*The Annals of Statistics*
1981, Vol. 9, No. 1, 45–58

# THE ROLE OF EXCHANGEABILITY IN INFERENCE[1]

BY D. V. LINDLEY AND MELVIN R. NOVICK

*University College London and The University of Iowa*

This paper is concerned with basic problems of statistical inference. The thesis is in three parts: (1) that inference is a procedure whereby one passes from a population (or sample) to a new individual; (2) that this connection can be established using de Finetti's idea of exchangeability or Fisher's concept of a subpopulation; (3) in making the connection use must be made of the appropriate probability. These three principles are used in a variety of situations and the topics discussed include analysis of variance and covariance, contingency tables, and calibration. Some comments on randomization are also included.

**1. Introduction.** This paper presents what we believe to be a useful way of looking at problems of statistical inference. The thesis is in three parts. First, it is argued that inference is a process whereby one passes from data on a set of units to statements about a further unit. Standard procedures concentrate on the data and tend to ignore the connection with the case to which the inference is to be applied. Second, we show how this connection can be established using either de Finetti's idea of exchangeability or Fisher's concept of a subpopulation. Third, in making the connection it is important to use the appropriate probability, since there are many instances where statisticians have used what we argue is the wrong value.

The paper begins with some striking examples. There then follows a section developing some technical ideas which are applied to resolve the paradoxes raised by the examples. Topics discussed include analysis of variance and covariance, contingency tables, and calibration. Some comments on randomization are also included. Although we do not enter into controversies over statistical methods of inference, the paper does, we believe, give support to the personalistic view by demonstrating its usefulness.

**2. Simpson's paradox.** Consider the data in Table 1 where 40 patients were given a treatment, $T$, and 40 assigned to a control, $\bar{T}$. The patients either recovered, $R$, or did not, $\bar{R}$. We are not considering small-sample problems so that the reader can if he wishes imagine all the numbers multiplied by 10,000, say. It is then clear that the recovery rate for patients

TABLE 1
*Recovery rates under treatment and control*

| | R | $\bar{R}$ | | Recovery Rate |
|---|---|---|---|---|
| T | 20 | 20 | 40 | 50% |
| $\bar{T}$ | 16 | 24 | 40 | 40% |
| | 36 | 44 | 80 | |

Received January, 1978; revised August, 1979.

[1] Research supported, in part, by the Office of Naval Research under contract N-00014-77-C-0428, Melvin R. Novick, Principal investigator.

*AMS* 1970 *subject classifications.* 62A15, 62F15.

*Key words and phrases.* Exchangeability, exchangeable populations, probability, propensity, randomization, analysis of variance and covariance, regression, Simpson's paradox, contingency tables, random quantities, information.

receiving the treatment at 50% exceeds that for the control at 40% and the treatment is apparently to be preferred. However, the sex of the patients was also recorded and Table 2 gives the breakdown of the same 80 patients with sex, $M$ male or $\bar{M}$ female, included. It will now be seen that the recovery rate for the control patients is 10% higher than that for the treated ones, both for the males and the females. Thus, what is good for the men is good for the women, but bad for the population as a whole. We refer to this as Simpson's (1951) paradox, though it occurs in Cohen and Nagel (1934). In Appendix 1 we describe the situation mathematically and show that the paradox can only arise if, $R$ and $T$ being positively associated, $M$ is positively associated both with $R$ and with $T$. This is exactly what has happened here: The males have been mostly assigned to the treated group, the females to the control; perhaps because the doctor distrusted the treatment and so was reluctant to give it to the females where the recovery rate is much lower than for males. Alternatively expressed, treatment and sex have been confounded. Nevertheless it comes as a surprise to most people to learn that confounding can actually reverse an effect; here from +10% to −10%.

An important problem posed by the paradox is this: Given a person of unknown sex would you expect the control or the treatment to be the more effective? (If having an unknown sex seems odd replace $M$ and $\bar{M}$ by a dichotomy that is difficult to determine, such as a genetic classification.) The answer seems clear that, despite Table 1, the control is better. If so, then this warns us to be very careful in using results like those in Table 1 to draw the opposite conclusions for could there not exist a factor, here sex, which reversed the conclusion? But is the answer so clear? Keeping the numbers the same, imagine data with $T$ and $\bar{T}$ replaced by white and black varieties of a plant respectively, and $R$ and $\bar{R}$ corresponding to high and low yields; the confounding factor being whether the plant grew tall, $M$, or short, $\bar{M}$. The white variety is 10% better overall, but 10% worse among both tall and short plants. In this case the white variety, $T$, seems the better one to plant; whereas $\bar{T}$, the control, was intuitively preferred in the medical situation.

The problem addressed in this paper is that of providing a formal framework within which such problems can systematically be resolved. In the next section we describe some mathematical ideas that are used in subsequent sections to discuss Simpson's paradox and related problems.

**3. Exchangeability and recognizable subpopulations.** Throughout the paper we shall use probability in the sense of a number which a person, conveniently called You, would attach to the truth of an event, $A$, were he to be informed of the truth of another event. $B$. It is termed the probability of $A$ given $B$ and written $p(A|B)$. Sometimes, reference to the conditioning event $B$, being understood, is omitted and we refer to Your probability of $A$, $p(A)$.This is not a frequency concept but its relation to relative frequencies will be considered later.

Let $X$ and $Y$ be two random variables, each of which may be multidimensional. Recovery

TABLE 2
*Recovery rates under treatment and control with sex as an added variable*

| Males | R | $\bar{R}$ | | Recovery Rate |
|---|---|---|---|---|
| T | 18 | 12 | 30 | 60% |
| $\bar{T}$ | 7 | 3 | 10 | 70% |
| | 25 | 15 | 40 | |

| Females | R | $\bar{R}$ | | Recovery Rate |
|---|---|---|---|---|
| T | 2 | 8 | 10 | 20% |
| $\bar{T}$ | 9 | 21 | 30 | 30% |
| | 11 | 29 | 40 | |

and yield are two examples from Section 2. Consider next a number of similar things termed units; in the examples of Section 2 they are patients and plants. For the $i$th unit, let the random variables $X$ and $Y$ assume values $X_i$, $Y_i$. While these are unknown to You, they will be referred to as random quantities and You will have probabilities for them. As soon as You observe them they become numbers $x_i$, $y_i$ and the randomness (and hence the probability notion) disappears.

