Published in final edited form as:
Stat Med. 2014 November 30; 33(27): 4734–4742. doi:10.1002/sim.6288.

Connections between Permutation and t-Tests: Relevance to Adaptive Methods

Michael Proschan[1], **Ekkehard Glimm**[2], and **Martin Posch**[3]

[1]National Institute of Allergy and Infectious Diseases, Bethesda, Maryland [2]Novartis Pharmaceuticals, Basel, Switzerland [3]Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria.

Abstract

A permutation test assigns a p-value by conditioning on the data and treating the different possible treatment assignments as random. The fact that the conditional type I error rate given the data is controlled at level α ensures validity of the test even if certain adaptations are made. We show the connection between permutation and t-tests, and use this connection to explain why certain adaptations are valid in a t-test setting as well. We illustrate this with an example of blinded sample size re-calculation.

Keywords

adaptive methods in clinical trials; blinded sample size re-calculation; p-value combination functions; permutation tests; asymptotic distribution; complete; sufficient statistic

1 Introduction

Randomized controlled trials are considered the gold standard for determining whether a new treatment is superior to a control. Assigning treatments at random tends to produce comparable arms that can then be compared in an unbiased manner. Consistent with the high degree of rigor in clinical trials is an analysis plan requiring as few assumptions as possible to avoid making subsequent changes in response to data not fitting the assumed model. One such plan in a fixed design setting uses a permutation test, whose validity is guaranteed under the strong null hypothesis that the experimental treatment has no effect compared to the control. Under this hypothesis, the observed data should be equally plausible regardless of the treatment labels. We can generate a valid reference distribution by 1) fixing the data at their observed values, 2) re-generating treatment labels, 3) re-computing the test statistic corresponding to those labels, and 4) repeating steps 1-3 until we exhaust the possibilities.

Because permutation tests condition on the observed data, they are also attractive in adaptive settings [1-2]. For example, consider a two-stage design in which we look at the first stage data blinded to treatment assignment and decide to increase the second stage sample size because the variance was larger than expected. The permutation distribution of the between-arm difference of the first stage is not marred by having looked at the data because a permutation test already conditions on all data other than the treatment assignments. This suggests that we might be able to combine the first stage p-value with a p-value from the second stage in a way that preserves the overall type I error rate.

Inferences from permutation tests are often extremely close to those of t-tests because permutation distributions are closely approximated by normal distributions. This is well-recognized (e.g., see Sections 4.1 and 4.2 of [3] for the two-sample and one-sample cases, respectively), and formal proofs date back many decades [4-6]. If we can construct a valid two-stage adaptive procedure using permutation tests, and permutation tests are asymptotically equivalent to t-tests, it seems reasonable that we might be able to use adaptive methods in a t-test setting as well. We show one such adaptation, sample size change, in a two-stage adaptive t-test setting.

We show the close connection between permutation tests and t-tests in both paired (Section 2) and unpaired (Section 3) settings. The permutation test is particularly simple in the paired setting because it is equivalent to conditioning on the absolute value of paired differences, as recognized by O'Brien and Fleming [7], among others. We exploit the asymptotic equivalence of permutation and t-tests to show why, under certain conditions, valid adaptations in a permutation test setting are also valid in a t-test setting. Section 4 explores the usefulness of these results in adaptive clinical trials with continuous outcomes.

2 One-Sample and Paired Settings

2.1 The Permutation Distribution of the Test Statistic

Paired data can arise in different ways in clinical trials, such as crossover designs, pair-matched community randomized trials, and trials in which the experimental treatment is applied to one eye or ear, etc., and a control to the other. The outcome is a difference D_i between the treatment (T) and control (C) observations on pair i. The permutation test is based on the idea that under the null hypothesis, D_i is symmetric about 0. Once we condition on $|D_i| = |d_i|$, or equivalently, $D_i^2=d_i^2$, D_i is equally likely to be ±$|d_i|$. That is, the conditional distribution of D_i given $D_i^2=d_i^2$ is the distribution of $Z_i d_i$, where d_i is fixed and

$$Z_i=\begin{cases} -1 & \text{with probability} & 1/2 \\ +1 & \text{with probability} & 1/2. \end{cases} \quad (1)$$

Because the D_i were independent before conditioning on $D_i^2=d_i^2$, $i = 1, \ldots, n$, the Z_i are also independent. Even though exactly half of the participants are assigned to the order (T,C) and half to (C,T), we need not impose the additional constraint that $\sum_{i=1}^n Z_i=0$. The symmetry assumption alone ensures that the conditional distribution of $\sum_{i=1}^n D_i$ given $D_1^2=d_1^2, \ldots, D_n^2=d_n^2$ is the distribution of

$$\sum_{i=1}^n Z_i d_i, \quad (2)$$

where the d_i are fixed constants and Z_i are iid with distribution (1). We consider all 2^n possibilities, $Z_i = +1$ or -1, $i = 1, \ldots, n$, and calculate $\sum_{i=1}^{n} Z_i d_i$ for each. We then see where the observed value $\sum_{i=1}^{n} d_i$ lies with respect to this permutation distribution. This is consistent with the analysis in Section 4.2 of [3]. For instance, for a 1-tailed test rejecting the null hypothesis for a large sum of differences, the p-value is the proportion of sums (2) that are at least as large as $\sum_{i=1}^{n} d_i$.

