Statistical Papers (2019) 60:395–404
https://doi.org/10.1007/s00362-018-01070-y

REGULAR ARTICLE

Randomization-based inference and the choice of randomization procedures

Yanying Wang[1] · William F. Rosenberger[1] · Diane Uschner[2]

Received: 10 September 2018 / Revised: 5 December 2018 / Published online: 1 January 2019

Abstract

In testing the significance of treatment effects in randomized clinical trials (RCTs), randomization-based inference is distinguished from population-based parametric and nonparametric inference, such as the *t*-test or permutation tests, taking into account three properties: preservation of type I error rate, relation of power to the randomization procedure, and flexibility in choosing the test statistic. In this paper, we revisit rationale of the properties and provide justification through simulations. We propose that the choice of randomization procedures and the analysis of RCTs can be facilitated by the application of randomization-based inference.

Keywords Randomization tests · Population tests · Randomization procedure · Type I error rate · Statistical power

1 Randomization model

Randomization tests have been a method of inference for analyzing data from randomized experiments since the days of Fisher (Anscombe 1948). Due to computational limitations, statisticians in the early days relied normal distribution theory to approximate randomization tests (Anscombe 1948):

✉ William F. Rosenberger
wrosenbe@gmu.edu

Yanying Wang
ywang69@gmu.edu

Diane Uschner
duschner@bsc.gwu.edu

[1] Department of Statistics, George Mason University, 4400 University Drive MS4A7, Fairfax, VA 22030, USA

[2] The George Washington University Biostatistics Center, 6110 Executive Boulevard Suite 750, Rockville, MD 20852, USA

> On this hypothesis [null hypothesis A that there is no difference in the effect of treatments], the set of $2n$ observations has been divided merely at random into two sets of n labeled P and Q [treatments compared in the experiment], and any other such subdivisions would have been equally likely to occur.
>
> We can test hypothesis A by normal-law theory if we postulate that the aggregate of $2n$ observations is a random sample from a normal distribution…For by the virtue of randomization, the P and Q observations are independent random samples from the same normal distribution.

The randomization mentioned in the above quote is a *random allocation rule* (RAR) in the language of modern RCTs. Today, randomization tests for a variety of randomization procedures can be efficiently computed via Monte Carlo simulation. Such a computing procedure is called re-randomization test (Rosenberger and Lachin 2016).

Under the null hypothesis of no treatment effect, the randomization model states that patient responses are unaffected by any treatment received. More formally, the hypothesis involves no parameters and is, in essence, a statement that the treatment assignents are independent of the patient outcomes. Therefore, given a measure of treatment difference, the randomization test computes the significance of observed test statistic by adding the frequency probabilities of treatment assignment sequences that lead to a test statistic value at least as extreme as the observation with reference to P_ϕ, the probability of treatment assignment sequences derived from the randomization procedure employed.

By the construction of the randomization model, the randomization distribution of a test statistic is always correct for the set of observations. As such, the p-value as a random variable is uniformly distributed. Hence the type-I error rate of randomization tests are preserved at the designated significance level as long as it is an achievable size and the study is not biased. In this sense, the randomization model is robust. Additionally, because of the statistical validity, any meaningful measure of the treatment difference gives a correct test for the null hypothesis. Further, the power of a randomization test and the randomization procedure employed are interrelated since the randomization distribution of a test statistic is calculated from P_ϕ. These three properties distinguish randomization tests from population-based tests in analyzing data from RCTs, whose validity is conditioned on the fulfillment of (distributional) assumptions.

2 Randomization procedures and Monte Carlo re-randomization tests

Consider a RCT comparing two treatment arms, A and B, among n patients. A randomization procedure is the probability model employed in a RCT to assign treatments to patients. For example, the *random allocation rule* can be compared to an urn model. "Suppose an urn contains $n/2$ balls of type A and $n/2$ balls of type B. Each time a patient is ready to be randomized, a ball is drawn and not replaced, and the corresponding treatment is assigned. This continues until the urn is depleted" (Rosenberger and Lachin 2016). Thus P_ϕ ensures balanced treatment allocation if n is even. When the *truncated binomial design* (TBD) is used, P_ϕ corresponds to a binomial distribution with success probability 1/2 until half of the patients are allocated to one of the treatments, and the remaining patients are assigned to the other treatment. The *permuted blocked design* divides the n treatment assignments into blocks, each with equal sizes containing an even number of assignments. Within each block, a forced balance design, such as the *random allocation rule*, is used, P_ϕ therefore ensures balanced assignments in each block. More randomization procedures will be mentioned in Sect. 4.

