# PERMUTATION TESTS IN CLINICAL TRIALS

DAVID M. ZUCKER
Hebrew University of Jerusalem
Department of Statistics

## 1 RANDOMIZATION INFERENCE—INTRODUCTION

Randomization—allocation of study treatments to subjects in a random fashion—is a fundamental pillar of the modern controlled clinical trial. Randomization bolsters the internal validity of a trial in three major respects:

1. It prevents possible investigator bias (which may otherwise exist even unintentionally) in the allocation of subjects to treatments.
2. It generates study groups that, on average, are balanced with respect to both known and unknown factors.
3. It provides a probability structure whereby the study results may be evaluated statistically without exogenous statistical modeling assumptions.

The purpose of this article is to elaborate on the last of these three points.

To begin, we note that many, probably most, statistical analyses in empirical research, especially in nonexperimental studies, involve some statistical modeling assumptions. For example, to take a simple case, the classical two-sample $t$-test for comparing two groups makes the following assumptions:

1. Each observation in each group is equal to the sum of a fixed group-specific population mean value plus a mean zero random error term.
2. All random error terms, both within and between groups, are statistically independent.
3. The random error terms have a normal distribution.
4. Within each group, the error terms all have the same variance. (On occasion it is also assumed that the variance is the same in the two groups, although this assumption can be avoided.)

Obviously, the more complex the statistical analysis, the more assumptions are involved. In a thorough statistical analysis, efforts are made to check the validity of the assumptions and to examine the sensitivity of the results to departures from the assumptions. In routine analyses, however, these checks are often skipped. Moreover, and more critically, the checks are not foolproof, and even after the checks, some unquantifiable element of uncertainty remains.

In an experiment in which treatments have been assigned to subjects by randomization, however, it is possible to apply a statistical method for testing treatment effect that does not require any statistical assumptions about the data beyond those inherently satisfied due to the randomization itself. This method is called a permutation test (or randomization test). The ability thus afforded to compute a $P$-value for testing treatment effect without relying on uncertain statistical assumptions is a major strength of the randomized trial.

## 2 PERMUTATION TESTS—HOW THEY WORK

The classic exposition of the randomized experiment in general, and the permutation test in particular, was given by R. A. Fisher (1). Rosenberger and Lachin (2) give an up-to-date comprehensive exposition. This article will describe how a permutation test works in the simple context of comparing two groups, a treatment group and a control group, but the idea applies more generally to multi-arm trials. The classical formulation of the permutation test is as a test of the null hypothesis that the treatment has no effect whatsoever on the subject's response; that is, each subject would exhibit exactly the same response whether given the study treatment or the control regimen. This null hypothesis is referred to by some authors as the "strong"

null hypothesis. Some remarks will be made later in this article on an alternative form of null hypothesis and the performance of permutation tests in that context. For now, however, we will stick with the null hypothesis that each subject would exhibit exactly the same response irrespective of the regimen given.

We will illustrate how the test works in the context of an example presented in Reference 3. We first must review the concept of a *P*-value in general. Given a statistic aimed at comparing the two groups—the difference between the means, for example—the *P*-value is defined as the probability that the statistic would be equal to or more extreme than the value actually observed if the null hypothesis were in fact true. A small *P*-value means that the observed value of the statistic is so extreme that it is unlikely to have arisen if the null hypothesis were true, and is thus an (indirect) indicator that the null hypothesis is false. It is conventional to say that there is "statistically significant" evidence of a treatment difference if the two-sided *P*-value (i.e., considering extremes in both directions) is less than 0.05.

With this background, we may now turn to the example. Suppose eight subjects were randomly assigned to either treatment or control on an equal basis (four per group). By "randomly assigned," we mean that the researchers chose the actual allocation at random from among the 70 possible ways of dividing eight subjects into two groups of four subjects each, with each possible allocation having an equal 1/70 chance of being employed. Suppose further that the final results with respect to some response variable were follows:

**Table 1. Illustrative Experimental Data**

| Subject | Group | Response |
|---|---|---|
| A | Control | 18 |
| B | Control | 13 |
| C | Control | 3 |
| D | Control | 17 |
| E | Treatment | 9 |
| F | Treatment | 16 |
| G | Treatment | 17 |
| H | Treatment | 17 |

The groups are to be compared by examining the difference in sample means between the treatment and control groups. The observed sample means are 14.75 for treatment and 12.75 for control, so that the observed mean difference is 2.00. We wish to evaluate the statistical significance of this result. We argue as follows. As noted, there are 70 possible ways of dividing the eight subjects into two groups of four subjects each. Under the null hypothesis that the treatment has no effect whatsoever on the response, the responses of all eight subjects would be the same no matter what the allocation was. The only thing that differs from allocation to allocation is who received which regimen.

