PM R 9 (2017) 306-310

www.pmrjournal.org

Statistically Speaking

Randomization Test: An Alternative Analysis for the Difference of Two Means

Regina L. Nuzzo, PhD

Introduction

Randomization tests, also known as approximate permutation tests or Monte Carlo permutation tests, are gaining popularity among statisticians and researchers. They rely on fewer assumptions than do common parametric tests (such as the *t*-test) and so can be used when requirements for parametric tests are not satisfied, and they can sometimes be more powerful than common rank-based nonparametric tests (such as the Mann-Whitney *U* test) and so also can be used when typical nonparametric tests are not desired. They are also quite intuitive and flexible and often can be implemented easily with free software packages. Using an example from previously published data, in this article I introduce readers to a simple randomization test based on reallocation of group assignments for comparing the means of 2 groups in a randomized trial.

Data Example

Randomization tests are part of a broader framework of powerful computer-based methods that rely on permuting or resampling observed data to draw conclusions [1-3]. Although the concept of randomization tests for analyzing continuous data was introduced by pioneering statistician R.A. Fisher in the early 1930s, the methods required calculations that were cumbersome to do by hand, and these tests never gained much favor among researchers. Modern computing power, however, has made randomization tests much more attractive, and some researchers have since argued that these class of methods often are superior to traditional tests for biomedical research [4]. Moreover, recent statistics education reform has embraced randomization tests, along with resampling- and simulation-based approaches in general, as a more transparent and intuitive path for hypothesis testing [5,6].

Consider this data example to illustrate a randomization test of 2 samples for a continuous measure: Frazzitta et al [7] reported on a study in which patients in the intensive care unit (ICU) with severe acquired brain injury were randomized either to conventional physiotherapy or to an early stepped verticalization protocol. Researchers collected Disability Rating Scale (DRS) scores at 3 time points (baseline, discharge from ICU, and discharge from rehabilitation) for 15 patients in the treatment group and 16 patients in the control group. For the sake of simplicity, we will focus here only on DRS improvement from baseline to ICU discharge for the 2 groups, using data made publicly available by the authors.

Figure 1 shows a box plot with improvements in DRS scores for patients in the 2 groups (prepared with the free BoxPlotR application [8,9]). We can see that most patients made modest improvements of 5 points or less, but in each group a few patients made much greater improvements relative to the rest of their group, which results in a strong right skew. Results seem to indicate greater improvements with the treatment than would otherwise occur under standard care; for example, nearly one-half of the patients in the treatment group surpassed even the most improved patient in the control group, and both mean and median improvements were greater in the treatment group (8.5 points versus 3.1 points, and 5 points versus 1.5 points, respectively). Yet the data were not unequivocally in favor of the treatment, because there was still a substantial overlap of scores between the 2 groups, especially among the least responsive patients, with 4 of the control group and 3 of treatment group patients showing no improvement at all.

Randomization Test Null Hypothesis and Test Statistic

Did the treatment protocol have a different effect on patients' improvement on DRS compared with receiving standard care? To investigate this with a randomization test, we will set up a reference position that we will attempt to falsify: What if being assigned to either the treatment group or to the control group had absolutely no differential effect on a patient's improvement between these 2 time points? If so, then a patient would have had the same improvement no matter to which group they

1934-1482/$ - see front matter
http://dx.doi.org/10.1016/j.pmrj.2017.02.001

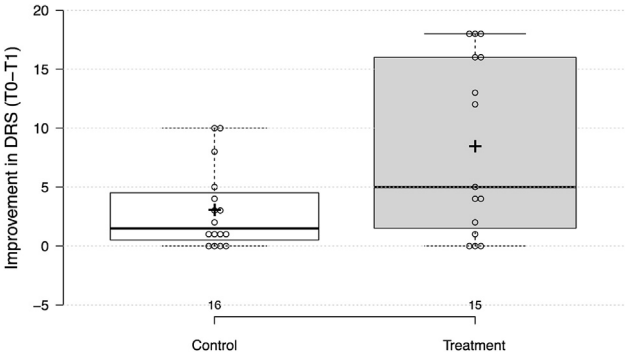

Figure 1. Box plot of original data from Frazzitta et al [7]. Note that values are computed here so that positive scores indicate a decrease (improvement) in disability between baseline and discharge from the intensive care unit. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend to minimum and maximum values; crosses represent sample means; width of the boxes is proportional to the square root of the sample size; data points are plotted as open circles. n = 16, 15 sample points. DRS, Disability Rating Scale.