The Monte Carlo re-randomization test is applied to estimate the p-value of an exact randomization test. The algorithm for a re-randomization test is as follows. For given patient response data, the treatment assignment sequence is regenerated L times, and the test statistic is re-computed each time. The p-value is then determined as the proportion of the L simulations that results in equally or more extreme test statistic value S_l than the observed test statistic $S_{obs.}$, $1 \leq l \leq L$. The two-sided Monte Carlo p-value estimator is defined as (Plamadeala and Rosenberger 2012):

$$\hat{p} = \frac{\sum_{l=1}^{L} I(|S_l| \geq |S_{obs.}|)}{L}.$$

The choice of test statistic is arbitrary and is chosen to reflect the information on the treatment difference.

3 Validity of randomization tests, permutation tests, and the *t*-test

We distinguish, especially in a RCT context, randomization tests from permutation tests (Pesarin 2001). At an early stage of the development of randomized experiments, the distinction between the terms *permutation test* and *randomization test* may not be evident, when the experimental result is tested by referring to all possible *permutations* of the treatment arrangement; for instance, in blocks of agricultural land when treatment arrangements are equiprobable. Nonetheless, when the permutation test omits the probabilities of the treatment arrangements and superimposes a distribution structure on the observations [exchangeability (Pesarin 2001)], the two methods are very different. In testing the null hypothesis, permutation tests start with the assumption that the permutations of patient observations with respect to treatments are equally likely to occur on the basis of a postulation that the permutations are random samples from some population distribution satisfying exchangeability. For randomization tests, however, this is unnecessary since P_ϕ induces the randomization distribution of the permutations.

Below we include a table (Table 1) in a recent paper by Rosenberger et al. (2018) to illustrate the difference. Consider the randomization of treatments in a block of size 4, with patient outcome x_1, x_2, x_3, x_4. The first column lists the possible randomization of treatments for the four patients. The second column shows another way of listing the randomizations in the first column by looking at the possible permutations of patient responses in the two treatment groups. These data permutations are equally likely if the RAR is used but are not equiprobable if the TBD is used. Therefore, the permutation test and the randomization test would yield same p-value only if the RAR were the randomization procedure.

Table 1 Four treatment assignments under the random allocation rule (RAR) and truncated binomial design (TBD) (Rosenberger et al. 2018)

Randomization sequence	Data permutation		Probability	
x_1, x_2, x_3, x_4	A	B	RAR	TBD
$AABB$	x_1, x_2	x_3, x_4	1/6	1/4
$ABAB$	x_1, x_3	x_2, x_4	1/6	1/8
$ABBA$	x_1, x_4	x_2, x_3	1/6	1/8
$BAAB$	x_2, x_3	x_1, x_4	1/6	1/8
$BABA$	x_2, x_4	x_1, x_3	1/6	1/8
$BBAA$	x_3, x_4	x_1, x_2	1/6	1/4

The distinction between the t-test and randomization test we would like to highlight is that the former targets an asserted population distribution instead of the patient observations themselves (Anscombe 1948). In parametric methods, considerable attention is spent in adjusting the population distribution function to fit the observations, which mitigates but does not solve the problem of mismatched distributional models as randomization tests do. Because population-based inference focuses exclusively on the population distribution model separate from the randomized experiment, its validity in analyzing the data from the experiment is not guaranteed.