By way of analogy, imagine a deck of eight cards, one corresponding to each subject, with the subject's eventual end of study response written on the face of the card. At the beginning, all cards are face down, corresponding to the fact that at the beginning we do not know the response values. We shuffle the cards well, and then split the deck into two packs of four cards each, one pack corresponding to treatment and the other to control. We wait some period of time, during which the numbers on all cards remain the same, corresponding to the fact that the treatment has no effect. Finally, we are allowed to turn the cards face up and see the response values. We can then can compute the mean difference between the groups. We wish to determine the probability of obtaining a mean difference as extreme or more so than that observed through mere "luck of the draw" in splitting the deck into the two packs.

Given the final observed response values—the numbers on each of the eight cards—we may enumerate easily all the possible realizations that could have eventuated in this experiment with the given subjects under the null hypothesis. The list of realizations consists simply of the list of the 70 possible ways of dividing the eight subjects into the two groups, with the response value associated with each subject being in every case the value observed in fact in the actual experiment—because, again, under the null hypothesis, the treatment does not affect the response at all. Each of these possible realizations has an equal chance of 1/70 of having arisen, because the allocation was chosen at

**Table 2. List of Possible Realizations of the Experiment Under the Null Hypothesis**

| Case | Subjects On Treatment | Observations On Treatment | Observations On Control | Mean Treatment | Response Control | Mean Diff |
|---|---|---|---|---|---|---|
| 1 | EFGH | 9, 16, 17, 17 | 18, 13, 3, 17 | 14.75 | 12.75 | 2.00 |
| 2 | DEFG | 17, 9, 16, 17 | 18, 13, 3, 17 | 14.75 | 12.75 | 2.00 |
| 3 | DFGH | 17, 16, 17, 17 | 18, 13, 3, 16 | 16.75 | 10.75 | 6.00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 70 | ABCD | 18, 13, 3, 17 | 9, 16, 17, 17 | 12.75 | 14.75 | −2.00 |

**Table 3. Null Hypothesis Distribution of the Mean Difference**

| | | | | | | |
|---|---|---|---|---|---|---|
| Mean Diff | −7 | −6.5 | −6 | −5 | −4.5 | −4 |
| Probability | 1/70 | 3/70 | 1/70 | 3/70 | 4/70 | 3/70 |
| Mean Diff | −3 | −2.5 | −2 | −1 | −0.5 | 0 |
| Probability | 3/70 | 4/70 | 3/70 | 3/70 | 4/70 | 6/70 |
| Mean Diff | 0.5 | 1 | 2 | 2.5 | 3 | 4 |
| Probability | 4/70 | 3/70 | 3/70 | 4/70 | 3/70 | 3/70 |
| Mean Diff | 4.5 | 5 | 6 | 6.5 | 7 | |
| Probability | 4/70 | 3/70 | 1/70 | 3/70 | 1/70 | |

random. For each possible realization (case) listed, we may compute the mean difference between treatment and control that arises under that realization. We obtain a list of the form shown in Table 2.

Keeping in mind that each of these cases has a 1/70 chance of having occurred, we may obtain from this list the null hypothesis probability distribution—called the "permutation distribution"—for the mean difference. Table 3 displays this null distribution: the table provides a list of the various possible values of the mean difference, along with the probability of obtaining each value.

The mean difference observed in the actual experiment was 2.00. We can compute the probability of observing a difference of this magnitude or greater under the null hypothesis, by pure chance, from the probability distribution in Table 3. For a two-sided test we work with the absolute value of the difference. The probability is found to be 50/70, which equals 0.71. Thus, the permutation test two-sided $P$-value for testing the null hypothesis in this experiment is $P = 0.71$.

