5-1-2005

Teaching Random Assignment: Do You Believe It Works?

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 221-226

1538 – 9472/04/$95.00

Teaching Random Assignment: Do You Believe It Works?

Shlomo S. Sawilowsky
Educational Evaluation & Research
Wayne State University

Textbook authors admonish students to check on the comparability of two randomly assigned groups by conducting statistical tests on pretest means to determine if randomization worked. A Monte Carlo study was conducted on a sample of n = 2 per group, where each participant's personality profile was represented by 7,500 randomly selected and assigned scores. Independent samples t tests were conducted and the results demonstrated that random assignment was successful in equating the two groups on 7,467 variables. The students' focus is redirected from the ability of random assignment to create comparable groups to the testing of the claims of randomization schemes.

Key words: Random assignment, Monte Carlo, comparable groups

Introduction

Random assignment is one of the more difficult concepts in designing experiments. Researchers harbor considerable distrust in the ability of random assignment to create comparable groups. Interestingly, the seeds of distrust in random assignment are sown in statistics and research textbooks. For example, in a pretest-posttest treatment vs control group design, Tuckman (1994) noted, "It is not uncommon to assign Ss randomly to groups and then to check on the distribution of control variables by comparing the groups to assess their equivalence on these variables" (p. 130). Students are told to check on the comparability of the two groups by conducting statistical tests on the pretest means, as Krathwohl (1993) stated, "The pretest tells us whether randomization worked and the groups are really comparable" (p. 452).

This article is based on a presentation delivered in 1999 to the American Educational Research Association, Special Interest Group Educational Statisticians, Montreal, Canada. The author gratefully acknowledges discussions with Drs. Lori Rothenberg and Randy Lattimore on earlier versions of this article. Email the author at: shlomo@wayne.edu.

This problem is exacerbated when researchers consider the typical small samples available for research in applied fields. For example, Gall, Borg, and Gall (1996) stated,

> The probability that random assignment will produce initially equivalent treatment groups increases as the size of the sample in each group increases. For example, equivalent groups will more likely result if 100 individuals are randomly assigned to two treatment groups (n = 50 per group) than if 10 individuals are assigned to those to groups (n = 5 per group). p. 489.

Similar statements are found in Cook and Campbell (1979), Crowl (1996), Vockell and Asher (1995), and others.

A Previous Demonstration

Strube (1991) noted "small samples cause other problems that argue against their routine use". Indeed, small samples present difficulties with regard to the generalizability of results. Strube (1991) endeavored to show that "the probability of an erroneous inference" is "generally no greater than the nominal Type I error rate" (p. 346). In this respect, Strube's (1991) article was convincing.

However, Strube's (1991) demonstration may not have been the most

effective approach in convincing researchers of the ability of random assignment to produce baseline equality among two groups. First, Strube (1991) used a relatively complex design: a 2 x 2 (treatment vs control × nuisance variable) with samples sizes from N = 8 to 100. Second, Strube (1991) modeled the presence of effect size from .25 to 4.00, such as Cohen's d, where

$$d = \frac{\left(\overline{x}_1 - \overline{x}_2\right)}{s_p}, \qquad (1)$$

and s_p is the pooled standard deviation. These effect sizes were treated as small to very large nuisance parameters. They were deliberately introduced as the error terms in the simulation, as opposed to studying the behavior of random assignment on random fluctuations. A simpler demonstration is clearly warranted.

Methodology

In order to explicate the effects of random assignment, and to demonstrate to researchers that it indeed works with a sample as small as N= 4 or n = 2 per group, a Monte Carlo study was conducted. A program was written in Fortran 90/95. Façade, a personality profile, was created by dimensioning four arrays. Each of the four façade arrays, representing a participant's profile, contained 7,500 values.

These values were comprised of 1,250 scores obtained from each of six real data sets described by Micceri (1989) as being representative of the most prolific shapes of data set distributions in psychology and education research. (For histograms and descriptive statistics on these data sets, see Sawilowsky & Blair, 1992). The six data sets were:

- smooth symmetric (from an achievement instrument)
- extreme asymmetry (from a psychometric instrument)
- extreme asymmetry (achievement)
- digit preference (achievement)
- discrete mass a zero with gap (achievement)
- multimodal lumpy (achievement)

The personality profile for each participant was created as follows. Scores were sampled of size N = 4, independently and with replacement from the data sets. Next, the scores were randomly assigned to two groups, with $n_1 = n_2 = 2$. This process was repeated 1,250 times for each data set. Then, an independent samples t test was conducted on each of these variables for a total of 7,500 t tests.

