ED 237 515 TM 830 049

AUTHOR	Holland, Paul W.; Rubin, Donald B.
TITLE	On Lord's Paradox. Program Statistics Research.
INSTITUTION	Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
REPORT NO	ETS-PSRP-TR-82-34; ETS-RR-82-36
PUB DATE	21 May 82
NOTE	47p.; Prepared for the Festschrift in honor of Frederic M. Lord, May 22-23, 1982.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Estimation (Mathematics); *Influences; *Mathematical Models; *Research Methodology; *Statistical Analysis; Statistics
IDENTIFIERS	Causal Inferences; *Causal Models; *Lord (Frederic M); Lords Paradox

ABSTRACT

Lord's Paradox is analyzed in terms of a simple mathematical model for causal inference. The resolution of Lord's Paradox from this perspective has two aspects. First, the descriptive, non-causal conclusions of the two hypothetical statisticians are both correct. They appear contradictory only because they describe quite different aspects of the data. Second, the causal inferences of the statisticians are neither correct nor incorrect since they are based on different assumptions that our mathematical model makes explicit, but neither assumption can be tested using the data set that is described in the example. We identify these differing assumptions and show how each may be used to justify the differing causal conclusions of the two statisticians. In addition to analyzing the classic "diet" example which Lord used to introduce his paradox, we also examine three other examples that appear in the three papers where Lord discusses the paradox and related matters. (Author)

RR-82-36

On Lord's Paradox

Paul W. Holland
and
Donald B. Rubin

ETS

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 82-34

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

On Lord's Paradox

Paul W. Holland

and

Donald B. Rubin

Program Statistics Research
Technical Report No. 82-34

Research Report RR-82-36

May 21, 1982

Prepared for the Festschrift in honor of Frederic M. Lord, May 22-23, 1982. The preparation of this paper was supported in part by the Program Statistics Research Project, Educational Testing Service, Princeton, New Jersey, 08541.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the statisticians in the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

ABSTRACT

Lord's Paradox is analyzed in terms of a simple mathematical model for causal inference. The resolution of Lord's Paradox from this perspective has two aspects. First, the descriptive, non-causal conclusions of the two hypothetical statisticians are both correct. They appear contradictory only because they describe quite different aspects of the data. Second, the causal inferences of the statisticians are neither correct nor incorrect since they are based on different assumptions that our mathematical model makes explicit but neither assumption can be tested using the data set that is described in the example. We identify these differing assumptions and show how each may be used to justify the differing causal conclusions of the two statisticians. In addition to analyzing the classic "diet" example which Lord used to introduce his paradox, we also examine three other examples that appear in the three papers where Lord discusses the paradox and related matters.

Table of Contents

Page

1. Introduction 1
2. A Model for Causal Inference 2
   - 2.1 The Elements of the Model 2
   - 2.2 Discussion of the Elements of the Model . . . 3
   - 2.3 Three Kinds of Studies 8
   - 2.4 Causal Effects and Related Quantities in Causal Studies 9
3. Lord's Examples 10
   - 3.1 Example 1: Two Subpopulations Receiving One Treatment 11
   - 3.2 Example 2: A Descriptive Study 20
   - 3.3 Example 3: Contemplating New Treatments . . . 23
   - 3.4 Example 4: Two Explicit Treatments 27
4. Discussion 29

Appendix: Randomization and Inference for Causal Effects 32

- A.1 The Completely Randomized Experiment 32
- A.2 Causal Effects in Subpopulations 32
- A.3 Randomization Within Subpopulations 33
- A.4 Randomization Based on a Covariate 34
- A.5 Randomization Based on a Covariate Within Subpopulations 36
- A.6 Descriptive Studies 37

References 40

1. Introduction

Lord's Paradox first appeared in a short, two-page article (Lord, 1967) in Psychological Bulletin. This article presents a remarkable contrast between two statisticians who draw widely different conclusions from the same set of data. The culprit appears to be that the analysis of covariance cannot be counted on to make a proper allowance for uncontrolled preexisting differences between natural groups. Much to the dismay of the editor of Psychological Bulletin, Lord did not resolve his paradox. This fact increases the interest in the questions it raises. The impact of the paper has been an extensive analysis and criticism of the use of the analysis of covariance that still continues (e.g., Games, 1976 and Lindley and Novick, 1981). Lord wrote two additional short pieces on the paradox -- a second article in Psychological Bulletin, (Lord, 1968) and an entry in the Encyclopedia of Educational Evaluation (Lord, 1973 in Anderson, et al., 1973, page 233). We base our discussions on these three articles by Lord.

Lord uses examples to illustrate his points, and there are four examples discussed in the three papers. Our approach differs from Lord's in that we first present a mathematical framework that is complex enough to accommodate what we regard to be the important features of the examples treated by Lord, and we then apply this framework to each of his examples. As will become evident, we believe that there are several different issues that arise in these examples, and we feel that our mathematical framework provides the structure for a precise analysis.

Our paper is organized as follows. In Section 2 we describe the general mathematical framework or model for causal inference. In Section 3 we apply this general framework to each of the examples appearing in Lord's three papers. Section 4 gives our general conclusions regarding the nature of Lord's Paradox. We include an Appendix which indicates various related results that follow from our model.

2. A Model for Causal Inference

In this section we describe our model for causal inference and derive the results from it that we need for the examples that Lord discussed. More technical consequences of the model are derived in the Appendix.

2.1 The Elements of the Model

The chief issue that is of concern in Lord's Paradox is the attribution of cause. Much has been written about causation but our point of departure is the analysis of causal effects given in Rubin (1974, 1977, 1978, 1980). However, it will be sufficient for our purposes to deal with a simplified, population-level, version of Rubin's model. We have used this simplified model elsewhere (Holland and Rubin, 1980) to analyze causal inference in retrospective, case-control studies often used in medical research.

Our model is similar to those used to describe many simple statistical problems. However, we are absolutely explicit about certain distinctions and elements that are usually left implicit in other discussions. We believe that it is impossible to give a coherent analysis of causal inference without being at least as explicit as we are here.