When the number of permutations is small, we can enumerate all values in the permutation distribution. When the number of pairs is large, we can approximate the p-value by simulation, generating a large number of vectors $(Z_1, \ldots, Z_n)$ and computing $\sum_{i=1}^{n} Z_i d_i$ for each. The approximate 1-tailed p-value is the proportion of simulated values of $\sum_{i=1}^{n} Z_i d_i$ that are at least as large as $\sum_{i=1}^{n} d_i$. An alternative method of approximating the p-value when n is large is presented in the next subsection.

2.2 Approximating The Permutation Distribution for Large *n*

We approximate the permutation distribution using a normal distribution with the same mean and variance. The mean of Z_i is, from (1), 0, so the mean of the permutation distribution of $\sum_{i=1}^{n} D_i$, namely the mean of Expression 2, is also 0. The variance of Z_i is just $\mathrm{E}\left(Z_i^2\right)=\mathrm{E}\left(1\right)=1$, so the permutation variance of the sum of differences is $\sum_{i=1}^{n} d_i^2$. Approximating the permutation distribution by a normal with mean 0 and variance $\sum_{i=1}^{n} d_i^2$ can be justified rigorously using the Lindeberg-Feller version of the central limit theorem because we are dealing with a sum of independent, but not identically distributed, random variables with mean 0 and respective variances d_i^2, $i = 1, 2, \ldots$ (see Section 27 of [8]).

Instead of starting with the unstandardized statistic $\sum_{i=1}^{n} D_i$, we could have begun with the standardized test statistic,

$$\tilde{T}=\frac{\sum_{i=1}^{n} D_i}{\sqrt{\sum_{i=1}^{n} D_i^2}}=\frac{\sum_{i=1}^{n} D_i}{\sqrt{n\tilde{\sigma}^2}}, \quad (3)$$

where $\tilde{\sigma}^2$ is the *total variance* $\sum_{i=1}^{n} D_i^2/n$. The only difference between (3) and the usual t-statistic is the use of the total variance $\tilde{\sigma}^2$ instead of the usual variance estimate $s^2=\sum_{i=1}^{n}\left(D_i-\bar{D}\right)^2/\left(n-1\right)$. The same reasoning as before shows that the permutation distribution of $\tilde{T}$, namely the conditional distribution of $\tilde{T}$ given $D_1^2=d_1^2, \ldots, D_n^2=d_n^2$, is the distribution of

$$\frac{\sum_{i=1}^{n} Z_i d_i}{\sqrt{\sum_{i=1}^{n} d_i^2}}. \quad (4)$$

Here again the d_i are fixed constants and the Z_i are the only source of randomness.

We can summarize these findings as follows.

1. The permutation distribution of $\sum_{i=1}^{n} D_i$ is the conditional distribution of $\sum_{i=1}^{n} D_i$ given $D_1^2 = d_1^2, \ldots, D_n^2 = d_n^2$, which is the distribution of Expression (2).
2. For large n, this conditional distribution depends on $d_1^2, \ldots, d_n^2$ only through $\tilde{\sigma}^2 \sum_{i=1}^{n} d_i^2/n$. Thus, $\tilde{T} | \left(D_1^2 = d_1^2, \ldots, D_n^2 = d_n^2\right) \approx \tilde{T} | \tilde{\sigma}^2$.
3. For large n, the conditional distribution of the standardized statistic $\tilde{T}$, given $\tilde{\sigma}^2$, is approximately N(0, 1).

Item 3 shows that the asymptotic conditional distribution of $\tilde{T}$ given $\tilde{\sigma}^2$ does not depend on $\tilde{\sigma}^2$. In other words, $\tilde{T}$ must be asymptotically independent of $\tilde{\sigma}^2$.

2.3 Deducing An Exact Result for Normal Random Variables

The preceding subsection showed that the conditional distribution of $\tilde{T}$ is, for large n, approximately normal and independent of $\tilde{\sigma}^2$. This suggests that for iid normal data with mean 0, $\tilde{T}$ might be independent of $\tilde{\sigma}^2$ for *any* sample size n. We will show the equivalent result that $\sum_{i=1}^{n} D_i / \left(\sum_{i=1}^{n} D_i^2\right)^{1/2}$ is independent of $\sum_{i=1}^{n} D_i^2$ for any n if the D_i are iid N(0, σ^2).