A number $n$ of units is termed exchangeable in $X$ if the joint probability distribution $p(X_1, X_2, \cdots, X_n)$ is invariant under permutation of the units. A further unit is exchangeable in $X$ with the set if all $(n + 1)$ units are so exchangeable. In the medical example the $n = 40$ patients who received the treatment might be judged exchangeable in recovery, and a further patient might be judged exchangeable with the 40 were he to receive the treatment.

A number of units is termed exchangeable in $X$, given $Y = y$, if the joint conditional distribution $p(X_1, X_2, \cdots, X_n | Y_i = y, \text{all } i)$ is invariant under permutation of the units. If this holds for all $y$, we refer to exchangeability in $X$, given $Y$. A further unit is exchangeable in $X$ given $Y = y$ with the set if the enlarged set of all $(n + 1)$ units is so conditionally exchangeable. In the medical example, the 40 patients who were male $(y)$ might be judged exchangeable in recovery (conditional on their sex). If the same holds for the 40 females then the 80 patients are exchangeable in recovery given sex. This is not the only possible definition of conditional exchangeability; another form is discussed in Appendix 2. The form given here is adequate for the applications considered in the present paper.

Consider the case where $X$ refers to an event such as recovery, and so only takes two values, $R$ and $\bar{R}$. Suppose $n$ units in the data and a further unit are exchangeable in $X$. Then for inference purposes, we may be interested in the possible value of $X$ in unit $(n + 1)$ given the values of $X$ in the $n$ units. If $n$ is large, the probability that the event will occur in the new unit is simply the frequency with which the event has occurred in the $n$ units. A rigorous demonstration of this requires de Finetti's theorem on the structure of exchangeable sequences; however the result is intuitively obvious. It is important because it provides a link between the view of probability adopted here and the frequency viewpoint. We shall use the term *propensity* (or *chance*) to describe the frequency and write $P(A)$ for the propensity of an event $A$. The result just mentioned may be abbreviated to $p(A) = P(A)$, equating the probability and propensity. Notice that the condition of exchangeability has been omitted from the notation. The concept extends to conditional exchangeability in $X$ given $Y$ when we will have $p(A | Y = y) = P(A | Y = y)$, the propensity of $A$ among those units having $Y = y$. In the medical example, if a judgment of exchangeability in recovery, given sex and treatment, is made then the probability that another male will recover, given the treatment, is, from Table 2, the propensity $^{18}/_{30} = 0.6$.

The ideas of probability and exchangeability just mentioned are due to de Finetti (1974). Fisher uses the term, probability, somewhat differently and in conjunction with the concept of a population. We pause to consider these ideas and their interconnections. Fisher's ideas are most clearly expressed in his last book (1956), in particular in this section from page 33:

> "This fundamental requirement [of no recognizable subset] for the applicability to individual cases of the concept of classical probability shows clearly the role of subjective ignorance, as well as that of objective knowledge in a typical probability statement. It has often been recognized that any probability statement, being a rigorous statement involving uncertainty, has less factual content than an assertion of certain fact would have, and at the same time has more factual content than a statement of complete ignorance. The *knowledge* required for such a statement refers to a well-defined aggregate, or population of possibilities within which the limiting frequency ratio must be exactly known. The necessary *ignorance* is specified by our inability to discriminate any of the different subaggregates having different limiting frequency ratios, such as must always exist."

The concept of a population of units is close to saying that those units are exchangeable. The identification of a subaggregate, or subpopulation, is related to conditional exchangeability

(of $X$, given $Y = y$), the discrimination Fisher refers to being effected by $Y$. Even the apparently dissimilar notions of probability, of Fisher and de Finetti, are not unrelated: the limiting-frequency (or propensity) being relevant as a probability statement by You whenever exchangeability is present. Moreover, the relevant frequency is determined by recognizing the appropriate subpopulation, or type of conditional exchangeability. On the other hand, there are two important differences between the concepts. First, exchangeability of units refers explicitly to a random quantity, whereas a population does not. Thus the units might be exchangeable in $X$, but not in $Y$, or even in $(X, Y)$. Second, no guidance seems to be given on how to recognize whether an individual unit belongs to a population, whereas exchangeability, being a statement about units, does: one unit cannot disturb a limiting frequency, whereas it can affect exchangeability. We attach considerable importance to this last point because of our view of inference as a passage from data to a unit and not, except as an intermediary, to a parameter.

Perhaps the most important difference between the two notions is that de Finetti gives us a precise definition that we can operate with; whereas Fisher conveys only a brilliant suggestion that suffers from vagueness in individual applications. The way we prefer to regard the situation is that exchangeability makes precise the concepts of populations and subpopulations; and we will often find it convenient to use Fisherian language. A major task in inference, as discussed below, is the identification of the appropriate population to which an individual belongs. Thus in the medical example we can recognize a subpopulation of treated patients, namely that defined by sex; whereas, to anticipate the argument below, the total population is relevant in the agricultural example. To simplify: practitioners seem to prefer the language of populations; theoreticians, that of exchangeability.

Fisher used the concept of subpopulation in inference through the concept of an ancillary statistic. This may be a misuse of the concept—it certainly is from the Bayesian view—but that does not affect the validity and usefulness of the basic notion. He also used probability to mean "a fraction of a set" and denied, despite the above quotation, some aspects of it as a limiting frequency. This point is discussed by Savage (1976, Section 4.3). Again, this does not invalidate our arguments that follow.

We now apply these ideas. In Section 4 we deal with two random variables only, starting with the special case of events and later generalizing. In Sections 5 and 6 we discuss three random variables, where new phenomena enter, and Simpson's paradox.

