Noname manuscript No.
(will be inserted by the editor)

Applications of Conditional Power Function of Two-Sample Permutation Test

Monjed H. Samuh · Fortunato Pesarin

Received: date / Accepted: date

Abstract Permutation or randomization test is a nonparametric test in which the null distribution (distribution under the null hypothesis of no relationship or no effect) of the test statistic is attained by calculating the values of the test statistic overall permutations (or by considering a large number of random permutation) of the observed dataset. The power of permutation test evaluated based on the observed dataset is called *conditional power*. In this paper, the conditional power of permutation tests is reviewed. The use of the conditional power function for sample size estimation is investigated. Moreover, reproducibility and generalizability probabilities are defined. The use of these probabilities for sample size adjustment is shown. Finally, an illustration example is used.

Keywords Generalizability probability · Permutation test · Reproducibility probability · Sample size adjustment · Sample size estimation

1 Introduction

The permutation test was first studied by Fisher (1934, 1935). He investigated the permutation approach for exact inference within the conditionality and sufficiency principles of inference. For example, he introduced the permutation test as the exact test for the relationship between two binary variables when the frequencies in some cells are small; that is, when the chi-square test fails in the sense that its asymptotic distribution can be quite far away from the exact. It is also useful for one sided testing. In addition, Fisher introduced the exact test for testing differences between means of two populations when

M. H. Samuh
Applied Mathematics & Physics Department, Palestine Polytechnic University, Palestine;
E-mail: monjedsamuh@ppu.edu

F. Pesarin
Department of Statistical Sciences, University of Padova, Italy

the assumptions of the two-sample t-test are not realized. He pointed out that the type I error probability for the two-sample permutation test in quite mild conditions is closely approximated by the normal theory.

Pitman (1937a,b, 1938) developed permutation tests in agreement with the F-test for the comparison of $k \geq 2$-samples and for bivariate correlation. For two-sample design, Pitman introduced a test statistic which is a monotonic increasing function of the square of the t-test statistic.

Permutation tests constitute a subclass of nonparametric tests (Lehmann and Romano, 2005; Pesarin and Salmaso, 2010). They are computationally intensive, but modern computational powerful tools make permutation tests feasible if, in place of complete enumeration of permutation sample space, a random sample is obtained so as to satisfy any desirable accuracy in computing p-values (See Algorithm 1 below). Nonparametric test statistics do not depend on any particular distribution. In fact, they are distribution free since pooled observed data are always sets of sufficient statistics for the underlying unknown distribution, assuming the null hypothesis is true (See Pesarin and Salmaso, 2010, Sec. 2.1.3)). Some minimal assumptions are required to the data (e.g. exchangeability in the null hypothesis, often referred to as equality in distribution). The exchangeability assumption is generally assured by randomly assigning treatments to experimental units in experimental designs. In case of observational study, exchangeability in the null hypothesis shall be assumed in order to obtain exact testing solutions. If this assumption cannot be justified, then approximate permutation solutions are obtained in the same way as the nonparametric Behrens-Fisher testing (Pesarin, 2001).

The theory of optimal permutation tests is developed by Lehmann and Stein (1949). Hoeffding (1952) studied the behavior of asymptotic power of permutation tests. He found that for the randomized block design and for the two-sample designs they are asymptotically as powerful as their corresponding tests based on the parametric approach when these are working within their ideal conditions. For instance, the permutation test for the randomized block design is asymptotically as powerful as F-test, and the two-sample permutation test is asymptotically as powerful as student's t-test.

Permutation tests are widely used in many research fields such as agriculture, clinical trials, educational statistics, business statistics and industrial statistics, etc. For more works on permutation test and its variations see Edgington (1995), Salmaso (2003), Good (2005), Basso et al (2009), Pesarin and Salmaso (2010), Samonenko and Robinson (2015), McDonald et al (2016), Amro and Pauly (2017), and the references therein.

2 Two-sample permutation test

In this paper, testing problems for one-sided alternative hypotheses, as produced by treatments with non-negative effect size δ, are considered. Particularly, the fixed additive effects model is considered. This is written as

$$X_{1i} = \mu + \delta + \sigma E_{1i},\ i = 1, \ldots, n_1; \quad X_{2i} = \mu + \sigma E_{2i},\ i = 1, \ldots, n_2, \qquad (1)$$

where μ is a common location parameter, E_{ji} are random error deviates (supposed to be exchangeable) with location parameter zero and scale parameter one, σ is a scale parameter independent on experimental units and treatment levels, and δ is the effect size (treatment effect) which is typically unknown. In practice, for comparing $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n_1})$ to $\mathbf{X}_2 = (X_{21}, \ldots, X_{2n_2})$, and, without any loss of generality, since no specific assumptions on nuisance parameters μ and σ are required by permutation tests, $\mu = 0$ and $\sigma = 1$ are assumed. Therefore, the dataset can be also written as $\mathbf{X}(\boldsymbol{\delta}) = (\mathbf{E}_1 + \boldsymbol{\delta}, \mathbf{E}_2)$ where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{n_1})$ (For simplicity, set $\delta_i = \delta > 0$, for $i = 1, \ldots, n_1$), $\mathbf{E}_1 = (E_{11}, \ldots, E_{1n_1})$, and $\mathbf{E}_2 = (E_{21}, \ldots, E_{2n_2})$. The hypothesis of interest is

$$H_0 : \boldsymbol{\delta} = 0 \text{ versus } H_1 : \boldsymbol{\delta} > 0. \tag{2}$$

A suitable test statistic should be chosen such that, without any loss of generality, large values of it are considered to be against H_0. For more details about the choice of statistic in the permutation framework see Page 84 of Pesarin and Salmaso (2010). In our setting, $T = \bar{X}_1 - \bar{X}_2$ is used as a test statistic.

Now, for determining the p-value, an appropriate reference distribution is needed which is called the permutation distribution. For the two-sample design, the permutation test is carried out as follows.

