying structure of 7.2 is somewhat richer.

In my paper I have deliberately ignored technicalities. It has taken me several attempts to develop a suitable mathematical framework for conditional independence, which allows parameters as well as random variables, and I am not entirely satisfied with the outcome (developed in my 1980 *Annals of Statistics* paper). The natural approach is in terms of σ-fields rather than random variables, but even then Lemma 4.1 is not quite empty, as claimed by Professor Dickey.

At one stage I considered an approach similar to that of Professor Mouchart, in which parameters are given arbitrary prior distributions, although I cannot agree with Mouchart that a symmetrical view of independence is best. (Section 9 of my present paper contains traces of this approach, and also indicates its limitations: one would not normally consider distributions over Λ, the parameter-space for $\mathscr{Q}$.) However, I got into technical trouble with null sets, and chose a different development, using "statistical operations". Professor Kudō has considered these problems carefully, and I believe that recent work of his largely reconciles Mouchart's approach and mine. Mouchart and Rolin have done some further technical work which attacks DeGroot's problem of finding partitions Π of $\mathscr{X}$ for which $X \perp\!\!\!\perp Y \mid \Pi$.

Mr Rizvi and Professor Barnard note that I have not resolved all the philosophical problems of disentangling association and causation. I hope I may be forgiven for this. Perhaps I am biased in my belief that the language of conditional independence may provide a suitable matrix for further studies of this important question.

Mr Galbraith and Professor Geisser suggest that other notation may be more useful than mine. Certainly Mr Galbraith's translation of his own argument seems helpful in its context, but, as he points out, it can sometimes prove a hindrance by obscuring symmetry. I feel, too, that it might be more liable to lead to errors, such as in the example of Basu quoted by Mr McLaren. The perfect notation would render all general results obvious on sight, and I agree that mine is not perfect.

Professor Geisser complains that the expression $X \perp\!\!\!\perp Y \mid (Z, W)$ is not well defined. This is not so: it is equivalent to his first attempted translation. His second translation can be properly expressed by adding on the further conditions $X \perp\!\!\!\perp W \mid Z$ and $Y \perp\!\!\!\perp Z \mid W$. While his notation looks neater here, it appears to have misled him in his study of Section 7.2, since his implied relationship $Q \perp\!\!\!\perp S \mid G$ is *not* a consequence of my assumptions: it is the very essence of the problem that Q *will* depend on S, through S_0.

Professor Mardia expands on my analysis of the Fisher–von Mises distribution, highlighting its value as a test-case for ideas about inference. Although Dickey and I have used conditional independence to investigate problems involving nuisance parameters, we have not yet considered the kind of question he raises, which certainly deserves further attention.

A referee of the original version of my paper complained that there was little to interest "the practical man". It was as a result of this suggestion that Section 5 was added. In the event, it seems to have drawn most of the fire of the discussion, which clearly vindicates the referee's judgement. However, I am left even more confused by Simpson's paradox than I ever was before. Most discussants seem unhappy about my treatment of it, and I myself agree that there remains much scope for improvement. However, several discussants appear to feel that the whole problem is illusory, and with this I cannot agree. Causal inference is one of the most important, most subtle, and most neglected of all the problems of Statistics.

Randomization is an obvious possibility that was suggested more than once: according to Lindley and McLaren, an appropriate requirement is $X \perp\!\!\!\perp I \mid Z$. I too gave some thought to randomization, but did not reach any conclusion convincing enough to put in writing. I have conceptual difficulties with the essential use made of the randomization distribution for inference, and find it hard to accept as "appropriate" any probability which does not condition on individuals and treatment assignments. Nevertheless, although I do not understand randomization, I am convinced of its importance, and long for a deeper insight into its true role.

Professor Rubin is one of the small brave band who are beginning to chart the murky depths of causal inference. I differ from him on some matters of personal taste, the most important being his willingness to assign a *joint* distribution to all the conceptual responses of an individual under all applicable treatments, when in fact only one such response can ever be observed. I dislike this because I consider it "non-phenomenological" (there must be a shorter word!) and can only register surprise that this does not bother him too. As Quantum Theory discovered long ago, it is meaningless to assign probabilities to the joint occurrence of events which cannot occur jointly.