The geometric way to verify that $\sum_{i=1}^{n} D_i / \left(\sum_{i=1}^{n} D_i^2\right)^{1/2}$ is independent of $\sum_{i=1}^{n} D_i^2$ for normal data is to note that the distribution of a random sample from N(0, σ^2) is radially symmetric, meaning that the conditional distribution of $D_1, \ldots, D_n$ given $\sum_{i=1}^{n} D_i^2 = r^2$ is uniform on the hypersphere $\sum_{i=1}^{n} D_i^2 = r^2$ of radius r (see Figure 1 for the case of n = 2). Accordingly, the conditional distribution of

$$\frac{D_1}{\sqrt{\sum_{i=1}^{n} D_i^2}}, \ldots, \frac{D_n}{\sqrt{\sum_{i=1}^{n} D_i^2}}$$

given $\sum_{i=1}^{n} D_i^2 = r^2$ is uniform on the unit hypersphere. It follows that the conditional distribution of $D_1 / \left(\sum_{i=1}^{n} D_i^2\right)^{1/2} + \ldots + D_n / \left(\sum_{i=1}^{n} D_i^2\right)^{1/2}$ given $\sum_{i=1}^{n} D_i^2$ is the sum of components of a uniform random vector on the unit hypersphere, and therefore does not depend on $\sum_{i=1}^{n} D_i^2$. This also follows from Theorem 2.4.1 of [9] that $\left(D_1^2 / \sum_{i=1}^{n} D_i^2, \ldots, D_n^2 / \sum_{i=1}^{n} D_i^2\right)$ has a Dirichlet distribution D(1/2, . . . , 1/2) and is independent of $\sum_{i=1}^{n} D_i^2$.

The geometric method of proof suggests an alternative method of constructing a conditional test if we know the data are normal. When the number of observations is very small, e.g., n

= 2, the coarseness of the permutation distribution makes it impossible to achieve statistical significance. An alternative test based on rotation symmetry conditions on $\sum_{i=1}^{n} D_i^2 = r^2$ and treats all rotations of the data as equally likely. For example, when $n = 2$, we can 1) generate a uniform deviate θ on $[0, 2\pi)$, 2) set $D_1 = r\cos(\theta)$ and $D_2 = r\sin(\theta)$, 3) compute $\tilde{T}$ of (3), and repeat these three steps many times to get its reference distribution. The same procedure can be used for any n, but constructing rotations is somewhat more complicated [10]. For iid data data, all rotations are equally likely if and only if the D_i are $N(0, \sigma^2)$ for some σ^2 (see problem 4, page 53 of [11]). Thus, for iid data, the rotation method makes the same assumption as the t-test. Therefore, a more direct alternative is to simulate standard normal observations a very large number of times and estimate the $(1 - \alpha)$th quantile of $\tilde{T}$. This is a valid conditional test because, as we have seen, the conditional distribution of $\tilde{T}$ given $\sum_{i=1}^{n} D_i^2 = r^2$ is the same as the unconditional distribution of $\tilde{T}$.

A second way to verify the claim that $\sum_{i=1}^{n} D_i / \left(\sum_{i=1}^{n} D_i^2\right)^{1/2}$ is independent of $\sum_{i=1}^{n} D_i^2$ actually shows a more general result. It is based on Basu's theorem. Recall that a statistic $\underline{S}$ (which could be a vector) is said to be *sufficient* for a parameter $\underline{\theta}$ if the conditional distribution of the data given $\underline{S}$ does not depend on $\underline{\theta}$. $\underline{S}$ is called *complete* if $E\{f(\underline{S})\} = 0$ for all $\underline{\theta}$ implies that $f(\underline{S}) = 0$ with probability 1, where $f: R^k \mapsto R$ is any Borel function. A statistic $\underline{A}$ is called *ancillary* if its distribution does not depend on $\underline{\theta}$.

Theorem 1—Basu (1955) [12]. If $\underline{S}$ is sufficient and complete and $\underline{A}$ is ancillary, then $\underline{S}$ and $\underline{A}$ are independent.

In the setting of iid normal data $D_1, \ldots, D_n$ with known mean 0 and unknown variance σ^2, $\tilde{\sigma}^2 = \sum_{i=1}^{n} D_i^2 / n$ is sufficient and complete, while $\tilde{T}$ is ancillary because it is invariant to division of each D_i by the same constant. By Basu's theorem, $\tilde{T}$ and $\sum_{i=1}^{n} D_i^2$ are independent under the null hypothesis that $E(D_i) = 0$. They are not independent under the alternative hypothesis that $E(D_i) \neq 0$.