Note that in the calculation above, no assumptions whatsoever were made about the behavior of the response value. The calculation depended only on the following basic facts.

1. The treatment allocation followed was chosen at random, on an equally likely basis, from the 70 possible allocations of eight subjects to two groups of four subjects each.
2. The null hypothesis says that the treatment does not affect the response values in any way.
3. In light of Fact 2, we can—after the fact, given the observed data—construct a list of all possible realizations of the experiment with the given subjects under the null hypothesis, with each realization having the same 1/70 chance of having eventuated.
4. Given the list of realizations, we can, for each possible value of the observed mean difference, calculate the probability of observing that value with the given subjects under the null hypothesis.

In the case of $2 \times 2$ table analysis with a dichotomous response variable, the permutation test is known as the Fisher exact test and is available in many standard statistical routines such as SAS PROC FREQ (SAS Institute Inc., Cary, NC).

It is possible to adapt the permutation test method to more complex statistical models.

See References 2 and 4 for a discussion on how this adaptation is carried out.

In an experiment with a extremely small number of subjects, such as in the above example, the range of possible *P*-values that can arise is limited, and so the capacity to test the null hypothesis is limited. But with a moderate-to-large sample size, the range of possible *P*-values is typically large.

Similarly, in an experiment with a very small number of subjects, it is possible to list all possible outcomes and compute the *P*-value in a direct fashion. But as the sample size increases, the calculation quickly becomes formidable. For example, in an experiment with 5 subjects per group, there are 252 possible allocations; with 6 subjects per group, there are 924 possible allocations; with 10 subjects per group, there are 184,756 possible allocations; and with 20 subjects per group, there are about 138 billion possible allocations. With modern computing power and special algorithms, however, it is possible to perform permutation test calculations easily for experiments with 20–30 subjects or more per group. A well-known software package for such calculations is the StatXact package (Cytel Inc., Cambridge, MA). The calculations can also be performed in SAS PROC NPAR1WAY (SAS Institute Inc., Cary, NC), using the SCORES=DATA and EXACT options. At a certain point, however, the calculations become unmanageable, and a large sample approximation is employed. Section (3) discusses this approximation.

## 3 NORMAL APPROXIMATION TO PERMUTATION TESTS

Generally speaking, for a sample size sufficiently large, the permutation distribution of a statistic for testing the null hypothesis can be well approximated by a normal distribution. This result is a generalized version of the familiar classical central limit theorem of probability, which says that the distribution of the average of a large number of independent observations tends to a normal bell-curve distribution. The setup here is somewhat different from that of the classical theorem, but the nature of the final result is essentially the same. The normal limit result for permutation tests was proven mathematically around 40–50 years ago (5–8).

This limit theorem provides justification for the application of classical statistical tests, such as the two-sample *t*-test described in Section 1, to randomized experiments without relying on the classical assumptions underlying the tests. The classical test simply serves as an approximation—with theoretical backing—to the exact permutation test. Essentially, when the sample size is large, the two tests are approximately equivalent (9). Often, as in the two-sample *t*-test, the *t*-distribution is used as an approximating distribution rather than the normal distribution, which improves the approximation to some extent by matching the first two moments (10–12).

In practice, for a continuous response variable such as blood pressure, with 30–40 subjects per group the normal-theory test is quite adequate unless the distribution of the response variable is very peculiar. For $2 \times 2$ table analysis for a dichotomous response variable, the standard recommendation is to use the normal-theory chi-square test (the continuity correction usually improves the approximation) if the expected cell size, given the table margins, is five or more in all four cells, and otherwise to fall back to the Fisher exact test.

Thus, in general, it is common practice to use the normal-theory tests unless the study involves a small sample size or small number of events. Again, the normal-theory tests are justified as approximations to a permutation test. When the sample size or number of events is small, the proper course is to use an exact permutation test.