The basic elements of our model are:

(a) a population of units, P,

(b) an "experimental manipulation", with levels t or c, and its associated indicator variable, S,

(c) a subpopulation indicator variable, G,

(d) an outcome variable, Y, and

(e) an concomitant variable, X.

Each of these components to the model needs further specification and we do this in the next subsection. Figure 1 summarizes this framework.

Figure 1 about here

2.2 Discussion of the Elements of the Model

The population P of units underlies the rest of the model. Typical examples of "units" are human subjects, or rats, or households, or corn seeds. All variables are assumed to be functions that are defined on every unit in P. All probabilities, distributions and expected values are computed over P. A probability will mean nothing more nor less than a proportion of units in P. The expected value of a variable is merely its average value over all of P. Conditional expected values are subgroup averages where the subgroups are defined by the conditioning statement. In Figure 1 there are N units in P.

The "experimental manipulation" is the focus of all causal inference in our model. It is important to realize that by using the term "experimental manipulation" we do not mean to limit our discussion to the activities within a controlled randomized laboratory study. We do

Figure 1: A Framework for Causal Inference

				Y		X	
		S	G	Y_t	Y_c	X_t	X_c
Population P of Units	1	t or c	1 or 2				
	2						
	⋮						
	N						

mean to include any sort of well-defined experience to which each of the units in P may or may not be exposed. The key notion is the potential for exposing each unit to any one of the experimental conditions in the study. For causal inference, it is critical that each unit be potentially exposable to any one of the experimental conditions. As an example, the schooling a student receives is an experimental manipulation in our sense, whereas the student's race or gender is not.

For simplicity, in this paper we shall assume that there are just two different experimental conditions or levels of treatment, denoted t (treatment) and c (control). We will let S be a variable that indicates the experimental condition to which each unit in P is exposed, that is, S = t indicates the unit is exposed to t, while S = c indicates exposure to c. In a controlled study, S is constructed by the experimenter. In an uncontrolled study S is determined to some extent by factors beyond the experimenter's control. In either case, the critical feature of the "experimental manipulation" is that the value of S for each unit could have been different.

We will make the simplifying assumption that S is defined on all of P so that for each unit either S = t or S = c. In two of the examples in Section 3, S is a constant over P (i.e., there is only one treatment to which units are actually exposed). In one example, there is no S since there is no identified treatment. Our model is at the "population level" because we do not consider the inference problems associated with the sampling of units in P for study. The model described by Rubin (1974, 1977, 1978, 1980) deals with the added complexity of the sampling of units.

The term "independent variable" is often used to refer to a treatment indicator variable like S, but it is also applied more loosely to include an entirely different type of variable. In our model this second type of variable is the subpopulation indicator variable G. Evans and Anastasio (1968), among others, distinguish clearly between "genuine independent variables, treatments that can be manipulated" and "classifications or other variables which describe the intact groups." Lord is quite aware of the distinction and, in fact, describes his paradox as "a problem that arises in interpreting data on preexisting groups." He also refers to the impossibility of random assignment in the "comparison of the educational achievements of different racial groups." In our model we have two different variables S and G, in order to represent both cases. The variable G indicates the subpopulation membership of each unit, such as race or gender of students, "varieties" of corn, etc. Unlike S, it is not possible for the value of G for each unit to have been other than what it is. For the purposes of this paper, we have a single subpopulation indicator variable G which has only two possible values (G = 1,2) indicating, for example, male and female students, as shown in Figure 1.

By the "outcome variable" Y, we mean to convey the usual notion of dependent or criterion variable, with one important extension. When there is an experimental manipulation, there are multiple versions of Y, one for each treatment condition. In our case, these are denoted by Y_t and Y_c. The interpretation of these two values of Y for a given unit is that Y_t is the value of Y that would be observed if the unit were exposed to t while Y_c is the value of Y that would be observed *on the same unit* if it were exposed to c. The basic notion that a treatment *influences* the dependent variable is formalized in the model by the two values, Y_t and Y_c. If t influences Y, then the effect of the experimental manipulation is to make the value Y_t different from the value Y_c for each unit. The null hypothesis of "no treatment effect" (in its strongest form) corresponds to $Y_t = Y_c$ for all units in P. In studies where there is no experimental manipulation, there is only one version of Y. In such cases, we do not put a subscript on Y, nor do we subscript Y when we are referring to it without reference to the treatment conditions.

Central to Lord's Paradox is the availability of a variable X that is auxiliary to the outcome variable Y. We will call X a *concomitant variable* to distinguish it from Y. However, in Lord's examples there are two distinct types of concomitant variables that arise -- those that *are* and those that *are not* potentially influenced by the experimental manipulation. This state of affairs can be expressed as follows. Let X_t and X_c denote the value of X that would be observed if the unit were exposed to t or to c, respectively. If $X_t = X_c$ for all units, then X is not influenced by the treatment, and in this special case we shall call X a *covariate*. If $X_t \neq X_c$ for some units, then it is not a covariate, but we will still use the more general term *concomitant variable* to describe X in this case. By definition, the subpopulation indicator variable G is an example of a covariate. In real-life research designs, the question of whether or not $X_t = X_c$ can be quite serious and difficult to answer. Under the usual circumstances of educational research, pretests <u>are</u> covariates because they are recorded prior to the exposure of units to the treatment conditions, and they are, hence, not affected by exposure to one treatment or another.

2.3 Three Kinds of Studies

The primary purpose of our model is to allow an explicit description of the quantities that arise in three types of studies that we shall refer to as:

A) Descriptive studies,

B) Uncontrolled causal studies, and

C) Controlled causal studies.

Although this is an oversimplified categorization of research studies, we believe that it captures important distinctions that are germane to our analysis of Lord's Paradox.

A <u>descriptive study</u> has no experimental manipulation so that there is only one version of Y and of X and no treatment indicator variable S.