In this section, we demonstrate these distinctions by verifying the validity of randomization tests in comparison to permutation tests and the t-test via simulations. Type I error probability and power are simulated under a linear time trend using difference in means statistic under two randomization procedures: The *truncated binomial design* (TBD) and *permuted block design* (PBD) with block size $m = 4$ where the *random allocation rule* (RAR) is used within each block. Patient responses are sampled from $N(\Delta, 1)$, a normal distribution with a mean of Δ, $\Delta \in \{0, 0.1, \ldots, 2\}$, plus a time trend ranging linearly on the interval $(-2.2]$. The parameter values $\Delta = 0$ and $\Delta > 0$ correspond to the null and the alternative hypothesis respectively. Each simulation is based on 10,000 tests where sample size $n = 50$. Monte Carlo simulation is applied to compute permutation and randomization tests. For the permutation test, the number of permutations of patient outcomes is 15,000. For the randomization test, the number of re-generated treatment assignment sequences is also 15,000.

The assumptions of the permutation test (exchangeability) and the t-test (normality and homogeneity of variance) are violated due to the time trend. When the TBD is employed, in addition to the influence of a time trend, the treatment assignment sequences are not equally likely. It is seen in the results (see Fig. 1) that only the randomization test preserves nominal type I error rate (rejection rate at $\Delta = 0$) consistently. For the permutation test and the t-test, however, type I error rates are deflated under the PBD and inflated under the TBD. The normal-law test is no longer valid in the situation, not to mention further approximating randomization tests. Note also that under the PBD, the power of randomization tests is higher than those of the other two. For example, at $\Delta = 0.9$, the power given by the randomization test reaches 0.82, whereas those given by the t-test and the permutation test are 0.52 and 0.49 respectively. If the randomization test is classified as nonparametric, this contradicts

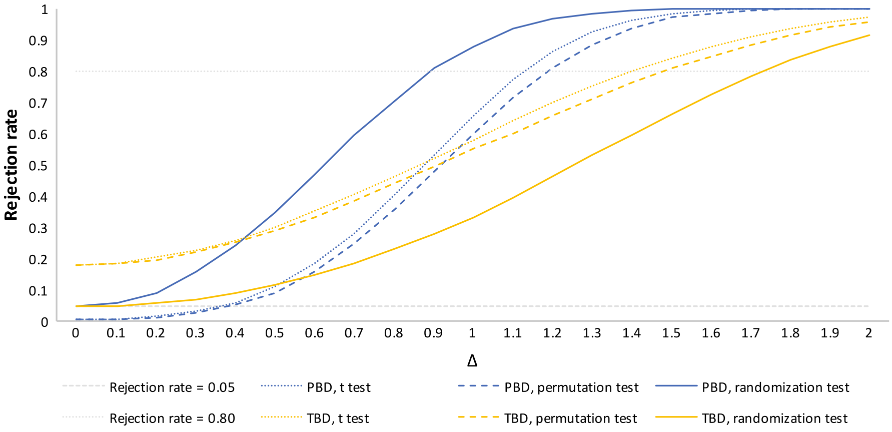

Fig. 1 Power curves of the randomization test, the permutation test, and the *t*-test under a linear time trend model and two randomization procedures

the usual sentiment that nonparametric tests have less power than parametric tests. It can further be seen that the performance of the permutation test is close to the *t*-test yet dissimilar to the randomization test, which emphases gap between population-based and randomization-based inference in the analysis of RCTs.

4 Power of randomization tests and choice of randomization procedure and test statistic

Though randomization tests are statistically valid by construction, which randomization procedure or test statistic would be most useful for effectively verifying treatment differences depend on the nature of variability in patient responses and how large the treatment differences are likely to be (Anscombe 1948). Consider a RCT with two treatment arms, A and B. In this section, we discuss the impact of randomization procedure and test statistic on power under three background models of variability in patient responses: Time trend, outliers, and heavy-tailed outcome data.

In each model, we compare the performance of randomization tests under two test statistics, difference in means and simple rank statistic, and eight randomization procedures: *complete randomization* (CR), RAR, TBD, PBD with block size $m = 4$, random block design (RBD) with maximal block size $B_{\max} = 3$, Wei's *urn design* with parameters $\alpha = 0, \beta = 1$ (UD(0,1)), Efron's *biased coin design* (BCD) with parameter $p = 2/3$, the *big stick design* (BSD) with imbalance intolerance parameter $b = 3$. The RAR is used within each block of the PBD and RBD. The randomization procedures can be categorized into four groups: CR, forced balanced designs (RAR, TBD), forced balanced design within blocks (PBD, RBD), and other designs that are developed to balance treatment assignments (UD(0,1), BCD, BSD). Details of these procedures can be found in Rosenberger and Lachin (2016, Chapter. 3). Each simulation is based on $10,000$ re-randomization tests with sample size $n = 50$. In each test, treatment assignment sequence is re-generated 15,000 times.