## 4 ANALYZE AS YOU RANDOMIZE

For a permutation-based analysis (either an exact permutation test or a corresponding normal approximation) to be valid without added assumptions, as described above, the form of analysis must match the randomization scheme. This fact is known as the "analyze as you randomize" principle. Thus, if a matched-pairs design has been used, with randomization carried out within each pair, then the pairing must be accounted for in the

analysis. Similarly, if stratified randomization has been carried out, then the analysis must account for the stratification. Failure to include matching or stratification factors in the analysis can lead to inaccurate *P*-values. The level of error tends to be small in trials with a continuous endpoint and a large sample size, but it can be more substantial in very small trials or dichotomous endpoint trials with a low event rate.

Another case calling for emphasis on the need to "analyze as you randomize" is the cluster randomization design (or group randomization design). Here the unit of randomization is some aggregate of individuals. This design is common in community-based and school-based trials, which often involve aggregate-level interventions. In cluster randomization trials, the unit of analysis must be the cluster rather than the individual. This is necessary to preserve the validity of the permutation-based analysis, as indicated above, and also to take proper account of between-cluster variation and thereby avoid serious type I error inflation (13, 14).

To provide statistically rigorous results, a cluster randomization trial must include an adequate number of clusters. Many cluster randomization studies involve a very small number of clusters per arm, such as two or four. In such a study, it is almost impossible for permutation-based analysis to yield a statistically significant result. A normal-theory procedure such as the *t*-test has no justification in this case. With a trial of this small size, the normality assumption cannot be effectively checked, and the central limit theorem argument presented in the preceding section to justify the use of a normal-theory analysis as an approximation to a permutation-based analysis does not apply. A trial of such a small size may be useful as a pilot study, but it cannot yield statistically definitive conclusions. By contrast, studies such as CATCH (15, 16) (96 schools) and COMMIT (17, 18) (11 matched pairs of communities) included enough units for meaningful statistical analysis. In COMMIT, because of the relatively small number of units, an exact permutation test was used rather than a normal-theory test. In References 16 and 19, methods are described that allow individual-level explanatory variables to be accounted for while maintaining the cluster as the primary unit of analysis. An article by Donner and Klar in this encyclopedia provides further discussion on statistical analysis of group-randomization trials.

## 5 INTERPRETATION OF PERMUTATION ANALYSIS RESULTS

When a permutation-based analysis yields a statistically significant result, this is evidence against the null hypothesis that the treatment has no effect whatsoever on any subject's response. If the result is in the positive direction, the inference to be made is that there are at least *some* individuals for whom the treatment regimen is better than the control regimen. It cannot necessarily be inferred that treatment is better than control for all individuals or even that treatment is better than control on an "average" basis; the only truly definitive statement that can be made is that some individuals do better with treatment than with control. This conclusion is admittedly one of limited scope, but it is nonetheless meaningful and is achieved with a high degree of certainty, in that it does not rely on any outside assumptions.

Note, however, that along with the conclusion that the treatment is better than control for some people comes the proviso that treatment might be the same or worse than control for other people, so that "on average" there might be no difference between the two regimens.

Usually clinical trial investigators like to go further and try to infer that treatment is better than control on some "average" basis. It must first be understood how this objective can be formulated from a formal statistical standpoint. In a typical clinical trial, patients are recruited on a volunteer, "catch-as-catch can" basis, and the set of patients entering the trial does not represent a formal random statistical sample from any particular defined population. Accordingly, we must suppose the trial patients behave like a random sample from some hypothetical superpopulation. This supposition has some plausibility in some circumstances, but it must be realized that it is essentially a statistical modeling assumption.

Given that we are willing to make the above-described supposition, in trials with large sample sizes, the classical normal-theory procedures generally will provide valid tests of the "population-level" null hypothesis that the superpopulation mean response is the same for treatment as for control, against the alternative that it is better on treatment (or worse, or different). There remains the question of what happens in trials with small sample sizes, and in particular how permutation tests perform in relation to the "population-level" null hypothesis. This question has been investigated in References (20–22). Overall, these investigations have shown that permutation tests generally can have inflated type I error relative to the "population-level" null hypothesis, but if the number of subjects (or units, in a cluster randomized trial) is the same in the two experimental arms, then the type I error level is typically close to the desired level.

If the "population-level" null hypothesis has been rejected with results in the positive direction, the inference to be made is that treatment is superior to control on an "average" basis *for individuals similar to those who participated in the trial*. Generalization of the results to other types of individuals requires careful judgment.