<u>Controlled</u> and <u>uncontrolled causal studies</u> both have an experimental manipulation and differ only in the degree of control that the experimenter has over the treatment indicator, S. In a controlled causal study, the values of S are determined by the experimenter and can depend on numerous aspects of each unit, e.g., subpopulation membership, values of covariates, but not on the value of Y_t or Y_c since the value of the outcome variable is observed after the values of S are determined by the experimenter. In an uncontrolled causal study the values of S are determined by factors that are beyond the experimenter's control. Of critical importance is the fact that, in a controlled causal study S can be <u>made</u> to be statistically independent of Y_c and Y_t whereas in an

uncontrolled causal study this is not true. All of Lord's examples concern either descriptive studies or uncontrolled causal studies; these are the types of studies that commonly arise in the behavioral sciences and involve "preexisting groups".

2.4 Causal Effects and Related Quantities in Causal Studies

The causal effect of t on Y (relative to c) for each unit in P is given by the difference*, $Y_t - Y_c$. This is the amount that t has increased (or decreased) the value of Y (relative to c) on each unit. The expected value $E(Y_t - Y_c)$ is the average causal effect of t versus c on Y in P. Since the expected value of a difference is the difference in expected values, i.e.

$$E(Y_t - Y_c) = E(Y_t) - E(Y_c), \qquad (2.1)$$

we see that the unconditional means of Y_t and Y_c over P have direct causal interpretations.

In a causal study, whether controlled or uncontrolled, the value of Y that is observed on each unit is Y_S, so that when S = t, Y_t is observed and when S = c, Y_c is observed. Hence the expected value of Y for the "treatment group" is the following conditional expectation:

$$\underline{\text{treatment}}\ \underline{\text{group}}\ \underline{\text{mean}} = E(Y_t | S = t). \qquad (2.2)$$

The mean of Y for the "control group" is

$$\underline{\text{control}}\ \underline{\text{group}}\ \underline{\text{mean}} = E(Y_c | S = c). \qquad (2.3)$$

*In a more general setting the definition of a causal effect at the unit level would not require that the subtraction, $Y_t - Y_c$, be meaningful. This is beyond the scope of this paper but is discussed in Holland and Rubin (1980).

In general, there is no reason why $E(Y_t)$ and $E(Y_t|S = t)$ should be equal. Similarly for $E(Y_c)$ and $E(Y_c|S = c)$. Hence, in general, neither $E(Y_t|S = t)$ nor $E(Y_c|S = c)$ has a direct causal interpretation.

However, $E(Y_t)$ and $E(Y_t|S = t)$ are always related through this basic equation:

$$E(Y_t) = E(Y_t|S = t)\ P(S = t) + E(Y_t|S = c)\ P(S = c). \tag{2.4}$$

Similarly,

$$E(Y_c) = E(Y_c|S = c)\ P(S = c) + E(Y_c|S = t)\ P(S = t). \tag{2.5}$$

Note that equation (2.4) involves the average value of Y_t among those units exposed to c. Similarly, equation (2.5) involves the average value of Y_c among those units exposed to t. But, $E(Y_t|S = c)$ and its companion $E(Y_c|S = t)$ can never be directly measured except when Y_t and Y_c can both be observed on all units. This is the fundamental problem of causal inference. In the Appendix we show how experimental randomization resolves this problem by making (2.1) equal to the difference between (2.2) and (2.3).

3. Lord's Examples

Lord uses four principal examples over the course of his discussion in Lord (1967), Lord (1968), and Lord (1973). Each example is a ficticious research study that could describe a real-life investigation. Example 1 is usually referred to as Lord's Paradox. The other three examples amplify the issues that arise there. In this section, we analyze each example in terms of the model given in Section 2.

3.1 Example 1: Two Subpopulations Receiving One Treatment

Lord's famous paradox is the centerpiece of both Lord (1967) and Lord (1973) and a variant of it is mentioned briefly in Lord (1968). It is introduced in Lord (1967) with this short paragraph:

> A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded.

There is no other information describing this hypothetical study in the three papers; but other information is given describing the observed data values. Nevertheless, from this short description we can identify all of the relevant elements of the model. Table 1 summarizes this identification.

- - - - - - - - - -

Table 1 goes here

- - - - - - - - - -

The question mark (?) in Table 1 is due to the fact that although the dining hall diet is clearly the treatment, t, whose effect on student weight is sought by the study; there is no control diet, c, even hinted at in the three papers. In our model the influence of t on Y is always relative to some other condition c. The fact that, in this example, Y_c is vaguely defined and not observed directly plays a crucial role in our analysis of the paradox. It should be remembered

TABLE 1

Identification of the Elements of the Model in Example 1

Study Design

P : The students at the university in the specified school year,

t : The dining hall diet,

c : ?

S : $S = t$ for all units.

Variables Measured

G : Student gender (1 = male, 2 = female),

X : The weight of a student in September,

Y : The weight of a student in June.

that Y_c represents the weight in June of a student exposed to the control diet. Since no one is exposed to c, anyone analyzing the data will be forced to make untestable assumptions about the value of Y_c in order to obtain numerical answers to causal questions.

There is only one version of X in this example since it is measured in September, prior to the onset of the treatment; hence, X is a covariate. Finally, since all students are exposed to t and none to c we have S = t for all students.

Lord frames his paradox in terms of the analyses of two hypothetical statisticians who come to quite different conclusions from the data in this example. The samples are all assumed to be large so that the focus is on the interpretation of the values of parameters that have been estimated with high precision. We shall summarize all statistical analyses in terms of the parameters that are estimated. The effect of the dining hall diet on a student's weight is given by the difference $Y_t - Y_c$, so that the average causal effect of the diet on student weight is the expected value of this difference. However, one of the features of this study is an expressed interest in "... any sex differences in these effects." Thus, the average causal effects for males and for females need to be separately estimated. The parameters of interest are the average causal effects for males and for females

$$\Delta_i = E(Y_t - Y_c | G = i), \quad i = 1,2, \tag{3.1}$$

and the difference of average causal effects,

$$\Delta = \Delta_1 - \Delta_2. \tag{3.2}$$

In terms of the individual subpopulation averages, Δ may be expressed either as

$$\Delta = [E(Y_t | G = 1) - E(Y_c | G = 1)] - [E(Y_t | G = 2) - E(Y_c | G = 2)], \tag{3.3}$$

or as

$$\Delta = [E(Y_t | G = 1) - E(Y_t | G = 2)] - [E(Y_c | G = 1) - E(Y_c | G = 2)]. \tag{3.4}$$

Equation (3.4) is especially useful in this example since it separates the observed Y_t from the unobserved Y_c.