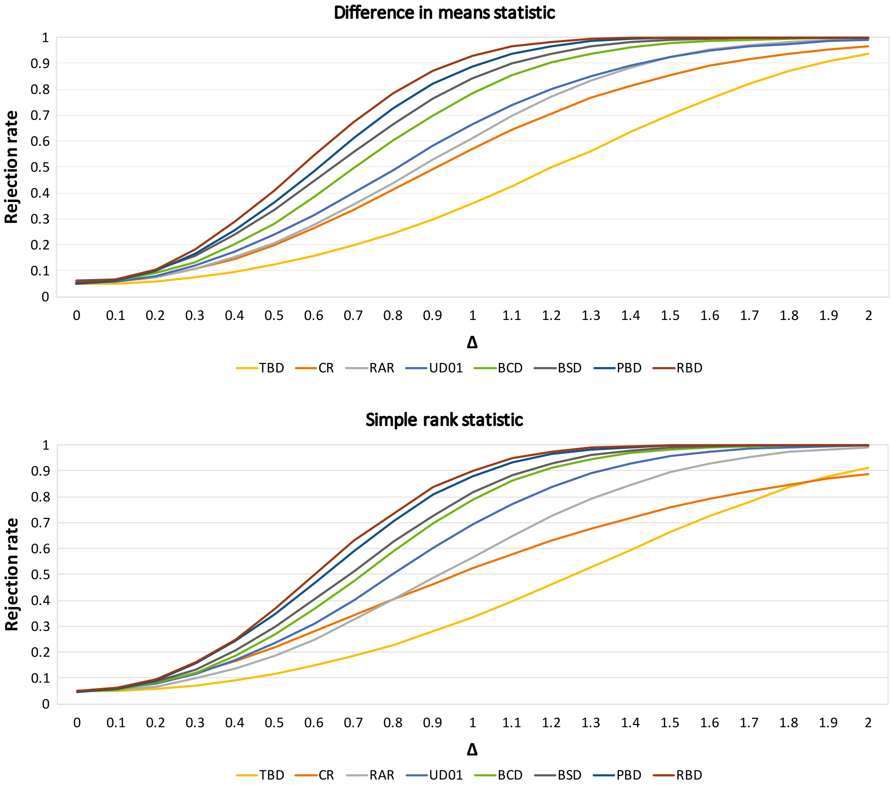

Fig. 2 Power curves of randomization tests under a linear drift and eight randomization procedures

4.1 Time trend

The linear drift (see Sect. 3) is applied to model time trend. The results shown in Fig. 2 elucidate the differentiation of power with regard to randomization procedures and, again, the preservation of type I error rate. The lowest power occurs when the TBD is employed, whereas PBD and RBD achieve the highest power at all Δ values. Recall that block designs ensure balanced assignment within every block of patients thereby balance treatment assignments throughout the course of a trial. On the contrary, the TBD can result in serious imbalance at some point in the trial since treatment assignments would end with either A's or B's with non-negligible probability. It is also observed in this situation that the power curves are not sensitive to the change of test statistic.

Since the highest power is attained by block designs, we further examine the relation of block size to power. With the employment of the PBD, the power curves are simulated under a series of block sizes; $m = 2, 4, 8, \ldots, 44, 48, 50$. The result (see Fig. 3) shows a consistent improvement of power with the gradual decrease of block size for both test statistics. This illustrates that the confounding impact of time trend can be governed by adjusting block size with respect to the expected amount of heterogeneity in patient responses.