**4. Two random variables.** The discussion in this section owes much to Meehl and Rosen (1955) and is included as an introduction to the ideas that are then used in Sections 5 and 6 for the three-variable situation. It is convenient to work in terms of examples and we begin with one involving two events. Patients were classified according to whether or not they had a disease, $D$, and whether their reaction to a test was positive or negative. Possible results on $n = 100$ patients are given in Table 3. To a statistician, this is a 2 by 2 contingency table and he could employ many of the techniques devised for such tables. In particular, he might regard $D$ and $\bar{D}$ as two hypotheses and + and − as data appropriate for distinguishing between the two. We argue that typically the inference problem is not confined to the $n = 100$ units (patients) in the data base but extends to include other units: For example, those who have responded positively to the test but are not known to have the disease. Connection between

TABLE 3
*Possible results on* n = 100 *patients*

| | D | $\bar{D}$ | |
|---|---|---|---|
| + | 16 | 24 | 40 |
| − | 4 | 56 | 60 |
| | 20 | 80 | 100 |

the new patient and the data base can, we argue, conveniently be described in terms of exchangeability or populations and we explore various possibilities.

One possibility is to regard the new patient as exchangeable in both variables, disease and test, with those in the table. Alternatively expressed, the 101 patients are a random sample from a population in both events. If so, we can relate probabilities for the new patient to propensities in the data base and, for example, declare $p(D|+) = P(D|+) = 0.4$ so that a patient responding positively has a probability of 0.4 of having the disease.

Another possibility is to judge the new patient exchangeable in test result given the disease classification. This might be appropriate if, for example, the 100 patients in the data base were from one city, the new patient from another city, and it was felt that the disease propensity might differ between cities, but the test behaved similarly in the two places. In this case all one can infer by exchangeability is $p(+|D) = 0.8$ and $p(+|\bar{D}) = 0.3$, the corresponding propensities. Alternatively expressed, two subpopulations can be recognized, corresponding to $D$ and $\bar{D}$. A new point now arises. In the instance of a patient with known test result but unknown disease classification we require $p(D|+)$. This cannot be obtained from the data alone. With Bayes rule

$$p(D|+) \propto p(+|D)p(D) = 0.8p(D),$$

but $p(D)$ is still required. Without the judgment of exchangeability in $D$ this cannot be obtained from the data. Statisticians have bypassed this problem by confining their attention to the values obtainable from the data. These, we argue, are the wrong probabilities. The patient has been tested; has he got the disease? The test result is given; the disease is in doubt. Let us explore this further.

The positive result and the disease are positively associated so that + favors $D$; −, $\bar{D}$. The statistician's error rates are therefore $p(-|D) = 0.2$ and $p(+|\bar{D}) = 0.3$ and the test appears quite useful as a diagnostic for the disease. Write $p(D) = p$, then by Bayes rule in odds form

$$\frac{p(D|+)}{p(\bar{D}|+)} = \frac{8}{3}\frac{p}{1-p}$$

and, for a negative test result,

$$\frac{p(D|-)}{p(\bar{D}|-)} = \frac{2}{7}\frac{p}{1-p}.$$

However for $p < 3/11 = 0.27$, both these expressions are less than unity so that if $p$ satisfies this inequality the test result, on its own, is useless. Equally for $p > 7/9 = 0.78$ both are greater than one and again the test is of no value. With the judgment of full exchangeability and no subpopulation identification $p = 0.2$, the propensity for the disease, $P(D)$, and the test is useless. The statistician's argument is therefore incomplete because it ignores the disease probability and uses the wrong probabilities: for example, $p(+|D)$ instead of $p(D|+)$.

Another possible exchangeability judgment is that of exchangeability in disease given test result. This seems an unlikely one but Dawid (1977) has given a careful discussion of how this might happen. When it does, the required $p(D|+)$ can be equated directly to $P(D|+) = 0.4$ and Bayes rule does not have to be invoked.

There are other possibilities. For example, You may judge the new patient exchangeable in test result given $D$, but not given $\bar{D}$. A possible reason for this is that You may feel that the disease is the same in the two cities but that patients without the disease have different disease patterns in the two places; say an alternative to $D$ being $D_1$, which is common in one city but rare in the other. In this case only $p(+|D) = 0.8$ can be found from the propensities in the data base, so that both $p(+|\bar{D})$ and $p(D)$ are required from elsewhere before $p(D|+)$ can be assessed Bayes rule; alternatively it may be assessed directly.

We learn three things from this study. First, that there are various forms of connection between the data and a new unit for which an inference is to be made. Second, that these connections may be made using exchangeability. Third, that care needs to be exercised in using the appropriate probability.

In those cases where exchangeability does not provide adequate connection for the required

probability to be equated to a propensity it will be necessary to assess probabilities using additional information beyond that in the data base. For example, we saw above that if exchangeability in test result given disease class is all that is assumed, Bayes rule required $p(D)$ to be assessed. We may have data on the disease propensity in the city from which the new patient comes: if so, that may provide $p(D)$. Alternatively, we may merely feel that the disease is more common in his city so that $p(D) > P(D) = 0.2$ and some judgment will have to be used in default of data.

Few additional points arise when we pass from two events to two general random variables, $X$ and $Y$. Again there are various forms of exchangeability assumption: in both $X$ and $Y$, in $X$ given $Y$, or in $X$ given $Y = y$ for some $y$. Another terminology is sometimes used in this context besides that of exchangeability or subpopulation, namely to describe a random variable as either random or fixed. Thus in regression of $Y$ on $X$, Kendall and Stuart (1969) discuss the cases of $X$ random, and the more common case of $X$ fixed. These correspond to the joint exchangeability of $X$ and $Y$, and to that of $Y$ given $X$, respectively. An interesting case is that of calibration, where $X$ is a precise measurement—perhaps the true value—and $Y$ a simple but less precise one. The usual judgment is that of exchangeability in $Y$ given $X$ but the required probability is $p(X|Y=y)$—from the imprecise measurement it is required to evaluate the true value and hence calibrate the measurement. Bayes rule has then to be invoked and it is necessary to obtain $p(X)$ from sources other than the data. In Kendall and Stuart's terminology, $X$ is fixed yet has to be estimated. Similar problems arise in discrimination and classification problems were $X$ describes the class of, and $Y$ the measurements on, the units.