The t test is a widely used procedure for the statistical comparison of the means of two groups. The null hypothesis is H_o: $\mu_1 = \mu_2$, which is tested against the alternative hypothesis H_a: $\mu_1 \neq \mu_2$, by the formula

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sum x_1^{\,2} + \sum x_2^{\,2}}{n_i(n_i - 1)}}}, \qquad (2)$$

where $n_i = 2$ in this example. Essentially, the difference in the means of the raw scores of the two groups are standardized when divided by an estimate of the pooled population variance, which is the error term. Then, the obtained statistic is compared with the critical value given in t tables (as found in most statistics textbooks or statistics software packages) for the nominal α level of 0.01 and the degrees of freedom (df) of $(n_1 + n_2 - 2)$, or 2 df in the current example.

The α level indicates that if the obtained t statistic exceeds the tabled value, the difference in means between the two groups is likely to have occurred by chance with a probability of less than one out of 100 under the truth of the null hypothesis. Thus, the proposition that the two groups are equal on that construct of the personality profile for the four participants (and random assignment equalized the two groups) would be rejected. However, if the obtained t statistic is less than the critical value, then the hypothesis that the two groups are equal in term of their respective means for

that façade variable (and random assignment equalized the two groups) would be retained.

Results

Table 1 contains a compilation of façade variables where statistically $\bar{x}_1 \neq \bar{x}_2$ at the α = 0.01 level, despite random assignment. The variable numbers refer to different characteristics presented by each participant in the experiment. For example, Variable 373 from the Extreme Asymmetry (Psychological Scale) data set might refer to a score from a standardized measure of depression. Indeed, there were 15 variables where $\bar{x}_1 > \bar{x}_2$ and 18 variables where $\bar{x}_2 > \bar{x}_1$, for a total of only 33 variables out of 7,500 where random assignment failed to make the two groups comparable when sample size was as small as $n_1 = n_2 = 2$.

The failure rate of random assignment in producing a comparable group depends on nominal α. Setting nominal α to 0.05 (probability of one out of twenty) will produce more variables where statistically $\bar{x}_1 \neq \bar{x}_2$, and setting nominal α to 0.001 (probability of one out of 1,000) will eliminate many of the variables listed in Table 1. A tangential statistical issue is discussed in the Appendix.

A Classroom Experiment

An experiment was conducted with three sections of a graduate level introductory research course to assess the effectiveness of the methodology in this article for teaching random assignment. The number of participants was N = 56 ($n_1 = 20$, $n_2 = 18$, $n_3 = 18$). Informed consent was not required of the participants because this was part of the regular curriculum.

The students were surveyed at the beginning of the semester with the following question: “Do you believe that random assignment of subjects in an experiment into a treatment and a control group can produce comparable groups?”. The forced response format was “Yes”, “Maybe”, or “No”. If students answered “Maybe”, they were asked to explain under what conditions they believed that random assignment does not work

Two of the three classes were arbitrarily selected to receive the material in this article as part of their course pack (Treatment One), without identifying the author of the article as their instructor. Later in the semester, at the usual point in the curriculum where random assignment was assigned to be discussed, students in the Treatment One classes were referred to the materials in the course pack. (There was no reading assignment for the textbook.) The students in the Treatment Two class, who did not have this article in their course pack, were directed to their version of the syllabus which assigned the textbook chapter on random assignment. The textbook is a current, popular offering with a discussion similar to that found in many research textbooks.

After the students completed the reading assignment one week later, but prior to class discussion on random assignment, they were asked to respond again to the survey question. The pretest (i.e., beginning of the semester) and posttest (i.e., after reading this article or the textbook chapter) responses are recorded in Table 2. An analysis of the posttest scores for the Treatment One classes and Treatment Two class were conducted with a stratified 2 x 3 singularly ordered categorical design, with the pretest scores serving as the covariate. The data analysis was conducted with *StaxXact* (Mehta & Patel, 1999).