Notice that the more commonly used test statistic

$$T = \frac{\bar{D}}{\sqrt{s^2/n}} = \frac{\sum_{i=1}^{n} D_i}{\sqrt{n \sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2 / (n-1)}} = \frac{\sum_{i=1}^{n} D_i}{\sqrt{n s^2}}$$

is also ancillary for the same reason, where $s^2 = (n-1)^{-1} \sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2$ is the usual sample variance. Basu's theorem implies that the usual t-statistic is also independent of $\sum_{i=1}^{n} D_i^2$ under the null hypothesis that $E(D_i) = 0$ (page 412 of [13]).

3 Two-sample Settings

3.1 The Permutation Distribution of the Test Statistic

Consider a two sample setting with *n*/2 observations from treatment (T) and *n*/2 from control (C). The unstandardized statistic is

$$\sum_{i \in T} X_i - \sum_{i \in C} X_i. \quad (5)$$

Under the null hypothesis, conditioned on $X_1 = x_1, \ldots, X_n = x_n$, the treatment observations are equally likely to be any subset of size *n*/2 from $(x_1, \ldots, x_n)$. Therefore, the permutation distribution of the unstandardized statistic (5), namely its conditional distribution given $X_1 = x_1, \ldots, X_n = x_n$, is the distribution of

$$\sum_{i=1}^{n} Z_i x_i, \quad (6)$$

where the x_i are fixed constants and each Z_i has distribution given by (1); however, unlike the paired setting, the two-sample setting requires the imposition of the constraint that $\sum_{i=1}^{n} Z_i = 0$ because the distribution of the difference in sample means changes if the numbers assigned to T and C change. Note the similarity between (6) and (2). The only difference is that the Z_i in (2) are independent, whereas the Z_i of (6) are not because they sum to 0. This small deviation leads to slightly different variances for (6) and (2). For small *n*, we can enumerate all possible *Z* vectors of ±1 that sum to 0, and compute (6) for each. The one-tailed p-value is the proportion of statistics that are at least as large as the observed value. For large *n*, we can simulate or use the method of the next subsection.

3.2 Approximating The Permutation Distribution for Large *n*

As in the paired setting, we approximate the permutation distribution with a normal distribution with the same mean and variance (a formal proof of the asymptotic normality of the permutation distribution is in [14]). The mean of (6) is clearly 0. From expression 7.13 of [15],

$$var\left(\sum_{i=1}^{n} Z_i x_i\right) = \frac{n}{n-1} \sum \left(x_i - \bar{x}\right)^2 = n\tilde{\sigma}^2, \quad (7)$$

where $\tilde{\sigma}^2 = (n-1)^{-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$ is the *total variance*, the sample variance of all observations. Thus, the conditional distribution of (5) given $X_1 = x_1, \ldots, X_n = x_n$, namely the distribution of Expression (6), depends on $x_1, \ldots, x_n$ only through $\tilde{\sigma}^2$.

Suppose that we had begun with the standardized statistic

$$\tilde{T} = \frac{\sum_{i \in T} X_i - \sum_{i \in C} X_i}{\sqrt{n\tilde{\sigma}^2}}. \quad (8)$$

Because the denominator is constant once we condition on $X_1 = x_1, \ldots, X_n = x_n$, the permutation distribution of (8) is the distribution of

$$\frac{\sum_{i=1}^{n} Z_i x_i}{\sqrt{n\tilde{\sigma}^2}},$$

which is approximately standard normal. Again we are asserting that the *conditional* distribution of $\tilde{T}$ given $\tilde{\sigma}^2$ is approximately standard normal. Recapitulating, we have:

1. The permutation distribution of $\sum_{i\in T}^{n} X_i - \sum_{i\in C} X_i$ is the distribution of $\sum_{i=1}^{n} Z_i x_i$, namely the conditional distribution of $\sum_{i=1}^{n} Z_i X_i$ given $X_1 = x_1, \ldots, X_n = x_n$.
2. For large n, this conditional distribution depends on $x_1, \ldots, x_n$ only through the total variance $\tilde{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. Thus, $\tilde{T}|\left(X_1 = x_1, \ldots, X_n = x_n\right) \approx \tilde{T}|\tilde{\sigma}^2$.
3. For large n, the conditional distribution of the standardized statistic $\tilde{T}$ given $\tilde{\sigma}^2$ is approximately N(0, 1).

Because the conditional distribution of $\tilde{T}$ given $\tilde{\sigma}^2$ is approximately the same for every $\tilde{\sigma}^2$, $\tilde{T}$ is asymptotically independent of $\tilde{\sigma}^2$.