Calibration, discrimination, and classification all are fields in which the wrong probability has often been used, particularly by statisticians. It is perhaps worth pointing out that the correct approach has for long been standard practice in some fields. Thus in educational testing with $X$ the true score and $Y$ the observed score, exchangeability is invoked for $Y$ given $X$, the propensity being described by test error. The distribution of true score, $X$, in the population is then used to derive the required distribution of $X$, given $Y$. The appropriate regression formula is due to Kelley (1923). Similar early examples occur in actuarial science in connection with claim frequencies; see, for example, Whitney (1918) or Longley-Cook (1962) who provides a survey. Similar approaches are used in electrical engineering, particularly in signal discrimination, as numerous papers in the proceedings of IEEE testify.

**5. Three random events.** We first apply the lessons learned in Sections 3 and 4—namely the connection of the data with a new unit, the judgment of exchangeability, and calculation of the appropriate probability—to the two examples, medical and agricultural, of Section 2. Consider in the first case a new patient, male, about whom a decision has to be made as to whether to give him the treatment or not. A possible judgment might be of exchangeability in recovery, given sex and treatment, in which case $p(R|TM) = 0.6$ and $p(R|\bar{T}M) = 0.7$ are available by equating the probabilities and propensities, and consequently the treatment should be withheld. Alternatively four subpopulations are identifiable as $TM$, $T\bar{M}$, $\bar{T}M$, and $\bar{T}\bar{M}$. The same conclusion would hold for a female. We mentioned in Section 1 the possible case of someone of unknown sex (or perhaps unknown genetic makeup). We would then need $p(R|T)$ which, with exchangeability of the type just assumed, is not available from observed propensities in the data. However, by extending the conversation to include sex,

$$\begin{aligned} p(R|T) &= p(R|TM)p(M|T) + p(R|T\bar{M})p(\bar{M}|T) \\ &= 0.6p(M|T) + 0.2p(\bar{M}|T), \end{aligned}$$

and only $p(M|T)$ is required to complete the analysis. Without an assumption of exchangeability in sex given treatment this cannot be derived from the propensities of the data. (This was $P(M|T) = 0.75$.) Instead You might judge that the decision to use the treatment or the control is not affected by the unknown sex, so that $M$ and $T$ are independent. In default of other knowledge You might judge the new patient to be exchangeable in sex with the rest of the population, where the propensity to be male is about ½. Hence $p(M|T) = 0.5$ and $p(R|T) = 0.4$. A similar calculation for the control gives $p(R|\bar{T}) = 0.5$ and the control is preferred for

a person of unknown sex. (Once $M$ and $T$ have been judged independent, the male propensity is irrelevant to the 10% drop in recovery rate if the treatment is applied.)

The above judgment of exchangeability—in $R$, given treatment and sex—or the identification of the appropriate four subpopulations, is an expression of Your belief that treatment and sex cause the recovery rate to have a certain value. In this view, cause is a judgment by You, that if this happens then that will randomly follow. In the agricultural example the causation pattern is likely to be different. (Remember, treatments are replaced by varieties; sex by height; and recovery by yield.) There the yield and height are a result of the variety planted, so that the exchangeability is in yield and height, given variety. Hence, the propensities of Table 2 now provide $p(RM|T)$, etc., the joint distributions of yield and height, given variety. In particular, You have the margins $p(R|T)$ and $p(R|\bar{T})$ direct from Table 1: their values are respectively 0.5 and 0.4 and the white variety, $T$, is preferred. Here only two subpopulations are identified.

In the last paragraph the concept of a "cause" has been introduced. One possibility would be to use the language of causation, rather than that of exchangeabillity or identification of populations. We have not chosen to do this; nor to discuss causation, because the concept, although widely used, does not seem to be well-defined. (There the emphasis is on *definition*: there is, of course, an extensive philosophical literature that does not produce a mathematical definition. The admirable monograph by Suppes (1970) is the best reference: a more recent discussion is by Toda (1977).) One definition, that is used in experimental design, is stated by Rubin (1974, 1978):

> "The causal effect of one treatment relative to another for a particular experimental unit is the difference between the result if the unit had been exposed to the first treatment and the result if, instead, the unit had been exposed to the second treatment."

This is fine as far as it goes but, as Rubin points out, it cannot be tested directly since a unit typically cannot be exposed to two treatments. A way to test it is to use "similar" units, some having one treatment, some another; but then a judgment of similarity is involved. Such a judgment is conveniently expressed in terms of exchangeability, as Rubin does. There is a link between our ideas and causation but we have chosen not to explore them in this paper, partly because it would make the paper overlong, but more importantly because of formidable difficulties of definition. We hope that our suggestions involving exchangeability and populations will be of some help in formalizing and understanding causation.

Another way of looking at Simpson's paradox is through correlation ideas. Thus, it might be said that the correlation between treatment and recovery is "spurious" in the medical case; but that between their agricultural parallels, variety and yield is "real". This view is usefully explored by Simon (1954) who distinguishes between the two types of correlation using linear models relating the three variables; models which would have different structures in the medical and agricultural cases. These will be considered below when discussing variables rather than events. Exchangeability has the advantage over correlation ideas in applying to nonlinear situations.

The contrast between the medical and agricultural examples shows that there can be no unique method of analyzing the data of Table 2. The inferences in the two cases are completely different: $\bar{T}$ is better in the medical, $T$ in the agricultural, case. Our argument is that the reason for the difference, and hence the choice of the appropriate analysis, can easily be appreciated using the notion of exchangeability, or equivalently that of subpopulations. Another advantage, carefully discussed by Rubin (1978), is that the Bayesian argument is considerably simplified when the treatment allocation is performed using a random mechanism.

It has been pointed out in Section 1 (and in the appendix) that the paradox arises in the medical example because treatment and sex have been confounded. However, this confounding does not affect the agricultural example, where the obvious interpretation of Table 1 is, as we have seen, the correct one. These ideas are connected with the role of randomization in experimental design. It would be argued that had the treatment in the medical case been

assigned at random the paradox could not have arisen. This is in agreement with the view adopted here. A mechanism is judged random by You if, among other things, You consider that the mechanism is unconnected with any other factor. With such a judgment no other factor such as sex would be expected to disturb the basic interpretation of Table 1. We therefore see that randomization can play an important role even in the personalistic, Bayesian view of inference. This is contrary to the opinion resulting from the basic theorem in decision theory, that for any randomized decision procedure there exists a nonrandomized one which is not worse than it, to the effect that randomization is unnecessary in the Bayesian approach. The reason for the difference is that the use of a random mechanism is not necessary, it is merely useful. What is needed is a judgment of nonexistence of an effect confounded with treatment. It would be quite sensible in this view to allocate the treatments deliberately and thoughtfully so that the allocation appeared to possess no confounding characteristics. One advantage of a random mechanism is that most people, and not just You, will believe it to be random and hence without connection to another effect such as sex.