Statistician 1 bases his conclusion about the effect of the diet on the difference between the distributions of Y_t and of X in each subpopulation -- i.e., males and females. In terms of the means of these distributions, the corresponding parameters are the average differences

$$D_i = E(Y_t - X | G = i), \quad i = 1,2. \tag{3.5}$$

The quantity D_i is the mean weight gain in subpopulation i. The difference of the gains is

$$D = D_1 - D_2. \tag{3.6}$$

From the description of the pattern of data values given by Lord in this example Statistician 1 *observes* that there are no differences between the beginning and ending weight distributions for either males or females. Thus the D_i in (3.5) are both zero. From this observation, Statistician 1 *concludes* that:

> ... as far as these data are concerned, there is no evidence of any interesting effect of diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change.

This causal inference is not true without making additional assumptions. The D_i in (3.5) are not average causal effect parameters. In drawing his conclusion, Statistician 1 is making an assumption about the numerical values of the unobserved variable Y_c. There are several possible assumptions he could make to justify his conclusion. One of the simplest is to assume that the response to the control diet, whatever it might be, is given by the student's weight in September, i.e.

$$Y_c = X. \tag{3.7}$$

Under this entirely untestable assumption, the D_i in (3.5) are equal to the average causal effects parameters Δ_i in (3.1).

In Lord (1968), Lord makes a brief reference that is related to the assumption (3.7). He refers to critics of Lord (1967) who suggest that "the obvious procedure to use" is the "gain score", $Y_t - X$. We would interpret such critics as attempting to obtain an estimate of the causal effect of the dining hall diet on each student by making assumption (3.7). Since assumption (3.7) cannot be tested with the available data, acceptance or criticism of it must be based on intuition and/or subject-matter experience.

Statistician 2 computes a covariance adjusted difference of the two subpopulation means. This corresponds to computing the following two conditional expectations (i.e., within-group regression functions):

$$E(Y_t | X, G = i), \quad i = 1,2. \tag{3.8}$$

The mean, conditional, weight gain in group i at X is

$$D_i(X) = E(Y_t - X | X, G = i), \quad i = 1,2. \tag{3.9}$$

The difference in these conditional weight gains at X is

$$D(X) = D_1(X) - D_2(X). \tag{3.10}$$

For simplicity, Lord assumes that the conditional expectations in (3.8) are both linear and parallel. Thus we can write

$$E(Y_t | X, G = i) = a_i + bX, \quad i = 1,2. \tag{3.11}$$

Hence, $D_i(X)$ simplifies to

$$D_i(X) = a_i + (b-1)X, \quad i = 1,2, \tag{3.12}$$

and D(X) simplifies to

$$D(X) = a_1 - a_2. \tag{3.13}$$

Thus, D(X) is independent of the value of X. Statistician 2 correctly interprets D(X) as the average amount more that a male (G = 1) will weigh in June than will a female (G = 2) of equal initial weight, X. Although correct, this statement about D(X) bears no direct relevance to the differential causal effect of the dining hall diet on the June weights of male and female students. This is because D(X) in (3.10) is not directly related to the causal effect parameters Δ_1, Δ_2 and Δ given in (3.1) and (3.2).

However, under an untestable assumption that is akin to but different from (3.7), D(X) equals Δ and consequently does measure the differential causal effect of interest. To see this we generalize the assumption (3.7) to

$$Y_c = \alpha + \beta X. \tag{3.14}$$

Assumption (3.14) asserts that a student's weight in June under the control diet, Y_c, is a deterministic linear function of the value of the student's weight in September, X. Furthermore, the same linear function applies to all students regardless of gender. The assumption of Statistician 1 is that of no weight change under the control diet: i.e., $\alpha = 0$, $\beta = 1$. If Statistician 2 makes the alternative assumption that $\beta = b$ where b is the common slope of the two within-groups regression lines in (3.11), then he may interpret D(X) in (3.10) as the difference in causal effects, Δ defined in (3.2). We omit the straightforward algebra that shows this. These results are summarized in Table 2:

- - - - - - - - -

Table 2 goes here

- - - - - - - - -

We wish to emphasize that the assumptions that lead to the formulas used by the two statisticians in Table 2 are not the only ones, nor are they the most general ones. For example, Statistician 1 could make the weaker assumption that $E(Y_c|G = i) = E(X|G = i)$ instead of (3.7). Any assumption about Y_c must be untestable in this example and yet will lead to a formula for Δ. The plausibility of any particular assumption about Y_c must be argued from considerations external to the data,

TABLE 2

A Summary of Two Sets of Assumptions That Lead to the Conclusion of Each Hypothetical Statistician in Lord's Paradox

Both assume $Y_c = \alpha + \beta X$ for all units in P.

	Statistician 1	Statistician 2
Testable Assumptions		$E(Y_t \mid X, G = i) = a_i + bX$
Untestable Assumptions	$\alpha = 0$ $\beta = 1$	$\beta = b$
Formula for causal effects Δ_i	$\Delta_i = E(Y_t - X \mid G = i)$	$\Delta_i = E(Y_t - \alpha - bX \mid G = i)$
Formula for differential causal effect Δ	$\Delta = E(Y_t \mid G = 1) - E(Y_t \mid G = 2) - [E(X \mid G = 1) - E(X \mid G = 2)]$	$\Delta = E(Y_t \mid G = 1) - E(Y_t \mid G = 2) - b[E(X \mid G = 1) - E(X \mid G = 2)]$
	= difference in mean weight gains	= covariance adjusted mean difference in June weights

and in many cases particular assumptions may be perfectly reasonable. There are statements in Lord (1967) and Lord (1973) that suggest that Lord would be willing to accept the assumption that justifies Statistician 1 rather than the one that justifies Statistician 2. Our view is slightly different. To paraphrase Lord, there is no statistical procedure that can be counted on to make untestable assumptions that are correct. In the case of the diet example, neither assumption seems obviously appropriate.

