A PARADOX IN THE INTERPRETATION OF GROUP COMPARISONS

FREDERIC M. LORD

Educational Testing Service

Attention is called to a basic source of confusion in the interpretation of certain types of group comparison data.

It is common practice in behavioral research, and in other areas, to apply the analysis of covariance in the investigation of preexisting natural groups. The research worker is usually interested in some criterion variable \( y \) and would like to make allowances for the fact that his groups are not matched on some important independent variable or "control" variable \( x \). The situation is such that observed differences in the dependent variable might logically be caused by differences in the independent variable, and the research worker wishes to rule out this possibility.

It is widely recognized that ideally the research worker should assign cases or individuals at random to the groups that are to be studied by analysis of covariance. In behavioral research and in many other areas, such random assignment is usually difficult or impossible—as, for example, in a comparison of the educational achievements of different racial groups. The research worker usually uses analysis of covariance regardless, or he may try to resort to a simple and direct interpretation of group means.

The present note points out a type of problem that arises in interpreting data on preexisting groups. The difficulty can most easily be pointed out with the help of a hypothetical illustrative example.

A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference 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.

At the end of the school year, the data are independently examined by two statisticians. Both statisticians divide the students according to sex. The first statistician examines the mean weight of the girls at the beginning of the year and at the end of the year and finds these to be identical. On further investigation, he finds that the frequency distribution of weight for the girls at the end of the year is actually the same as it was at the beginning.

He finds the same to be true for the boys. Although the weight of individual boys and girls has usually changed during the course of the year, perhaps by a considerable amount, the group of girls considered as a whole has not changed in weight, nor has the group of boys. A sort of dynamic equilibrium has been maintained during the year.

The whole situation is shown by the solid lines in the diagram. Here the two ellipses represent separate scatterplots for the boys and the girls. The frequency distributions of

![Diagram](image-url)
initial weight are indicated at the top of the diagram and the identical distributions of final weight are indicated on the left side. People falling on the solid 45° line through the origin are people whose initial and final weight are identical. The fact that the center of each ellipse lies on this 45° line represents the fact that there is no mean gain for either sex.

The first statistician concludes that as far as these data are concerned, there is no evidence of any interesting effect of the school 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.

The second statistician, working independently, decides to do an analysis of covariance. After some necessary preliminaries, he determines that the slope of the regression line of final weight on initial weight is essentially the same for the two sexes. This is fortunate since it makes possible a fruitful comparison of the intercepts of the regression lines. (The two regression lines are shown in the diagram as dotted lines. The figure is accurately drawn, so that these regression lines have the appropriate mathematical relationships to the ellipses and to the 45° line through the origin.) He finds that the difference between the intercepts is statistically highly significant.

The second statistician concludes, as is customary in such cases, that the boys showed significantly more gain in weight than the girls when proper allowance is made for differences in initial weight between the two sexes. When pressed to explain the meaning of this conclusion in more precise terms, he points out the following: If one selects on the basis of initial weight a subgroup of boys and a subgroup of girls having identical frequency distributions of initial weight, the relative position of the regression lines shows that the subgroup of boys is going to gain substantially more during the year than the subgroup of girls.

The college dietician is having some difficulty reconciling the conclusions of the two statisticians. The first statistician asserts that there is no evidence of any trend or change during the year for either boys or girls, and consequently, a fortiori, no evidence of a differential change between the sexes. The data clearly support the first statistician since the distribution of weight has not changed for either sex.

The second statistician insists that wherever boys and girls start with the same initial weight, it is visually (as well as statistically) obvious from the scatterplot that the subgroup of boys gains more than the subgroup of girls.

It seems to the present writer that if the dietician had only one statistician, she would reach very different conclusions depending on whether this were the first statistician or the second. On the other hand, granted the usual linearity assumptions of the analysis of covariance, the conclusions of each statistician are visibly correct.

This paradox seems to impose a difficult interpretative task on those who wish to make similar studies of preformed groups. It seems likely that confused interpretations may arise from such studies.

What is the "explanation" of the paradox? There are as many different explanations as there are explainers.

In the writer's opinion, the explanation is that with the data usually available for such studies, there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled pre-existing differences between groups. The researcher wants to know how the groups would have compared if there had been no pre-existing uncontrolled differences. The usual research study of this type is attempting to answer a question that simply cannot be answered in any rigorous way on the basis of available data.

(Received May 10, 1965)