UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIV

Some Advances in Permutation Testing

Direttore della Scuola: Ch.ma Prof.ssa ALESSANDRA SALVAN

Supervisore: Ch.mo Prof. FORTUNATO PESARIN

Co-supervisore: Ch.mo Prof. FRIEDRICH LEISCH

Dottorando: MONJED H. M. SAMUH

December 12, 2011

"The difference between a successful person and others is not a lack of strength, not a lack of knowledge, but rather in a lack of will"

Vincent T. Lombardi

To my family ...

Acknowledgements

During the period of my study, it has been my good fortune to encounter many people who have given me more of their time, companionship, professional and personal help.

I would first of all like to express my deepest gratitude to my supervisors, Prof. **Fortunato Pesarin** (Padova University, Italy) and Prof. **Friedrich Leisch** (University of Natural Resources and Life Sciences, Austria). Their encouragement, supervision and support enabled me to grow up as a Ph.D for independently carrying out research.

I am grateful to thank Prof. **Siegfried Kropf** (Magdeburg University, Germany) and Prof. **Dieter Rasch** for being the external reviewers. Their comments and suggestions were very constructive for improving this thesis.

I would like to gratefully acknowledge the director of the Ph.D school, Prof. **Alessandra Salvan**, for the perfect organization of the doctoral program and for the readiness shown during my study.

I am greatly indebted to *Cassa di Risparmio di Padova e Rovigo* (**CARIPARO**) foundation who funded my Ph.D study.

I cannot forget in this acknowledgement my colleagues and friends of the Ph.D program, Antonio Canale, Riccardo De Bin, Marlies Ranieri, Nicola Lunardon, Davide Risso and Francesca Solmi, for their warm friendship during my stay in Italy. Many thanks to my friends in Copernico ESU residence. It was fun having you around.

I would like to thank Prof. **Mohammad Fraiwan Al-Saleh** (Yarmouk University, Jordan), Dr. **Abdulhakeem Eideh** (Al-Quds University, Palestine) and my colleagues at Palestine Polytechnic University, in particular, the president of the university, Dr. **Ibrahim Al-Masri**, for their encouragement.

I don't think the words are enough to express my gratitude to my family. Without their encouragement, I would not have a chance to continue my study at Padova University.

Monjed H. Samuh

Padova
December 12, 2011

Abstract

The main objective of this Ph.D thesis is to provide some advances in permutation testing within different fields of statistics. Mainly, the thesis is divided into four parts.

First, the two notions of power function of permutation tests (conditional and unconditional) are reviewed. The use of empirical conditional power function for sample size estimation is investigated. Then, the notions of reproducibility probability and generalizability probability are defined within the permutation framework. It is shown that the reproducibility and generalizability probabilities are important tools for sample size adjustment.

Second, permutation tests with ranked set sampling are investigated. The effectiveness of ranked set sampling on the power of permutation tests is studied. Two-sample permutation test is considered as a guide. The power of the two-sample permutation test is computed for ranked set and simple random samples. It is shown that the test for ranked set sample is more powerful than for simple random sample. Moreover, the effectiveness of the set size and number of cycles of ranked set sample is studied. It is shown that the power increased by the set size and/or the number of cycles. In addition, two test statistics are proposed for ranked set sample and investigated under different kind of distributions (symmetric and asymmetric).

Third, permutation tests in linear mixed model are investigated. Some tests for a zero random effect variance component are reviewed and a new permutation test is proposed. Random intercept model is considered as a guide. The proposed permutation test has the correct nominal level of significance and is more powerful than the usual tests based on a mixture of χ^2 distributions. Moreover, the proposed permutation test is the fastest, according to computing time, approach among those resampling-based test approaches.

Finally, permutation tests in cluster analysis is investigated. Tests for random agreement between two sets of clusters of a dataset are discussed. The adjusted Rand index is proposed as a test statistic. Two testing methods are proposed. The first method is based on the χ^2 distribution assuming the cluster sizes within each set of clusters are equal. The second method is based on the permutation approach. Comparison between these proposed methods is carried out in terms of empirical level of significance.

Riassunto

L'obiettivo principale di questa tesi di Dottorato è di conseguire alcuni sviluppi nell'analisi di permutazione nell'ambito di diversi campi della statistica. La tesi è suddivisa in quattro parti.

La prima parte prende in esame due nozioni relative alla potenza del test di permutazione (condizionata e incondizionata). E' stato anche indagato l'uso della potenza empirica condizionata per la valutazione della dimensione del campione. Quindi, vengono definite all'interno dell'approccio di permutazione, le nozioni di probabilità di riproducibilità e di probabilità di generalizzabilità. Viene mostrato che le probabilità di riproducibilità e generalizzabilità sono strumenti importanti nell'aggiornamento della dimensione del campione.

Nella seconda parte vengono studiati test di permutazione nel ranked set sampling. Quindi viene anche studiato l'effetto di questo tipo di campionamento sulla potenza dei test. Un test di permutazione per due campioni è stato preso come guida. L'efficienza del test di permutazione per due campioni viene calcolata per il ranked set sampling e quello casuale semplice. Viene anche esaminata l'efficienza relativa del ranked set sampling rispetto al campione casuale semplice nella condizione di uguaglianza delle numerosità campionarie effettivamente osservate. Viene inoltre esaminata l'efficienza rispetto alla dimensione delle unità e il numero dei cicli del ranked set sampling; ne risulta che l'efficienza aumenta a seconda del set size e/o il numero dei cicli. Inoltre, vengono proposti due test statistici di permutazione esaminati sotto diversi tipi di distribuzione degli errori (simmetrica e asimmetrica).

Nella terza parte, vengono esaminati test sul modello lineare misto. Viene in particolare proposto un test di permutazione per l'ipotesi nulla che la componente di varianza sia pari a zero contro l'alternativa che sia positiva. Fa da guida il modello dell'intercetta casuale. Il test di permutazione proposto ha il corretto livello di significatività ed è più efficiente dei test basati sulla mistura di distribuzioni χ^2. Inoltre, il test proposto è anche l'approccio più veloce in termini di tempi di calcolo fra quelli basati sul ricampionamento.

Infine, vengono esaminati test di aggregazione casuale fra due gruppi cluster di un medesimo set di dati. L'adjusted Rand index viene adottato come test statistico. Vengono proposti due metodi di analisi. Il primo è basato sulla distribuzione χ^2 tramite l'uso della relazione tra la statistica di Pearson e l'adjusted Rand index. Il secondo è basato sull'approccio permutazionale. Il confronto tra i due metodi proposti è svolto in termini di livello empirico di significatività.

Contents

Chapter 1

Introduction

Contents

1.1 Overview

Traditional parametric tests such as t-tests and F-tests are not always robust to violation of its assumptions of normally distributed errors, homoscedasticity and random sampling from a target population. However, the normality assumption may not always be reasonable. In the analysis of univariate data, often someone try avoiding the problem of non-normal data by finding suitable transformations while maintaining the homoscedasticity assumption in the null hypothesis. Note that this assumption is not generally attained if the monotonic transformations are not linear (Box and Tiao, 1964; Posten, 1978; Rasch and Guiard, 2004). An alternative approach is to use permutation tests, where errors are not assumed to be normally distributed and/or homoscedastic in the alternative, while maintaining dominance in distribution.

The use of permutation tests has received renewed attention in recent years with the advent of much faster and more accessible computer power. In general, for an exact test by permutation, the reference distribution of a relevant test statistic under the null hypothesis is constructed by calculating its value for all possible rearrangements (permutations) of the observations (or by a large random samples of such rearrangements). A p-value is then calculated as the proportion of the values of the statistic obtained under permutation that are equal to or more extreme than the observed value.

All simple and many relatively complex parametric tests have a corresponding permutation test version that is defined by using the same test statistic as the parametric test, but obtains the p-value from the sample-specific permutation distribution of that statistic, rather than from the theoretical distribution derived from the parametric assumption. Fisher (1934, 1935) introduced the permutation test as the exact test for the association between two binary variables when the expected number of cells is less than 5; that is, when the chi-square test fails. Also it is useful for one sided testing if at least one variable is ordered categorical. In addition, he introduced the exact test for testing differences between means of two populations

when the assumptions of the two-sample t-test were not met. Pitman (1937a,b, 1938) developed exact permutation methods consistent with the Neyman-Pearson approach 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 are used in different fields of statistics. For examples, Sun and Sherman (1996) used permutation tests in survival analysis, Mehta and Patel (1997) used permutation tests in categorical data analysis, Anderson and Robinson (2001) used permutation tests for linear models, and Fitzmaurice et al. (2007) used permutation tests for generalized linear mixed models. In this thesis, empirical conditional power analysis of permutation tests is investigated. Permutation tests are studied in ranked set sampling, linear mixed model and cluster analysis. New tests are proposed and compared with some available parametric and nonparametric tests.

1.2 Main Contributions of the Thesis

The main contributions of this Ph.D thesis are:

- In accordance with Goodman (1992), Shao and Chow (2002) and De Martini (2008) the notions of reproducibility probability and generalizability probability are defined within the permutation framework and their use for sample size adjustment is addressed. Moreover, the use of empirical conditional power approach for sample size estimation is studied.

- Ranked set sampling (RSS) is a sampling scheme which can successfully replace simple random sampling (SRS) in experimental settings where measuring the units of interest is difficult, expensive, or time consuming, but ranking small subsets of units is relatively easy and inexpensive. The use of statistical methods based on RSS can lead to a substantial improvement over analogue methods associated with SRS schemes (Wolfe, 2004). In this thesis, particularly, in Chapter 4, the effectiveness of the ranked set sampling on the empirical power function of permutation tests is studied. Moreover, the effect of the set size and the number of cycles in ranked set sampling is addressed.

- In linear mixed models, testing for zero variance component is problematic. This is because the null hypothesis lies on the boundary of the parameter space. Some available tests for the variance component are reviewed and a new test within the permutation framework is presented. Comparisons between these tests are done in terms of empirical level of significance, empirical unconditional power and execution time.

- In cluster analysis, it is of interest to measure the agreement (or similarity) between two sets of clusters created independently by two observers. Some measures of agreement can be found in the literature such as Rand index

(Rand, 1971) and Jaccard index (Jaccard, 1901). Usually large values of these measures indicate for a high agreement but not always; that is, we could have a high value of such an index for a random agreement. Therefore, instead of just measure the agreement, parametric and nonparametric tests for the null hypothesis of random agreement are proposed. Comparisons between these tests are done in terms of empirical level of significance.

Chapter 2

Permutation Tests

Contents

2.1 Brief History

The idea of permutation test dates back to Fisher (1934/35), and Pitman (1937/38) was next to consider permutation tests.

Fisher (1934, 1935) introduced the permutation approach for exact inference within the conditionality and sufficiency principles of inference. He introduced the permutation test as the exact test for the association between two binary variables when the expected number of cells is less than 5; that is, when the chi-square test fails. Also it is useful for one sided testing if at least one variable is ordered categorical. In addition, Fisher introduced the exact test for testing differences between means of two populations when the assumptions of the two-sample t-test were not met. He pointed out that the probability of a type I error (see Section 2.3) for the two-sample permutation test (Section 2.2) is closely approximated the normal theory probability of a type I error for the particular problem with which he dealt.

Pitman (1937a,b, 1938) developed exact permutation methods consistent with the Neyman-Pearson approach 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 are considered a subclass of nonparametric tests (Lehmann and Romano, 2005; Pesarin and Salmaso, 2010). They are computationally intensive,

but modern computational power makes permutation tests feasible. Nonparametric test statistics do not rely on a specific probability distribution that describes the underlying population. In fact, permutation tests are always distribution free since observed data are sufficient statistics in the null hypothesis (see Pesarin and Salmaso, 2010, Sec. 2.1.3). Some assumptions are required to the samples (e.g. exchangeability). The exchangeability assumption is generally assured by random allocation of treatments to units in experimental work. 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 accordance, for instance, with the nonparametric Behrens-Fisher testing.

The theory of optimal permutation tests is developed by Lehmann and Stein (1949). Hoeffding (1952) studied the asymptotic power behavior of permutation tests. He found that permutation tests for the randomized block design and for the two-sample designs are asymptotically as powerful as their related parametric tests. Thus, the permutation test for the randomized block design is asymptotically as powerful as the normal theory 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. For more works on permutation test and its variations see Edgington (1995), Pesarin (2001), Salmaso (2003), Good (2005), Basso et al. (2009) and Pesarin and Salmaso (2010) and the references therein.

2.2 Two-Sample Permutation Test

2.2.1 Main notation

Assume that a unidimensional nondegenerate variable of interest X takes values on sample space $\mathcal{X}$, and that associated with $(X, \mathcal{X})$ there are distributions P belonging to a nonparametric family $\mathcal{P}$. Each P gives the probability measure to events A belonging to a suitable σ-algebra $\mathcal{A}$. For quantitative variables defined on the real line, P is equivalent to the cumulative distribution function $F_P(x) := \int_{t \leq x} \mathrm{d}P(t)$, $x \in \mathcal{R}$. The notation $(X, \mathcal{X}, \mathcal{A}, P)$ summarizes the statistical model associated with the problem at hand.