3.3 Deducing An Exact Result for Normal Random Variables

Having seen that $\tilde{T}$ is asymptotically independent of $\tilde{\sigma}^2$, we naturally inquire whether this holds for any n if the data are normally distributed. In the two sample setting, assume the null hypothesis that the X_i are iid N(μ, σ^2). Then $\left(\bar{X}, \tilde{\sigma}^2\right)$, being complete and sufficient, is independent of any ancillary statistic by Basu's theorem. In particular, Basu's theorem implies that $\left(\bar{X}, \tilde{\sigma}^2\right)$ is independent of the ancillary statistic $\tilde{T}$ of (8). $\tilde{T}$ is ancillary because transforming each observation X_i by $X_i' = \left(X_i - a\right)/b$ does not change $\tilde{T}$. Furthermore, the usual t-statistic

$$T = \frac{\sum_{i\in T} X_i - \sum_{i\in C} X_i}{\sqrt{ns^2}}$$

where s^2 is the familiar pooled variance, is also ancillary for the same reason. Therefore, T is independent of $\left(\bar{X}, \tilde{\sigma}^2\right)$ (page 414 of [13]).

For simplicity, we have assumed equal sample sizes in the two arms. With unequal sample sizes, the same arguments prove that $\tilde{T}$ and T are each independent of $\left(\bar{X}, \tilde{\sigma}^2\right)$, where

$$\begin{aligned}\tilde{T}= & \frac{(1/n_T)\sum_{i\in T}X_i-(1/n_C)\sum_{i\in C}X_i}{\sqrt{\tilde{\sigma}^2(1/n_T+1/n_C)}} \quad \text{and} \\ T= & \frac{(1/n_T)\sum_{i\in T}X_i-(1/n_C)\sum_{i\in C}X_i}{\sqrt{s^2(1/n_T+1/n_C)}}.\end{aligned} \quad (9)$$

The derivation is similar to the equal sample size case except that

$$Z_i=\begin{cases} +1/n_T & \text{w.p. } \frac{n_T}{n_T+n_C} \\ -1/n_C & \text{w.p. } \frac{n_C}{n_T+n_C} \end{cases}$$

and the permutation variance of $\sum_{i=1}^{n} Z_i x_i$ is $\tilde{\sigma}^2 \,(1/n_T+1/n_C)$.

4 Relevance to Adaptive Methods

As mentioned in the Introduction, permutation tests are useful in adaptive settings because they already condition on data other than the treatment labels, including the data used to modify the trial [1,16]. We exploit the connection between results about permutation and t-tests to show how to construct an exact, level α test with adaptive sample size modification.

We first review sample size calculations in a non-adaptive, paired t-test setting. We must estimate the variance σ^2 of D_i. The sample size for power $1 - \beta$ to detect a difference of size δ in a 1-tailed test at level α is

$$n=n\left(\sigma^2\right)=\frac{\sigma^2(z_\alpha+z_\beta)^2}{\delta^2}, \quad (10)$$

where, for $0 < a < 1$, z_a is the $100(1 - a)$th percentile of the standard normal distribution. Pre-trial variance estimates are based on data from other studies that might not be completely comparable to the current trial. It is appealing to use the current trial data to revise the sample size should the original variance estimate be too small. Revising the sample size in a way that preserves the blinding is also desirable to avoid bias [17].

The following is one possible method of blinded sample size re-calculation. Before the trial begins, use the best available data to form a sample size estimate n_0. After $n_1 = n_0/2$ observations (stage 1), replace σ^2 by the total variance $\tilde{\sigma}^2=(n_0/2)^{-1}\sum_{i=1}^{n_0/2} D_i^2$ and compute a “new” sample size $\nu=max\left(n_0, n\left(\tilde{\sigma}^2\right)\right)$ from (10). The usual t-statistic T_1 from stage 1 is independent of ν because ν is a function of $\tilde{\sigma}^2$, and T_1 is independent of $\tilde{\sigma}^2$ from the results of Section 2.3. Likewise, the first stage p-value P_1 is independent of ν. Therefore, the conditional distribution of P_1 given ν is uniform [0, 1].

Now accrue $n_2 = \nu - n_1 = \nu - n_0/2$ additional observations in stage 2 and compute the usual t-statistic T_2 and its p-value P_2 using only stage 2 observations. Conditioned on ν, P_2 is uniform [0, 1] under the null hypothesis. Now combine P_1 and P_2 to preserve the overall type I error rate. Two possible p-value combination functions are Fisher's combination method and the inverse normal combination method:

$$\begin{aligned} f\left(P_1,P_2\right)= & \quad -2 \quad ln\left(P_1P_2\right) \quad \text{or} \\ g\left(P_1,P_2\right)= & \quad \sqrt{\tfrac{n_1}{\nu}}\Phi^{-1}\left(1-P_1\right)+\sqrt{\tfrac{n_2}{\nu}}\Phi^{-1}\left(1-P_2\right) \end{aligned} \quad (11)$$