In summary, we believe that the following views resolve Lord's Paradox. If both statisticians made only descriptive statements, they would both be correct. Statistician 1 makes the unconditional descriptive statement that the average weight gains for males and females are equal; Statistician 2 makes the conditional (on X) statement that for males and females of equal September weight, the males gain more than the females. In contrast, if the statisticians turned these descriptive statements into causal statements, neither would be correct or incorrect because untestable assumptions determine the correctness of the causal statements. These sets of assumptions are outlined in Table 2. In a sense then, Statistician 1 is wrong because he makes a causal statement without specifying the assumption needed to make it true. Statistician 2 is more cautious, since he makes only a descriptive statement. However, unless he too makes further assumptions, his descriptive statement is completely irrevelant to the campus dietician's interest in the effect of the dining hall diet.

3.2 Example 2: A Descriptive Study

This example is given at the beginning of Lord (1968) as an illustration of a type of situation in which the analysis of covariance is often applied. Lord gives only the following discussion of Example 2:

> ...a group of underprivileged students is to be compared with a control group on freshman grade-point average (y). The underprivileged group has a considerably lower mean grade-point average than the control group. However, the underprivileged group started with a considerably lower mean aptitude score (x) than did the control group. Is the observed difference between groups on y attributable to initial differences on x? Or shall we conclude that the two groups achieve differently even after allowing for initial differences in measured aptitude?

In attempting to identify the various elements of the model of Section 2 for this example we must decide whether this study is intended to be descriptive or causal. This decision hinges on the interpretation given to the "control group". "Underprivileged" refers to a vague mixture of social, nutritional, economic and educational circumstances and sometimes even to racial differences. In some unusual circumstances, such as with twins separated shortly after birth, it can be reasonable to consider "underprivileged" as an experimental manipulation; in such cases, both the mean aptitude score X and the freshman grade point average Y would be affected by exposure to this experimental manipulation, and both would be represented in our model by two versions, i.e., X_t, X_c, Y_t, Y_c. Although it is conceptually possible to regard "control" and "underprivileged" as two levels of an experimental manipulation, in practice it is often unreasonable to do so since the exposure essentially begins at birth. Hence we shall interpret this example simply as a descriptive study in which there are two subpopulations (i.e., "underprivileged" and the "control group") being compared. Table 3 identifies the elements of the model with the interpretation of Example 2 as a descriptive study.

Table 3 goes here

The concomitant variable X defines a subpopulation of P for each of its values, i.e. the subpopulation of P for which X = 75. In terms of our model, it is not possible to ask if the value of Y for a unit would be different had the value of X for that unit been different. This fact renders meaningless the question of whether or not an observed difference between two groups on Y is attributable to differences in the values of X for the two groups. In order to attribute cause to the values of a variable (i.e., to estimate a causal effect in our model), it is necessary for these values to indicate the levels of a treatment. Hence, causal statements involving the influence of a concomitant variable on a dependent variable are generally not meaningful. However, there are useful descriptive parameters that can be estimated in this

TABLE 3

Identification of the Elements of the Model in Example 2

Study Design

P : The freshman class at the university in a given year.

Variables Measured

G : Underprivileged status (1 = underprivileged, 2 = control).

X : Score on an aptitude test taken prior to college entrance.

Y : Freshman grade-point average.

type of study. The mean difference between the grade-point average of students in the two subpopulations with the same value of X is given by the difference between the two regression functions

$$E(Y|G = 1, X) - E(Y|G = 2, X). \tag{3.15}$$

This difference may be useful for predictive purposes, but it cannot be given a causal interpretation in our model.

3.3 Example 3: Contemplating New Treatments

Lord gives this example in Lord (1968). His description is as follows:

> ...Suppose an agronomist is studying the yield of various varieties of corn. He plants 20 flower pots with seeds of a "black" variety and 20 more pots with seeds of a "white" variety. For simplicity of illustration, suppose that he treats all 40 plants equally for several months, after which he finds that the white variety has yielded considerably more marketable grain than the black variety. However, it is a fact that black variety plants average only 6 feet high at flowering time, whereas white variety plants average 7 feet. He now asks the question, would the black variety produce as much salable grain if conditions were adjusted so that it averaged 7 feet in height at flowering time?

Table 4 identifies the elements of the model in this example.

- - - - - - - - -

Table 4 goes here

- - - - - - - - -

This example is like the first one in that only one level of the experimental manipulation occurs in the study. However, Lord is quite clear in this example as to the problems created by not having an explicitly defined alternative experimental condition. In fact, the question he raises in this example concerns the choice of t. In his words:

> In practice, the answer depends on what we do to secure black-variety plants averaging 7 feet in height. This could be done by destroying the shorter plants, by applying more fertilizer, or by stretching the plants at night while they are young, or by other means. *The answer depends on the means used.*

The role of the concomitant variable in this example is quite different from the previous ones. It is evident that the measured value of X will be affected by t since that would be the stated purpose of the treatment. Thus, there are two versions of X, X_t and X_c, and only X_c is measured in this study. Not only must one make untestable assumptions as to the value of Y_t, it is also necessary to make assumptions about the value of X_t. The parameter of interest in this example is the average causal effect on yield for the "black" variety,

TABLE 4

Identification of the Elements of the Model in Example 3

Study Design

P : Corn seeds.

t : ?

c : The "standard" treatment applied by the agronomist.

S : S = c for all units.

Variables Measured

G : Corn variety (1 = "black", 2 = "white").

X : Height at flowering time.

Y : Amount of marketable grain produced.

i.e.

$$E(Y_t - Y_c|G = 1) = E(Y_t|G = 1) - E(Y_c|G = 1). \qquad (3.16)$$

The value of $E(Y_c|G = 1)$ can be computed from the data, but the value of $E(Y_t|G = 1)$ is determined by whatever untestable assumptions we make.