In practice scientists do not allocate completely at random: instead they obtain a random allocation from the mechanism and then inspect it for any unusual features before using it. Thus, if in the random selection of a Latin square, one in which the treatments lay down the diagonal was obtained, it would be discarded and a new allocation selected. In other words, the scientist always thinks about the proposed allocation before using it; which is essentially the argument here—use an allocation which You think is free from confounding. In any case, it is better to avoid randomization, as far as possible, by blocking with respect to any factor thought to influence the results; randomization is only a last resort. Notice that in small samples, not discussed in this paper, an allocation found by a random mechanism will always be confounded with some effect: one can do no better than what the personalistic view suggests, use an allocation which You think is unlikely to have important confounding effects. In the agricultural example the confounding with height is irrelevant since the allocation (of variety) influences the height and joint exchangeability of yield and height is reasonable. Thus it is only necessary to consider effects, such as sex, which exist prior to allocation and not those, such as height, which are influenced by the assignment. As Lord (1969) points out, the agricultural experiment is noninformative about the yield of white plants made to grow tall.

Simpson's paradox is related to the sure-thing principle of Savage (1962), and the relation has been explored by Blyth (1972) and by others in the discussion to that paper. The principle says that if act $f$ is preferred to act $g$ when an event $A$ is true, and also when $A$ is false, then $f$ is preferred to $g$ when You are uncertain about $A$. The medical case is apt: $\bar{T}$ is preferred to $T$, both for $M$ and $\bar{M}$, and therefore for someone of unknown sex. The agricultural example appears to violate the principle. The resolution lies in the fact that there the choice of act—black or white variety—is no longer available to You if $A$, a tall plant, is true. Consequently the premises of the principle are not correct. The principle might apply in Lord's case of conditions in which plants are made to grow tall. Again the notion of exchangeability conveniently captures the essence of the distinction.

In Section 4 we discussed the choice of the appropriate probability. The same point arises with three events. In the two examples of Simpson's paradox the appropriate one, $p(R \mid T)$, is available directly. An extension of the disease/test example of Section 3 will illustrate the point more forcefully. Suppose, in addition to $D$, $\bar{D}$, and $+$, $-$, the sex was also recorded. Then the judgment of exchangeability might be in respect of test result given sex and disease class. Hence $p(+ \mid DM)$ etc. would be available from the data propensities, whereas the quantities required would be $p(D \mid +M)$ etc. for someone of known sex, or $p(D \mid +)$ if that is unknown. By Bayes formula

$$p(D \mid +M) \propto p(+ \mid DM)p(D \mid M)$$

and the first factor is available, but the second, $p(D \mid M)$ would need to be assessed by other methods. The evaluation of $p(D \mid +)$ would proceed as for $p(R \mid T)$ above by extension of the conversation to include sex.

Two further points are worth making before passing to more general random variables than events. First, it should be noted that even with the full data of Table 3 in the medical

example the treatment $T$ might still be preferable to the control $\bar{T}$, even with the exchangeability assumption already made. For example, there could exist another dichotomy, say rural and urban, which would reverse the difference again. Thus for any combination of sex and urbanization, the treatment might give the preferred recovery rate.

The second point leads on from this. Many sciences are observational and not experimental; sociology, for example. In these cases factors cannot always be selected in such a way that You expect no confounding. Consequently it is sometimes dangerous to make deductions from observational data and conclude that these will hold for controlled data. Fisher (1958) made this point in connection with lung cancer, arguing that the observed association with smoking might not hold if smoking was controlled because there might exist a factor, he suggested a genetic one which played the role of sex in our example, that created a spurious association. Another instance of this might be provided by the same data set as in Table 2 with varieties replaced by racial classification, yield by intelligence, and sex by social class. The white people would appear more intelligent than the black but this might be due to confounding with social class. Yet this might (or might not) be again reversed by confounding with some other, unknown factor. Observational materials are themselves inadequate in situations like this; some judgment of exchangeability is essential in such cases. The possibility of stronger judgment of exchangeability in the case of designed experiments as against observational data is one way of accounting for the superiority of the former type of data collection over the latter.

**6. Three random variables.** We now pass from the consideration of three events to look at situations where one or more of the events are replaced by general random variables. Consider first the agricultural example of Simpson's paradox with the high or low yields replaced by $Y$, a random variable measuring the yield in, say, tons per acre. Table 4 provides an example. It is derived from Table 3 by multiplying the propensities there by 40 (to avoid fractions) and calling them expectations. Thus, for $M$ and $T$, $p(R \mid MT) = 0.6$ giving $\bar{Y} = 24$. In each cell $n$ refers to the number of observations. Again $n$ might be multiplied by some large number and $\bar{Y}$ indentified with expectations, such as $E(Y \mid MT)$. The paradox arises since $E(Y \mid MT) < E(Y \mid M\bar{T})$, and similarly with $\bar{M}$, yet $E(Y \mid T) > E(Y \mid \bar{T})$; and is due to the confounding between $M$ and $T$. Merely displaying the result in this tabular form suggests analysis of variance techniques and in the language of that area: There are main effects of both factors and a pronounced interaction. In the agricultural version of Table 4, the judgment is of exchangeability in $Y$ (yield) and $M$ (height) given $T$ (variety), so that only the main effect of variety is important in considering a new plot. With the medical situation the exchangeability is in respect to $Y$ (which might be a measure of recovery, say increase in blood-cell count) given $M$ (sex) and $T$ (treatment). Here the interaction is relevant and the important feature that carries over to a new patient is the conditional distribution of $Y$ given $M$ and $T$, and the usual breakdown into main effects and interaction is of limited use. This emphasizes again the point made earlier that there can be no unique analysis of data without consideration of the new unit to which the inference is to be applied. Notice that had the design been balanced

TABLE 4
*Resulting expectations*

| | M | $\bar{M}$ | Total |
|---|---|---|---|
| T | n = 30<br>$\bar{Y}$ = 24 | n = 10<br>$\bar{Y}$ = 8 | n = 40<br>$\bar{Y}$ = 20 |
| $\bar{T}$ | n = 10<br>$\bar{Y}$ = 28 | n = 30<br>$\bar{Y}$ = 12 | n = 40<br>$\bar{Y}$ = 16 |
| Total | n = 40<br>$\bar{Y}$ = 25 | n = 40<br>$\bar{Y}$ = 11 | n = 80<br>$\bar{Y}$ = 18 |

with $n = 20$ in each cell the main effect of treatment would have agreed with that for each sex separately.