[18-19]. Whichever function is used, it must be specified before the trial. The conditional distribution of $f(P_1, P_2)$ given ν is chi-squared with 4 degrees of freedom. Therefore, if we reject the null hypothesis at the end of the trial if $f(P_1, P_2)$ exceeds the $100(1 - \alpha)$th percentile of a χ_4^2 distribution, the conditional type 1 error rate given ν is controlled at level α, hence so is the unconditional error rate. The same is true if we had pre-specified $g(P_1, P_2)$: if we reject if $g(P_1, P_2)$ exceeds the $100(1 - \alpha)$th percentile of the standard normal distribution, then the type 1 error rate is controlled both conditional on ν and unconditionally. Note that for the procedure to be valid, the recalculated sample size must be a function of $\sum_{i=1}^{n} D_i^2$; otherwise, the type I error rate need not be controlled. Also, one is not allowed to change the p-value combination function after looking at data.

Table 1 shows the simulated Type I error rate and power for four different adaptive methods: the one treating the recalculated sample size as if it had been fixed in advance, Fisher's combination of p-values, the inverse normal method with fixed and equal weights for the two stages, and the new inverse normal method with adaptive weights. The first stage sample size is $n_1 = 30$. Without loss of generality, we took the true variance to be 1. Notice that in these simulations of a million clinical trials with a first stage sample size of 30, the type I error rate was controlled at 0.05 even for the method treating the sample size as if it had been fixed in advance. However, the Appendix shows that type I error rate inflation is possible even when the sample size rule is based on the total variance. Our result is consistent with other findings of slight error rate inflation in some superiority and non-inferiority settings with unblinded or blinded sample size recalculation [20-22]. This is contrary to sentiments expressed in guidance documents issued by regulatory agencies such as the FDA that any blinded sample size re-calculation has no impact on the type 1 error rate. Nonetheless, for the sample size reassessment rule based on the standard sample size formulas for the two-sample t-test, no relevant inflation of the type 1 error rate was found for a wide variety of scenarios [23], consistent with sentiments expressed in [24]. Table 1 also shows that the new method (last column) has virtually the same power, but it is guaranteed to control the type I error rate.

We can do a similar thing in the two-sample setting, in which case the relevant total sample size formula is

$$n=n\left(\sigma^2\right)=\frac{4\sigma^2(z_\alpha+z_\beta)^2}{\delta^2}. \quad (12)$$

If n_0 is the pre-trial sample size estimate, stage 1 consists of the first $n_1 = n_0/2$ observations. Compute the total variance, $\tilde{\sigma}^2$, and the new sample size $\nu=max\left(n_0, n\left(\tilde{\sigma}^2\right)\right)$ using (12). At the end of the trial, combine the stages using the pre-specified p-value combination function and its null distribution (χ_4^2 or N(0, 1) for *f* or *g*, respectively). The type 1 error rate will be controlled both conditional on ν and unconditionally.

4.1 More General Asymptotic Results

Our argument in Section 2.2 actually shows that the conditional distribution of $\tilde{T}$ (and *T*) given $D_1^2=d_1^2, \ldots, D_n^2=d_n^2$ is approximately standard normal for large *n* regardless of whether the D_i are normal. This follows from the fact that, with probability 1, $\sum_{i=1}^{n} Z_i d_i / \left(\sum_{i=1}^{n} d_i^2\right)^{1/2}$ converges in distribution to a standard normal as $n \to \infty$ if the d_i are realizations from iid random variables from any distribution with finite variance (because the Lindeberg condition is satisfied). This argues for the asymptotic validity of adaptive t-tests even when the data come from a non-normal distribution (but symmetric under the null). In fact, the convergence of $\sum_{i=1}^{n} Z_i d_i / \left(\sum_{i=1}^{n} d_i^2\right)^{1/2}$ to a standard normal is relatively fast for symmetric distributions, which explains why there is no material inflation of the type I error rate even if we naively treat the recalculated sample size as if it had been prespecified (see Table 1). A similar comment applies in the two-sample setting.

5 Discussion

Permutation tests are very useful in adaptive clinical trials. Because they condition on all data other than treatment labels, they are valid under the strong null hypothesis even if we peek at data. This article exploited the close connection between permutation and t-tests to understand the validity of certain adaptive t-tests. Specifically, in the one-sample setting, we can peek at the data from the first stage of our trial in a blinded way, change the sample size, and still control the type 1 error rate both conditional on that sample size and unconditionally. In the two sample setting, we can peek at the overall mean and variance in a blinded way, change the sample size, and still control the type 1 error rate both conditionally and unconditionally. An important caveat is that we are testing the strong null hypothesis. In particular, for the t-test, we assume that under the null hypothesis, the data are normally distributed with equal variances in the two arms.