1. Randomly assign experimental units to one of the two groups with n_1 units assigned to the first group (treatment group) and n_2 units assigned to the second group (control or placebo group). Then, the observed datasets, $\mathbf{X}_1$ and $\mathbf{X}_2$, are obtained and the test statistic is evaluated, $T^o = T(\mathbf{X})$.
2. Permute the $n = n_1 + n_2$ observations between the two groups. Write down the set of all possible permutations, $\mathcal{X}$. The cardinality of this support is $n!$.
3. For each permutation $\mathbf{X}^* \in \mathcal{X}$, compute the test statistic, $T^* = T(\mathbf{X}^*)$. Since sample means are invariant with respect to the order of imputed data, the cardinality of related permutation space $\mathcal{X}$ reduces to

$$\binom{n}{n_1} = \frac{n!}{n_1! n_2!}.$$

4. Compute the p-value,

$$\lambda_T(\mathbf{X}) = \frac{\text{number of } T^*\text{'s} \geq T^o}{\binom{n}{n_1}}.$$

5. If a preassigned level of significance, α, has been set, declare the test to be significant if the p-value is not larger than this level.

Since it is tedious or even practically impossible to write down and enumerate the whole members of permutation sample space $\mathcal{X}$, conditional Monte Carlo simulation based on B random permutations from $\mathcal{X}$ (Algorithm 1) is used to approximate the p-value at any desired accuracy.

Algorithm 1 Conditional Monte Carlo (CMC)

1. For the given dataset $\mathbf{X}$, calculate the observed test statistic, $T^o = T(\mathbf{X})$.
2. From $\mathcal{X}$ take a random permutation $\mathbf{X}^*$ of $\mathbf{X}$, and calculate the corresponding test statistic $T^* = T(\mathbf{X}^*)$.
3. Independently repeat Step 2 a large number, e.g. B times, giving B values for T^*, say $\{T_b^*,\, b = 1, \ldots, B\}$.
4. The permutation p-value is estimated as

$$\hat{\lambda}_T(\mathbf{X}) = \frac{\sum_{b=1}^{B} \mathbb{I}(T_b^* \geq T^o)}{B},$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Note that $\hat{\lambda}_T(\mathbf{X})$ is an unbiased estimate of the true $\lambda_T(\mathbf{X})$ and, due to the Glivenko-Cantelli theorem (Shorack and Wellner, 1986), as B diverges it is strongly consistent. Moreover, the standard error for $\hat{\lambda}_T(\mathbf{X})$ is $\sqrt{\lambda_T(\mathbf{X})\left(1-\lambda_T(\mathbf{X})\right)/B}$. Therefore, a $100(1-\alpha)\%$ approximate confidence interval for $\lambda_T(\mathbf{X})$ is

$$\hat{\lambda}_T(\mathbf{X}) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\lambda}_T(\mathbf{X})\left(1-\hat{\lambda}_T(\mathbf{X})\right)}{B}}.$$

3 Conditional power function of permutation test

The statistical power of a hypothesis test is defined as

$$\Pr(\text{reject } H_0|\boldsymbol{\delta}) = \begin{cases} \alpha & \text{if } H_0 \text{ is true} \\ 1-\beta(\boldsymbol{\delta}) & \text{if } H_0 \text{ is false} \end{cases}$$

where α is the type I error probability and β is the type II error probability. The power of permutation tests calculated based on the observed dataset is called the *conditional power*. It is calculated as

$$W\left[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}\right] = \mathbb{E}\left\{\mathbb{I}[\lambda_T(\mathbf{X}^\dagger(\boldsymbol{\delta})) \leq \alpha]\big|\mathcal{X}\right\}, \tag{3}$$

which is a function of the *effect* $\boldsymbol{\delta}$ for a given sample size n, of preassigned level of significance α and of suitable test statistic T conditional on the permutation space $\mathcal{X}$ associated with the given pooled dataset $\mathbf{X}$. Note that $\lambda_T(\mathbf{X}^\dagger(\boldsymbol{\delta}))$ is evaluated based on the dataset $\mathbf{X}^\dagger(\boldsymbol{\delta}) = (\mathbf{E}_1^\dagger + \boldsymbol{\delta}, \mathbf{E}_2^\dagger)$, where $\mathbf{E}^\dagger \in \mathcal{E}$ is any reassignment (a permutation) of unobserved error deviates $\mathbf{E}$.

Apparently, the true value of the conditional power function is not only tedious but it is also difficult to attain exactly; since $[n!/(n_1!n_2!)]^2$ permutations have to be considered. Instead, empirical conditional power is evaluated based on iterated CMC simulation. Algorithm 2 is used for estimating it.

Algorithm 2 Empirical Conditional Power Function

1. Consider the set of error deviates $\mathbf{E} = \mathbf{E}_1 \uplus \mathbf{E}_2$ and the effect size $\boldsymbol{\delta}$, where $\uplus$ is the symbol for concatenating two vectors.
2. From $\mathbf{E}$ take a re-randomization $\mathbf{E}^\dagger$ of $\mathbf{E}$, and define $\mathbf{X}^\dagger(\boldsymbol{\delta}) = (\mathbf{E}_1^\dagger + \boldsymbol{\delta}, \mathbf{E}_2^\dagger)$.
3. Use the CMC algorithm to calculate $\hat{\lambda}_T(\mathbf{X}^\dagger(\boldsymbol{\delta}))$.
4. Independently repeat Steps 2 and 3 a large number, e.g. R times, giving R estimated p-values, say $\{\hat{\lambda}_T(\mathbf{X}_r^\dagger(\boldsymbol{\delta})),\ r = 1, \ldots, R\}$.
5. The empirical conditional power is evaluated as

$$\hat{W}[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}] = \frac{\sum_{r=1}^{R} \mathbb{I}[\hat{\lambda}_T(\mathbf{X}_r^\dagger(\boldsymbol{\delta})) \leq \alpha]}{R}.$$