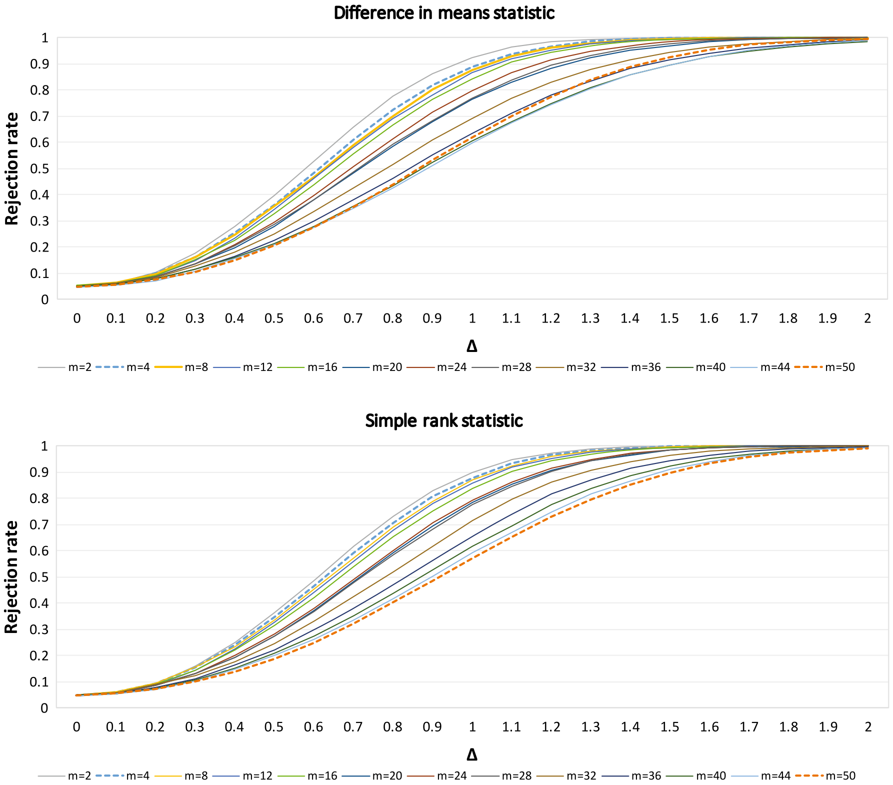

Fig. 3 Power curves of randomization tests under a linear drift and PBDs with different block size m

4.2 Outliers

The situation of outliers is modeled by Cauchy distribution $Cauchy(x_0, \gamma)$, where x_0, γ are the location and scale parameter respectively. Under H_0, patient outcomes are sampled from $Cauchy(0, 1)$. Under H_A, patient responses to treatment A are sampled from $Cauchy(\Delta, 1)$. Figure 4 shows that the differences of power curves between the randomization procedures are negligible, indicated by overlapped curves for both test statistics. But the test of difference in means has low power compared to the test using linear rank statistic for all randomization procedures. For example, at $\Delta = 2$, the powers given by difference in mean statistic aggregate at 0.29, while those given by simple rank statistic reach 0.88. Noticing that Cauchy distribution does not has a finite distributional mean, it is expected that difference in mean statistic cannot be effective in discerning the treatment group difference.

We further study a less extreme model of outliers. Under H_0, patient outcomes are sampled either from $N(0, 1)$ (with probability 0.8) or from $N(5, 1)$ (with probability 0.2). Under H_A, patient responses to treatment A are sampled either from $N(\Delta, 1)$ (with probability 0.8) or from $N(5 + \Delta, 1)$ (with probability 0.2). Figure 4 shows that, when the simple rank statistic is applied, the distinction between power curves