DeGroot and Nimmo-Smith both make important points about the case where we cannot assign treatments arbitrarily to individuals. I accept DeGroot's censure that performance should not have been omitted from the discussion of graduate-school admissions. In fact my own discussion following Theorem 3.1 introduces this in a satisfactory way, and it now seems to me that the correct expression of "no bias in admissions policy" is (5.4) with Z = performance. However, I am not sure how far this gets us.

In laying my ideas on conditional independence before the statistical community, my hope was that they would prove of value outside the particular topics that I have discussed. I am therefore particularly heartened by those contributions (by McLaren, Bickel, DeGroot, Galbraith, Mouchart, Novick and Wijsman) which describe further extensions and applications of the theory. When I originally circulated my paper, it drew a certain amount of correspondence in which yet other applications were pointed out. In particular, T. P. Speed has for a long time been making fruitful use of conditional independence in unpublished work on Markov random fields, with applications to contingency table analysis.

Although some discussants have their doubts that conditional independence can live up to my inflated claims, I have been very gratified by the overall constructive and stimulating response to my paper. My appreciation and thanks go out to all the contributors. If my notation and general theory find application in the future work of others, I shall at least have justified to the Society's Printers my insistence that, somehow or other, they should find a way to print that awkward symbol $\perp\!\!\!\perp$.

References in the Discussion

Akhmeteli, M. A. *et al.* (1977). Methods for the detection of haemophilia carriers: a memorandum. *Bull. W.H.O.*, **55**, 675–702.

Anderson, T. W. (1959). Some scaling models and estimation procedures in the latent class model. In *Probability and Statistics, the Harold Cramer Volume* (O. Grenander, ed.), pp. 9–38. New York: Wiley.

Basu, D. (1958). On statistics independent of sufficient statistics. *Sankhyā*, **20**, 3–4.

Birnbaum, A. (1968). Chapters 17–20 in *Statistical Theories of Mental Test Scores*, by F. M. Lord and M. R. Novick. Reading, Mass.: Addison-Wesley.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass.: M.I.T. Press.

Boruch, R. F. (1972). Relations among statistical methods for assuring confidentiality of social research data. *Soc. Sci. Res.*, **1**, 403–414.

Davis, W. W. and DeGroot, M. H. (1978). A new look at selecting regression models. Carnegie-Mellon University, Dept. of Statistics, Technical Report No. 149.

Dawid, A. P. (1979). Some misleading arguments involving conditional independence. *J. R. Statist. Soc.* B, **41**, No. 2 (in the press).

Edwards, W. (1968). Conservatism in human information processing. In *Formal Representation of Human Judgement* (B. Kleinmutz, ed.). New York: Wiley.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity*, **25**, 95–105.

Goodman, L. A. (1971). The partitioning of chi-square, the analysis of marginal contingency tables and the estimation of expected frequencies in multidimensional contingency tables. *J. Amer. Statist. Ass.*, **66**, 339–344.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Hinkley, D. V. (1979). Predictive likelihood. *Ann. Statist.*, **7** (in press).

Jeffreys, H. (1961). *Theory of Probability*. Oxford: University Press.

Kadane, J. B. and Dickey, J. M. (1979). Bayesian decision theory and the simplification of models. In *Evaluation of Econometric Models* (J. Kmenta and J. B. Ramsey, eds). New York: Academic Press (to appear).

Lange, K. and Elston, R. C. (1975). Extensions to pedigree analysis. *Human Heredity*, **25**, 95–105.

Linnik, Yu. V. (1975). *Problems of Analytical Statistics*. Calcutta: Statistical Publishing Society.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph* No. 7.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

McLaren, A. D. (1967). Appendix to Chown, S., Belbin, E. and Downs, S. Programmed instruction as a method of teaching paired associates to older learners. *J. Gerontol.*, **22**, 219.

Mardia, K. V. (1975). Statistics of directional data (with Discussion). *J. R. Statist. Soc.* B, **37**, 349–393.

Martin, F., Petit, J.-L. and Littaye, M. (1973). Indépendance conditionelle dans le modèle statistique Bayésien. *Ann. Inst. Poincaré*, IX, 19–40.

Mouchart, M. and Rolin, J.-M. (1978). A note on conditional independence with statistical applications. Mimeograph.

Navon, D. (1978). The importance of being conservative: some reflections on human Bayesian behaviour. *Brit. J. Math. Statist. Psychol.*, **31**, 33–48.

Pickard, D. K. (1977). A curious binary lattice process. *J. Appl. Prob.*, **14**, 717–731.

Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Ass.*, **60**, 63–69.