6. Repeat Steps 1-5 for different values of $\boldsymbol{\delta}$ to attain a function in $\boldsymbol{\delta}$.

3.1 Empirical post-hoc conditional power function

It is clear from Algorithm 2 that in order to evaluate the empirical conditional power $\mathbf{E}$ must be known, but $\mathbf{X} = \boldsymbol{\delta} + \mathbf{E}$ is observed and its components ($\mathbf{E}$ & $\boldsymbol{\delta}$) cannot be separately observed in general. Thus, the conditional power is just a *virtual notion*. However, in lieu of $W[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}]$, the so-called *empirical post-hoc conditional power* $\hat{W}[(\boldsymbol{\delta}; \hat{\boldsymbol{\delta}}, n, \alpha, T)|\mathcal{X}]$ may be achieved. The main point is to find an empirical estimate of $\mathbf{E}$, $\hat{\mathbf{E}}$, by subtracting a suitable estimate of the effect $\boldsymbol{\delta}$, $\hat{\boldsymbol{\delta}}$, from the observed dataset $\mathbf{X}$. Thus, the empirical pooled set of error deviates is given by $\hat{\mathbf{E}} = \hat{\mathbf{E}}_1 \uplus \mathbf{E}_2 = (\mathbf{X}_1 - \hat{\boldsymbol{\delta}}) \uplus \mathbf{X}_2$. Note that this gives rise to approximate solution because exchangeability condition is now approximate as $\hat{\boldsymbol{\delta}}$, being essentially calculated from data of first group $\mathbf{X}_1$, is not a permutation invariant quantity. It is worth noting that the *empirical post-hoc conditional power* $\hat{W}[(\delta; \hat{\delta}, n, \alpha, T)|\mathcal{X}]$ is essentially ruled by the given data set $\mathbf{X}$ only. Also it is important to note that, according to (Pesarin and Salmaso, 2010, p. 98), this can be viewed as a least squares consistent estimate of the unconditional power function specific of the same test T, $W_P(\delta; n, \alpha, T)$ say. Actually, such a property resides on that $W_P(\delta; n, \alpha, T)$ is nothing but the mean value of $\hat{W}[(\delta; \hat{\delta}, n, \alpha, T)|\mathcal{X}]$ with respect to the underlying population distribution P. We use it as the basic notion that provides a rational justification to our application proposals as discussed from Section 4 onwards.

There are different approaches to estimate $\boldsymbol{\delta}$ which depend on the design of study (Cooper and Hedges, 1997; Hedges and Olkin, 1985; Cohen, 1988) and on the test statistic T actually used. For instance, in the two sample design when the test statistic T is based on comparison of two sample means typically it is $\hat{\boldsymbol{\delta}} = \bar{X}_1 - \bar{X}_2$. Algorithm 3 is used to find the empirical post-hoc conditional power function.

Algorithm 3 Empirical Post-Hoc Conditional Power Function

1. For the given dataset $\mathbf{X}$, find an estimate of $\boldsymbol{\delta}$, $\hat{\boldsymbol{\delta}}$. Then consider the consequent empirical error deviates $\hat{\mathbf{E}} = (\mathbf{X}_1 - \hat{\boldsymbol{\delta}}) \biguplus \mathbf{X}_2$.
2. From $\hat{\mathbf{E}}$ take a random re-randomization $\hat{\mathbf{E}}^{\dagger}$ of $\hat{\mathbf{E}}$, and for any chosen $\boldsymbol{\delta}$ define $\hat{\mathbf{X}}^{\dagger}(\boldsymbol{\delta}) = (\hat{\mathbf{E}}_1^{\dagger} + \boldsymbol{\delta}, \hat{\mathbf{E}}_2^{\dagger})$.
3. Use the CMC algorithm to calculate the related p-value $\hat{\lambda}_T(\hat{\mathbf{X}}^{\dagger}(\boldsymbol{\delta}))$.
4. Independently repeat Steps 2 and 3 a large number, e.g. R times, giving R estimates of empirical p-values, say $\{\hat{\lambda}_T(\hat{\mathbf{X}}_r^{\dagger}(\boldsymbol{\delta})), r = 1, \ldots, R\}$.
5. The empirical post-hoc conditional power is given by

$$\hat{W}[(\boldsymbol{\delta}; \hat{\boldsymbol{\delta}}, n, \alpha, T)|\mathcal{X}] = \frac{\sum_{r=1}^{R} \mathbb{I}[\hat{\lambda}_T(\hat{\mathbf{X}}_r^{\dagger}(\boldsymbol{\delta})) \leq \alpha]}{R}.$$

6. Repeat Steps 2-5 for different values of $\boldsymbol{\delta}$ to attain a function in $\boldsymbol{\delta}$.

4 Some applications of empirical conditional power function

In general, the power of a particular test is affected by many factors (Kraemer and Thiemann, 1987; Lipsey, 1990; Hallahan and Rosenthal, 1996). In respect of a two-sample design, under mild regularity conditions, the main three factors are:

1. Sample size, n. Everything else being fixed, if test statistic T is consistent, the greater the sample size, the greater the power.
2. Significance level, α. Everything else being fixed, the greater the significance level, the greater the power.
3. (Standardized) effect size, $\Delta = \boldsymbol{\delta}/\sigma$. It is easier to expose a large effect than to expose a small effect; that is, the greater the effect size, the greater the power.

Power analysis is discussed in different fields of studies. Cohen (1988) studied power analysis for the behavioural sciences; he provided power tables for various common parametric statistical tests that can be consulted to determine the sample size for specified values of α, Δ and power. Moher et al (1994) studied power analysis in clinical trials and Markowski and Markowski (1999) studied power analysis in business researches.