It is assumed that for any statistical model $(X, \mathcal{X}, \mathcal{A}, P)$ there exists, possibly unknown, the density of P with respect to a dominating measure ζ on $(\mathcal{X}, \mathcal{A})$ and defined as $f_P := \mathrm{d}P/\mathrm{d}\zeta$. Moreover, let $\mathbf{X}_j = \{X_{ji},\, i = 1, \ldots, n_j\} \in \mathcal{X}^{n_j}$ be the independent and identically distributed (iid) sample data from $(X, \mathcal{X}, \mathcal{A}, P_j)$ of size n_j, $j = 1, 2$, and $n = n_1 + n_2$ is the total sample size. For datasets with two independent samples, one may write $\mathbf{X} = \{X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}\} \in \mathcal{X}^n$, whose related model is $(\mathbf{X}, \mathcal{X}^n, \mathcal{A}^{(n)}, P^{(n)})$, where $P^{(n)} = P_1^{(n_1)} P_2^{(n_2)}$. In the context of permutation tests, it may be convenient to use the unit-by-unit representation $\mathbf{X} = \mathbf{X}^{(n)} = (\mathbf{X}_1^{(n_1)}, \mathbf{X}_2^{(n_2)}) = \{X(i),\, i = 1, \ldots, n;\, n_1, n_2\}$ to denote datasets,

where it is intended that the first n_1 data in the list belong to the first sample (treatment group) and the rest to the second sample (control group). Indeed, if $\mathbf{u}^* = (u_1^*, \ldots, u_n^*)$ indicates a permutation of unit labels $\mathbf{u} = (1, \ldots, n)$, then $\mathbf{X}^* = \{X^*(i) = X(u_i^*),\ i = 1, \ldots, n;\ n_1, n_2\}$ is the related permutation of $\mathbf{X}$. And so, $\mathbf{X}_1^* = \{X_1^*(i) = X(u_i^*),\ i = 1, \ldots, n_1\}$ and $\mathbf{X}_2^* = \{X_2^*(i) = X(u_i^*),\ i = n_1+1, \ldots, n\}$ are the two permuted samples respectively. One may also use the same symbol $\mathbf{X}$ to denote the pooled dataset as obtained by $\mathbf{X} = \mathbf{X}_1 \uplus \mathbf{X}_2$, where $\uplus$ is the symbol for concatenating two vectors.

In this thesis and for two-sample design, testing problems for one-sided alternatives as generated by symbolic treatments with non-negative fixed shift effects δ are considered. In particular, the fixed additive effects model is considered, which is written as

$$X_{1i} = \mu + \delta + \sigma Z_{1i},\ i = 1, \ldots, n_1; \quad X_{2i} = \mu + \sigma Z_{2i},\ i = 1, \ldots, n_2, \tag{2.1}$$

where μ is a population constant, Z_{ji} are exchangeable random errors with null location and unit scale parameter, σ is a scale coefficient independent on units and treatment levels, and δ is the treatment effect (effect size) which is unknown even after data have been collected. In practice, without loss of generality, $\mu = 0$ (because it is a nuisance quantity common to all units and thus is not essential for comparing $\mathbf{X}_1$ to $\mathbf{X}_2$) and $\sigma = 1$ are chosen. Therefore, the dataset can be written as $\mathbf{X}(\boldsymbol{\delta}) = (\mathbf{Z}_1 + \boldsymbol{\delta}, \mathbf{Z}_2)$ where $\boldsymbol{\delta} = (\delta_i = \delta > 0,\ i = 1, \ldots, n_1)$. The hypotheses of interest are

$$H_0 : \{\boldsymbol{\delta} = 0\} \text{ against } H_1 : \{\boldsymbol{\delta} > 0\}. \tag{2.2}$$

It should be emphasized that $\{\boldsymbol{\delta} = 0\}$ is equivalent to $\{\mathbf{X}_1 \overset{d}{=} \mathbf{X}_2\}$, i.e. to the equality in distribution of treatment and control groups. The latter notation is in accordance with the notion that data of two groups are exchangeable, same as permutable, in the null hypothesis. The alternative is then consistent with the notion that distribution of treatment group ($\mathbf{X}_1$) stochastically dominates that of control group ($\mathbf{X}_2$).

A suitable test statistic, $T : \mathcal{X}^n \to \mathcal{R}^1$ should be chosen such that, without loss of generality, large values are evidence against H_0. Typically, $T(\mathbf{X}) = S_1(\mathbf{X}_1) - S_2(\mathbf{X}_2)$ for the comparison with two-sample permutation design, where functions S_j, $j = 1, 2$ are assumed to be:

1. symmetric, that is, invariant with respect to rearrangements of data input, i.e., their arguments;

2. strictly increasing, that is, $S_j(\mathbf{X} + \mathbf{Y}) \geq S_j(\mathbf{X})$, $j = 1, 2$, for any dataset $\mathbf{X}$ and nonnegative $\mathbf{Y} \overset{p}{\geq} 0$ so that large values of T are evidence against H_0.

The conditional support of T is given by

$$\mathcal{T}(\mathbf{X}) = \{T^* = T(\mathbf{X}^*),\ \mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}\},$$

where $\mathbf{X}^*$ is a permutation of $\mathbf{X}$, $\mathcal{X}_{/\mathbf{X}}$ is the collection of all permutations generated by $\mathbf{X}$ and it is called the permutation sample space or the conditional reference space.

For a given level of significance α, the critical value of the permutation test is T_α. For simplicity, the non-randomized version of permutation test is adopted. By indicating with $T^o = T(\mathbf{X})$ the observed value of T, H_0 is rejected if $T^o \geq T_\alpha$, and the test is given by

$$\varphi(\mathbf{X}|\mathcal{X}_{/\mathbf{X}}) = \begin{cases} 1 & \text{if } T^o \geq T_\alpha \\ 0 & \text{otherwise} \end{cases}$$

Due to the difficulty of expressing the permutation distribution of T^* in a closed form, the determination of $T_\alpha(\mathbf{X})$ is considered not convenient in practice. So, the p-value approach is considered. The p-value is defined as

$$\lambda_T(\mathbf{X}) = Pr\{T^* \geq T^o|\mathcal{X}_{/\mathbf{X}}\},$$

which is a non-increasing function of T^o, and hence, H_0 is rejected if $\lambda_T \leq \alpha$, for any fixed value of α. The non-randomized permutation test is then given by

$$\varphi(\mathbf{X}|\mathcal{X}_{/\mathbf{X}}) = \begin{cases} 1 & \text{if } \lambda_T(\mathbf{X}) \leq \alpha \\ 0 & \text{otherwise} \end{cases}$$

In practice, since the p-value $\lambda_T(\mathbf{X})$ is one-to-one with the test statistic $\varphi(\mathbf{X}|\mathcal{X}_{/\mathbf{X}})$, is itself used with the role of test statistic for which the critical value is α, because in the null hypothesis the distribution of $\lambda_T(\mathbf{X})$ is uniform over its support.

It is worthwhile to observe that the hypothetical frequency interpretation of such reported p-values is as follows. If we were to accept the available data as just decisive evidence against H_0, then we would reject the null hypothesis when true a long-run proportion $\lambda_T(\mathbf{X})$ of times.

2.2.2 Permutation test procedure

A two-sample 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 treatment group and n_2 units assigned to the control or placebo group. Then, the observed datasets, $\mathbf{X}_1$ and $\mathbf{X}_2$, are obtained and the test statistic is calculated, T^o.

2. Permute the $n = n_1 + n_2$ observations between the two groups so that there are n_1 observations for the treatment group and n_2 observations for the control group. Write down the set of all possible permutations, i.e. the permutation sample space $\mathcal{X}_{/\mathbf{X}}$. The cardinality of $\mathcal{X}_{/\mathbf{X}}$ is

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

3. For each permutation of the data, i.e. for each $\mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}$, compute the test statistic, $T^* = T(\mathbf{X}^*)$.

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 statistically significant if the p-value is not larger than this level.

Since it is tedious to write down the whole permutation sample space, conditional Monte Carlo algorithm (Algorithm 2.1) is used to estimate the p-value at any desired accuracy.

Algorithm 2.1 Conditional Monte Carlo (CMC)

1. For the given dataset $\mathbf{X}$, calculate the observed test statistic, T^o.

2. Take a random permutation $\mathbf{X}^* \in \mathcal{X}_{/\mathbf{X}}$ of $\mathbf{X}$, and calculate the corresponding test statistic $T^* = T(\mathbf{X}^*)$.

3. Independently repeat Step 2 a large number, say B, of times, giving B test statistics, 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 unbiased and strongly consistent due to Glivenko-Cantelli theorem (Shorack and Wellner, 1986).

2.3 Power Functions of Permutation Tests

Neyman and Pearson (1933) were the first to discuss the concepts of *type I error* and *type II error*. Type I error occurs when the researcher rejects the null hypothesis when it is true. Type I error probability is determined by the level of significance α. Hence, α is the probability of making a type I error when the null hypothesis is true. α is defined as the long-run relative frequency by which type I errors are made over independently repeated samples from the same population under the same null hypothesis, assuming the null hypothesis is true. Conversely, type II error occurs when the researcher accepts the null hypothesis when the alternative is true. The probability of making a type II error under the alternative is denoted by β.

In general, type I error is considered to be more serious, and then more important to avoid, than a type II error. Unfortunately, everything else being fixed, it is not

possible to decrease both errors at the same time; reduce the type I error leads to increase the type II error. Therefore, statisticians fix α and try to minimize β.

α and β can be calculated using the power function. The *power function* is defined as

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

The power of permutation tests may be generally thought of in two quite different ways (Box and Andersen, 1955): first, as a power conditional upon the observations which is considered in Section 2.3.1 as *conditional power*, and second, as what will be called an *unconditional power* which is discussed in Section 2.3.2 (Kempthorne et al., 1961; Collier and Baker, 1966; Pesarin and Salmaso, 2010).

2.3.1 Conditional power function

For testing the hypotheses in Equation 2.2 the conditional power function is defined as

$$\begin{aligned} W\left[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}\right] &= \mathbb{E}[\varphi(\mathbf{X}(\boldsymbol{\delta})|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})})] \\ &= Pr\left[\lambda_T(\mathbf{X}(\boldsymbol{\delta})) \leq \alpha|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}\right] \\ &= \mathbb{E}\left\{\mathbb{I}[\lambda_T(\mathbf{X}^\dagger(\boldsymbol{\delta})) \leq \alpha]\big|\mathcal{X}_{/\mathbf{X}^\dagger(\boldsymbol{\delta})}\right\}, \end{aligned} \tag{2.3}$$

It is worthwhile to observe that $W\left[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}\right]$ is a function of the *effect size* $\boldsymbol{\delta}$ for a given sample size n, preassigned level of significance α and suitable test statistic T conditional on the observed dataset which is a sufficient statistic for the underlying distribution P in the null hypothesis. One may write

$$W\left[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}\right] = \begin{cases} \alpha & \text{if } \boldsymbol{\delta} = 0 \\ 1-\beta_{/\mathbf{X}(\boldsymbol{\delta})} & \text{if } \boldsymbol{\delta} > 0 \end{cases}$$

It is also worth noting that $\lambda_T(\mathbf{X}^\dagger(\boldsymbol{\delta}))$ is the p-value calculated on the dataset $\mathbf{X}^\dagger(\boldsymbol{\delta}) = (\mathbf{Z}_1^\dagger + \boldsymbol{\delta}, \mathbf{Z}_2^\dagger)$, where $\mathbf{Z}^\dagger \in \mathcal{Z}_{/\mathbf{Z}}$ is a random permutation of unobservable deviates $\mathbf{Z}$. Indeed, the randomization principle essentially involves a random assignment of a subset $\mathbf{Z}_1^\dagger$ of deviates $\mathbf{Z}$ to treated units for which $\boldsymbol{\delta}$ is active and the rest to the untreated, so that $\mathbf{Z}_1^\dagger + \boldsymbol{\delta}$ are the data $\mathbf{X}_1^\dagger$ of the treatment group. From this point of view, the actual dataset $\mathbf{X}(\boldsymbol{\delta})$ is just one of the possible sets $\mathbf{X}_1^\dagger$ that can be obtained by a re-randomization of deviates to treatments. And so the notion of conditional power uses as many datasets $\mathbf{X}^\dagger$ as there are re-randomizations in $\mathcal{Z}_{/\mathbf{Z}}$ (Pesarin and Salmaso, 2010).

It is clear that the true value of the conditional power function is not only tedious but also virtual to attain. Hence, Algorithm 2.2 is used for evaluating it empirically.