were assigned, and it was just random occurrence that gave us our observed pattern of scores. Furthermore, if patients' group assignment had no effect on their scores, then any random shuffling of the 31 observations between 2 groups would be as likely as any other grouping. It would then be unlikely to have a situation such as the one we observed previously, in which the handful of patients with exceedingly high improvements were all in the treatment group, but it is still possible. Figure 2 shows a box plot with one such random shuffling of patient improvement scores. In this permutation, 4 of the patients who improved more than 15 points landed in the control group and only one in the treatment group, and now the group means and medians are closer together.

For our randomization test, the following reference position will be our null hypothesis: "Being assigned to

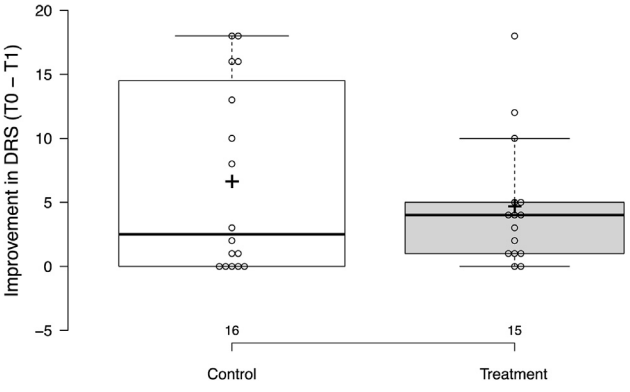

Figure 2. Box plot showing one possible sample under the null hypothesis, in which group membership is equally interchangeable. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend to minimum and maximum values; crosses represent sample means; width of the boxes is proportional to the square root of the sample size; data points are plotted as open circles. n = 16, 15 sample points. DRS = Disability Rating Scale.

the treatment group versus control group had no differential effect on the improvement score for any patient." This is a more general starting point than offered by most parametric null hypotheses (such as for a *t*-test), most notably in that it does not mention population means of any kind, or in fact a population at all. This reflects a subtle difference in approach between permutation-based randomization tests and traditional tests, the implications of which have been debated by researchers. In practice, it means that randomization tests do not require that the studied individuals be a representative random sample of a population, and that these tests can be used when assumptions of parametric methods are violated, such as when samples are small or drawn from skewed populations.

Notice also that the randomization test null hypothesis is so general that it does not specify upfront what test statistic the researcher will use. It simply hypothesizes that the effects of the treatment and control are equal, and the researcher is free to choose a meaningful quantitative measure to evaluate that statement. We might decide that if the treatment and control effects were not in fact equal, then the mean score in each group would be different. For this we would therefore build onto our original null hypothesis to ask, "If being assigned to either group had no effect on a patient's improvement, then how often would we see group means as far apart as we did for these patients (5.4 points)?" Yet that is not the only measure we could use; if the improvement of the middle-most patient in each group was most important for evaluating our treatment, then we would ask how often under the null hypothesis the group medians would be as far apart as what we observed. Nearly any quantitative statistic could be used: the difference between the 75th percentiles of the groups, the difference between the variances, the difference in skewness, and so on. The difference in group means is often a reasonable choice, and that is what we will use here.

Now that we have a null hypothesis and a test statistic, we want to compare our observed results to all the data permutations that are possible under the null hypothesis. Unfortunately, there are more than 300 million possible arrangements of our data (31 patients divided into groups of 16 and 15 = 31!/16!/15! = 300,540,195), which is computationally daunting even with today's computing power. Therefore, typical practice is to randomly sample an arbitrarily large number of these shuffled permutations, record the difference in means between the 2 groups for each shuffling, and then examine how rare our observed result appears to be.

Example Results and Comparison With Published Results

Figure 3 shows a histogram of the results of 10,000 such randomizations under the null hypothesis. We can see that only 80 times of 10,000 did the mean