For most common statistical tests, power is easily calculated from tables. For example, see Cohen (1988) for some parametric tests and Randles and Wolfe (1979) for some one- and two-sample nonparametric tests. Owen (1965) provided power tables for various tests which use the student t-distribution. Moreover, statistical computer software (e.g. **R**, **SPSS, etc.**) are used to evaluate the statistical power. For more complex tests, and for most nonparametric tests, ready tables are often not available and not easily expressed. In these cases, Monte Carlo simulations can be used to approximate the power. For example, Collings and Hamilton (1988) proposed a bootstrap method which does not require any knowledge of the underlying distribution P for estimating the power of the two-sample Wilcoxon test. See also Epstein (1955), Teichroew (1955) and Hemelrijk (1961). However, some authors derived the power func-

tions and/or tables but only in limited cases. For example, see Dixon (1954), Barton (1957), Bell et al (1966), Haynam and Govindarajulu (1966) and Milton (1970).

In the following sections, some applications of empirical conditional power function of permutation tests are investigated. In particular, the use of empirical conditional power for sample size estimation, reproducibility probability, generalizability probability, and sample size adjustment are investigated.

4.1 Sample size calculation

Sample size calculation is an important and often difficult step in planning a research design. Large sample size may waste time, resources and money, while small sample size may procure inaccurate results. There are different approaches for sample size calculation including confidence interval approach (McHugh, 1961) and Bayesian approach (Wang et al, 2005). One of the most popular approaches involves studying the power of a test (See for example Aguirre-Urreta and Rönkkö, 2015; Akobeng, 2016; Giraudeau et al, 2016). In our context, the empirical conditional power function of permutation test is used as an important tool for estimating a suitable and proper sample size for a particular study.

Consider the two-sample design in which $\mathbf{X}_1 = \{X_{11}, \ldots, X_{1n_1}\}$ are iid in $P(x + \Delta)$ and $\mathbf{X}_2 = \{X_{21}, \ldots, X_{2n_2}\}$ are iid in $P(x)$ and the two samples are independent of one another. Let $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$ be the hypotheses of interest. Under the normality assumption of the study distribution, the power of the test is evaluated as

$$1 - \beta = 1 - \Phi\left(z_\alpha - \Delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\right), \tag{4}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and z_α is the upper α critical value. It is worthwhile to observe that the power is monotonic non-decreasing in n_1 and/or n_2. Moreover, for fixed total sample size, the highest power is attained when $n_1 = n_2$.

For a preassigned level of significance α, the sample size required to expose an effect size Δ with a desired level of power $1 - \beta$ can be calculated from Equation 4 (See for example Chow and Liu, 2004, pages 445-451). Let $n_1 = \rho n$, where $0 < \rho < 1$ and $n = n_1 + n_2$, then

$$n = \frac{1}{\rho(1-\rho)}\left(\frac{z_\beta + z_\alpha}{\Delta}\right)^2. \tag{5}$$

See also Chow et al (2002) for sample size determination based on non-central t-distribution.

Noether (1987) studied sample size determination for some nonparametric tests. For the two-sample Wilcoxon test, the total sample size is given by

$$n = \frac{1}{12\rho(1-\rho)}\left(\frac{z_\beta + z_\alpha}{\Delta_{Noether} - 0.5}\right)^2, \tag{6}$$

where $\Delta_{Noether} = Pr(\mathbf{X}_1 > \mathbf{X}_2)$ is Noether's effect size. One possible way of estimating $\Delta_{Noether}$ is

$$\hat{\Delta}_{Noether} = \frac{4U}{n^2},$$

where U is the Mann-Whitney statistic. Simonoff et al (1986) showed that the maximum likelihood estimator of $\Delta_{Noether}$ is given by

$$\hat{\Delta}_{Noether} = \Phi\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2_{\mathbf{X}_1} + S^2_{\mathbf{X}_2}}}\right),$$

where $\bar{X}_1$ and $S^2_{\mathbf{X}_1}$ are the mean and variance of the first dataset $\mathbf{X}_1$ and $\bar{X}_2$ and $S^2_{\mathbf{X}_2}$ are the corresponding quantities for the second dataset $\mathbf{X}_2$. Hamilton and Collings (1991) used the results of Collings and Hamilton (1988) to suggest a procedure to determine sample size of the two-sample Wilcoxon test.

Within the permutation framework, De Martini (2002) studied the use of the estimated unconditional power of permutation tests for sample size estimation. In this section, the sample size is estimated by the use of conditional power function of permutation tests.

For a preassigned level of significance α, the sample size required to expose an effect size Δ with a desired level of power $\tilde{W} \in (\alpha, 1)$ can be obtained by solving

$$n = \arg\min_n \{\hat{W}[(\Delta; n, \alpha, T)|\mathcal{X}] = \tilde{W}\}.$$

Since it is generally not possible to write the conditional power function in closed form, the sample size cannot be exactly determined. Therefore, simulation study is considered to estimate it. Algorithm 4 is used for sample size estimation to expose an effect size Δ with a desired power $\tilde{W}$.

Algorithm 4 Sample Size Estimation

1. Start with a pilot sample of size $n = n_1 + n_2$; where n_1 to be drawn from the treatment population and n_2 from the control population, without assuming the knowledge of their distributions.
2. Calculate the empirical conditional power $\hat{W}$.
3. Adjust the sample size n to achieve desirable empirical conditional power $\tilde{W}$.
4. To obtain a function in n, Steps 1 and 2 are repeated for different values of n.

The required sample size n for exposing the effect size Δ with a desired power that is equal to the power at a given effect size $\tilde{\Delta}$ with a total sample size $\tilde{n}$ is derived in the following way.

$$\hat{W}[(\Delta; n, \alpha, T)|\mathcal{X}] = \tilde{W}[(\tilde{\Delta}; \tilde{n}, \alpha, T)|\mathcal{X}]$$

if and only if

$$\Delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \tilde{\Delta}\sqrt{\frac{\tilde{n}_1 \tilde{n}_2}{\tilde{n}_1 + \tilde{n}_2}}.$$

Let $n_1 = \rho n$ $(0 < \rho < 1)$ and $\tilde{n}_1 = \tilde{\rho}\tilde{n}$ $(0 < \tilde{\rho} < 1)$, then

$$n = \frac{\tilde{\rho}(1-\tilde{\rho})\tilde{n}}{\rho(1-\rho)} \left(\frac{\tilde{\Delta}}{\Delta}\right)^2. \tag{7}$$

It is worthwhile to observe that this equality is asymptotically true and approximation is good for relatively small sample sizes. This approximation is mainly due to differences on supports for the involved permutation distributions.