Empirical post-hoc conditional power function In order for Algorithm 2.2 to be effectively carried out, it is necessary, in the given dataset, to separate the

Algorithm 2.2 Empirical Conditional Power Function

1. Consider the pooled set of deviates $\mathbf{Z} = \mathbf{Z}_1 \biguplus \mathbf{Z}_2$ and the effects $\boldsymbol{\delta}$.
2. Take a re-randomization $\mathbf{Z}^\dagger$ of $\mathbf{Z}$ and the corresponding dataset $\mathbf{X}^\dagger(\boldsymbol{\delta}) = (\mathbf{Z}_1^\dagger + \boldsymbol{\delta}, \mathbf{Z}_2^\dagger)$.
3. Use the CMC algorithm to calculate the p-value $\hat{\lambda}_T(\mathbf{X}^\dagger(\boldsymbol{\delta}))$.
4. Independently repeat Steps 2 and 3 a large number, say R, of times, giving R p-values, say $\{\hat{\lambda}_T(\mathbf{X}_r^\dagger(\boldsymbol{\delta})), r = 1, \ldots, R\}$.
5. Finally, the empirical conditional power is given by
$$\hat{W}[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}] = \frac{\sum_{r=1}^{R} \mathbb{I}[\hat{\lambda}_T(\mathbf{X}_r^\dagger(\boldsymbol{\delta})) \leq \alpha]}{R}.$$
6. To obtain a function in $\boldsymbol{\delta}$, Steps 1-5 are repeated for different values of $\boldsymbol{\delta}$.

contributions of random deviates $\mathbf{Z}$ from those of effects $\boldsymbol{\delta}$. This is generally not possible in practice, because usually $\mathbf{X}$ is observed; its components $\mathbf{Z}$ and $\boldsymbol{\delta}$ are not separately observable. Thus, the conditional power is essentially a *virtual notion* in the sense that it is well defined but is not calculable exactly. However, in place of $\hat{W}[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}]$, the so-called *empirical post-hoc conditional power* $\hat{W}[(\boldsymbol{\delta}; \hat{\boldsymbol{\delta}}, n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}]$ may be achieved. The main idea is to find an empirical estimate of $\mathbf{Z}$, $\hat{\mathbf{Z}}$, by subtracting a suitable estimate of the effect size $\boldsymbol{\delta}$, $\hat{\boldsymbol{\delta}}$, from the observed dataset $\mathbf{X}$. Thus, the empirical pooled set of deviates is given by $\hat{\mathbf{Z}} = \hat{\mathbf{Z}}_1 \biguplus \mathbf{Z}_2 = (\mathbf{X}_1 - \hat{\boldsymbol{\delta}}) \biguplus \mathbf{X}_2$. Note that this gives rise to approximate solution because exchangeability condition is now approximate as $\hat{\boldsymbol{\delta}}$ is not a permutationally invariant estimate.

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). For two sample permutation design, the difference between sample means, $\hat{\boldsymbol{\delta}} = \bar{X}_1 - \bar{X}_2$, is considered.

To sum up, Algorithm 2.3 is used to find the empirical post-hoc conditional power function.

2.3.2 Unconditional power function

To define the unconditional power, the mean value of the conditional power, $W[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}}]$, with respect to the underlying distribution P, must be ob-

Algorithm 2.3 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 deviates $\hat{\mathbf{Z}} = (\mathbf{X}_1 - \hat{\boldsymbol{\delta}}) \biguplus \mathbf{X}_2$.
2. Take a random re-randomization $\hat{\mathbf{Z}}^\dagger$ of $\hat{\mathbf{Z}}$. Then for any chosen $\boldsymbol{\delta}$ the corresponding dataset $\hat{\mathbf{X}}^\dagger(\boldsymbol{\delta}) = (\hat{\mathbf{Z}}_1^\dagger + \boldsymbol{\delta}, \hat{\mathbf{Z}}_2^\dagger)$.
3. Use the CMC algorithm to calculate the p-value $\hat{\lambda}_T(\hat{\mathbf{X}}^\dagger(\boldsymbol{\delta}))$.
4. Independently repeat Steps 2 and 3 a large number, say R, of times, giving R p-values, say $\{\hat{\lambda}_T(\hat{\mathbf{X}}_r^\dagger(\boldsymbol{\delta})), r = 1, \ldots, R\}$.
5. Finally, the empirical post-hoc conditional power is given by
$$\hat{W}[(\boldsymbol{\delta}; \hat{\boldsymbol{\delta}}, n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\boldsymbol{\delta})}] = \frac{\sum_{r=1}^{R} \mathbb{I}[\hat{\lambda}_T(\hat{\mathbf{X}}_r^\dagger(\boldsymbol{\delta})) \leq \alpha]}{R}.$$
6. To obtain a function in $\boldsymbol{\delta}$, Steps 2-5 are repeated for different values of $\boldsymbol{\delta}$.

tained. That is:

$$\begin{aligned} W(\boldsymbol{\delta}; n, \alpha, T, P) &= \mathbb{E}_{\mathcal{X}^n \backslash \mathcal{X}_{/\mathbf{X}}} \left\{ \mathbb{E} \left[W((\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}}) \right] \right\} \\ &= \mathbb{E}_{\mathcal{X}} \left\{ W[(\boldsymbol{\delta}; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}}] \right\} \\ &= \int_{\mathcal{X}^n} \mathbb{I} \left[\lambda_T(\mathbf{X}(\boldsymbol{\delta})) \leq \alpha | \mathcal{X}_{/\mathbf{X}} \right] dP(\mathbf{X}(\boldsymbol{\delta})) \end{aligned}$$

Note that in order to properly define the unconditional power $W(\boldsymbol{\delta}; n, \alpha, T, P)$, the underlying population distribution P must be fully specified, that is, defined in its analytical form and all its parameters. Also note that averaging with respect to the whole sample space $\mathcal{X}^n$ implies taking the mean with respect to each conditional distribution over $\mathcal{X}_{/\mathbf{X}}$ and then taking the mean of these with respect to the distribution over $\mathcal{X}^n \backslash \mathcal{X}_{/\mathbf{X}}$.

In practice, the unconditional power is based upon random sampling from some population. The p-value of the permutation test is conditional upon the observations for each sample, but the power is the proportion of p-values that are less than or equal α over repeated sampling from the underlying population. Algorithm 2.4 is used for evaluating the unconditional power based on a standard Monte Carlo simulation.

If the true effect size is unknown, one may attain the empirical post-hoc unconditional power function, denoted by $\hat{W}(\boldsymbol{\delta}; \hat{\boldsymbol{\delta}}, n, \alpha, T, P)$.

Algorithm 2.4 Empirical Unconditional Power Function

1. Choose a virtual value of the effect size $\boldsymbol{\delta}$.
2. From the given population distribution P draw one set of n deviates $\mathbf{Z}$, and then add $\boldsymbol{\delta}$ to the first n_1 errors to define the dataset $\mathbf{X}(\boldsymbol{\delta}) = (\mathbf{Z}_1 + \boldsymbol{\delta}, \mathbf{Z}_2)$.
3. Use the CMC algorithm to calculate the p-value $\hat{\lambda}_T(\mathbf{X}(\boldsymbol{\delta}))$.
4. Independently repeat Steps 2 and 3 a large number, say R, of times, giving R p-values, say $\{\hat{\lambda}_T(\mathbf{X}_r(\boldsymbol{\delta})), r = 1, \ldots, R\}$.
5. Finally, the empirical unconditional power is given by
$$\hat{W}(\boldsymbol{\delta}; n, \alpha, T, P) = \frac{\sum_{r=1}^{R} \mathbb{I}[\hat{\lambda}_T(\mathbf{X}_r(\boldsymbol{\delta})) \leq \alpha]}{R}.$$
6. To obtain a function in $\boldsymbol{\delta}$, Steps 1-5 are repeated for different values of $\boldsymbol{\delta}$.

2.4 Illustration Examples

2.4.1 Degree of reading power

In his Ph.D thesis, Schmitt (1987) was interested to test 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 aspects of reading ability that the treatment is designed to improve. The DRP scores are reported in Table 2.1.

Table 2.1: 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 $H_0 : \{\mu_t = \mu_c\}$ versus $H_1 : \{\mu_t > \mu_c\}$, Algorithm 2.1 is used. The difference between the sample means is considered as a test statistic. The observed test statistic is $T^o = 9.954$ and the conditional p-value is $\hat{\lambda} = 0.015$. At $\alpha = 0.05$ the null hypothesis is rejected.

Figure 2.1(a) shows the permutation distribution of the difference of means based on 5000 iterations. The solid vertical line in the figure marks the location of the

statistic for the original sample, $T^o = 9.954$. Use the permutation distribution exactly as if it were the sampling distribution: the p-value is the probability that the statistic takes a value at least as extreme as 9.954 in the direction given by the alternative hypothesis.

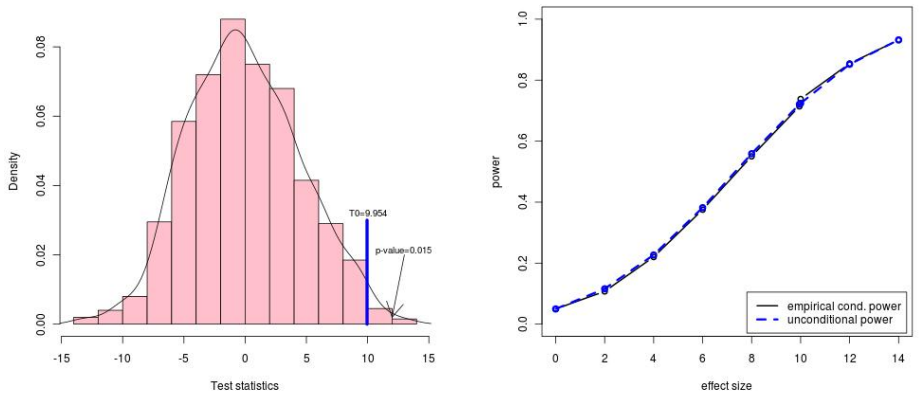
Figure 2.1(a) shows that the permutation distribution has a roughly normal shape. Because the permutation distribution approximates the sampling distribution, and hence the sampling distribution is close to normal. Therefore, the usual two-sample t-test can safely be applied. Using the t-test, the p-value is 0.013, which is very close to the p-value obtained using the permutation test.

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

$$W(\boldsymbol{\delta}; n, \alpha, T, P) = 1 - F_t(t_{df}^{1-\alpha}, df, ncp), \tag{2.4}$$

where F_t is the student t-distribution, $df = n_1 + n_2 - 2$ 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 $ncp = \delta \left(S_p^2 (\frac{1}{n_1} + \frac{1}{n_2}) \right)^{-1/2}$, $S_p^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}$ is the pooled variance.

Figure 2.1(b) shows the empirical post-hoc conditional power function together with the unconditional (parametric) power function.



(a) (b)

Figure 2.1: DRP data: (a) The permutation distribution. (b) The unconditional power and the empirical post-hoc conditional power functions.

2.4.2 Tawjihi exam 2009/2010

The Tawjihi exam is a school matriculation exam, part of education in Palestine, which is a prerequisite for graduation and university entrance. Palestine is divided into two geographic regions: the West Bank and Gaza Strip. Since June 2007, Gaza

Strip is under siege, and this impacted negatively on the schools' ability to proceed normally with a structured learning-teaching process. Therefore, it is expected that Tawjihi results in Gaza Strip are worse than in West Bank. So, it is of interest to test $H_0 : \{\mu_{WB} = \mu_{GS}\}$ versus $H_1 : \{\mu_{WB} > \mu_{GS}\}$.

Two samples are randomly chosen from these two regions, each of size 10. The data are reported in Table 2.2.

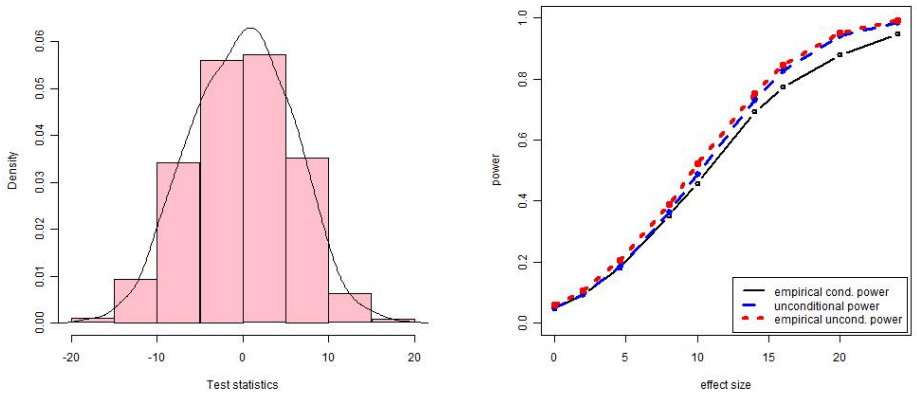
Table 2.2: Tawjihi results in Palestine, 2009/2010

West Bank, $\mathbf{X}_{WB}$					Gaza Strip, $\mathbf{X}_{GS}$				
57.4	70.1	92.9	93.4	66.0	73.3	50.1	71.8	56.5	68.4
58.1	55.5	79.8	51.5	84.2	55.9	59.6	81.3	58.5	69.7

Algorithm 2.1 is used and the difference between the sample means is considered as a test statistic. The observed test statistic is $T^o = 4.58$ and the conditional p-value is $\hat{\lambda} = 0.215$. At $\alpha = 0.05$, the null hypothesis is not rejected.

Figure 2.2(a) shows that the permutation distribution has a roughly normal shape. Applying the usual two-sample t-test, the p-value is 0.226.

Figure 2.2(b) shows three power curves; the empirical post-hoc conditional power curve (Algorithm 2.3), the empirical unconditional power curve (Algorithm 2.4) and the unconditional (parametric) power curve.



(a) (b)

Figure 2.2: Tawjihi data: (a) The permutation distribution. (b) Power functions.