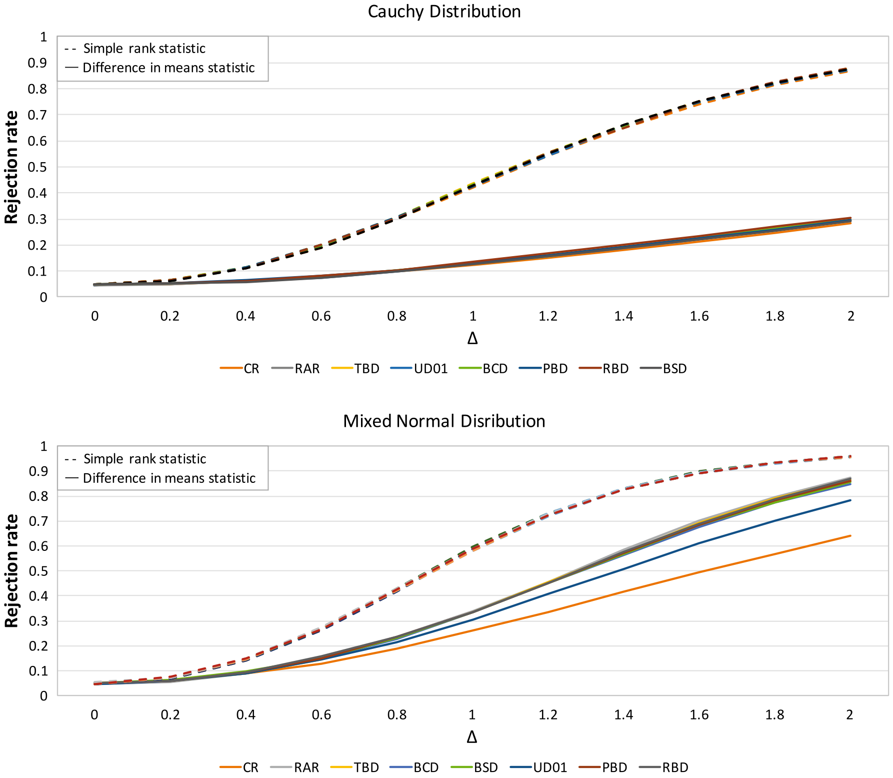

Fig. 4 Power curves of randomization tests under two outliers model and eight randomization procedures

under different randomization procedures is negligible. When the difference in means statistic is used, the impact of the choice of randomization procedure is displayed to some extent: the power curves of UD(0,1) and CR are below those given by other procedures. Since balancing treatment assignment is not a consideration in CR, and UD tends to CR asymptotically (Rosenberger and Lachin 2016), the result is not surprising. It thus implies that the impact of unbalanced treatment assignments on power in this situation is reduced by using the simple rank statistic. Moreover, it is again seen that the test of difference in means produces low power in comparison to the test using simple rank statistic for all randomization procedures. We therefore conclude that the choice of test statistic has a greater impact.

4.3 Heavy-tailed

Heavy-tailed outcome data is model by exponential distribution $Exp(\Delta)$ where Δ is the mean. Under H_0, patients outcomes are assumed to be $Exp(1)$. Under H_A, patient responses to treatment A are sampled from $Exp(1 + \Delta)$. From the results shown in Fig. 5, it is observed that, similar to the outliers model, the influence of randomization procedure on power varies according to test statistic. The impact of