Let $\mu_t(x)$ and $\mu_c(x)$ be defined by

$$\begin{aligned} \mu_t(x) &= E(Y_t|G = 1, X_t = x) \\ \mu_c(x) &= E(Y_c|G = 1, X_c = x), \end{aligned} \qquad (3.17)$$

so that $\mu_t(x)$ is the regression of Y on X under treatment t for the "black" variety, and $\mu_c(x)$ is this regression under treatment c. To obtain an "analysis of covariance" solution we may assume that these two regression functions are equal, i.e.

$$\mu_t(x) = \mu_c(x). \qquad (3.18)$$

Let us also suppose that this regression is linear, i.e.

$$\mu_c(x) = a_c + b_c\ x. \qquad (3.19)$$

Assumption (3.18) is untestable, but assumption (3.19) can be tested with the data. We may then compute the unknown quantity in (3.16), $E(Y_t|G = 1)$, by the formula

$$E(Y_t|G = 1) = E(\mu_c(X_t)|G = 1) = a_c + b_c\ E(X_t|G = 1). \qquad (3.20)$$

Since the mean of Y_c for G = 1 can be expressed as

$$E(Y_c|G = 1) = E(\mu_c(X_c)|G = 1) = a_c + b_c\ E(X_c|G = 1), \qquad (3.21)$$

the average increase in yield for the "black" variety is:

$$E(Y_t - Y_c|G = 1) = b_c[E(X_t|G = 1) - E(X_c|G = 1)] \qquad (3.22)$$

which is an "analysis of covariance" solution. However, we agree with Lord that the plausibility of the untestable assumption (3.18) depends on the choice of t. For example, it might be a plausible assumption if "additional fertilizer" is the new treatment, but "stretching the young plants at night" might only lengthen them with no corresponding change in yield or might kill them, and in either case (3.18) would not be appropriate.

3.4 Example 4: Two Explicit Treatments

Although the first three examples are intended to illustrate certain points and are not considered by Lord as indicative of real research studies, the final example, in Lord (1973), illustrates that "the paradox is not just an amusing statistical puzzle." Lord's statement of the example is as follows.

> ...consider the problem of evaluating federally funded special education programs. A group of disadvantaged children are pretested in September, then enrolled in a special program, and finally posttested in June. A control group of children are similarly pretested and posttested but not enrolled in the special program. Since the most disadvantaged children are selected for the special program, the control group...will typically have higher pretest scores than the disadvantaged group.

This is the first of these examples in which two levels of an experimental manipulation are explicitly present. Table 5 identifies the elements of the model in this example.

- - - - - - - - - -

Table 5 goes here

- - - - - - - - - -

TABLE 5

Identification of the Elements of the Model in Example 4

Study Design

P	:	The students in the specific schools in the given school year.
t	:	The special education program.
c	:	The standard educational program.
S	:	Treatment indicator.

Variables Measured

G	:	Disadvantaged indicator (1 = disadvantaged, 2 = control).
X	:	Pretest in September.
Y	:	Posttest in June.

Even though two treatments are explicitly defined, there is ambiguity as to how they are assigned and what the relationship between S and G is. The remark "Since the most disadvantaged children are selected for the special program" might be read as meaning that the selection of a unit into a treatment group is made on the basis of the pretest score, with the lower scoring children more likely to be enrolled in the special program. On the other hand, the description might be interpreted as implying that S = G and that G indicates a classification of children into "disadvantaged" and "control", not determined by X. The differences between these two possibilities are of fundamental importance.

First, suppose that assignment to t or c was based on the value of X, and that the values of G are just labels determined by the covariate X. If the regressions of Y_c and Y_t on X are linear and parallel, then, as we show in the Appendix, the usual covariance adjusted estimator estimates the causal effect, $E(Y_t - Y_c)$.

In contrast, suppose that S = G and that there are two existing subpopulations indicated by G, and that G is not a function of X alone. Now S and G are completely confounded, so that in order to estimate the effect of t vs. c on Y for each subpopulation, we must make assumptions about the values of Y_t and Y_c for the groups exposed to c and t respectively. These assumptions will be untestable and similar to those made in Example 1.

4. Discussion

We believe that Lord touched upon a number of important issues in the examples that surround his paradox. The blind use of complicated statistical procedures, like the analysis of covariance, is doomed to lead to absurd conclusions. On the other hand, the analysis of covariance is a useful tool that can often render an apparently intractable problem manageable. We think that the value of the model described in Section 2 is that it forces one to think carefully about the attribution of cause. Causal statements made in natural language are often vague and potentially misleading. The role of mathematics is to give precision to natural language statements, and we believe that this is an important aspect of our analysis of Lord's Paradox.

We believe that the appropriate way to resolve Lord's Paradox is to be absolutely explicit about the untestable assumptions that need to be made to draw causal inferences. These assumptions all involve the responses of units to a treatment to which they are unexposed and thereby turn observations about data (i.e., descriptive conclusions) into causal inferences. We only disagree with the tone of Lord's three articles that suggests the analysis of covariance cannot be trusted except under special experimental designs. We feel that our model shows that in most complex studies in which causal inferences are of concern, there are always both testable and untestable assumptions that must be made in order to draw causal conclusions. We believe that it is both scientifically necessary and pragmatically helpful to make these assumptions explicit.

The distinction between causal inference and descriptive inference is essential in many contexts, and this distinction is clarified by our framework. For example, questions such as "Is the new diet more effective for males or females?" are causal and imply a comparison of an outcome for the new diet with an outcome for the control diet. Similar sounding questions may not be causal and involve no attribution of cause. For example, "Who gained more under the new diet, males or females?" is not a causal question, but a purely descriptive one, and, as such, it can be answered without making the assumptions necessary for causal inferences. Descriptive questions differ from causal questions in that there is no implied comparison of the values of an outcome variable under different levels of an experimental manipulation.

As illustrated in the Appendix, the calculations required to answer descriptive questions may, in some cases, be identical to the calculations that are required to answer causal questions under specific assumptions. The scientific and practical interpretations of the results of the calculations are, however, dramatically different for descriptive and causal questions. The Appendix shows how experimental randomization can alleviate the problem of having to make untestable assumptions to draw causal inferences. This should not be interpreted as meaning that randomization is necessary for drawing causal inferences. In many cases, appropriate untestable assumptions will be well supported by intuition, theory, or past evidence. In such cases, we should not avoid drawing causal inferences and hide behind the cover of uninteresting descriptive statements. Rather we should make causal statements that explicate the underlying assumptions and justify them as well as possible.