Chapter 3

Empirical Conditional Power Analysis

Contents

In this chapter, reproducibility and generalizability probabilities are defined within the permutation framework. It is shown that these probabilities can be useful for sample size adjustment. Moreover, the use of empirical conditional power function of permutation tests for sample size estimation is investigated. Two-sample permutation design is considered as a guide and some real data applications are used.

3.1 Introduction

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

1. Sample size, n. Everything else being fixed, the greater the sample size, the greater the power of the test.

2. Significance level, α. Everything else being fixed, the greater the significance level, the greater the power of the test.

3. (Standardized) effect size, $\Delta = \boldsymbol{\delta}/\sigma$. It is easier to detect a large effect than it is to detect a small effect; that is, the greater the effect size, the greater the power of the test.

The most important component affecting statistical power is sample size in the sense that the most frequently asked question in practice is how many observations need to be collected.

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**) are used to calculate the power of the test. 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 estimate power. For example, Collings and Hamilton (1988) proposed a bootstrap method which does not require any knowledge of the underlying distribution to estimate the power of the two-sample Wilcoxon test. See also Epstein (1955), Teichroew (1955) and Hemelrijk (1961). However, some authors derived the power functions 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 this chapter, some applications of empirical conditional power function of permutation tests are investigated. In particular, the use of empirical conditional power for sample size estimation is investigated in Section 3.2.1, reproducibility probability is investigated in Section 3.2.2, generalizability probability is investigated in Section 3.2.3 and sample size adjustment is investigated in Section 3.2.4. Real data applications are presented in Section 3.3. Concluding remarks are contained in Section 3.4.

3.2 Applications of Empirical Conditional Power Function

3.2.1 Sample size calculation

Sample size calculation is an important and often difficult step in planning a research study. Samples that are too large may waste time, resources and money, while samples that are too small may lead to 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 of hypothesis. In our context, the empirical conditional power function of permutation test is used as an important

tool for estimating an appropriate sample size for a particular study.

Consider the two samples in which $\mathbf{X}_1 = \{X_{11}, \ldots, X_{1n_1}\}$ are iid $F(x+\Delta)$ and $\mathbf{X}_2 = \{X_{21}, \ldots, X_{2n_2}\}$ are iid $F(x)$ and the two samples are independent of one another. We shall focus on the null hypothesis $H_0 : \{\Delta = 0\}$ versus the alternative $H_1 : \{\Delta > 0\}$. If the underlying distribution is normal, using t-statistic, the power of the test is given by

$$1 - \beta = 1 - \Phi\left(z_\alpha - \Delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\right), \tag{3.1}$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution and z_α is the upper α critical value of the standard normal distribution. It is worthwhile to observe that the power is monotonic nondecreasing 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 detect an effect size Δ with a desired level of power $1-\beta$ can be calculated from Equation 3.1 (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

$$n = \frac{1}{\rho(1-\rho)}\left(\frac{z_\beta + z_\alpha}{\Delta}\right)^2. \tag{3.2}$$

See also Chow et al. (2002) for sample size calculation based on noncentral t-distribution.

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

$$n = \frac{1}{12\rho(1-\rho)}\left(\frac{z_\beta + z_\alpha}{\Delta_{Noether} - 0.5}\right)^2, \tag{3.3}$$

where $\Delta_{Noether} = Pr(\mathbf{X}_1 > \mathbf{X}_2)$ is Noether's effect size. There are several ways of estimating $\Delta_{Noether}$ under various assumptions, one possibility 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 detect an effect size Δ with a desired level of power $\tilde{W} \in (\alpha, 1)$ can be obtained by solving

$$n = \arg\min_{n} \{W[(\Delta; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\Delta)}] = \tilde{W}\}.$$

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 3.1 is used for sample size estimation to detect an effect size Δ with a desired power $\tilde{W}$.

Algorithm 3.1 Sample Size Estimation

1. Start with a pilot sample of size $n = n_1 + n_2$; 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 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 detecting 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 as follows.

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

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

$$n = \frac{\tilde{\rho}(1-\tilde{\rho})\tilde{n}}{\rho(1-\rho)} \left(\frac{\tilde{\Delta}}{\Delta}\right)^2. \tag{3.4}$$

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.

3.2.2 Reproducibility probability

Suppose that one study has been conducted and the result is significant. What is the probability that a second study will produce a significant result? In other words, what is the probability that the significant result from the first study is reproducible? Statistically, if the two studies are independent, the probability of observing a significant result from the second study is given by the power of the test, irrespective of whether the result from the first study was significant or not. However, such information from the first study should be useful in the evaluation of the probability of observing a significant result in the second study. This leads to the concept of reproducibility probability, which is different from the 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. In other words, the reproducibility probability is defined as the power with Δ replaced by its estimate $\hat{\Delta}_0$ based on the data from the 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, denoted by $\hat{W}[(\hat{\Delta}; \hat{\Delta}, n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\hat{\Delta})}]$. It is used to assess how reliable the testing inference associated with $(T, \mathbf{X})$ is, in the sense that if by chance the probability of obtaining the same inference with $(T, \mathbf{X}^{\dagger})$ as with $(T, \mathbf{X})$ is greater than (say) 0.50, then the actual inferential conclusion, given the set of units underlying $\mathbf{X}$, is reproducible more often than not.

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 $\hat{\Delta}$?*. That is, contemplate a future study exactly like the one we just did, with the same sample size; what is the probability of achieving statistical significance if the same effect is observed?

It is worthwhile to observe that the outcome (significance or non-significance) of a single test using adequate sample size in no way affects or alters the levels of power, α, and effect size set a priori by the researcher. These concepts relate to statistical tests in general and not 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).

3.2.3 Generalizability probability

As discussed in Section 3.2.2, the concept of reproducibility is used to evaluate whether results observed from the same population are reproducible from study site to study site. It is of interest to study how likely the results can be reproducible to a *different but similar* population. 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 is 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 Gaussian to 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 are two *different but similar* populations. In population A, the effect size is given by $\Delta = (\mu_1 - \mu_2)/\sigma$. Suppose that in population B the population mean difference is changed to $\mu_1 - \mu_2 + \eta$ and the population variance is changed to $C^2\sigma^2$, 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.

If the power of the current study (under population A) is $W[(\Delta; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\Delta)}]$, then the power of the future study (under population B) is $W[(D\Delta; n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\Delta)}]$. If D is known, then the generalizability probability is the reproducibility probability $\hat{W}[(D\hat{\Delta}; \hat{\Delta}, n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\hat{\Delta})}]$. When the value of D is unknown, a set of D-values may be considered.

3.2.4 Sample size adjustment

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

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

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

$$n = \tilde{n}\left(\frac{\tilde{\Delta}}{\hat{\Delta}}\right)^2. \tag{3.5}$$

Generalizability probability can be used for sample size adjustment. 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.

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

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

$$n = \tilde{n}\left(\frac{\tilde{\Delta}}{D\hat{\Delta}}\right)^2. \tag{3.6}$$

3.3 Illustration Examples

3.3.1 Degree of reading power (revisited)

Sample size calculation Algorithm 3.1 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 3.1. For example, if the desired power is $\tilde{W} = 0.90$, one may consider $n_1 = 13$ and $n_2 = 7$.

Table 3.2 reports the (parametric) unconditional power calculated using Equation 2.4 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 is occurred when $n_1 = 10$ and $n_2 = 10$. 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 3.1 or 3.2, the sample sizes to detect an effect size $\boldsymbol{\delta} = 10$ are calculated using Equation 3.4. 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 3.1: 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 3.2: 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 3.1 or 3.2, the required sample sizes to detect 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 $n_2 = 7$. From Section 2.4.1, it is found that 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 $\hat{W}[(\hat{\Delta}; \hat{\Delta}, n, \alpha, T)|\mathcal{X}_{/\mathbf{X}(\hat{\Delta})}] = 0.722$ (see Figure 2.1(b)). That is, the probability of getting a significance results to detect an effect size $\hat{\boldsymbol{\delta}} = 9.954$ at level of significance $\alpha = 0.05$ is high, 72.2%.

Sample size adjustment Hence, in order to have a reproducibility probability equals to 0.85, one may adjust the sample size using Equation 3.5. 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 detect an effect size 9.954 with a desired reproducibility probability of 0.85, the sample sizes should be $n_1 = n_2 = 14$.

3.3.2 Tawjihi exam 2009/2010 (revisited)

Sample size calculation Algorithm 3.1 is used to calculate the required sample sizes to detect an effect size $\boldsymbol{\delta} = \mu_{WB} - \mu_{GS} = 10$. The results are reported in Table 3.3. For example, if the desired power is $\tilde{W} = 0.80$, one may consider $n_1 = 15$ and $n_2 = 15$.

Table 3.4 reports the (parametric) unconditional power calculated using Equation 3.1 as a function with the sample sizes. It is assumed that the true standard deviation is $\sigma = 13.03$ and hence $\Delta = 10/13.03 \approx 0.77$. It is clear that balanced designs are more powerful than unbalanced. Moreover, the power is not affected by whether the size of the treatment group is greater or smaller than the size of the control group.

Now, given the information reported in Table 3.3 or 3.4, the sample sizes to detect an effect size $\boldsymbol{\delta} = 5$ are calculated using Equation 3.4. Assuming $\rho = \tilde{\rho} = 0.5$ and $\tilde{n} = 30$, then $n = 120$. Hence, $n_1 = n_2 = 60$.

Table 3.3: Tawjihi Example: empirical conditional power and sample sizes, $\boldsymbol{\delta} = 10$

		n_2					
		5	10	15	20	25	30
n_1	5	0.49	0.56	0.59	0.57	0.62	0.67
	10	0.59	0.71	0.71	0.71	0.75	0.81
	15	0.67	0.78	0.80	0.80	0.83	0.89
	20	0.68	0.81	0.84	0.83	0.89	0.92
	25	0.73	0.85	0.88	0.89	0.92	0.96
	30	0.72	0.86	0.89	0.91	0.94	0.97

Table 3.4: Tawjihi Example: parametric unconditional power and sample sizes, $\boldsymbol{\delta} = 10$

		n_2					
		5	10	15	20	25	30
n_1	5	0.53	0.59	0.63	0.66	0.68	0.70
	10	0.59	0.68	0.73	0.76	0.78	0.81
	15	0.63	0.73	0.78	0.82	0.84	0.88
	20	0.66	0.76	0.82	0.86	0.88	0.91
	25	0.68	0.78	0.84	0.88	0.91	0.94
	30	0.70	0.81	0.88	0.91	0.94	0.96

Reproducibility probability The required sample sizes to detect the virtual effect size $\tilde{\boldsymbol{\delta}} = 10$ at level of significance $\alpha = 0.05$ with a desired level of power $\tilde{W} = 0.70$ are $n_1 = 10$ and $n_2 = 10$. From Section 2.4.2, it is found that the observed effect size is $\hat{\boldsymbol{\delta}} = 4.58$ or $\hat{\Delta} = \hat{\boldsymbol{\delta}}/\sigma = 0.35$. Therefore, the reproducibility

probability is given by $\hat{W}[(\hat{\Delta};\hat{\Delta},n,\alpha,T)|\mathcal{X}_{/\mathbf{X}(\hat{\Delta})}] = 0.186$ (see Figure 2.2(b)). That is, the probability of getting a significance results to detect an effect size $\hat{\boldsymbol{\delta}} = 4.58$ at level of significance $\alpha = 0.05$ is very low, 18.6%.

Sample size adjustment Hence, in order to have a high reproducibility probability, e.g. 70%, one may adjust the sample size using Equation 3.5. Let $\tilde{\boldsymbol{\delta}} = 10$, $\tilde{n} = 20$ and $\hat{\boldsymbol{\delta}} = 4.58$, then $n = 95.34525 \approx 96$ and hence $n_1 = n_2 = 48$. That is, in order to detect an effect size 4.58 with a desired reproducibility probability of 0.70, the sample sizes should be $n_1 = n_2 = 48$.

Generalizability probability A sample of size $n = 96$ (48 from each region) is taken form the first population (students attended Tawjihi exam 2009/2010) and the observed effect size is $\hat{\boldsymbol{\delta}} \approx 6.17$ and the p-value is 0.016 which is significant. Now, given these information, one may ask what is the probability of obtaining a significance result if one would draw a sample of size $n = 96$ from students attended Tawjihi exam 2010/2011 (different but similar population). Assume $D \approx 1.34$ (in fact it is, otherwise a set of D-values are considered), the generalizability probability evaluated at $D\hat{\boldsymbol{\delta}} = 1.34 \times 6.17 \approx 8.27$ or equivalently $D\hat{\Delta} \approx 0.59$ is given by $\hat{W}[(D\hat{\Delta};\hat{\Delta},n,\alpha,T)|\mathcal{X}_{/\mathbf{X}(\hat{\Delta})}] = 0.885$. That is, if one would draw a sample of size $n = 96$ from students attended Tawjihi exam 2010/2011, then in order to detect an effect size $D\hat{\Delta} \approx 0.59$ the probability of getting a significant result is 88.5%. Consider $D = (0.2, 0.4, 0.8, 1.2, 1.34, 1.5, 2)$, then the generalizability probabilities are reported in Table 3.5.