Acknowledgments

We would like to thank the referees for very helpful comments that improved the clarity and quality of the manuscript.

Martin Posch was supported by the Austrian Science Fund (FWF):P23167

Appendix: Potential Alpha Inflation for Sample Size Recalculation with t-tests and Small Samples

Consider a one-sample setting with a fixed sample size of three paired differences D_1, D_2, D_3. The t-statistic is

$$T_3 = \frac{\bar{D}_i}{\sqrt{\frac{\left(\sum_{i=1}^{3} D_i^2\right) - 3\bar{D}^2}{6}}}. \quad (13)$$

Suppose that $R_2 = r$, where $R_2 = \sqrt{D_1^2 + D_2^2}$. As $r \rightarrow 0$, both D_1 and D_2 converge to 0, and $|T_3|$ converges to 1. It follows that for $c > 1$, $P(T_3 > c \mid R_2 = r) \rightarrow 0$ as $r \rightarrow 0$. The same holds for $P(T_3 > c \mid R_2 \leq r)$. If c_3 is the level α critical value for the 1-tailed t-statistic T_3, then

$$\alpha = Pr\left(T_3 > c_3\right) = \lambda_r Pr\left(T_3 > c_3 \mid R_2 \leq r\right) + \left(1 - \lambda_r\right) Pr\left(T_3 > c_3 \mid R_2 > r\right),$$

where $\lambda_r = \Pr(R_2 \leq r)$. Assuming that α is small enough that $c_3 > 1$, $\Pr(T_3 > c_3 \mid R_2 \leq r) < \alpha$ for small r (because $|T_3|$ converges to 1 as $r \rightarrow 0$). Therefore, to offset this deficit and give level α for the t-test with a fixed sample size of 3, we must have

$$Pr\left(T_3 > c_3 \mid R_2 > r\right) > \alpha. \quad (14)$$

Now consider a two-stage procedure based on observing $R_2 = D_1^2 + D_2^2$ and deciding whether to stop at two observations or add a third. We know from the fact that T_2 is independent of R_2 that

$$Pr\left(T_2 > c_2 \mid R_2 \leq r\right) = \alpha, \quad (15)$$

where T_2 is the t-statistic based on just D_1 and D_2, and c_2 is its critical value. Putting (14) and (15) together, we find that the overall type I error rate

$$\lambda_r P\left(T_2 > c_2\right) + \left(1 - \lambda_r\right) \quad Pr\left(T_3 > c_3\right)$$

must exceed $\lambda_r \alpha + (1 - \lambda_r)\alpha = \alpha$ for sufficiently small r. That is, a procedure that chooses a sample size of 2 if $R_2 \leq r_0$ and 3 if $R_2 > r_0$ is guaranteed to have type I error rate inflation if r_0 is sufficiently small.

This same argument can be used for general n because if $R_{n-1} = \sqrt{D_1^2 + \ldots D_{n-1}^2}$, then $|T_n| \rightarrow 1$ as $R_{n-1} \rightarrow 0$. Any procedure that chooses sample size $n - 1$ if $R_{n-1} \leq r_0$ and n if $R_{n-1} > r_0$ is guaranteed to have type I error rate inflation if r_0 is sufficiently small.

References

1. Edwards D. On model prespecification in confirmatory randomized studies. Statistics in Medicine. 1999; 18:771–785. [PubMed: 10327526]
2. Posch M, Proschan MA. Unplanned adaptations before breaking the blind. Statistics in Medicine. 2012; 31:4146–4153. [PubMed: 22736397]
3. Box, GE.; Hunter, WG.; Hunter, JS. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley and Sons; New York: 1978.

4. Wald A, Wolfowitz J. Statistical tests based on permutations of the observations. Annals of Mathematical Statistics. 1944; 15:358–372.

5. Noether GE. On a theorem by Wald and Wolfowicz. Annals of Mathematical Statistics. 1949; 20:455–458.

6. Hoeffding W. A combinatorial central limit theorem. Annals of Mathematical Statistics. 1951; 22:558–566.

7. O'Brien PC, Fleming TR. A paired Prentice-Wilcoxon test for censored paired data. Biometrics. 1987; 43:169–180.

8. Billingsley, P. Probability and Measure Anniversary. John Wiley & Sons; New York: 2012.

9. Fang, KT.; Zhang, YT. Generalized Multivariate Analysis. Springer; New York: 1990.

10. Langsrud O. Rotation tests. Statistics and Computing. 2005; 15:53–60.

11. Arnold, SF. The Theory of Linear Models and Multivariate Analysis. Wiley; New York: 1981.

12. Basu D. On statistics independent of a complete sufficient statistic. Sankhya. 1955; 15:377–380.

13. Shao, J. Mathematical Statistics. Springer; New York: 2003.

14. van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; Cambridge: 1998.

