METHODS & DESIGNS

Not-quite-random assignments[1]

DONALD MAINLAND,[2] *NEW YORK UNIVERSITY MEDICAL CENTER, 550 First Avenue, New York, New York 10016*

A problem of random assignment is discussed.

A Bit of Matching

If we are about to compare two treatments, such as drugs, on different Ss and if we know, or have good reason to believe, that different classes of Ss (e.g., males and females, or children, young adults, the middle-aged, and the aged) will respond differently, or to a different degree, to one or both treatments, we know that we ought to apply the following rules: (1) Separate the Ss into the appropriate classes. (2) Randomly assign the treatments separately within each class.

In some instances we can subject all the resulting data to one rather complex analysis and then seek for interclass differences afterward; but in many cases we have to analyze each class separately and then we may adopt various ways of combining the results. This is, however, not the subject of this note, which springs from the following comment:

"While I agree in principle with an overall randomized assignment procedure, I practice stratified randomization when I expect to have small groups and want to be sure to have, for example, the age-sex sub-groups balanced. If I have only two young females I would prefer to have one in each test group than both in the same group, though I don't expect to be able to spot even a 'trend.' However, age and sex are usually such important factors in so many of the studies I have made that I think it is better to balance first and then study the data to see whether most of the difference found between the two groups is concentrated in the one sex or the one age level (young, middle or old)."

I must confess that, even after years of following the twofold rule just quoted, I have an instinctive leaning toward the reader's method of trying to match the samples. However, I have had too much experience in digging singed moths out of candle wax to feel much confidence in an instinctive impulse without a critical look at its possible consequences. In trying to take a critical look at the effect of such an attempt to improve on random assignment, I have at various times worked with fictitious figures to compare the probabilities obtained from completely random assignment and the not-quite-random type. I now believe that we do not need such a numerical effort in order to see the objection to the attempted improvement.

An Experiment Followed by a Random-Frequency Test

Let us suppose that we are testing two treatments, A and B, that we have two females, and a larger, but even, number of males. We assign one female to A and one to B, strictly at random; then we assign an equal number of males to each treatment, again by a strictly random method. We assess the outcome in each individual as X or Not-X, and after the experiment we put the assessment from each individual into the appropriate cell of a fourfold contingency table. We can then apply a random-frequency test—either Fisher's "exact" test or the fourfold-table chi-square test (with Yates's correction), which gives a close approximation to the probability P obtained by the "exact" test. (The same principles would apply if the outcome of the experiment were assessed by measurement and we applied an appropriate random-frequency test, a randomization test, a rank-sums test, or the t test. Therefore this variant need not be discussed separately.)

The random-frequency test answers the question: "If the total numbers of Xs and Not-Xs were randomly divided, many times, between Samples A and B, how often would the A-B differences be as large as, and larger than, the difference found in the actual experiment?" The point to note is that in some of these randomizations both the females would be assigned to A, in others both would be assigned to B, and in others one would be in A and the other in B. Obviously, we cannot assume that the frequencies of the various A-B differences would be the same in this overall randomization as it would be if we reproduced by randomization the actual design of our original experiment, because in the latter we always put one female in each group.

The discrepancy between the two sets of frequencies might not be great, but we could only find it out by a special analysis after each particular experiment in which we had done such a partial matching. It seems very doubtful if such an effort would be worthwhile, because, if we place no restriction on the randomization in the original experiment and then apply a standard test, if there is no difference between the effect of A and B, we set our own risk of being misled (e.g., P = 0.05). We are thereby making allowance for all the various ways in which random assignment can produce differences, including the placing of the two females in the same treatment group.

This case exemplifies a very general rule—that the analysis of the data should be determined by the design of the experiment.

If the investigator felt uncomfortable with the unrestricted randomization because he believed that females were likely to differ greatly from males in their response to either A or B or both treatments, it would seem that he ought to leave the two females out of the experiment, because whatever results he obtained from the bisexual samples he could not, in view of his belief, logically assume that it would apply to both sexes.

We have considered here an extreme case, two females among a larger number of males; but the same considerations would apply if the disproportion were not quite as great. The picture would be even more complex if an attempt were made to match on two attributes, such as sex and age. If it is really desirable to match Ss, the best way, in general, is to analyze each subgroup separately and then combine the information. However, we ought at the beginning to decide whether or not we will include a few Ss of various kinds that differ from the majority.

A Few Odd Ones?

Relevant to this question is a statement made in an editorial on "The Controlled Therapeutic Trial" (British Medical Journal, 2:791-792, 1948) written by Sir A. Bradford Hill. The statement is as follows:

"In such trials it is often tempting to add little groups of patients of differing types here, there and everywhere with the object of learning rather more. Though with a statistical design it will certainly sometimes pay to do so, very often the rather more becomes the rather less. Such a trial gives doubtful answers to the many points but not decisive answer to any."

Population Estimates

Let us suppose now that an investigator with two female Ss

and many more males has been influenced by the emphasis on population estimates in place of random-frequency tests. He intends, after the experiment, to see what the A-sample and the B-sample would tell him about the frequencies of Xs and Not-Xs in their respective populations if they were random samples of those populations. He might say: "Would it not be desirable to make the populations as alike as I can in sex ratio by putting a female in each treatment group before I do the experiment?"

It seems to me that he would thereby be creating an artificial similarity of the two populations. Whatever the actual sex ratios were in either population, unless he had a very large random sample, it would be more likely to differ from the population sex ratio than to be identical with it, and for this identity to occur simultaneously in samples from two populations would be very rare. As with the random-frequency test, if he really believed that sex made a big difference in outcome (Xs vs Not-Xs), he would be better advised to exclude the females from the experiment.

NOTES

1. Reprinted from Note 32 which appears in Mainland, D. Notes in Biometry in Medical Research. V.A. Monograph 10-1, Supplement 5, October 1968, pp. 15-18. Published by the Veterans Administration Department of Medicine and Surgery.

2. Medical Statistics Unit and Rheumatic Diseases Study Group, New York University Medical Center, Room 1106, 112 East 19th St., New York, N.Y. 10003.