Appendix: Randomization and Inference for Causal Effects

We now shall show how randomization and related topics can be brought into the model and how they allow causal inferences to be drawn using standard statistical methods.

A.1 The Completely Randomized Experiment

Randomization has a powerful effect and a special place in our model. In a completely randomized study, great effort is made to insure that S is statistically independent of all other variables in the study. In particular S is made to be independent of Y_t and Y_c. Hence we have

$$E(Y_t) = E(Y_t | S = t) = E(Y_t | S = c) \tag{A.1}$$

and

$$E(Y_c) = E(Y_c | S = c) = E(Y_c | S = t). \tag{A.2}$$

The crucial consequence of randomization in our model is that it forces the equality of the average causal effect and the treatment-control-group mean difference:

$$E(Y_t - Y_c) = E(Y_t | S = t) - E(Y_c | S = c). \tag{A.3}$$

A.2 Causal Effects in Subpopulations

When subpopulations have been defined using G, it is natural to want to estimate a causal effect in each subpopulation. By analogy with equation (4.3), the average causal effect in subpopulation i is:

$$E(Y_t - Y_c | G = i) = E(Y_t | G = i) - E(Y_c | G = i). \tag{A.4}$$

Thus, the unconditional means of Y_t and Y_c for the units with G = i have direct causal interpretations. However, the expected values of Y for treated and control units with G = i is given by, in analogy with (2.2) and (2.3),

$$\underline{\text{treatment group mean for } G = i \text{ units}} = E(Y_t | S = t, G = i) \qquad (A.5)$$

and

$$\underline{\text{control group mean for } G = i \text{ units}} = E(Y_c | S = c, G = i). \qquad (A.6)$$

The quantities in (A.4) are related to the quantities in (A.5) and (A.6) by the following equations which are analogous to equations (2.4) and (2.5):

$$E(Y_t | G = i) = E(Y_t | S = t, G = i) P(S = t | G = i) + E(Y_t | S = c, G = i) P(S = c | G = i), \qquad (A.7)$$

$$E(Y_c | G = i) = E(Y_c | S = c, G = i) P(S = c | G = i) + E(Y_c | S = t, G = i) P(S = t | G = i). \qquad (A.8)$$

Note that equation (A.7) involves the mean of Y_t for units exposed to c with G = i and equation (A.8) involves the mean of Y_c for units exposed to t with G = i, i = 1,2. But $E(Y_t | S = c, G = i)$ and $E(Y_c | S = t, G = i)$ can never be directly measured. As with causal effects in the population, randomization plays a special role when estimating causal effects in subpopulations.

A.3 <u>Randomization Within Subpopulations</u>

Suppose that within each subpopulation, S is independent of (Y_t, Y_c). This will hold, for example, in completely randomized experiments and in "randomized block" experiments, where different randomization rules might be used within each subpopulation. For example, when $G = 1$, the probability of being treated is .4 whereas when $G = 2$, the probability of being treated is .6. If S is conditionally independent of (Y_t, Y_c) given G, then

$$E(Y_t|G = i) = E(Y_t|S = t, G = i) = E(Y_t|S = c, G = i),$$

and

$$E(Y_c|G = i) = E(Y_c|S = c, G = i) = E(Y_c|S = t, G = i).$$

Thus randomization within subpopulations forces the within subpopulation equality of the average causal effect and the treatment-control-group mean difference, i.e.

$$E(Y_t - Y_c|G = i) = E(Y_t|S = t, G = i) - E(Y_c|S = c, G = i).$$

A.4. Randomization Based on a Covariate

Suppose the concomitant X is a covariate so that $X = X_t = X_c$. When a covariate is observed before treatment conditions are selected, it can be used to select units into treatment conditions. For example, let X be a pretest, and suppose students with low scores of X are assigned with high probability to take a special educational program, those with middle scores are assigned with equal probability to the special and regular programs, and those with high scores are assigned with high probability to the regular program.

In such a situation, the randomization is a function of the observed value of X, and it follows that S and Y_1, Y_2 are conditionally independent given X. Hence,

$$E(Y_t|X) = E(Y_t|S = t, X) = E(Y_t|S = c, X) \qquad (A.9)$$

and

$$E(Y_c|X) = E(Y_c|S = c, X) = E(Y_c|S = t, X). \qquad (A.10)$$

The importance of equations (A.9) and (A.10) is that from the observed data (Y_S, S, X) we may estimate these regressions:

$$E(Y_t|S = t, X) \quad \text{and} \quad E(Y_c|S = c, X).$$

From (A.9) and (A.10) it follows that these regressions equal $E(Y_t|X)$ and $E(Y_c|X)$, respectively. Now suppose that $E(Y_t|X)$ and $E(Y_c|X)$ are linear, say

$$E(Y_t|X) = \alpha_t + \beta_t X \qquad (A.11)$$

and

$$E(Y_c|X) = \alpha_c + \beta_c X. \qquad (A.12)$$

Then the least squares regression of Y_t on X for the treatment group units estimates equation (A.11), and the least squares regression of Y_c on X for the control group units estimates equation (A.12). (Of course, there are other ways to estimate these conditional expectations when they are linear and more generally, when they are not (e.g., see Rubin, 1977)).

Suppose that we have estimated $E(Y_t|X)$ and $E(Y_c|X)$; how can we estimate the average causal effect $E(Y_t - Y_c)$ in P? Let P(X) represent the distribution of X in P. Then

$$E(Y_t - Y_c) = \sum_X [E(Y_t|X) - E(Y_c|X)] \, P(X). \qquad (A.13)$$

That is, the average causal effect of t versus c on Y in P is simply the average value of the difference between the conditional expectations of Y_t and of Y_c at X, where the average over X is weighted to reflect the proportion of units at each value of X. If

$$E(Y_t|X) - E(Y_c|X) = K \quad \text{for all } X, \qquad (A.14)$$

then the causal effect of t versus c is the same for all X, and equals the causal effect of t versus c in P. When (A.14) holds, the averaging in (A.13) is irrevelant. Assumption (A.14) (i.e., parallel regressions) when combined with the linearity assumptions (A.11) and (A.12) yields the model underlying the usual covariance adjusted estimator. That is, if

$$E(Y_t|X) = \alpha_t + \beta X$$

and

$$E(Y_c|X) = \alpha_c + \beta X$$

then

$$E(Y_t - Y_c) = \alpha_t - \alpha_c.$$

Thus, the standard analysis of covariance estimator is appropriate when (a) assignment into treatment group is based on X, and (b) the t and c regressions of Y on X are linear and parallel. Rubin (1977) discusses this case and more complicated ones.