The Mann-Whitney statistic for the data in Table 2 was 979.5, and the exact one-sided p-value = 0.0011. An inspection of the entries in the table indicates that there was a statistically significant difference between the two curricular approaches for these 56 students. The material in this article was superior to the discussion in a typical graduate level research textbook in persuading students on the effectiveness of random assignment in research and experimental design.

Table 1. Situations Where $\bar{x}_1 \neq \bar{x}_2$ Despite Randomization For 1,250 Variables From Each Of 6 Real Achievement and Psychology Populations, n = 2, α = 0.01.

Population	Variable	$\bar{x}_1 > \bar{x}_2$	$\bar{x}_2 > \bar{x}_1$
Smooth Symmetric (Achievement Scale)	370		
Smooth Symmetric (Achievement Scale)	1066		✓
Smooth Symmetric (Achievement Scale)	1100		✓
Discrete Mass At Zero (Achievement Scale)	625		✓
Discrete Mass At Zero (Achievement Scale)	831	✓	
Discrete Mass At Zero (Achievement Scale)	959	✓	
Extreme Asymmetry (Achievement Scale)	291		✓
Extreme Asymmetry (Achievement Scale)	336	✓	
Extreme Asymmetry (Achievement Scale)	667	✓	
Extreme Asymmetry (Achievement Scale)	701	✓	
Extreme Asymmetry (Psychological Scale)	190		✓
Extreme Asymmetry (Psychological Scale)	373		✓
Extreme Asymmetry (Psychological Scale)	1089	✓	
Digit Preference (Achievement Scale)	17		✓
Digit Preference (Achievement Scale)	45		✓
Digit Preference (Achievement Scale)	156	✓	
Digit Preference (Achievement Scale)	172		✓
Digit Preference (Achievement Scale)	492		✓
Digit Preference (Achievement Scale)	641		✓
Digit Preference (Achievement Scale)	693	✓	
Digit Preference (Achievement Scale)	810		✓
Multimodal Lumpy (Achievement Scale)	23	✓	
Multimodal Lumpy (Achievement Scale)	281		✓
Multimodal Lumpy (Achievement Scale)	301	✓	
Multimodal Lumpy (Achievement Scale)	323		✓
Multimodal Lumpy (Achievement Scale)	441	✓	
Multimodal Lumpy (Achievement Scale)	504		✓
Multimodal Lumpy (Achievement Scale)	564	✓	
Multimodal Lumpy (Achievement Scale)	835	✓	
Multimodal Lumpy (Achievement Scale)	841		✓
Multimodal Lumpy (Achievement Scale)	851		✓
Multimodal Lumpy (Achievement Scale)	929	✓	
Multimodal Lumpy (Achievement Scale)	1025	✓	
Total/7,500		15/7,500	18/7,500

Table 2. Responses (Percent) Of 56 Students To The Question, "Do you believe that random assignment of subjects in an experiment into a treatment and a control group can produce equal groups?"

	Pretest Scores		Posttest Scores	
	Intervention			
Response	This Article	Textbook Chapter	This Article	Textbook Chapter
Yes	2 (5.3%)	1 (5.5%)	29 (76.3%)	3 (16.7%)
Maybe	13 (34.2%)	7 (38.9%)	7 (18.4%)	8 (44.4%)
No	23 (60.5%)	10 (55.6%)	2 (5.3%)	7 (38.9%)
Total	38	18	38	18

An interesting topic of classroom discussion centered on the reasons why some students responded "Maybe" or "No". At the pretest stage, the reasons given by the students for "Maybe" were random assignment only worked if (a) there was a large sample size, (b) the data collection instruments were reliable, or (c) the researcher was lucky. These reasons were maintained by the students in Treatment Two class at the posttest stage.

It was also interesting to note that the two respondents in the Treatment One class who responded "No" at the posttest stage indicated that, as members of an ethnic minority, they remained suspicious of any methodology that purports to equalize the characteristic or traits of participants assigned to two groups in an experiment.

Conclusion

Return to the initial question on the advice of textbook authors to check on random assignment to see if randomization worked with a statistical test on the pretest scores. The current study is a demonstration of the ability of randomization to create comparable groups. Therefore, the focus of the researcher's concern should not be on the ability of random assignment. Instead, it should pertain to the validity of the scheme implemented by the researcher to randomly assign participants to groups.