A sample of size $n = 96$ is drawn from students attended Tawjihi exam 2010/2011 and the empirical post-hoc conditional power is reported in Table 3.5. It is clear that the generalizability probability obtained by the use of the information based on a sample from students attended Tawjihi exam 2009/2010 is very close to the empirical post-hoc conditional power obtained by a sample from students attended Tawjihi exam 2010/2011.

Table 3.5: Tawjihi Example: Generalizability, $\hat{W}[(D\hat{\Delta};\hat{\Delta},n,\alpha,T)|\mathcal{X}_{/\mathbf{X}(\hat{\Delta})}]$

	$D\hat{\boldsymbol{\delta}}$						
	1.23	2.47	4.94	7.41	8.27	9.26	12.34
GP	0.12	0.21	0.53	0.84	0.89	0.95	0.99
PHP	0.12	0.23	0.60	0.89	0.95	0.98	0.99

GP: The generalizability probability calculated based on a sample drawn from students attended Tawjihi exam 2009/2010. PHP: The empirical post-hoc conditional power calculated based on a sample from students attended Tawjihi exam 2010/2011.

3.4 Concluding Remarks

In this chapter:

- Sample size is estimated by the use of empirical conditional power function of permutation tests. 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 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 is defined within permutation framework. It is an important tool for sample size adjustment and is used to measure the reliability of the test.
- Generalizability probability is defined within permutation framework. It is also used for sample size adjustment.

Chapter 4

Permutation Tests with Ranked Set Sampling

Contents

In this chapter, the permutation test is studied in the context of Ranked Set Sampling (RSS). The RSS version of the test statistic is defined and the power is compared with its counterpart in Simple Random Sampling (SRS). The effect of the set size and the number of cycles in RSS is also addressed. The two-sample permutation design is considered as a guide.

4.1 Introduction

Ranked Set Sampling (RSS), a sampling technique, was first introduced by McIntyre (1952, 2005) as an efficient alternative to Simple Random Sampling (SRS) for estimating the expected pasture yields in agricultural experimentation. It is obviously applicable in other situations as well. Dell and Clutter (1972) used RSS in ecological and environmental studies. Samawi (1999) and Samawi and Al-Sagheer (2001) used RSS in medical studies.

RSS can be useful when measurements are expensive (in terms of time, money, or other) but units from the population can be easily ranked. In McIntyre's case, measuring the plots of pasture yields requires mowing and weighting crop yields, which is time consuming. However, a small number of plots can be even though sufficiently well ranked by eye without measurement. McIntyre's goal was to develop a sampling technique to reduce the number of necessary measurements to be made, maintaining the unbiasedness of the SRS mean and reducing the variance of the mean

estimator by incorporating the outside information provided by visual inspection. Therefore, since the ranking of the plots could be done very cheap, he developed a technique to implement this advantage (Rey, 2004).

RSS can be used in certain medical studies. For instance, it can be used in the determination of normal ranges of certain medical measures, which usually involves expensive laboratory tests. Samawi (1999) considered using RSS for the determination of normal ranges of bilirubin level in blood for new born babies. To establish such ranges, blood sample must be taken from the sampled babies and tested in a laboratory. But, on the other hand, the ranking of the bilirubin levels of a small number of babies can be done by observing whether their face, chest, lower parts of the body and the terminal parts of the whole body are yellowish, since, as the yellowish color goes from face to the terminal parts of the whole body, the level of bilirubin in blood goes higher.

For discussions of some other settings where ranked set sampling techniques have found applications, see Patil (1995), Barnett and Moore (1997) and Chen et al. (2004).

Algorithm 4.1 described the original form of RSS conceived by McIntyre.

Algorithm 4.1 Ranked Set Sampling Technique

1. Randomly select m sets, each of size m elements from the population of interest.

2. The elements of each set in Step 1 are ranked with respect to the variable of interest, say $\mathbf{X}$, visually or by any negligible cost method that does not require actual measurements.

3. Identify by judgment the i^{th} minimum from the i^{th} set, $i = 1, 2, \ldots, m$. The set of the m elements obtained is called a ranked set sample.

4. Independently repeat Steps 1-3 h times (cycles), if necessary, to obtain an RSS of size $n = mh$.

Figure 4.1 describes each step in the process of RSS (Algorithm 4.1) in terms of matrices. Let $Y_i = \{X_{(ii)}, i = 1, \ldots, m\}$; that is, the obtained RSS, $\{X_{(11)}, X_{(22)}, \ldots, X_{(mm)}\}$, is denoted by $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$. If the process is repeated h cycles, then the RSS can be represented as a matrix of size $n = h \times m$ as it is shown in Step 4 of Figure 4.1.

To understand the structure of RSS and its variation from SRS, consider the simple case of a single cycle ($h = 1$) with set size m. Let $X_1, \ldots, X_m$ be a SRS of size m from a continuous distribution with probability density function (pdf) $f(x)$ and cumulative distribution function (cdf) $F(x)$ and let $Y_1, \ldots, Y_m$ be a RSS of size m obtained as described in Algorithm 4.1 from m independent random samples of m elements each.

In the case of a SRS, the m observations are iid $f(x)$. However, there is

Step 1: $\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mm} \end{bmatrix}$ Step 2: $\begin{bmatrix} X_{(11)} & X_{(12)} & \cdots & X_{(1m)} \\ X_{(21)} & X_{(22)} & \cdots & X_{(2m)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{(m1)} & X_{(m2)} & \cdots & X_{(mm)} \end{bmatrix}$

Step 3: $\{X_{(11)}, X_{(22)}, \ldots, X_{(mm)}\}$ Step 4: $\begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{h1} & Y_{h2} & \cdots & Y_{hm} \end{bmatrix}$

Figure 4.1: Ranked set sampling procedure

no additional structure imposed on their relationship to one another. Letting $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$ be the order statistics associated with these SRS observations. Note that they are dependent random variables with joint pdf given by

$$f_{X_{(1)},\ldots,X_{(m)}}(x_1, \ldots, x_m) = m! \prod_i f(x_i) \mathbb{I}_{\{-\infty<x_1<\cdots<x_m<\infty\}}(x_1, \ldots, x_m).$$

In the case of a RSS, additional information and structure has been provided through the judgement ranking process involving a total of m^2 sample elements. The m measurements $Y_1, \ldots, Y_m$ are also order statistics but in this case they are *independent* observations and each of them provides information about a different aspect of the population. The joint pdf for $Y_1, \ldots, Y_m$ is given by

$$f_{Y_1,\ldots,Y_m}(y_1, \ldots, y_m) = \prod_i f_{Y_i}(y_i),$$

where

$$f_{Y_i}(y_i) = \frac{m!}{(i-1)!(m-i)!}[F(y_i)]^{i-1}[1-F(y_i)]^{m-i} f(y_i)$$

is the pdf for the i^{th} order statistic for a SRS of size m from the population with pdf $f(x)$ and cdf $F(x)$ (David and Nagaraja, 2003). This extra structure in RSS make it to be more efficient (in terms of variance of estimates of the mean) than comparable procedures based on a SRS with the same number of measured observations. However, these extra structure make the theory of RSS more difficult than their SRS counterparts.

It is worthwhile to emphasize that in RSS m^2 elements are selected at no cost and m of them are identified at no extra cost. The m identified elements make up the RSS. Then, measurements on these m elements are made and the needed information is obtained. The information in this carefully selected sample is more than the information in a SRS of m elements. Thus, comparing a RSS of size m with SRS of size m^2 does not make any sense. However, if measurements are made on all m^2 units, then all of them should be used not only the m units.

The mathematical theory of RSS established by Takahasi and Wakimoto (1968). They showed that the mean of the RSS is an unbiased estimator of the population

mean, and has smaller variance than the mean of a SRS. Stokes and Sager (1988) used RSS to estimate distribution functions. They showed that the empirical distribution function (edf) of a RSS is an unbiased estimator of the distribution function and has a smaller variance than that from a SRS.

In the context of statistical hypothesis, Kotia and Babua (1996) derived the exact distribution of the RSS sign test. They showed that the test is more powerful than the counterpart SRS sign test. Liangyong and Xiaofang (2010) proposed the sign test based on RSS for testing hypotheses concerning the quantiles of a population characteristic.

In particular, the two-sample design has been approached by collecting two independent RSS. Several procedures have been developed to make inference on a location shift between two populations. Bohn and Wolfe (1992, 1994) proposed the RSS analogue of the usual two-sample Wilcoxon test and studied its relative properties both under perfect and imperfect judgement. Ozturk (1999) studied the effect of the RSS on two-sample sign test statistic. Ozturk and Wolfe (2000) presented an optimal RSS allocation scheme for a two-sample RSS median test. They derived the exact distribution of the ranked set two-sample median test and tabulated for selected sample and set sizes. For more work on RSS and its variations see Al-Saleh and Al-Omari (2002), Al-Saleh and Samuh (2008), Samuh and Al-Saleh (2011) and Drikvandi et al. (2011).

It is worthwhile to emphasize that when the judgement rankings for obtaining a RSS are done perfectly, the sample consists of independent order statistics from the original underlying distribution of the data. If judgement rankings are not done perfectly, then the cdf of the i^{th} judgement order statistic will no longer be the cdf of the i^{th} order statistic. In this chapter, perfect judgement rankings are assumed. Moreover, the empirical conditional and unconditional power functions of the two-sample RSS permutation test are computed and compared with their counterparts in SRS.

This chapter is organized as follows. The construction of the two-sample RSS design is described in Section 4.2. Permutation test with two proposed test statistics is discussed in Section 4.3. Simulation study that document the benefits of permutation approach of the two-sample RSS is provided in Section 4.4. Real data application is considered in Section 4.5. Finally, Section 4.6 is devoted for concluding remarks.

4.2 Two-Sample Ranked Set Samples

Consider the two samples in which $\mathbf{X}_1 = \{X_{11}, \ldots, X_{1n_1}\}$ are iid $F(x + \boldsymbol{\delta})$ and $\mathbf{X}_2 = \{X_{21}, \ldots, X_{2n_2}\}$ are iid $F(x)$ and the two samples are independent of one another. In the corresponding RSS design, the treatment sample $\mathbf{Y}_t$ of h_1 cycles and m samples is drawn from $F(x + \boldsymbol{\delta})$ and the control sample $\mathbf{Y}_c$ of h_2 cycles and m samples is drawn from $F(x)$. The two samples, $\mathbf{Y}_t$ and $\mathbf{Y}_c$, are independent of one another. The measured data are displayed in Figure 4.2. It is worthwhile to

observe that the data within each column are iid while the data within each row are independent. That is, for each $i = 1, \ldots, m$, $Y_{t1i}, \ldots, Y_{th_1 i}$ are iid $f_{Y_i}(x + \boldsymbol{\delta})$ and $Y_{c1i}, \ldots, Y_{ch_2 i}$ are iid $f_{Y_i}(x)$, where $f_{Y_i}(\cdot)$ is the distribution of the i^{th} order statistic. And for each $j = 1, \ldots, h_1$, $Y_{tj1}, \ldots, Y_{tjm}$ are independent and for each $j' = 1, \ldots, h_2$, $Y_{cj'1}, \ldots, Y_{cj'm}$ are independent.

$$\mathbf{Y}_t = \begin{bmatrix} Y_{t11} & Y_{t12} & \cdots & Y_{t1m} \\ Y_{t21} & Y_{t22} & \cdots & Y_{t2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{th_1 1} & Y_{th_1 2} & \cdots & Y_{th_1 m} \end{bmatrix} \qquad \mathbf{Y}_c = \begin{bmatrix} Y_{c11} & Y_{c12} & \cdots & Y_{c1m} \\ Y_{c21} & Y_{c22} & \cdots & Y_{c2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{ch_2 1} & Y_{ch_2 2} & \cdots & Y_{ch_2 m} \end{bmatrix}$$

Figure 4.2: Two-sample RSS design, $\mathbf{Y}_t$: treatment group and $\mathbf{Y}_c$: control group

4.3 Permutation Test

In this section, permutation approach for testing $H_0 : \{\boldsymbol{\delta} = 0\}$ versus $H_1 : \{\boldsymbol{\delta} > 0\}$ is used. Note that under the null hypothesis, the exchangeability assumption holds within columns and hence exact permutation solution may exist. Permutation should be applied to the data column by column; the first column from $\mathbf{Y}_t$ by the first column from $\mathbf{Y}_c$, the second column from $\mathbf{Y}_t$ by the second column from $\mathbf{Y}_c$, and so forth. In other words, a new matrix $\mathbf{Y}$ of size $(h_1 + h_2) \times m$ is created by concatenating the two matrices $\mathbf{Y}_t$ and $\mathbf{Y}_c$. The permutation $\mathbf{Y}^*$ of $\mathbf{Y} = \mathbf{Y}_t \biguplus \mathbf{Y}_c$ is obtained by permuting the data points within each column of $\mathbf{Y}$ so as to preserve diversity of distributions. The permutation sample space $\mathcal{Y}_{/\mathbf{Y}}$ contains all permutations of $\mathbf{Y}$.