The assumption of exchangeability on its own is not enough for valid inferences. For example, in a randomized block design with treatments $T_i$ and blocks $B$ giving yields $Y$, the exchangeability is for $Y$ given $T$ and $B$. This, by itself, gives no guide to treatments, $p(Y \mid T_i)$. Usually one assumes that yield differences $\Delta Y$ for two treatments, $T_i$ and $T_j$, are independent of $B$ so that $p(\Delta Y \mid T_i, T_j, B)$, available by exchangeability, reduces to the required $p(\Delta Y \mid T_i, T_j)$. This is the asssmption of additivity.

Suppose next that in addition to $Y$, the nuisance factor, sex or height, is also a continuous random variable, $X$ say. The agricultural situation again provides an example with $X$ as height. The paradox arises whenever

$$E(Y \mid X, T) < E(Y \mid X, \bar{T})$$

for all $X$, and yet

$$E(Y \mid T) > E(Y \mid \bar{T}).$$

This is clearly possible even within the restricted context of linear regression with fixed slopes. For suppose

$$E(Y \mid X, T) = \alpha + \beta X \quad \text{and} \quad E(Y \mid X, \bar{T}) = \bar{\alpha} + \beta X$$

and $\alpha < \bar{\alpha}$. We then have

$$E(Y \mid T) = \alpha + \beta\mu \quad \text{and} \quad E(Y \mid \bar{T}) = \bar{\alpha} + \beta\bar{\mu}$$

with $\mu = E(X \mid T)$, $\bar{\mu} = E(X \mid \bar{T})$. The paradox arises if $(\mu - \bar{\mu}) > (\bar{\alpha} - \alpha)/\beta$, assuming $\beta > 0$.

Just as the previous situation was concerned with the analysis of variance, this case is handled using covariance ideas. There is a substantial literature, see for example, Lord (1967) and Elashoff (1969) on when analysis of covariance is appropriate. Again considerations of exchangeability clarify the picture. If exchangeability in $Y$ given $X$ and treatment is appropriate as in the medical situation, the propensities provide for $p(Y \mid X, T)$ and in particular $E(Y \mid X, T)$. The required expectation is

$$\begin{aligned} E(Y \mid T) &= \int E(Y \mid X, T) p(X \mid T)\, dX \\ &= \alpha + \beta E(X \mid T) \end{aligned}$$

on assuming linearity. But $E(X \mid T)$ is not available from the data and hence the covariance adjustment is essential. On the other hand, in the agricultural case $Y$ and $X$ are exchangeable given the variety and $p(Y, X \mid T)$ is available from the data. In particular so is the marginal expectation $E(Y \mid T)$ and the covariance adjustment is unnecessary. It is often said that the covariate must not be associated with the treatment. The examples show that this is false. Notice also that the discussion does not involve considerations of normality, etc.

We now discuss an example from educational testing where the need for a covariance adjustment is none too clear but where exchangeability resolves the issue. Indeed, it was this case that started us on the whole discussion. An experiment was designed to investigate the effectiveness of one instructional method $T$ in comparison with the standard method $\bar{T}$. Two groups were chosen, one taught by $T$, and other by $\bar{T}$. The students were then given a test (called the posttest) and their scores $Y$ were recorded. Since the two groups may have had different abilities, a pretest was also given resulting in scores $X$. The problem would appear to be essentially the same as the medical one in which $X$, replacing sex, and $T$ influence $Y$ so that exchangeability is in $Y$ given $X$ and $T$, and the covariance adjustment for $x$ is necessary. We suggest this is not reasonable. For suppose You had pretest value $X = x$, would You consider yourself exchangeable with those who took part in the test and had score $x$? We suggest not, because $X$ is well known to depend on the group to which the student belongs: a value $x$ in a strong group is probably more indicative of ability than $x$ in a weak group. What You might do is to consider Yourself exchangeable with those students in the test having the same pretest

true-score as Yourself. But true scores have not been measured and so are not available from the data. The analysis can proceed as follows, all expectations being for the unit, You, and $\tau$ denoting the true score.

$$\begin{aligned} E(Y \mid T, X) &= E\{E(Y \mid T, X, \tau) \mid T, X\} \\ &= E\{E(Y \mid T, \tau) \mid T, X\} \\ &= E\{\alpha + \beta\tau \mid T, X\} \\ &= \alpha + \beta E(\tau \mid T, X) \end{aligned}$$

assuming linearity of regression. Similarly

$$E(Y \mid \bar{T}, X) = \bar{\alpha} + \beta E(\tau \mid \bar{T}, X).$$

Now $E(\tau \mid T, X) = \mu + \rho X$ and $E(\tau \mid \bar{T}, X) = \bar{\mu} + \rho X$ where $\rho$ is the reliability of the test. Hence

$$E(Y \mid T, X) = \alpha + \beta\mu + \beta\rho X$$

and

$$E(Y \mid \bar{T}, X) = \bar{\alpha} + \beta\bar{\mu} + \beta\rho X.$$

Hence the test is preferred if $\alpha + \beta\mu > \bar{\alpha} + \beta\bar{\mu}$; not necessarily if $\alpha > \bar{\alpha}$. Hence the covariance interpretation using $\alpha$ and $\bar{\alpha}$ is not the correct one. The same conclusion persists without a pretest since presumably $E(X \mid T) = E(X \mid \bar{T})$, the method being applied after the pretest.