4.2 Reproducibility probability

Suppose that one study has been carried out and the result of the test is significant. One may ask this question: What is the probability that a second study (using the same study population) will also generate a significant result? In other words, what is the probability that the result of the first study is reproducible? Statistically, if the two studies are independent, the probability of having a significant result in the second study is given by the power of the test, irrespective of the result of the first study whether was significant or not. However, such information from the first study may be useful in predicting the result of the second study. This results in getting the notion of reproducibility probability, which is different from the unconditional power of the test.

Shao and Chow (2002) defined the reproducibility probability as a person's subjective probability of observing a significant result from a future study, when significant results from one or several previous studies are observed. Goodman (1992) defined the reproducibility probability as an estimated power of the future study using the data from the previous study. With other terms, the reproducibility probability is defined as the power evaluated at $\Delta = \hat{\Delta}_0$, where $\hat{\Delta}_0$ is the estimated effect size of the first study (or a previous study).

Within the permutation framework, Pesarin and Salmaso (2010) defined the reproducibility probability or the *actual* post-hoc conditional power as the power with Δ replaced by its estimate $\hat{\Delta}$ obtained before randomization, in our notation it is denoted by $\hat{W}[(\hat{\Delta}; \hat{\Delta}, n, \alpha, T)|\mathcal{X}]$. Onwuegbuzie and Leech (2004) and Lenth (2007) pointed out that such reproducibility probability can provide useful information for replication studies. Brewer and Sindelar (1988) argued that this is merely a rephrasing of the a priori problem, namely, " *What would the power be if I used my α, n and post-hoc (observed) effect size $\hat{\Delta}$?*"'. It is worthwhile to observe that the outcome (significance or non-significance) of a test using suitable sample size definitely will not affect the power, α, and chosen effect size. These in genaral are statistical test concepts and are not devoted or related to a single study. Moreover, p-value and reproducibility probability are not equivalent notions in the sense that the later implies re-randomization whereas the former does not. However, they are quite closely related (Thomas, 1997; Levine and Ensom, 2001; Onwuegbuzie and Leech, 2004). The most recent work on reproducibility probability for nonparametric tests is done by De Capitani and De Martini (2016).

4.3 Generalizability probability

As discussed in the previous section, the reproducibility concept is used to evaluate whether the obtained results are reproducible from study to study of **one specific population**. One may be concerned in evaluating how likely the obtained results of a single study from a particular population can be reproduced to a *different but similar* population, where similar means that they are provided with about the same standardized distribution. For example, in clinical development (See Shao and Chow, 2002), after the investigational drug product has been shown to be effective and safe with respect to a target patient population (e.g. adults), it is often of interest to study a similar but different patient population (e.g. elderly patients with the same disease under study or a patient population with different ethnic factors) to see how likely the clinical result is reproducible in the different population. This information can be useful in regulatory submission for supplement new drug application (for example, when generalizing the clinical results from adults to elderly patients) and regulatory evaluation for bridging studies (for example, when generalizing clinical results from a European to an Asian patient population). For this purpose, the concept of generalizability probability is proposed. It is simply the reproducibility probability in a different population.

Let A and B be two *different but similar* populations. In population A, the effect size is given by $\Delta = (\mu_1 - \mu_2)/\sigma$. Suppose now, in population B, the difference in means is $\mu_1 - \mu_2 + \eta$ and the population variance is $C^2\sigma^2$, so the new effect size is given by

$$\frac{\mu_1 - \mu_2 + \eta}{C\sigma} = \frac{D(\mu_1 - \mu_2)}{\sigma},$$

where

$$D = \frac{1 + \eta/(\mu_1 - \mu_2)}{C}$$

is a measure of change in the effect size for the population difference. In practice, $\eta < (\mu_1 - \mu_2)$ and, thus, $D > 0$.

If the power of the current study (under population A) is $\hat{W}[(\Delta; n, \alpha, T)|\mathcal{X}]$, then the power of the future study (under population B) is $\hat{W}[(D\Delta; n, \alpha, T)|\mathcal{X}]$. If D is known, then the generalizability probability is just the reproducibility probability $\hat{W}[(D\hat{\Delta}; \hat{\Delta}, n, \alpha, T)|\mathcal{X}]$, that works as a least square estimate. If the true value of D is unknown, a set of D-values may be considered.

4.4 Sample size adjustment

If the sample size of a previous study was calculated based on conditional power function with a priori effect size $\tilde{\Delta}$ and preassigned level of significance α, then it is reasonable to adjust the sample size for the current study using the information of the previous study. The concept of reproducibility probability is useful in providing important information for adjusting the sample size. If

the reproducibility probability is smaller than the desired power level of the current study, then sample size should be increased. Otherwise, the sample size may be decreased to avoid wasting resources.