For example, consider the well-documented tumult over the 1970 United States military draft lottery conducted by the Selective Service under the auspices of Executive Order No. 11497 to Part 1631.5 of the Selective Service Regulations signed on November 26, 1969 by President Richard M. Nixon. Fienberg (1971; see also Notz, Staw, & Cook, 1971) raised questions regarding the process of that lottery, where slips of paper containing birth dates were placed in capsules and subsequently into a box. There was a proclivity for dates from December (μ= 121.5), November (μ = 148.7), October (μ = 182.5), and September (μ= 157.3) to be selected from the box, rather than January (μ = 201.2), February (μ = 203.0), March (μ = 225.8), and April (μ = 203.7).

Perhaps, this occurred because capsules bearing these dates were placed in the box last. Alternatively, the capsules for the earlier months were well mixed in the box because there was room to do so. However, as the capsules for the latter months were placed in the box, the lack of room limited the ability to mix the capsules. In either case, the slips of paper were not sufficiently mixed in the box, and hence, birth dates at the end of the year were more likely to be selected. The lack of non-randomness of this scheme would have been easily detected if a statistical test been conducted.

As noted by Cook and Campbell (1979), "the equivalence achieved by random

assignment is probabilistic. Thus it is not inevitable that a correctly implemented randomization procedure will result in groups that do not differ" (p. 341). Indeed, this study showed that for 33 of the 7,500 variables, random assignment resulted in differences between the two groups. Random assignment is probabilistic; it is not a guarantee. However, "Without randomization, the possibility of bias due to prior differences on an uncontrolled third variable can seldom, if ever, be ruled out as an alternative explanation of the results"(Linn, 1986). Textbook authors should more clearly distinguish between the probabilistic nature of randomization and the limitations or failure of some schemes to achieve randomization, because poorly conceived randomization schemes do create distrust in the ability of random assignment.

References

Blair, R. C. (1987). *RANGEN*. Boca Raton, FL: IBM Corporation.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago, IL: Rand McNally.

Crowl, T. K. (1996). *Fundamentals of educational research*. (2nd ed.). Madison, WI: Brown & Benchmark.

Fienberg, S. E. (1971). Randomization and social affairs: The 1970 draft lottery. *Science, 171*, 255-261.

Gall, M. D., Borg, W. R., Gall, J. P. (1996). *Educational research: An introduction*. (6th ed.). White Plains, NY: Longman.

Krathwohl, D. R. (1993). *Methods of educational and social science research: An integrated approach*. NY: Longman.

Linn, R. L. (1986). Quantitative methods in research on teaching. (3rd ed.). In M C. Wittrock (Ed.), *Handbook of research on teaching*. NY: Macmillan.

Mehta, C., & Patel, N. (1999). *StatXact 4 for windows: statistical software for exact nonparametric inference: User's manual*. Cambridge, MA: Cytel Software Corporation.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Notz, W. W., Staw, B. M., & Cook, T. D. (1971). Attitude toward troop withdrawal from Indochina as a function of draft number: Dissonance or self-interest? *Journal of Personality and Social Psychology*, *20*, 118-126.

Sawilowsky, S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin*, *111*, 353-360.

Strube, M. J. (1991). Small sample failure of random assignment: A further examination. *Journal of Consulting and Clinical Psychology*, *59*, 346-350.

Tuckman, B. W. (1994). *Conducting educational research*. (4th ed.). Fort Worth, TX: Harcourt Brace.

Vockell, E. L., & Asher, J. W. (1995). *Educational research*. (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Appendix

Theoretically, there should have been 75 Type I errors, instead of the 33 obtained in the study. Nevertheless, these results are consistent with the literature, as Monte Carlo studies (e. g., Sawilowsky & Blair, 1992) noted that the t test generally becomes conservative when sample sizes are low and the underlying assumption of normality is violated. In fact, data sampled from the deMoivre (normal) distribution produced 37 variables where $\overline{x}_1 > \overline{x}_2$, and 35 variables where $\overline{x}_2 > \overline{x}_1$, for a total of 72 Type I errors, which is excellent agreement with the theoretical value.

This article relates to the validity of statistical findings, but not the statistical power of a test or the generalizability of results. The purpose of this demonstration is to show random assignment works even if $n_i = 2$. The use of the randomized two group experimental design with only N = 4 is not suggested. It should also be noted that the t test is the only statistic available that can be used with N = 4 and $\alpha = 0.01$.