Nothing essentially new happens when the third factor, previously $T$ and $\bar{T}$, becomes continuous, $Z$ say. We can have $E(Y \mid X, Z)$ say increasing in $Z$ for all $X$, so that large values of $Z$ are to be preferred, and yet $E(Y \mid Z)$ is decreasing in $Z$ suggesting small values. Again linear multiple regression provides an example with

$$E(Y \mid X, Z) = \alpha X + \beta Z$$

and $\beta > 0$, yet

$$E(Y \mid Z) = \alpha E(X \mid Z) + \beta Z = \alpha(\mu + \sigma Z) + \beta Z$$

with $\alpha\sigma + \beta < 0$. The consideration of exchangeability and calculation of the appropriate probability together resolve the problem.

With $X$, $Y$ and $Z$ continuous and linear relations obtaining, the analysis of Simon (1954) previously referred to may be employed. In the medical example, the sex, $X$, affected the treatment, $Z$, both of which affected the recovery, $Y$. In the agricultural situation, the variety $Z$, affected both the height, $X$, and the yield $Y$. The respective linear models are (medical)

$$\begin{aligned} a_{11}X &= u_1 \\ a_{21}X + a_{22}Y + a_{23}Z &= u_2 \\ a_{31}X + a_{33}Z &= u_3 \end{aligned}$$

and (agricultural)

$$\begin{aligned} a_{11}X + a_{31}Z &= u_1 \\ a_{22}Y + a_{23}Z &= u_2 \\ a_{33}Z &= u_3. \end{aligned}$$

Here the $u$'s are error terms and the $a$'s, nonzero constants. Direct calculation shows that the paradox can obtain and the relevant inferences made. Our approach avoids the restriction to linearity.

**7. Conclusions.** We have argued that the basic process of inference is the passage of a data set to uncertainty statements about another unit, as exemplified by "the probability that John will recover if he is given treatment $T$ is 0.6". The introduction of parameters, the usual subject of inference statements, may often be a most useful device, but is not, in our view, essential. Once this view of inference is adopted, one sees that an important aspect of the

inference is the linkage between the data and the new unit. We have argued that this linkage can be formulated in terms of judgments of exchangeability between the unit and the data; or, alternatively expressed, judgments of which subpopulation the unit belongs to. (In this paper we have confined ourselves to large data sets, and hence to large populations. Additional complexities arise with smaller data sets and considerations of finite exchangeability that it would need another paper to explore. Nevertheless the consideration of what one might do with the population is a prerequisite to considerations of inference with a small data set. Our discussion of covariance analysis illustrates this.) Once the linkage is established, frequencies (or propensities) in the identified subpopulations may be equated with the corresponding probabilities for the new unit. A final point is that the required probability may not be obtainable directly in this way and that other information besides that in the data may be needed to combine with that originating from the data to make the final inference.

Once it is recognized that inference involves the passage from a data set to a new unit, it is clear that there is no unique analysis of a data set; for it is possible to imagine two units, linked in quite different manners, to the set. Thus the data of Table 3, supposed from a city, might be applied in one way (joint exchangeability of disease and test result) to another person from the same city; but otherwise (exchangeability in test result given $D$) for someone from a different environment.

In applying the ideas of this paper it is first necessary to consider the unit about which inferences are to be made. What do You know about the unit? Its sex, $M$, for example. What features of the unit can be controlled? The treatment $T$, say. What feature is of interest? Its recovery, $R$, perhaps. Then You need to calculate Your probability of what is of interest, given what you know and can control: here $p(R \mid M, T)$. The only tool available is the probability calculus, principally Bayes theorem and extension to include other variables, relating the required probability to others. Which others are used depends on the connection with the data.

In our experience, it is generally fairly easy to make the appropriate judgments of exchangeability, or to recognize the relevant populations. Sometimes it is necessary to include other variables, for example, true score in the educational example of Section 6. A useful guide is the notion of causality, of which another useful guide is the temporal order: Varietal choice later produces height and yield, but sex and treatment later effect recovery. The important point to recognize is that exchangeability is a judgment by You, not a property of the external world. In this view, causation is a reflection of our judgment about the world and not a truth about it. In the present state of knowledge we may say smoking causes lung cancer, yet later we may revise this to say that a genetic factor causes both.

It is important to recognize that the methods described in this paper do not only apply to situations in which it has been possible to take random samples from a population. Of course, if this has been done then complete exchangeability is available and propensities may be identified with probabilities. But if not, then recognition of subpopulations enables some partial identifications of probabilities and propensities to be made, and the remaining probabilities—about which the data is uninformative—have to be assessed directly. We had an example of this in Section 4 where $p(D)$ had to be found from sources other than the data. It is a useful contribution to our understanding of a situation to be able to spell out clearly just what it is that the data tell us, and what has to be inferred by other means, in order to make the final inference.

## APPENDIX 1.

SIMPSON'S PARADOX. Consider the paradox in the notation of the paper referring to events $R$, $T$, and $M$. Without loss of generality suppose $T$ and $R$ are positively associated; that is, $p(R \mid T) > p(R \mid \bar{T})$, and write $T \sim R$. Similarly suppose, again without loss of generality, $R \sim M$. We prove the following result which does not seem to be available elsewhere:

THEOREM. *If Simpson's paradox holds (with $T \sim R$ and $R \sim M$), then $T \sim M$.*

(Table 2 provides an illustration of this. In words it says that the new factor $M$, must, with the conventions here adopted, be positively associated both with $R$ and $T$.)

The proof is most easily appreciated using the figure, the upper unit interval gives probabilities conditional on $T$, the lower on $\bar{T}$. The arrows connect probabilities having the same conditions except for $\bar{T}$ replacing $T$. The essence of the paradox is that those arrows that involve sex go to the right; those that do not, go to the left. (We have supposed, again without loss of generality that $p(R \mid TM) > p(R \mid T)$.) The key point is that $p(R \mid T)$ is a weighted average of $p(R \mid TM)$ and $p(R \mid T\bar{M})$ with weights $p(M \mid T)$ and $p(\bar{M} \mid T) = 1 - p(M \mid T)$. For the reversal of direction of the arrows to take place when sex is excluded the weights in the upper interval, given $T$, must differ from those in the lower, given $\bar{T}$. In the figure $p(R \mid T)$ is nearer to the upper right-hand $p(R \mid TM)$ than $p(R \mid \bar{T})$ is greater than that on $p(R \mid \bar{T}M)$; that is, $p(M \mid T)$ exceeds $p(M \mid \bar{T})$ as was required.

