A Conundrum in the Analysis of Change

Garrett Fitzmaurice, ScD

From the Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA

Most of us believe that we have a good intuition for how to measure change: simply compare what happened before with what happened after. Therefore, it may come as a surprise to learn that the statistical analysis of change has been mired in controversy for many years. Unfortunately, much of the discussion and debate concerning the appropriate analysis of change has generated more heat than light. As a result, researchers are often confused about how best to analyze change in even the most rudimentary circumstances. In this article, I consider the analysis of change in one of the simplest longitudinal study designs: the parallel-groups repeated-measures design. In this study design, there are two or more groups of subjects, each measured repeatedly on two or more occasions. For simplicity, I focus on two groups of subjects measured on just two occasions, resulting in before and after measures of some outcome variable of interest. The primary goal of this study design is to compare the changes that the two groups experience, usually in response to some intervention administered during this period. In essence, the statistical analysis of data from this study design attempts to compare the groups after removing any initial differences between them. However, the manner in which these initial differences are eliminated can have subtle and significant effects on the analysis, especially on the interpretation of the results.

To highlight some of the conceptual difficulties that can arise in the analysis of change, consider the following (hypothetical) example of a study designed to assess the efficacy of a new diet pill, Diagra. In this study, investigators enrolled equal numbers of male and female subjects. Before receiving a 3-mo supply of Diagra, each subject's body weight was recorded. Three months later, subjects returned to the clinic to have a second measure of body weight recorded. The goal of the study was to examine the effects of Diagra and compare the change in weight experienced by men and women. When classified by sex, the following descriptive statistics for body weight were obtained: women had initial and final mean body weights of 130 and 120 lb, respectively (mean $\delta = 10$ lb), and men had initial and final mean body weights of 160 and 150 lb, respectively (mean $\delta = 10$ lb).

Overall, there appeared to be an average reduction in weight of 10 lb. However, to confirm her impression that the reduction in weight had been the same for men and women, one investigator conducted the following analysis of the data. She constructed a simple change score, say $\delta = (Y_2 - Y_1)$, where Y_1 is the initial measure of body weight and Y_2 is the postintervention measure of body weight. The investigator reasoned that any initial differences between the groups were eliminated by measuring changes from the initial weights. The investigator then conducted a two-sample *t* test[1] to compare men and women in terms of their average weight changes. The resulting *t* test confirmed her initial impression and she concluded that "there was no statistically discernible difference between men and women in terms of their average weight loss over 3 mo."

However, her coinvestigator, with the content of a recent course on regression and analysis of variance still fresh in his mind, considered an alternative analysis of the same data. He also reasoned that any initial differences between groups should be eliminated but decided to remove these initial differences with analysis of covariance. That is, he performed a regression analysis of the postintervention weight Y_2, with initial weight Y_1 and sex as covariates. In graphic terms, this analysis amounted to fitting two parallel lines for the relation between initial and final weights, one for males, the other for females. Although the slopes for the two regression lines were assumed to be the same (hence, the lines were parallel), the intercepts differed significantly.* These results showed an interesting differential effect of Diagra on the sexes. Unlike his colleague, the investigator concluded that "when differences in initial weight between men and women were properly accounted for, women showed a significantly greater decline in weight than men."

When the two investigators discussed their respective results, there was much confusion and consternation. The first investigator was convinced that their study had not demonstrated any differential effects of the new diet pill on the sexes. She argued that men and women had the same average weight loss of 10 lb. The second investigator was equally convinced that their study had provided evidence that women showed a significantly greater decline in weight than men, once any differences in initial weight had been "properly accounted for." Exasperated by their paradoxic conclusions, they agreed to seek counsel from their local statistician. To their surprise, the statistician pointed out that there was really no contradiction in the two sets of findings. However, he cautioned that much greater care had to be taken in how the results were interpreted because the two methods for analyzing the data were directed toward answering different scientific questions.

How did our two investigators arrive at apparently conflicting conclusions about the subject matter? In both cases, the investigators attempted to eliminate initial differences between the groups, but in quite different ways. The first investigator subtracted the initial weight from the final weight, producing a change score; the second investigator eliminated initial differences within an analysis of covariance, the latter producing an adjusted change score. As a result, the two analyses address somewhat different questions. Specifically, the first investigator conducted an analysis appropriate for answering the following scientific question: Is there any difference between the average weight loss of men and women? This might be thought of as an "unconditional" question, in the sense that it compares the average (or unconditional mean) weight loss in one population (say, males) with the average weight loss in another population (say, females). In this particular example, the correct answer to the

*Of note, regression analysis of the change score δ, with initial weight Y_1 and gender as covariates would yield identical regression coefficients for the gender effects. That is, an analysis of the covariance of δ (or an analysis of adjusted change) yields estimated intercepts for males and females that are identical to those from an analysis of covariance of Y_2. Thus, the difference in intercepts represents not only gender differences in the adjusted Y_2 means but also gender differences in the adjusted change score means.