To solve the testing problem, a suitable test statistic $T : \mathcal{X}^n \to \mathcal{R}^1$ should be chosen such that, without loss of generality, large values are evidence against H_0. Two test statistics are proposed. First, the difference between grand means of the two groups; that is,

$$T^1_{RSS} = \bar{Y}_t - \bar{Y}_c,$$

where $\bar{Y}_t = \frac{1}{h_1 m} \sum_i \sum_j Y_{tij}$ and $\bar{Y}_c = \frac{1}{h_2 m} \sum_i \sum_j Y_{cij}$. Second, the sum of the studentized statistics for all columns of the two matrices; that is,

$$T^2_{RSS} = \sum_{i=1}^{m} \left(\frac{\bar{Y}_{ti} - \bar{Y}_{ci}}{\hat{\sigma}_i} \right),$$

where $\hat{\sigma}_i^2 = \frac{1}{h_1 + h_2 - 2} \left[\sum_{j=1}^{h_1} (Y_{tji} - \bar{Y}_{ti})^2 + \sum_{j'=1}^{h_2} (Y_{cj'i} - \bar{Y}_{ci})^2 \right]$, $\bar{Y}_{ti} = \frac{1}{h_1} \sum_{j=1}^{h_1} Y_{tji}$, and $\bar{Y}_{ci} = \frac{1}{h_2} \sum_{j'=1}^{h_2} Y_{cj'i}$.

To obtain the p-value for testing H_0, Algorithm 4.2 is used.

Algorithm 4.2 Two-sample RSS permutation test

1. For the given two-sample RSS, $\mathbf{Y}_t$ and $\mathbf{Y}_c$, calculate the observed test statistic, T^o.
2. Concatenate $\mathbf{Y}_t$ and $\mathbf{Y}_c$ row-wise to get $\mathbf{Y} = \mathbf{Y}_t \biguplus \mathbf{Y}_c$.
3. Take a random permutation $\mathbf{Y}^* \in \mathcal{Y}_{/\mathbf{Y}}$ of $\mathbf{Y}$.
4. Split $\mathbf{Y}^*$ into two matrices such that $\mathbf{Y}_t^*$ containing the same number of rows as in $\mathbf{Y}_t$ and $\mathbf{Y}_c^*$ containing the rest.
5. Calculate the corresponding test statistic, $T^* = T(\mathbf{Y}^*)$.
6. Independently repeat Steps 3 to 5 a large number, say B, of times, giving B test statistics, say $\{T_b^*, b = 1, \ldots, B\}$.
7. The permutation p-value is estimated as

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

4.4 Simulation Study

This section looks at the empirical conditional and unconditional power of the proposed permutation testing procedure under different sampling schemes. The simulation study considers simple random samples and ranked set samples. The power of permutation test based on two-sample RSS with set size m and number of cycle h in each sample is computed and it is compared with the power of permutation test based on another two-sample SRS of size $h \times m$ in each sample. So comparisons are made considering the same numbers of really observed data since in this way costs of two sampling schemes are the same. Moreover, the two proposed test statistics, T_{RSS}^1 and T_{RSS}^2, are also compared.

In the simulation, the set sizes are taken as $m = \{2, 3, 4\}$ and the number of cycles with balanced designs are taken as $h_1 = h_2 = \{5, 10\}$. The nominal level of significance is taken as 0.05. In order to evaluate the empirical power of the test, the treatment groups are shifted by adding the shift parameters $\boldsymbol{\delta} = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The choice $\boldsymbol{\delta} = 0$ to check the empirical level of significance. The empirical power of the test is computed for ranked set and simple random samples conditionally and unconditionally. A simulation study based on 5000 datasets are performed. The considered permutations are $B = 1000$ on each dataset. Moreover, four different probability distributions were considered for the error terms in $\mathbf{Y} = (\mathbf{Z}_t + \boldsymbol{\delta}, \mathbf{Z}_c)$: normal distribution $N(0, 1)$; uniform distribution $U(-\sqrt{3}, \sqrt{3})$; skew normal distribution $SN(0, 1, -5)$; and exponential distribution $Exp(1)$.

4.4.1 Empirical unconditional power

The empirical unconditional power of permutation tests for the aforementioned configurations for a two-sample procedure at the 5% level is reported in Tables 4.1-4.4. It is clear that the unconditional power is improved using RSS. It is increased as m and/or h increased. For fixed total sample size, the power is increased by m much better than by h. For example, consider Table 4.1, for a sample of size $n = 20$ where $h = 5$ and $m = 4$, the power for detecting an effect of size $\boldsymbol{\delta} = 0.4$ using SRS is 0.340 and using RSS (let say, T^1_{RSS}) is 0.612, so the power is improved by 0.272. While if $h = 10$ and $m = 2$, then the power for detecting the same effect ($\boldsymbol{\delta} = 0.4$) using SRS is 0.358 and using T^1_{RSS} is 0.450, so the power is improved only by 0.092. In fact increasing m makes the data more representative to the underlying population (for more details see Samuh and Al-Saleh, 2011). Moreover, the proposed test statistic T^1_{RSS} is more powerful than T^2_{RSS} for symmetric distributions, while T^2_{RSS} is more powerful than T^1_{RSS} for asymmetric distributions.

4.4.2 Empirical conditional power

Considering the same aforementioned configurations, the empirical conditional power is reported in Tables 4.5-4.8. It is clear that the use of RSS does not affect the conditional power, whatever the set size and the number of cycles. Of course this is unstrange because conditional power use the observed dataset irrespective of their underlying distributions. Moreover, the proposed test statistic T^1_{RSS} seems to be more powerful than T^2_{RSS} for symmetric distributions, while T^2_{RSS} is more powerful than T^1_{RSS} for asymmetric distributions.

4.5 Illustration Example

4.5.1 Tawjihi exam 2009/2010 (revisited)

In this example, the empirical conditional and unconditional powers are calculated under SRS and RSS. For two-sample RSS, different set sizes, $m = \{2, 3, 4\}$, and different number of cycles, $h_1 = h_2 = \{5, 10\}$, with balanced designs are considered. For two-sample SRS, a sample of size $m \times h$ is drawn for each sample. Moreover, the two proposed test statistics, T^1_{RSS} and T^2_{RSS}, are considered. The results are reported in Tables 4.9 and 4.10. It is clear that the empirical unconditional power is improved using RSS and the two proposed test statistics, T^1_{RSS} and T^2_{RSS}, have the same level of power (see Figure 4.3(a)). Moreover, powers are increased as m and/or h increased. For fixed total sample size, the power is increased by m much better than by h (see Figure 4.3(b)).

4.6 Concluding Remarks

The effectiveness of RSS for improving the power of the test has been investigated conditionally and unconditionally. Since the conditional power does not require the

information of the underlying populations then it does not improve by the use of RSS. While the unconditional power has a clear improvement. It is recommended to increase the set size than increasing the number of cycles. Moreover, two test statistics are proposed for the RSS. The first proposed statistic is the difference between the two grand means of the two-sample ranked set samples, which is recommended for symmetric distributions. The second proposed statistic is the sum of the studentized statistics of the two-sample ranked set samples and it is recommended for asymmetric distributions.

Table 4.1: Empirical unconditional power, $\alpha = 0.05$, normal distribution

h	m	Sampling design	δ					
			0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.050	0.114	0.216	0.355	0.543	0.704
		T^1_{RSS}	0.050	0.130	0.285	0.481	0.670	0.838
		T^2_{RSS}	0.050	0.129	0.272	0.462	0.645	0.821
	3	SRS	0.052	0.141	0.290	0.473	0.689	0.843
		T^1_{RSS}	0.053	0.187	0.441	0.712	0.908	0.977
		T^2_{RSS}	0.055	0.183	0.427	0.691	0.895	0.973
	4	SRS	0.047	0.166	0.340	0.583	0.800	0.934
		T^1_{RSS}	0.051	0.256	0.612	0.885	0.986	0.999
		T^2_{RSS}	0.053	0.239	0.589	0.869	0.985	0.999
10	2	SRS	0.054	0.154	0.358	0.584	0.796	0.932
		T^1_{RSS}	0.057	0.188	0.450	0.722	0.911	0.978
		T^2_{RSS}	0.054	0.186	0.448	0.712	0.906	0.976
	3	SRS	0.048	0.197	0.456	0.737	0.917	0.982
		T^1_{RSS}	0.053	0.284	0.681	0.939	0.994	0.999
		T^2_{RSS}	0.053	0.286	0.673	0.938	0.994	0.999
	4	SRS	0.050	0.223	0.536	0.856	0.968	0.998
		T^1_{RSS}	0.052	0.397	0.859	0.993	0.999	0.999
		T^2_{RSS}	0.052	0.393	0.857	0.992	0.999	0.999

Table 4.2: Empirical unconditional power, $\alpha = 0.05$, uniform distribution

h	m	Sampling design	δ 0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.051	0.107	0.205	0.352	0.505	0.669
		T^1_{RSS}	0.051	0.137	0.269	0.470	0.666	0.841
		T^2_{RSS}	0.051	0.132	0.260	0.446	0.640	0.811
	3	SRS	0.052	0.132	0.274	0.472	0.681	0.841
		T^1_{RSS}	0.049	0.192	0.444	0.731	0.916	0.986
		T^2_{RSS}	0.047	0.186	0.428	0.704	0.900	0.978
	4	SRS	0.055	0.052	0.149	0.335	0.581	0.799
0.937								
		T^1_{RSS}	0.056	0.252	0.620	0.897	0.989	0.999
		T^2_{RSS}	0.057	0.248	0.617	0.888	0.985	0.999
10	2	SRS	0.056	0.150	0.339	0.575	0.803	0.930
		T^1_{RSS}	0.053	0.188	0.444	0.730	0.914	0.983
		T^2_{RSS}	0.054	0.182	0.432	0.721	0.907	0.981
	3	SRS	0.055	0.179	0.448	0.736	0.928	0.988
		T^1_{RSS}	0.050	0.279	0.701	0.944	0.997	0.999
		T^2_{RSS}	0.052	0.275	0.699	0.939	0.995	0.999
	4	SRS	0.052	0.221	0.538	0.855	0.976	0.998
		T^1_{RSS}	0.050	0.404	0.878	0.993	0.999	0.999
		T^2_{RSS}	0.054	0.403	0.880	0.993	0.999	0.999

Table 4.3: Empirical unconditional power, $\alpha = 0.05$, skew normal distribution

h	m	Sampling design	δ 0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.045	0.176	0.406	0.675	0.873	0.963
		T^1_{RSS}	0.049	0.212	0.519	0.797	0.949	0.991
		T^2_{RSS}	0.052	0.217	0.540	0.823	0.963	0.995
	3	SRS	0.054	0.230	0.545	0.827	0.960	0.996
		T^1_{RSS}	0.050	0.334	0.769	0.963	0.997	0.999
		T^2_{RSS}	0.050	0.352	0.810	0.979	0.999	0.999
	4	SRS	0.053	0.263	0.640	0.910	0.989	0.999
		T^1_{RSS}	0.052	0.451	0.911	0.998	0.999	0.999
		T^2_{RSS}	0.053	0.497	0.949	0.999	0.999	0.999
10	2	SRS	0.052	0.281	0.638	0.910	0.986	0.999
		T^1_{RSS}	0.053	0.323	0.775	0.974	0.999	0.999
		T^2_{RSS}	0.055	0.348	0.804	0.984	0.999	0.999
	3	SRS	0.052	0.337	0.801	0.979	0.999	0.999
		T^1_{RSS}	0.054	0.517	0.953	0.999	0.999	0.999
		T^2_{RSS}	0.053	0.573	0.974	0.999	0.999	0.999
	4	SRS	0.052	0.415	0.885	0.996	0.999	0.999
		T^1_{RSS}	0.050	0.693	0.996	0.999	0.999	0.999
		T^2_{RSS}	0.050	0.772	0.999	0.999	0.999	0.999

Table 4.4: Empirical unconditional power, $\alpha = 0.05$, exponential distribution

h	m	Sampling design	δ 0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.049	0.133	0.270	0.422	0.584	0.726
		T^1_{RSS}	0.055	0.142	0.298	0.507	0.683	0.815
		T^2_{RSS}	0.053	0.177	0.412	0.648	0.824	0.923
	3	SRS	0.053	0.156	0.325	0.530	0.717	0.852
		T^1_{RSS}	0.051	0.199	0.443	0.690	0.864	0.940
		T^2_{RSS}	0.046	0.299	0.685	0.912	0.982	0.999
	4	SRS	0.050	0.166	0.378	0.619	0.802	0.922
		T^1_{RSS}	0.054	0.251	0.580	0.830	0.954	0.986
		T^2_{RSS}	0.048	0.450	0.872	0.988	0.999	0.999
10	2	SRS	0.051	0.170	0.384	0.615	0.806	0.921
		T^1_{RSS}	0.059	0.193	0.477	0.717	0.891	0.964
		T^2_{RSS}	0.052	0.256	0.624	0.868	0.967	0.994
	3	SRS	0.056	0.202	0.473	0.750	0.911	0.978
		T^1_{RSS}	0.051	0.274	0.644	0.901	0.984	0.998
		T^2_{RSS}	0.048	0.455	0.888	0.994	0.999	0.999
	4	SRS	0.055	0.238	0.566	0.848	0.969	0.994
		T^1_{RSS}	0.042	0.358	0.799	0.971	0.997	0.999
		T^2_{RSS}	0.047	0.657	0.986	0.999	0.999	0.999