15. Rosenberger, WF.; Lachin, JM. Randomization in Clinical Trials: Theory and Practice. John Wiley & Sons; New York: 2002.

16. Zucker DM, Wittes JT, Schabenberger O, Brittain E. Internal pilot studies II: comparison of various procedures. Statistics in Medicine. 1999; 18:3493–3509. [PubMed: 10611621]

17. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics–Theory and Methods. 1992; 21:2833–2853.

18. Bauer P, Köhne K. Evaluations of experiments with adaptive interim analyses. Biometrics. 1994; 50:1029–1041. [PubMed: 7786985]

19. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics. 1999; 55:1286–1290. [PubMed: 11315085]

20. Wittes J, Schabenberger O, Zucker D, Brittain E, Proschan M. Internal pilot studies I: type I error rate of the naive t-test. Statistics in Medicine. 1999; 18:3481–3491. [PubMed: 10611620]

21. Friede T, Kieser M. Blinded sample size reassessment in non-inferiority and equivalence trials. Statistics in Medicine. 2003; 22:995–1007. [PubMed: 12627414]

22. Golkowski D, Friede T, Kieser M. Blinded sample size re-estimation in crossover bioequivalence trials. Pharmaceutical Statistics. 2014; 13:157–162. [PubMed: 24715672]

23. Glimm E, Läauter J. Some notes on blinded sample size re-estimation. arXiv preprint. 2013; arXiv: 1301–4167.

24. Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Statistics in Medicine. 2003; 22:3571–3581. [PubMed: 14652861]

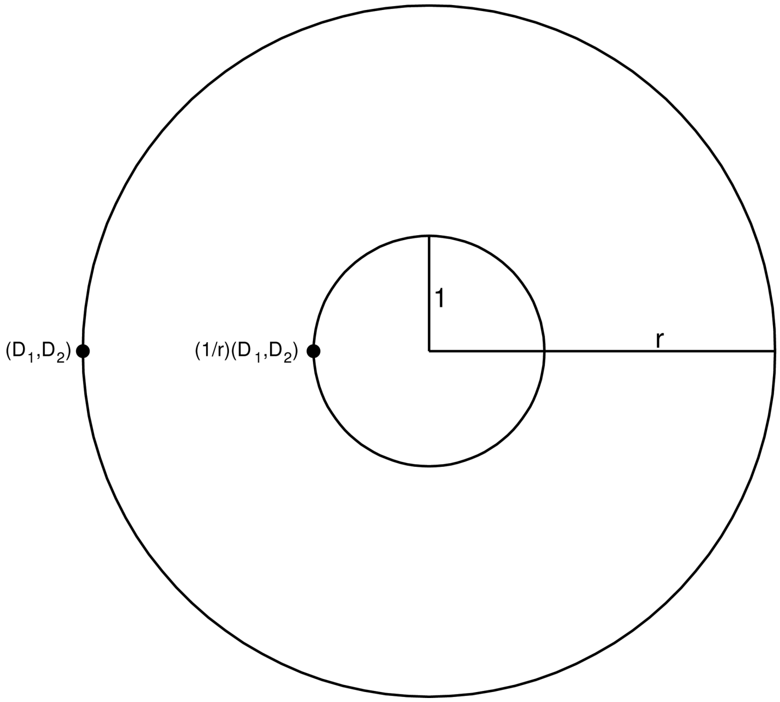

Figure 1.

If (D_1, D_2) are iid normals with mean 0, the conditional distribution of (D_1, D_2) given $D_1^2+D_2^2=r^2$ is uniform on the circle of radius *r* centered at (0, 0). Accordingly, given $D_1^2+D_2^2=r^2$, (1/*r*)(D_1,D_2 is uniform on the unit circle centered at (0, 0).

Table 1

Simulated type I error rate and power for four different adaptive tests described in the text. The first stage sample size is $n_1 = 30$, and a million trials were simulated.

True Mean	δ	fixed	Fisher	Inv Nor	New Inv Nor
0	0.5	0.050	0.050	0.050	0.050
0	1.0	0.050	0.050	0.050	0.050
0	1.5	0.050	0.050	0.050	0.050
0	2.0	0.050	0.050	0.050	0.050
0.3	0.3	0.982	0.974	0.971	0.982
0.3	0.45	0.884	0.863	0.875	0.883
0.3	0.6	0.775	0.748	0.773	0.773
0.5	0.5	0.997	0.995	0.996	0.997
0.5	0.75	0.975	0.967	0.974	0.974
0.5	1	0.946	0.930	0.938	0.944