A.5 Randomization Based on a Covariate Within Subpopulations

The argument of Section A.4 can be extended to cases with subpopulations. An example of such a study would be an evaluation of the effects of a special diet (S = t) versus a normal diet (S = c) for males (G = 1) and females (G = 2), in which the probability of assignment to treatment depends on initial weight (X) with different assignment rules being used for males and females (e.g., for X = weight in pounds, $P(S = 1|X, G = 1) = [1 + X/150]^{-1}$ and $P(S = 1|X, G = 2) =$ $[1 + X/120]^{-1}$). In such cases, S is conditionally independent of (Y_t, Y_c) given (G, X).

The entire argument of Section A.4 can be applied separately to each subpopulation indicated by G. Having obtained estimates of the causal effect of t versus c in each subpopulation, these estimates can be averaged (weighted by the relative frequency of the subpopulation) to obtain an estimate for the entire population. Alternatively, the difference between the subpopulation estimates can be computed in order to estimate the differential causal effect of t versus c in the two subpopulations.

It is important to note that this comparison of the sizes of the causal effects relies on the assumption of the conditional independence of S and (Y_t, Y_c) given (X, G) and involves the comparison of Y_t and Y_c, only one of which can be observed on each unit; this assumption has been called "strongly ignorable treatment assignment" in Rosenbaum and Rubin (1982) and plays a central role in causal inference.

A.6 Descriptive Studies

Descriptive studies are different from causal studies in that there is no experimental manipulation involved and therefore there is only one version of Y. The treatment indicator is not even defined in this case. For example, suppose G = 1 for males, G = 2 for females, Y = June weight in pounds and X = previous September weight in pounds.

One descriptive question is, "How much more do males weigh in June than do females?" The answer is given by the parameter:

$$E(Y|G = 1) - E(Y|G = 2).$$

Another descriptive question is, "How much more weight have males gained from September to June than have females?" It is answered by

$$\begin{aligned} &E(Y - X|G = 1) - E(Y - X|G = 2) \\ &= \Big[E(Y|G = 1) - E(X|G = 1)\Big] - \Big[E(Y|G = 2) - E(X|G = 2)\Big] \\ &= \Big[E(Y|G = 1) - E(Y|G = 2)\Big] - \Big[E(X|G = 1) - E(X|G = 2)\Big]. \end{aligned}$$

More complicated questions can be formulated by conditioning on X. For example: "How much more do males with September weight X weigh in June than do females with the same September weight, X?" It is answered by

$$E(Y|G = 1, X) - E(Y|G = 2, X). \qquad (A.15)$$

If the regressions of Y on X are linear and parallel in the subpopulations, i.e.

$$E(Y|G = i, X) = \alpha_i + \beta X, \quad i = 1,2,$$

then (A.15) equals $\alpha_1 - \alpha_2$ for all X, which is estimated by the standard analysis of covariance estimator. It is critical to realize, however, that the analysis of covariance estimator in this case is answering a purely descriptive question and not a causal question.

If the regressions of Y on X are not parallel in the subpopulation, i.e., if (A.15) is not constant for all X, then the answers to such descriptive questions as "How much more do males with September weight X weigh in June than do females with September weight X?" depend on the value of X. Sometimes, an average answer may be desired, and then the difference given by (A.15) will be averaged over the distribution of X in some standard population, say P:

$$\sum_{X} [E(Y|G = 1, X) - E(Y|G = 2, X)] P(X). \qquad (A.16)$$

Although (A.16) looks formally similar to (A.13), (A.13) is the answer to a causal question since it involves the comparison of Y_t and Y_c, whereas (A.16) is the answer to a descriptive question since it involves the comparison of the distribution of Y for two different values of G.

REFERENCES

Anderson, S.B. et al. (1973). Encyclopedia of Educational Evaluation. San Francisco, CA: Jossey-Bass.

Evans, S.H. and Anastasio, E.J. (1968). "Misuse of Analysis of Covariance when treatment effect and covariate are confounded." Psychological Bulletin, 69, 225-234.

Games, P.A. (1976). "Limitations of Analysis of Covariance on Intact Group quasi-experimental Designs." Journal of Experimental Education, 44, 51-54.

Holland, P.W. and Rubin, D.B. (1980). "Causal Inference in Case-control Studies". Jerome Cornfield Memorial Lecture, American Statistical Association Meetings, Houston, August.

Lindley, D.V. and Novick, M.R. (1981). "The role of exchangeability in inference." Annals of Statistics, 9, 45-58.

Lord, F.M. (1967). "A Paradox in the Interpretation of Group Comparisons." Psychological Bulletin, 68, 304-305.

Lord, F.M. (1968). "Statistical Adjustments When Comparing Preexisting Groups." Psychological Bulletin, 72, 336-337.

Lord, F.M. (1973). "Lord's Paradox." In Anderson, S.B. et al. Encyclopedia of Educational Evaluation. San Francisco, CA: Jossey-Bass.

Rosenbaum, P.R. and Rubin, D.B. (1982). "The Central Role of the Propensity Score in Observational Studies." To appear in Biometrika.

Rubin, D.B. (1974). "Estimating causal effects of treatments in randomized and non-randomized studies." Journal of Educational Psychology, 66, 688-701.

Rubin, D.B. (1977). "Assignment to treatment group on the basis of a covariate." Journal of Educational Statistics, 2, 1-26.

Rubin, D.B. (1978) "Bayesian inference for causal effects: The role of randomization." The Annals of Statistics, 7, 34-58.

Rubin, D.B. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test," by Basu. The Journal of the American Statistical Association, 75, 591-593.