The sample size $\tilde{n}$ can be adjusted to n according to the reproducibility probability as follows. The reproducibility probability is set to be equal to the a priori power $\tilde{W}$ which is evaluated at a virtual effect size $\tilde{\Delta}$ with total sample size $\tilde{n}$, then the new sample size n is derived. To this end,

$$\hat{W}[(\hat{\Delta};\hat{\Delta},n,\alpha,T)|\mathcal{X}] = \tilde{W}[(\tilde{\Delta};\hat{\Delta},\tilde{n},\alpha,T)|\mathcal{X}]$$

if and only if

$$\hat{\Delta}\sqrt{\frac{n_1 n_2}{n}} = \tilde{\Delta}\sqrt{\frac{\tilde{n}_1\tilde{n}_2}{\tilde{n}}}.$$

Let $n_1 = \rho n$, $0 < \rho < 1$ (one may consider $\rho = \tilde{\rho} = \tilde{n}_1/\tilde{n}$), then

$$n = \tilde{n}\left(\frac{\tilde{\Delta}}{\hat{\Delta}}\right)^2. \tag{8}$$

Also, the sample size can be adjusted using the generalizability probability. The new total sample size n to be drawn from the new population is derived as follows. The generalizability probability is set to be equal to the a priori power $\tilde{W}$ which is evaluated from the first population at a virtual effect size $\tilde{\Delta}$ with total sample size $\tilde{n}$, then the new sample size n to be drawn from the second population is derived.

$$\hat{W}[(D\hat{\Delta};\hat{\Delta},n,\alpha,T)|\mathcal{X}] = \tilde{W}[(\tilde{\Delta};\hat{\Delta},\tilde{n},\alpha,T)|\mathcal{X}]$$

if and only if

$$D\hat{\Delta}\sqrt{\frac{n_1 n_2}{n}} = \tilde{\Delta}\sqrt{\frac{\tilde{n}_1\tilde{n}_2}{\tilde{n}}}$$

Let $n_1 = \rho n$, $0 < \rho < 1$ (one may consider $\rho = \tilde{\rho} = \tilde{n}_1/\tilde{n}$), then

$$n = \tilde{n}\left(\frac{\tilde{\Delta}}{D\hat{\Delta}}\right)^2. \tag{9}$$

5 Illustration example: degree of reading power

In his Ph.D thesis, Schmitt (1987) was interested in testing whether directed reading activities in the classroom help elementary school students improve aspects of their reading ability. A treatment class of 21 third-grade students participated in these activities for eight weeks, and a control class of 23 third-graders followed the same curriculum without the activities. After the eight-week period, students in both classes took a Degree of Reading Power (DRP) test which measures the aspect of reading ability that the treatment is designed to improve. The DRP scores are reported in Table 1.

Table 1 Degree of reading power scores for third-graders.

Treatment Group, $\mathbf{X}_t$						Control Group, $\mathbf{X}_c$					
24	43	58	71	61	44	42	43	55	26	33	41
67	49	59	52	62	54	19	54	46	10	17	60
46	43	57	43	57	56	37	42	55	28	62	53
53	49	33				37	42	20	48	85	

For testing $H_0 : \Delta = \mu_t - \mu_c = 0$ versus $H_1 : \Delta = \mu_t - \mu_c > 0$, Algorithm 1 is used. The difference between the sample means is considered as a test statistic. The observed test statistic is $T^o = \bar{X}_1 - \bar{X}_2 = 9.954$ and the conditional p-value estimate, based on $B = 5\,000$ CMC replicates, is $\hat{\lambda} = 0.015$. At $\alpha = 0.05$ the null hypothesis is rejected.

Figure 1 shows the permutation distribution of T^*.

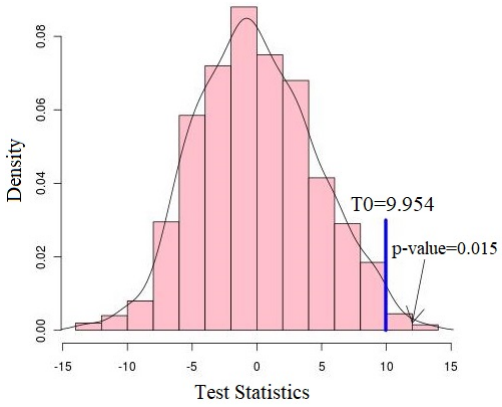

Fig. 1 DRP data: The permutation distribution.

Figure 1 shows that the distribution of T^* has almost normal shape. Since the permutation distribution approximates the unconditional sampling distribution of T, which is roughly normal. Therefore, the usual two-sample t-test can safely be applied provided that data are assumed homoschedastic both in H_0 and in H_1, i.e. by assuming that treatment does not influence the data variability. Using the usual t-test, the p-value is 0.013, which is very close to the p-value obtained using the permutation test.

Assuming the underlying distribution is normal, the estimated unconditional (parametric) power function can be obtained as follows.

$$\hat{W}(\boldsymbol{\delta}; n, \alpha, T, P) = 1 - F_t(t_{df}^{1-\alpha}, df, ncp), \tag{10}$$

where F_t is the cumulative distribution of the student t, $df = n_1 + n_2 - 2$ is the degrees of freedom, $t_{df}^{1-\alpha}$ is the $1-\alpha$ quantile of a student t-distribution with degrees of freedom df and non-centrality parameter $ncp = \delta \left(S_p^2 (\frac{1}{n_1} + \frac{1}{n_2}) \right)^{-1/2}$,

$S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2+\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2}$ is the pooled variance. Note that $\hat{W}(\cdot)$ is an estimate of $W(\cdot)$ since the *ncp* is a data dependent quantity.

Figure 2 shows the empirical post-hoc conditional power function together with the unconditional (parametric) power function.

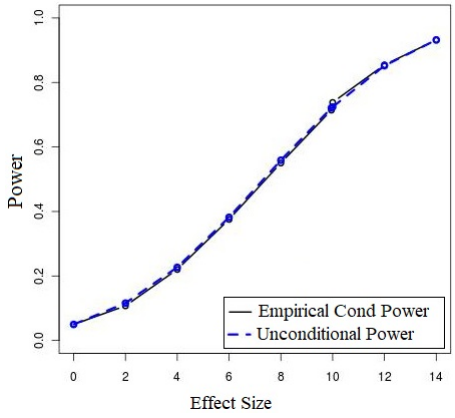

Fig. 2 DRP data: The parametric unconditional power and the empirical post-hoc conditional power functions.

Sample size calculation Algorithm 4 is used to calculate the required sample sizes to detect an effect size $\boldsymbol{\delta} = \mu_t - \mu_c = 14$. The results are reported in Table 2. For example, if the desired power is $\tilde{W} = 0.90$, one may consider $n_1 = 13$ and $n_2 = 7$.