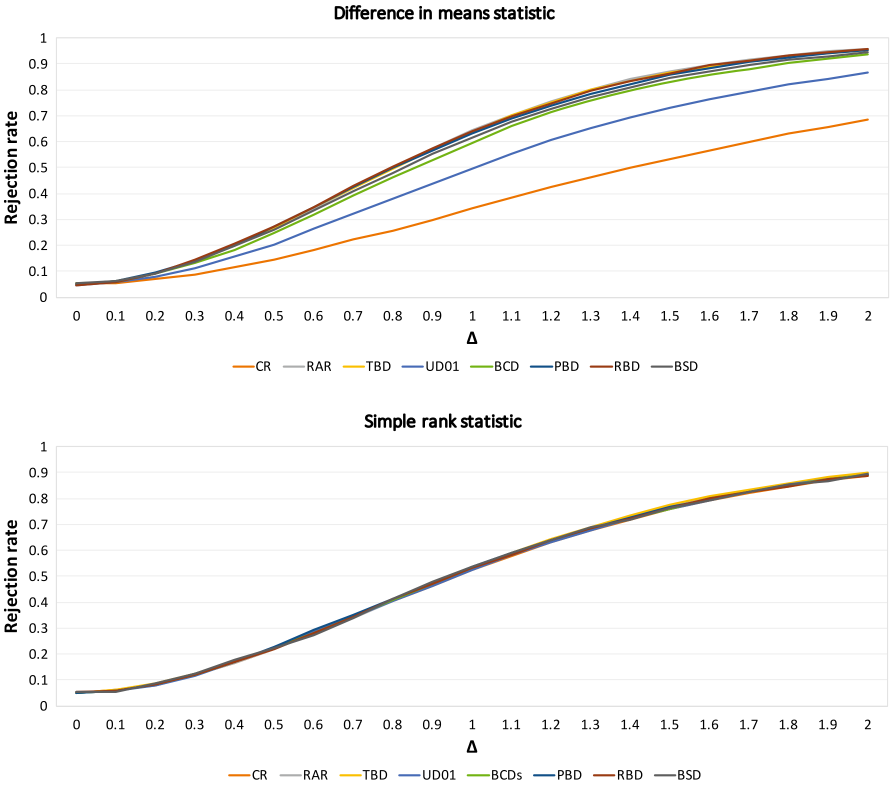

Fig. 5 Power curves of randomization tests under a heavy-tailed model and eight randomizations

randomization procedures is displayed when the difference in mean statistic is used. Specifically, the power of UD(0,1) and CR are much lower than those given by other randomization procedures. For instance, at $\Delta = 1$, the powers for CR, UD(0,1), and RBD are 0.34, 0.50, and 0.64, respectively. When simple rank statistic is employed, power of the test is unaffected by the choice of randomization procedures, represented by overlapped power curves. However, unlike the outliers model, improvement in power is not obvious.

4.4 Conclusion

We observed from the above three extreme examples the dynamic of selecting an appropriate randomization procedure with regard to the research interest (test statistic) and the clinical circumstance, which is made possible via studying the power of randomization tests. Periodic balance of treatment allocation in randomization helps mitigate the confounding influence of a time trend on the analysis of the treatment effect. When variability in patient responses is not related to the sequential order of treatment assignment (e.g., outliers, heavy-tailed), however, the power appears less contingent on the randomization procedures compared. In these circumstances, the choice of test statistic also matters. There may be other factors that weight an investigator's selection of a randomization procedure, including consideration of selection bias, balance, ethics, and presence of covariates. See (Rosenberger and Lachin 2016) for details.

5 Discussion

Discussions of other forms of analysis can be found in other literature, including covariate-adjusted regression models (Parhat et al. 2014) and confidence interval estimation (Rosenberger et al. 2018). Note that the model of data analysis (randomization model or population model) and the model of treatment effect (constant additive effect or covariate-associated effect) is not the same topic. This paper focuses only on the former.

In the study of type I error rate and power, one point that often causes confusion is that a completed RCT can by no means be replicated. In this sense, the practical meaning of type-I error rate and power, which are obtained under a population of replication, is ambiguous. To this we explain that, in calculating error rate and power of a randomization test, what is repeated is not a completed trial, but a thought experiment under a certain hypothesis. Recall that a RCT is conducted under equipoise; hypotheses of the treatment effect are applied in the analysis afterward. Because randomization tests integrate important information about the experiment and are statistically valid, the discussion of type I error rate and power becomes meaningful in the practical context: applying the method enables improved selection of randomization procedures in accordance with clinical circumstances to achieve the best possible scientific conclusion.

References

Anscombe FJ (1948) The validity of comparative experiments. J R Stat Soc Ser A 111:181–211

Parhat P, Rosenberger WF, Diao G (2014) Conditional monte carlo randomization tests for regression models. Stat Med 33:3078–3088

Pesarin F (2001) Multivariate permutation tests: with applications in biostatistics. Wiley, New York

Plamadeala V, Rosenberger WF (2012) Sequential monitoring with conditional randomization tests. Ann Stat 40:30–44

Rosenberger WF, Lachin JM (2016) Randomization in clinical trials: theory and practice, 2nd edn. Wiley, Hoboken

Rosenberger WF, Uschner D, Wang Y (2018) The 15th armitage lecture-randomization: the forgotten component of the randomized clinical trial. Stat Med 38:1–12