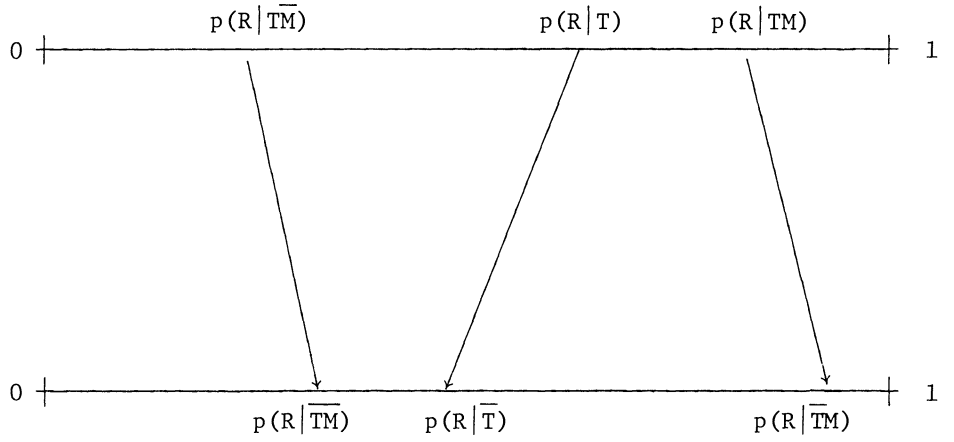


FIG. 1. *Simpson's paradox*

## APPENDIX 2

**A note on exchangeability in two variables.** If a set of units is exchangeable in $(X, Y)$ then it is both exchangeable in $Y$, and in $X$, given $Y$. This is clear from the defining relationship for exchangeability in $(X, Y)$, namely that $p(X_i = x_i, Y_i = y_i, \text{all } i)$ be invariant under relabelling of the units, by writing it as $p(X_i = x_i, \text{all } i \mid Y_i = y_i, \text{all } i)\, p(Y_i = y_i, \text{all } i)$, and considering the special case $Y_i = y$, all $i$. The converse of the statement in the first sentence is however not true. This is apparent since conditional exchangeability says nothing about probabilities of $X$-values given $Y$-values, except when the latter are all the same, and this is not enough to construct the defining relationship for exchangeability in $X$ and $Y$.

These considerations suggest an alternative definition of exchangeability in $X$, given $Y$, to that given in the body of the paper. This reads: A set of units is exchangeable in $X$, given $Y$, if $p(X_i = x_i, \text{all } i \mid Y_i = y_i, \text{all } i)$ is invariant under relabelling of the units. It is obvious on multiplying this by $p(Y_i = y_i, \text{all } i)$ that the converse is now true. We have used the (weaker) definition of conditional exchangeability because that is all that is needed to equate the propensity with the probability for a new unit. If one wishes to make inferences about several new units then the extended definition would be useful. To see this consider two new patients. H(enry) and J(ohn), who could be given either $T$ or $\bar{T}$. (We know them to be male and this condition is omitted from the notation.) To make inferences about their recovery we require probabilities exemplified by $p(HR, JR \mid HT, J\bar{T})$, where $HR$ means Henry recovers, etc. The new definition would enable this to be equated to a propensity. However we presumably judge

it to be equal to $p(HR \mid HT)p(JR \mid JT)$ and then the weaker form suffices. This condition is related to the assumption of "no interference between units" referred to by Rubin (1978).

## REFERENCES


[1] BLYTH, C. R. (1972). On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* **67** 364–366.

[2] COHEN, M. R. and NAGEL, E. (1934). *An Introduction to Logic and Scientific Method.* Harcourt Brace, New York.

[3] DAWID, A. P. (1976). Properties of diagnostic data distributions. *Biometrics* **32** 647–658.

[4] DE FINETTI, B. (1974). *Theory of Probability*, 2 vols. Wiley, London.

[5] ELASHOFF, J. D. (1969). Analysis of covariance: A delicate instrument. *Amer. Educ. Res. J.* **6** 383–401.

[6] FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.

[7] FISHER, R. A. (1958). Cigarettes, cancer and statistics. *Centennial Rev.* **2** 151–166.

[8] KELLEY, T. L. (1923). *Statistical Methods.* Harvard, Cambridge.

[9] KENDALL, M. G. and STUART, A. (1969). *The Advanced Theory of Statistics*, 3rd ed. Griffin, London.

[10] LONGLEY-COOK, L. H. (1962). An introduction to credibility theory. *Proc. Casualty Actuar. Soc.* **49** 194–221.

[11] LORD, F. M. (1967). A paradox in the interpretation of group comparisons. *Psych. Bull.* **68** 304–305.

[12] LORD, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psych. Bull.* **72** 336–337.

[13] MEEHL, P. E. and ROSEN, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psych. Bull.* **52** 194–216.

[14] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psych.* **66** 688–701.

[15] RUBIN, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6** 34–58.

[16] SAVAGE, L. J. (1962). *The Foundations of Statistical Inference.* Methuen, London.

[17] SAVAGE, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.* **4** 441–483.

[18] SIMON, H. A. (1954). Spurious correlation: a causal interpretation. *J. Amer. Statist. Assoc.* **49** 467–479.

[19] SIMPSON, E. H. (1951). The interpretation of interaction contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** 238–241.

[20] SUPPES, P. (1970). *A Probabilistic Theory of Causality.* North-Holland, Amsterdam.

[21] TODA, M. (1977). Causality, conditional probability and control. In *New Developments in the Applications of Bayesian Methods.* 109–124. (A. Aykac, and C. Brumat, eds.) North-Holland, Amsterdam.

[22] WHITNEY, A. W. (1918). The theory of experience rating. *Proc. Casualty Actuarial Soc.* **4** 274–292.



DEPARTMENT OF STATISTICS
UNIVERSITY COLLEGE
GOWER STREET
LONDON WC1E 6BT
ENGLAND

DIVISION OF EDUCATIONAL PSYCHOLOGY
MEASUREMENT AND STATISTICS
THE UNIVERSITY OF IOWA
IOWA CITY, IOWA 52242