Table 3 reports the (parametric) unconditional power calculated using Equation 10 as a function with the sample sizes. It is clear that balanced designs are more powerful than unbalanced. For example, consider the total sample size $n = 20$, then the highest power does occur when $n_1 = n_2 = 10$. Moreover, the power when $n_1 > n_2$ is higher than the power when $n_1 < n_2$, this is due to the sample variances; the sample variance of the treatment group is less than the sample variance of the control group.

Now, given the information reported in Table 2 or 3, the sample sizes to expose an effect size $\boldsymbol{\delta} = 10$ can be calculated using Equation 7. Assuming $\rho = \tilde{\rho} = 0.5$ and $\tilde{n} = 20$, then $n = 39.2 \approx 40$. Hence, $n_1 = 20$ and $n_2 = 20$.

Table 2 DRP Example: Empirical conditional power and sample sizes, $\boldsymbol{\delta} = 14$.

		n_2					
		5	7	10	13	16	20
n_1	5	0.54	0.71	0.80	0.69	0.61	0.48
	7	0.68	0.84	0.90	0.79	0.74	0.61
	10	0.78	0.91	0.96	0.89	0.87	0.75
	13	0.78	0.90	0.95	0.92	0.91	0.84
	16	0.78	0.90	0.95	0.94	0.93	0.88
	20	0.86	0.96	0.99	0.98	0.98	0.95

Table 3 DRP Example: parametric unconditional power and sample sizes, $\boldsymbol{\delta} = 14$.

		n_2					
		5	7	10	13	16	20
n_1	5	0.68	0.77	0.87	0.68	0.61	0.53
	7	0.80	0.87	0.94	0.81	0.74	0.65
	10	0.89	0.94	0.98	0.90	0.85	0.78
	13	0.87	0.93	0.98	0.93	0.90	0.84
	16	0.84	0.92	0.98	0.94	0.92	0.89
	20	0.91	0.96	0.99	0.98	0.97	0.94

Reproducibility probability According to Table 2 or 3, the required sample sizes to expose the virtual effect size $\tilde{\boldsymbol{\delta}} = 14$ at level of significance $\alpha = 0.05$ with a desired level of power $\tilde{W} = 0.85$ are $n_1 = 7$ and $n_2 = 7$. Recall the observed effect size is $\hat{\boldsymbol{\delta}} = 9.954$ or equivalently $\hat{\Delta} = \hat{\boldsymbol{\delta}}/S_p \approx 0.68$ based on sample sizes $n_1 = 21$ and $n_2 = 23$. Therefore, the reproducibility probability is given by $\hat{W}[(\hat{\Delta}; \hat{\Delta}, n, \alpha, T)|\mathcal{X}] = 0.722$ (see Figure 2). That is, with a new independent random experiment the probability is as high as 72.2% to obtain a significant result with same sample sizes, test T, at effect size $\delta = \hat{\delta} = 9.954$ and $\alpha = 0.05$.

Sample size adjustment Hence, in order to have a reproducibility probability equals to 0.85, one may adjust the sample size using Equation 8. Let $\tilde{\boldsymbol{\delta}} = 14$, $\tilde{n} = 14$ and $\hat{\boldsymbol{\delta}} = 9.954$, then $n = 27.6942 \approx 28$ and hence $n_1 = n_2 = 14$. That is, in order to expose an effect size 9.954 with a desired reproducibility probability of 0.85, the sample sizes should be $n_1 = n_2 = 14$.

6 Concluding Remarks

In this paper, two-sample permutation test procedure is discussed and the use of conditional Monte Carlo algorithm for evaluating the permutation (conditional) p-value is used. Then, the notion of conditional power function of permutation test is reviewed. The following are some applications of the empirical conditional power function that were discussed:

- Sample size estimation. A pilot sample with a reasonable size is drawn from the population of interest, without assuming the knowledge of its distribution, and then the empirical conditional power is calculated. The size is to be increased (or may be reduced) till a desired power is achieved. It is shown that two-sample balanced design is more powerful than unbalanced.
- Reproducibility probability. It is an important tool for sample size adjustment, and is used to measure the reliability of the test.
- Generalizability probability, which is also used for sample size adjustment.

Acknowledgements The authors would like to thank the associated editor and referees for their comments that contribute in improving the paper. We also greatly appreciate Dr. Ibrahim Almasri, Department of Applied Mathematics and Physics - Palestine Polytechnic University, for being kind enough to read and improve the language of the paper.

References

Aguirre-Urreta M, Rönkkö M (2015) Sample size determination and statistical power analysis in pls using r: an annotated tutorial. Communications of the Association for Information Systems 36(3):33–51

Akobeng AK (2016) Understanding type I and type II errors, statistical power and sample size. Acta Paediatrica 105(6):605–609

Amro L, Pauly M (2017) Permuting incomplete paired data: a novel exact and asymptotic correct randomization test. Journal of Statistical Computation and Simulation 87(6):1148–1159

Barton DE (1957) A comparison of two sorts of test for a change of location applicable to truncated data. Journal of the Royal Statistical Society 19:119–124

Basso D, Pesarin F, Salmaso L, Solari A (2009) Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications in R. Springer, New York

Bell CB, Moser JM, Thompson R (1966) Goodness criteria for two-sample distribution-free tests. The Annals of Mathematical Statistics 37:133–142

Brewer JK, Sindelar PT (1988) Adequate sample size: A priori and post hoc considerations. The Journal of Special Education 21:74–84

Chow SC, Liu JP (2004) Design and Analysis of Clinical Trials: Concepts and Methodologies, 2nd Edition. Wiley-Blackwell, New York

Chow SC, Shao J, Wang H (2002) A note on sample size calculation for mean comparisons based on non-central t-statistics. Journal of Biopharmaceutical Statistics 12:441–456

Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey

Collings BJ, Hamilton MA (1988) Estimating the power of the two-sample Wilcoxon test for location shift. Biometrics 44:847–860