## 6 SUMMARY

The permutation test is a method for analyzing randomized trials through which the null hypothesis that the treatment has no effect whatsoever on the response may be assessed statistically without statistical distribution assumptions beyond those arising from the randomization process itself. With a sufficiently large sample size, the permutation test can be approximated satisfactorily by a classical normal-theory test. This result provides clear justification for application of normal-theory tests in trials with moderate-to-large sample sizes. In very small trials, it is preferable to perform an exact permutation test.

## REFERENCES


1. R. A. Fisher, *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935. (8th Ed. New York: Hafner, 1966).
2. W. F. Rosenberger and J. M. Lachin, *Randomization in Clinical Trials: Theory and Practice*. New York: Wiley, 2002.
3. O. Kempthorne, *The Design and Analysis of Experiments*. New York: Wiley, 1952.
4. M. H. Gail, W. Y. Tan, and S. Piantadosi, Tests for no treatment effect in randomized clinical trials. *Biometrika* 1988; **75**: 57–64.
5. J. Hajek, Some extensions of the Wald-Wolfowitz-Noether theorem. *Ann. Math. Stat.* 1961; **32**: 506–523.
6. W. Hoeffding, A combinatorial central limit theorem. *Ann. Math. Stat.* 1951; **22**: 558–566.
7. G. E. Noether, On a theorem of Wald and Wolfowitz. *Ann. Math. Stat.* 1949; **20**: 455–458.
8. A. Wald and J. Wolfowitz, Statistical tests based on permutations. *Ann. Math. Stat.* 1944; **15**: 357–372.
9. W. Hoeffding, The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.* 1952; **23**: 169–192.
10. E. J. G. Pitman, Significance tests which may be applied to samples from any populations. *J. R. Stat. Soc.* 1937; **4** (suppl.): 119–130.
11. E. J. G. Pitman, Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *J. R. Stat. Soc.* 1937; **4** (suppl.): 225–232.
12. E. J. G. Pitman, Significance tests which may be applied to samples from any populations. III. The analysis of variance test. The analysis of variance test. *Biometrika*, 1937; **29**: 322–335.
13. G. V. Glass and J. C. Stanley, *Statistical Methods in Education and Psychology*. Englewood Cliffs, NJ: Prentice-Hall, 1970.
14. D. M. Zucker, An analysis of variance pitfall: the fixed effects analysis in a nested design. *Educ. Psychol. Measure.* 1990; **50**: 731–738.
15. R. V. Luepker, C. L. Perry, S. M. McKinlay, P. R. Nader, G. S. Parcel, E. J. Stone, L. S. Webber, J. P. Elder, H. A. Feldman, C. C. Johnson, S. H. Kelder, and M. Wu, Outcomes of a field trial to improve children's dietary patterns and physical activity: The Child and Adolescent Trial for Cardiovascular Health (CATCH). *JAMA* 1996; **275**: 768–776.

16. D. M. Zucker, E. Lakatos, L. S. Webber, D. M. Murray, S. M. McKinlay, H. A. Feldman, S. H. Kelder, and P. R. Nader, Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH). *Controlled Clinical Trials* 1995; **16**: 96–118.
17. COMMIT Research Group, Community Intervention Trial for Smoking Cessation (COMMIT): I. Cohort results from a four-year community intervention. *Amer. J. Public Health* 1995; **85**: 183–192.
18. M. H. Gail, D. P. Byar, T. F. Pechacek, and D. K. Corle, Aspects of the statistical design for the Community Health Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials* 1992; **13**: 6–21.
19. D. M. Zucker, Cluster randomization. In: N. Geller (ed.), *Contemporary Biostatistical Methods in Clinical Trials*. New York: Marcel Dekker, 2003.
20. T. Braun and Z. Feng, Optimal permutation tests for the analysis of group randomized trials. *J. Amer. Stat. Assoc.* 2001; **96**: 1424–1432.
21. M. H. Gail, S. D. Mark, R. J. Carroll, and S. B. Green, On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* 1996; **15**: 1069–1092.
22. J. P. Romano, On the behavior of randomization tests without a group invariance assumption. *J. Amer. Stat. Assoc.* 1990; **85**: 686–692.