Correspondence to: Garrett Fitzmaurice, ScD, Department of Biostatistics, Harvard School of Public Health, 655 Huntingdon Avenue, Boston, MA 02115, USA. E-mail: fitzmaur@hsph.harvard.edu

0899-9007/01/$20.00
PII S0899-9007(00)00593-1

question is a resounding no. The first investigator quite rightly concluded that there were no differences between men and women in their average change or decline in weight. The second investigator presented the results of an analysis that addressed a somewhat different scientific question, namely the "conditional" question: Is there any difference between the expected weight loss of a man and woman who have the same initial weight? That is, it answers the question: Is a female expected to lose more weight than a male, given that they both have the same initial weight? In this particular example, the correct answer is a resounding yes.

Why is that so? When the "conditional" question was posed in this way, we expected that women would lose more weight than men. The reasoning is as follows. If a man and a woman had the same initial weight, then 1) the woman was initially overweight and, as a result, was expected to lose weight (even if there was no effect due to the diet pills), or 2) the man was initially underweight and, as a result, was expected to gain weight (in the absence of any effect due to the diet pills).

The reason for these changes, apart from any putative effect due to the diet pills, is simply due to the phenomenon known as *regression to the mean*. Recall from a previous article[2] that regression to the mean necessarily occurs whenever there is less-than-perfect correlation between two variables (e.g., the less-than-perfect correlation between two measurements of weight taken 3 mo apart). Whenever two variables have a correlation less than 1 (and greater than −1), individuals with extreme values on one variable will, on average, have less extreme values on the other variable. This means that, when the same variable (e.g., body weight) is measured on two occasions, individuals who are extreme on the first occasion will be somewhat less extreme on the second occasion. That is, women who were overweight on the first occasion should regress toward their respective population mean (i.e., lose weight) on the second measurement occasion. By the very same token, initially underweight men should regress toward their respective population mean (i.e., gain weight) on the second measurement occasion. When this effect of regression to the mean is combined with any putative weight-loss effect that can be directly attributed to the diet pills, women should lose more weight than men (who have the same initial body weight). As a result, the second investigator was correct in his conclusion of a differential effect of the new diet pill on the sexes. However, it might be argued that he provided the correct answer to a question that might not have been of real scientific interest in this particular example. That is, although the conclusion is correct, the underlying premise of the question is somewhat incongruous.

In summary, the statistical analysis of change has generated heated debate among statisticians and researchers alike. In even the simplest of settings, there are subtle issues concerning the interpretation of alternative analytic methods that need to be considered very carefully. In my hypothetical example, two methods for analyzing change were considered: change-score analysis and analysis of covariance (or adjusted change-score analysis). Although these methods led to apparently conflicting results, the paradox lay in the interpretation of the analyses. This paradox, also known as *Lord's paradox* (named after Frederic Lord[3] who eloquently brought the issue to light), has led many researchers astray over the years. The paradox is resolved by noting that these alternative methods of analysis answer somewhat different scientific questions. In general, the choice between methods should be made on substantive grounds. That is, the design of the study and the research question should guide the choice of analytic method. The analysis of change scores is appropriate when the primary goal is to compare distinct populations in terms of their average changes over time. The analysis of change scores answers the question: Do the populations differ in terms of their average change? In general, analysis of covariance is preferred in cases where individuals have been assigned to groups at random (e.g., in a randomized clinical trial) or where the population distributions of the initial scores can reasonably be assumed to be equal (even though the sample means of the initial scores may differ across groups). In cases where the population distributions of the initial scores are equal, it is meaningful to ask the question: Is the expected change the same in all groups, when we compare individuals having the same initial score? Furthermore, analysis of covariance will provide a more powerful test of group differences. The latter has often been touted as the main reason analysis of covariance is preferred over analysis of change scores. This faulty rationale has blinded many researchers to the potential difficulties in interpreting the results of analysis of covariance when the assumption of equal population distributions of initial scores is not tenable. In conclusion, it is the study design and the scientific question of interest and not issues of statistical precision and power that should determine the choice between methods for analyzing change.

REFERENCES

1. Gauvreau K, Pagano M. Student's *t* test. Nutrition 1993;9:38
2. Fitzmaurice G. Regression to the mean. Nutrition 2000;16:81
3. Lord F. A paradox in the interpretation of group comparisons. Psychol Bull 1967;68:304