Cooper H, Hedges LV (1997) The Handbook of Research Synthesis. Russell Sage Foundation, New York

De Capitani L, De Martini D (2016) Reproducibility probability estimation and RP-testing for some nonparametric tests. Entropy 18(4):142
De Martini D (2002) Pointwise estimate of the power and sample size determination for permutation tests. Statistica 62:779–790
Dixon WJ (1954) Power under normality of several nonparametric tests. The Annals of Mathematical Statistics 25:610–614
Edgington ES (1995) Randomization Tests, 3rd Edition. Marcel Dekker, New York
Epstein B (1955) Comparison of some non-parametric tests against normal alternatives with an application to life testing. Journal of the American Statistical Association 50:894–900
Fisher RA (1934) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh
Fisher RA (1935) The Design of Experiments. Oliver and Boyd, Edinburgh
Giraudeau B, Higgins J, Tavernier E, Trinquart L (2016) Sample size calculation for meta-epidemiological studies. Statistics in medicine 35(2):239–250
Good P (2005) Permutation, Parametric and Bootstrap Tests of Hypotheses, 3rd Edition. Springer-Verlag, New York
Goodman S (1992) A comment on replication, p-values and evidence. Statistics in Medicine 11:875–879
Hallahan M, Rosenthal R (1996) Statistical power: Concepts, procedures, and applications. Behaviour Research and Therapy 34:489–499
Hamilton MA, Collings BJ (1991) Determining the appropriate sample size for nonparametric tests for location shift. Technometrics 33:327–337
Haynam GE, Govindarajulu Z (1966) Exact power of the Mann-Whitney test for exponential and rectangular alternatives. The Annals of Mathematical Statistics 37:945–953
Hedges LV, Olkin I (1985) Statistical Methods for Meta-Analysis. Academic Press, New York
Hemelrijk J (1961) Experimental comparison of Student's and Wilcoxon's two sample test. Quantitative Methods in Pharmacology pp 118–133
Hoeffding W (1952) The large-sample power of tests based on permutations of observations. The Annals of Mathematical Statistics 23:169–192
Kraemer HC, Thiemann S (1987) How Many Subjects? Statistical Power Analysis in Research. Sage Publications, Newbury Park, CA
Lehmann EL, Romano JP (2005) Testing statistical hypotheses, 3rd edn. Springer, New York
Lehmann EL, Stein C (1949) On the theory of some non-parametric hypotheses. Annals of Mathematical Statistics 20:28–45
Lenth RV (2007) Post hoc power: Tables and commentary. Tech. Rep. 378, The University of Iowa - Department of Statistics and Actuarial Science
Levine M, Ensom MHH (2001) Post hoc power analysis: An idea whose time has passed? Pharmacotherapy 21:405–409
Lipsey MW (1990) Design Sensitivity: Statistical Power for Experimental Research. Sage Publications, Newbury Park, CA

Markowski EP, Markowski CA (1999) Practical uses of statistical power in business research studies. Journal of Education for Business 75:122–125

McDonald J, Gerard PD, McMahan CS, Schucany WR (2016) Exact-permutation-based sign tests for clustered binary data via weighted and unweighted test statistics. Journal of Agricultural, Biological and Environmental Statistics 21(4):698–712

McHugh RB (1961) Confidence interval inference and sample size determination. The American Statistician 15:14–17

Milton RC (1970) Rank Order Probabilities: Two-Sample Normal Shift Alternatives. John Wiley & Sons Inc, New York

Moher D, Dulberg CS, Wells GA (1994) Statistical power, sample size, and their reporting in randomized controlled trials. Journal of the American Medical Association 272:122–124

Noether GE (1987) Sample size determination for some common nonparametric tests. Journal of the American Statistical Association 82:645–647

Onwuegbuzie AJ, Leech NL (2004) Post hoc power: A concept whose time has come. Understanding Statistics 3:201–230

Owen DB (1965) The power of Student's t-test. Journal of the American Statistical Association 60:320–333

Pesarin F (2001) Multivariate Permutation Tests: With Application in Biostatistics. John Wiley & Sons, Ltd., Chichester

Pesarin F, Salmaso L (2010) Permutation Tests for Complex Data: Theory, Application and Software. John Wiley & Sons, Ltd., Chichester

Pitman EJG (1937a) Significance tests which may be applied to samples from any population. Journal of the Royal Statistical Society Series B, 4:119–130

Pitman EJG (1937b) Significance tests which may be applied to samples from any population. II. the correlation coefficient test. Journal of the Royal Statistical Society Series B, 4:225–232

Pitman EJG (1938) Significance tests which may be applied to samples from any population. III. the analysis of variance test. Biometrika 29:322–335

Randles RH, Wolfe DA (1979) Introduction to the Theory of Nonparametric Statistics. John Wiley & Sons, New York

Salmaso L (2003) Synchronized permutation tests in 2^k factorial designs. Communication in Statistics - Theory and Methods 32:1419–1437

Samonenko I, Robinson J (2015) A new permutation test statistic for complete block designs. The Annals of Statistics 43(1):90–101

Schmitt MC (1987) The effects on an elaborated directed reading activity on the metacomprehension skills of third graders. PhD thesis, Purdue University

Shao J, Chow SC (2002) Reproducibility probability in clinical trials. Statistics in Medicine 21:1727–1742

Shorack GR, Wellner JA (1986) Empirical Processes with Applications to Statistics. Wiley Series in Probability & Mathematical Statistics, New York

Simonoff JS, Hochberg Y, Reiser B (1986) Alternative estimation procedures for $P_r(X < Y)$ in categorized data. Biometrics 42:895–907

Teichroew D (1955) Empirical power functions for nonparametric two-sample tests for small samples. The Annals of Mathematical Statistics 26:340–344

Thomas L (1997) Retrospective power analysis. Conservation Biology 11:276–280

Wang H, Chow SC, Chen M (2005) A Bayesian approach on sample size calculation for comparing means. Journal of Biopharmaceutical Statistics 15:799–807