Table 4.5: Empirical conditional power, $\alpha = 0.05$, normal distribution

h	m	Sampling design	δ 0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.047	0.107	0.180	0.299	0.441	0.613
		T^1_{RSS}	0.052	0.105	0.189	0.309	0.446	0.610
		T^2_{RSS}	0.051	0.099	0.174	0.275	0.408	0.565
	3	SRS	0.056	0.132	0.285	0.501	0.718	0.875
		T^1_{RSS}	0.050	0.144	0.305	0.540	0.761	0.913
		T^2_{RSS}	0.052	0.138	0.294	0.514	0.736	0.891
	4	SRS	0.052	0.164	0.375	0.641	0.867	0.969
		T^1_{RSS}	0.045	0.167	0.349	0.606	0.830	0.953
		T^2_{RSS}	0.043	0.156	0.321	0.565	0.794	0.927
10	2	SRS	0.052	0.172	0.418	0.714	0.905	0.989
		T^1_{RSS}	0.057	0.178	0.420	0.693	0.898	0.981
		T^2_{RSS}	0.054	0.172	0.414	0.689	0.891	0.978
	3	SRS	0.057	0.186	0.472	0.773	0.936	0.989
		T^1_{RSS}	0.049	0.177	0.424	0.699	0.902	0.981
		T^2_{RSS}	0.050	0.174	0.414	0.685	0.890	0.977
	4	SRS	0.055	0.206	0.520	0.822	0.961	0.997
		T^1_{RSS}	0.052	0.234	0.594	0.891	0.988	1.000
		T^2_{RSS}	0.050	0.228	0.579	0.877	0.987	1.000

Table 4.6: Empirical conditional power, $\alpha = 0.05$, uniform distribution

h	m	Sampling design	δ 0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.051	0.115	0.223	0.374	0.545	0.736
		T^1_{RSS}	0.055	0.106	0.203	0.338	0.515	0.680
		T^2_{RSS}	0.056	0.098	0.181	0.289	0.467	0.627
	3	SRS	0.050	0.138	0.290	0.501	0.723	0.895
		T^1_{RSS}	0.051	0.135	0.307	0.512	0.737	0.887
		T^2_{RSS}	0.055	0.123	0.268	0.449	0.673	0.843
	4	SRS	0.052	0.150	0.324	0.563	0.782	0.922
		T^1_{RSS}	0.054	0.148	0.320	0.538	0.758	0.917
		T^2_{RSS}	0.056	0.136	0.281	0.471	0.690	0.869
10	2	SRS	0.056	0.179	0.426	0.731	0.926	0.991
		T^1_{RSS}	0.047	0.134	0.300	0.518	0.731	0.885
		T^2_{RSS}	0.048	0.129	0.292	0.498	0.709	0.869
	3	SRS	0.046	0.184	0.450	0.732	0.926	0.988
		T^1_{RSS}	0.050	0.194	0.470	0.785	0.946	0.994
		T^2_{RSS}	0.049	0.186	0.453	0.765	0.938	0.990
	4	SRS	0.055	0.226	0.556	0.841	0.975	0.998
		T^1_{RSS}	0.051	0.208	0.542	0.832	0.967	0.996
		T^2_{RSS}	0.053	0.196	0.520	0.811	0.960	0.995

Table 4.7: Empirical conditional power, $\alpha = 0.05$, skew normal distribution

h	m	Sampling design	δ 0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.048	0.221	0.557	0.864	0.991	1.000
		T^1_{RSS}	0.050	0.195	0.440	0.768	0.947	0.999
		T^2_{RSS}	0.052	0.179	0.410	0.725	0.927	0.998
	3	SRS	0.055	0.174	0.384	0.651	0.855	0.975
		T^1_{RSS}	0.051	0.170	0.433	0.717	0.925	0.992
		T^2_{RSS}	0.050	0.170	0.426	0.702	0.913	0.988
	4	SRS	0.049	0.253	0.615	0.907	0.995	1.000
		T^1_{RSS}	0.051	0.330	0.772	0.980	1.000	1.000
		T^2_{RSS}	0.050	0.348	0.807	0.985	1.000	1.000
10	2	SRS	0.050	0.318	0.759	0.977	1.000	1.000
		T^1_{RSS}	0.052	0.256	0.634	0.930	0.997	1.000
		T^2_{RSS}	0.054	0.256	0.636	0.925	0.996	1.000
	3	SRS	0.045	0.380	0.860	0.996	1.000	1.000
		T^1_{RSS}	0.056	0.297	0.732	0.960	1.000	1.000
		T^2_{RSS}	0.053	0.294	0.727	0.958	1.000	1.000
	4	SRS	0.057	0.364	0.826	0.993	1.000	1.000
		T^1_{RSS}	0.057	0.387	0.865	0.996	1.000	1.000
		T^2_{RSS}	0.060	0.369	0.848	0.993	1.000	1.000

Table 4.8: Empirical conditional power, $\alpha = 0.05$, exponential distribution

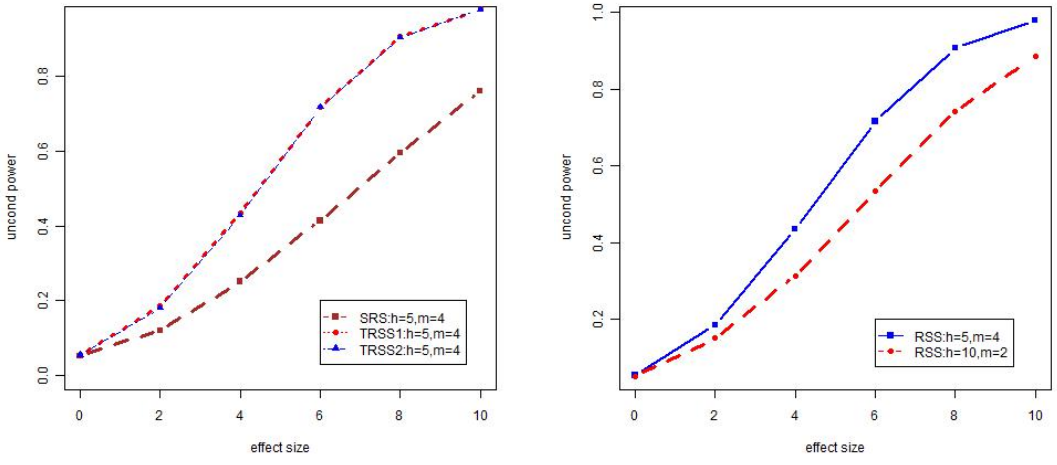
h	m	Sampling design	δ 0.00	0.20	0.40	0.60	0.80	1.00
5	2	SRS	0.055	0.107	0.194	0.302	0.436	0.569
		T^1_{RSS}	0.049	0.102	0.170	0.244	0.355	0.475
		T^2_{RSS}	0.047	0.103	0.197	0.296	0.429	0.565
	3	SRS	0.055	0.149	0.323	0.563	0.795	0.931
		T^1_{RSS}	0.054	0.170	0.416	0.712	0.911	0.988
		T^2_{RSS}	0.053	0.166	0.411	0.702	0.898	0.985
	4	SRS	0.048	0.133	0.275	0.451	0.648	0.829
		T^1_{RSS}	0.052	0.144	0.307	0.513	0.716	0.881
		T^2_{RSS}	0.051	0.171	0.376	0.627	0.842	0.953
10	2	SRS	0.049	0.149	0.342	0.592	0.802	0.941
		T^1_{RSS}	0.051	0.175	0.409	0.689	0.899	0.978
		T^2_{RSS}	0.052	0.174	0.411	0.680	0.899	0.977
	3	SRS	0.048	0.218	0.560	0.844	0.978	0.998
		T^1_{RSS}	0.046	0.158	0.369	0.624	0.847	0.953
		T^2_{RSS}	0.047	0.171	0.421	0.690	0.885	0.971
	4	SRS	0.051	0.302	0.753	0.973	1.000	1.000
		T^1_{RSS}	0.049	0.238	0.556	0.855	0.979	0.999
		T^2_{RSS}	0.050	0.256	0.592	0.878	0.984	0.999

Table 4.9: Tawjihi Example: Empirical unconditional power, $\alpha = 0.05$

h	m	Sampling design	δ 0.00	2.00	4.00	6.00	8.00	10.00
5	2	SRS	0.056	0.103	0.165	0.257	0.366	0.488
		T^1_{RSS}	0.052	0.111	0.196	0.317	0.469	0.628
		T^2_{RSS}	0.055	0.110	0.195	0.312	0.463	0.617
	3	SRS	0.048	0.105	0.217	0.342	0.490	0.646
		T^1_{RSS}	0.051	0.149	0.307	0.523	0.749	0.892
		T^2_{RSS}	0.056	0.148	0.308	0.522	0.741	0.884
	4	SRS	0.052	0.121	0.251	0.414	0.595	0.760
		T^1_{RSS}	0.054	0.186	0.435	0.717	0.907	0.980
		T^2_{RSS}	0.055	0.181	0.428	0.718	0.904	0.979
10	2	SRS	0.056	0.129	0.238	0.398	0.606	0.768
		T^1_{RSS}	0.053	0.151	0.314	0.534	0.743	0.885
		T^2_{RSS}	0.055	0.152	0.320	0.538	0.739	0.885
	3	SRS	0.053	0.149	0.324	0.539	0.743	0.893
		T^1_{RSS}	0.049	0.206	0.507	0.796	0.951	0.993
		T^2_{RSS}	0.053	0.206	0.511	0.807	0.953	0.993
	4	SRS	0.054	0.169	0.381	0.652	0.852	0.957
		T^1_{RSS}	0.053	0.282	0.687	0.937	0.996	1.000
		T^2_{RSS}	0.049	0.287	0.692	0.941	0.997	1.000

Table 4.10: Tawjihi Example: Empirical conditional power, $\alpha = 0.05$

h	m	Sampling design	δ 0.00	2.00	4.00	6.00	8.00	10.00
5	2	SRS	0.052	0.081	0.145	0.228	0.300	0.404
		T^1_{RSS}	0.048	0.091	0.136	0.202	0.405	0.527
		T^2_{RSS}	0.052	0.086	0.123	0.180	0.355	0.475
	3	SRS	0.048	0.104	0.184	0.282	0.513	0.684
		T^1_{RSS}	0.055	0.119	0.220	0.402	0.542	0.701
		T^2_{RSS}	0.054	0.110	0.201	0.357	0.499	0.653
	4	SRS	0.057	0.118	0.203	0.337	0.623	0.798
		T^1_{RSS}	0.052	0.121	0.225	0.365	0.613	0.803
		T^2_{RSS}	0.052	0.115	0.199	0.322	0.553	0.742
10	2	SRS	0.053	0.115	0.243	0.420	0.589	0.774
		T^1_{RSS}	0.049	0.122	0.212	0.377	0.597	0.766
		T^2_{RSS}	0.049	0.119	0.211	0.357	0.582	0.755
	3	SRS	0.056	0.141	0.271	0.451	0.736	0.885
		T^1_{RSS}	0.049	0.154	0.316	0.535	0.798	0.937
		T^2_{RSS}	0.047	0.153	0.301	0.516	0.784	0.923
	4	SRS	0.053	0.161	0.369	0.623	0.861	0.962
		T^1_{RSS}	0.053	0.142	0.411	0.674	0.803	0.932
		T^2_{RSS}	0.053	0.139	0.398	0.651	0.779	0.918



(a) (b)

Figure 4.3: Tawjihi 2009/2010: Unconditional power (a) SRS versus RSS. (b) RSS: $h = 5$ and $m = 4$ versus $h = 10$ and $m = 2$.

Chapter 5

Tests for Variance Components in Linear Mixed Models

Contents

Standard asymptotic χ^2 distribution of the likelihood ratio statistic under the null hypothesis does not hold when the parameter value is on the boundary of the parameter space. In mixed models, it is of interest to test for a zero random effect variance component. Some available tests for the variance component are reviewed and a new test within the permutation framework is presented. The unconditional power and level of significance of the different tests are investigated by means of a Monte Carlo simulation study.

5.1 Introduction

Mixed models (e.g. Verbeke and Molenberghs, 2000), hierarchical models (e.g. Raudenbush and Bryk, 2002) or multilevel regression models (e.g. Snijders and Bosker, 1999) are an extension of regression models in which data have a hierarchical structure with units nested in clusters. A common application is on individuals nested in institutions or organizations (e.g. students in schools, employees in firms, or patients in hospitals). Another kind of application is on repeated measures where measurement occasions are nested in individuals.

Mixed models are widely used in many research fields such as social sciences (Afshartous and de Leeuw, 2004), econometrics (Swamy, 1970) and political science (Garner and Raudenbush, 1991).

To facilitate calculations and clarify ideas, the simplest case of linear mixed models, random intercept model, involving two levels of analysis is considered as a guide. Level one units are referred to as *subjects* and level two units as *clusters*. A model with one level-1 predictor, which is observable and has a linear relationship with the level-1 dependent variable, is considered.

Let the random variable Y_{ij} denote the response of interest for the i^{th} subject in the j^{th} cluster, X_{ij} denote the related observed covariate, β_1 is a fixed parameter or regression coefficient, γ_{0j} is the cluster intercept, β_0 is the average intercept across the clusters, ε_{ij} is the level-1 residual, and ξ_j is the level-2 residual. The level-1 model, which relates the response variable to the covariate, is written as

$$Y_{ij} = \gamma_{0j} + \beta_1 X_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n_j;\ j = 1, \ldots, J, \tag{5.1}$$

while the level-2 model, describing the variation between clusters, is written as

$$\gamma_{0j} = \beta_0 + \xi_j, \quad j = 1, \ldots, J. \tag{5.2}$$

Combining Equations 5.1 and 5.2 into a single equation gives one that looks like a common regression equation with an extra error term ξ_j. This error term indicates that the mean intercepts can randomly differ across clusters. The combined model is written as

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \xi_j + \varepsilon_{ij}, \quad i = 1, \ldots, n_j;\ j = 1, \ldots, J. \tag{5.3}$$

For fixed X_{ij}, the essential assumptions for the random intercept model are that:

1. ξ_j are iid normal with mean $\mathbb{E}(\xi_j) = 0$ and variance $\mathbb{V}(\xi_j) = \sigma^2_\xi$;
2. ε_{ij} are iid normal with mean $\mathbb{E}(\varepsilon_{ij}) = 0$ and variance $\mathbb{V}(\varepsilon_{ij}) = \sigma^2_\varepsilon$;
3. ξ_j and ε_{ij} are independent.

It is of interest to test whether the random effects should be included in the model. This is equivalent to testing if the between-cluster σ^2_ξ is zero. That is,

$$H_0 : \{\sigma^2_\xi = 0\} \quad \text{versus} \quad H_1 : \{\sigma^2_\xi > 0\}. \tag{5.4}$$

This problem is nonstandard because the parameter value under H_0 is on the boundary of the parameter space $[0, \infty)$. Therefore, the likelihood ratio and score statistics no longer have the standard asymptotic χ^2 distribution (Self and Liang, 1987; Stram and Lee, 1994; Verbeke and Molenberghs, 2003).

This chapter is organized as follows. Likelihood ratio tests and their asymptotic distributions are reviewed in Section 5.2. Simulation-based tests (exact likelihood ratio tests, parametric bootstrap tests and permutation tests) are reviewed in Section 5.3. A new permutation test is proposed in Section 5.4. Simulation study that document the benefits of the new permutation test is provided in Section 5.5. Concluding remarks are contained in Section 5.6.

5.2 Likelihood Ratio Tests

Suppose we wish to test

$$H_0 : \{\sigma_\xi^2 \in \Theta_0\} \quad \text{versus} \quad H_1 : \{\sigma_\xi^2 \in \Theta_1\}, \quad \Theta = \Theta_0 \cup \Theta_1.$$

Let $\ell_{\Theta_0}^{ML}$ and ℓ_{Θ}^{ML} be the log likelihood functions maximised over Θ_0 and Θ, respectively. Then the likelihood ratio test (LRT) statistic is given by

$$LRT = -2\left[\ell_{\Theta_0}^{ML} - \ell_{\Theta}^{ML}\right].$$

Using the restricted likelihood functions, the restricted likelihood ratio test ($RLRT$) statistic is given by

$$RLRT = -2\left[\ell_{\Theta_0}^{REML} - \ell_{\Theta}^{REML}\right].$$

It follows from the classical likelihood theory (see e.g. Pace and Salvan, 1997, Sec. 3.4) that under some regularity conditions LRT and $RLRT$ follow, asymptotically under H_0, a χ^2 distribution with degrees of freedom equal to the difference between the number of parameters in Θ and Θ_0. One of the regularity conditions under which the χ^2 approximation is valid is that the parameter value under the null hypothesis is not on the boundary of the parameter space Θ, such as in hypothesis 5.4. Self and Liang (1987) and Stram and Lee (1994) showed that the LRT statistic in this case has an asymptotic null distribution that is a mixture of χ_0^2 and χ_1^2 distributions, each having an equal weight of 0.5. χ_0^2 denotes the distribution with all probability mass at zero, so the correct p-value is obtained by halving the p-value obtained from the χ_1^2 distribution. This result also applies for $RLRT$, as shown by Morrell (1998) (see also Verbeke and Molenberghs, 2000).

5.3 Simulation-Based Tests in the Literature

5.3.1 Finite sample distribution of LRT and $RLRT$

In linear mixed models with one variance component, finite sample distributions of the LRT and $RLRT$ are derived by Crainiceanu and Ruppert (2004). They considered the spectral representations of the LRT and $RLRT$ as the basis of efficient simulation algorithms of their null distributions. They provided an algorithm for simulating the null finite distribution of LRT (and $RLRT$). For more details, see Crainiceanu and Ruppert (2004), page 168.

Crainiceanu and Ruppert's algorithm is implemented in R by Scheipl (2010) in the package "RLRsim". The Function "exactLRT" is used for finite sample LRT, and "exactRLRT" for finite sample $RLRT$.

In R, the function "lmer" in the package "lme4" produced by Bates (2010) can be used to fit the linear mixed models. It is worthwhile to observe that the "exactLRT" function is not working properly with "lmer" function. This is due to some modifications done on "lmer" function after Scheipl has been implemented his package.

5.3.2 Parametric bootstrap tests

A parametric bootstrap test (Efron and Tibshirani, 1993; Davison and Hinkley, 1997) for variance components is proposed by Sinha (2009) in generalized linear mixed models based on the score test (Silvapulle and Silvapulle, 1995). Via simulation, Sinha (2009) showed that the empirical level of significance of the parametric bootstrap test is much closer to the nominal level and it is more powerful than the usual asymptotic score test based on a mixture of χ^2 distributions. Bootstrap tests are more commonly based on LRT or $RLRT$ (see Faraway, 2006, Sec. 8.4).

To obtain a parametric bootstrap estimate of the LRT statistic's p-value, Algorithm 5.1 is used.

Algorithm 5.1 Parametric Bootstrap Method

1. For the given dataset, calculate the LRT statistic, denoted by LRT^o.
2. Generate a bootstrap sample from the model under H_0 and calculate the corresponding bootstrap LRT^* statistic.
3. Independently repeat Step 2 a large number, say B, of times, giving B test statistics, say $\{LRT_b^*, b = 1, \ldots, B\}$.
4. The bootstrap p-value is obtained as the proportion of samples with LRT_b^* greater than or equal to LRT^o.

5.3.3 Permutation tests

Fitzmaurice et al. (2007) proposed a permutation test for variance components in generalized linear mixed models based on the LRT statistic. Their results are compared with the asymptotic $50:50$ χ^2 distribution of the LRT and with the LRT distribution proposed by Crainiceanu and Ruppert (2004). The proposed permutation test has the correct nominal level under the null hypothesis, and it is more powerful than the usual tests based on a mixture of χ^2 distributions. Although their results were obtained for the case of LRT, the same procedure can be used for $RLRT$.

Algorithm 5.2 is used for obtaining a permutation estimate of the LRT statistic's p-value.

5.4 A New Permutation Test

Fitzmaurice et al. (2007) considered the LRT as a test statistic in their algorithm and this require the underlying distribution to be known. In this section, a new permutation algorithm is proposed which does not require any knowledge of the underlying distribution.

Algorithm 5.2 Fitzmaurice et al. (2007) Approach

1. For the given dataset, calculate the LRT statistic, denoted by LRT^o.
2. Randomly permute the cluster indices while maintaining a fixed number of subjects within a cluster and calculate the corresponding permutation LRT^* statistic.
3. Independently repeat Step 2 a large number, say B, of times, giving B test statistics, say $\{LRT_b^*,\, b = 1, \ldots, B\}$.
4. The permutation p-value is obtained as the proportion of samples with LRT_b^* greater than or equal to LRT^o.

Let us consider the random intercept model (5.3), repeated here as a guide:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \xi_j + \varepsilon_{ij}, \quad i = 1, \ldots, n_j; j = 1, \ldots, J.$$

Normality assumptions for the random error components are not required. The hypotheses of interest are given by

$$H_0 : \{\sigma_\xi^2 = 0\} \quad \text{versus} \quad H_1 : \{\sigma_\xi^2 > 0\}.$$

Under H_1, the cluster-specific regression lines have different intercepts but the same slope. The testing problem can be treated as permutation ANOVA by removing the effect of the covariate(s). To this end, the least square estimators of β_0 and β_1 under H_0 are computed then the empirical deviates $R_{ij} = Y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{ij}$ are obtained. The R_{ij} are exchangeable, so the resulting problem is equivalent to permutation ANOVA. In terms of the population deviates $(\xi_j + \varepsilon_{ij})$, the testing problem is:

$$H_0 : \{\xi_1 = \cdots = \xi_J\} \equiv \{\sigma_\xi^2 = 0\} \quad \text{versus} \quad H_1 : \{H_0 \text{ is false}\}.$$

The usual F-test statistic is

$$F = \frac{N - J}{J - 1} \frac{\sum_{j=1}^{J} n_j (\bar{R}_j - \bar{R})^2}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (R_{ij} - \bar{R}_j)^2}, \tag{5.5}$$

where $\bar{R}_j = \frac{1}{n_j} \sum_i R_{ij}$ and $\bar{R} = \frac{1}{N} \sum_j n_j \bar{R}_j$. The F-statistic (5.5) is permutationally equivalent to the following T-statistic (see Pesarin and Salmaso, 2010, Sec. 2.4)

$$T = \sum_{j=1}^{J} n_j \bar{R}_j^2.$$

Steps for obtaining a conditional Monte Carlo estimate of the permutation p-value are summarized in Algorithm 5.3.

It is worthwhile to observe that the least square estimators of β_0 and β_1 and hence the empirical deviates R_{ij} are derived only once, which make our proposed algorithm a bit faster than others.

Algorithm 5.3 A New Permutation Test Approach

1. For the given dataset, under H_0, compute the least square estimates of β_0 and β_1 and calculate the empirical deviates $R_{ij} = Y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{ij}$.
2. Calculate the observed test statistic, T^o.
3. Randomly permute the cluster indices while maintaining the same number of subjects within a cluster and calculate the corresponding test statistic, T^*.
4. Independently repeat Step 3 many times, say B times, giving B test statistics, say $\{T_b^*, b = 1, \ldots, B\}$.
5. The permutation p-value is obtained as the proportion of samples with T_b^* greater than or equal to T^o.

5.5 Simulation Study

A simulation study is conducted to assess the level of significance and the power of the proposed permutation test for variance components and to compare it with the aforementioned available tests. In the simulation, different number of clusters, $J = \{10, 50\}$, and different number of observations within a cluster, $n_j = n = \{5, 25, 100\}$, $j = 1, \ldots, J$ (balanced designs), are considered. Several other combinations are performed, not reported here, and the results follow the same behavior. A simulation study based on 2000 datasets are performed. The permutation and the bootstrap are based on $B = 500$ replications. Moreover, $\sigma_\xi^2 = 0$ is chosen to examine the level of significance of the tests, and $\sigma_\xi^2 = \{0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.60, 0.80, 1.00\}$ are chosen to investigate the power behavior. The nominal level of significance was set to $\alpha = 0.05$. In the simulation, the model in equation (5.3) is considered, where $\xi_j \sim N(0, \sigma_\xi^2)$, $\varepsilon_{ij} \sim N(0, 1)$, $X_{ij} \sim N(0, 1)$, $\beta_0 = 0$ and $\beta_1 = 1$.

In the following, *LRT* is abbreviated for the likelihood ratio approach $(0.5\chi_0^2 + 0.5\chi_1^2)$, *ERLRT* for the finite sample restricted likelihood ratio approach, *Boot* for the parametric bootstrap approach, *Fitz* for Fitzmaurice et al. (2007) approach and *PT* for the proposed permutation approach.

The execution times taken for a single computation of each test, using a PC with a single CPU and considering a design where $n = 100$ and $J = 50$, are reported in Table 5.1. Of course, the *LRT* and *ERLRT* methods are faster than the others because they do not require resampling process. The proposed permutation test *PT* is largely the fastest among the resampling tests.

The empirical level of significance for all the tests are reported in Table 5.2. The empirical level of significance of the bootstrap approach in the simulation configurations is between 0.049 and 0.055, which is much closer to the nominal 0.05 level than the other tests. Our proposed *PT* is the second preferable test in terms of

empirical level of significance.

To investigate the power of the proposed permutation test, some configurations are reported in Table 5.4. It is clear that *PT* is more powerful than the *LRT* and *ERLRT* methods and it is a good competitor of the *Boot* and *Fitz* methods.

One configuration with an unbalanced design is investigated, $J = 10$ clusters with average cluster size equal to 25 (half cluster of size 10 and half clusters of size 40). The empirical level of significance and power of the tests are reported in Table 5.3. The power of the *LRT* method is the worst. The *PT* method is a good competitor of the *ERLRT*, *Boot* and *Fitz*. In addition, *Boot* and *Fitz* have an empirical level of significance much closer to the nominal level than the others.

The power of the proposed permutation test when the distributions of the random error components are misspecified is investigated. Specifically, the model of Equation 5.3 is considered but a gamma distribution is assumed for the random error components ξ_j and ε_{ij}; i.e. $\xi_j = \sigma_\xi(\xi_j^* - 1)$ where ξ_j^* is distributed as gamma with location and scale parameters equal to 1. A similar distribution is used to generate the errors ε_{ij}. The empirical level of significance of the tests are reported in Table 5.5. The proposed permutation test *PT* and the *Boot* test have an empirical level of significance between 0.045 and 0.051 which are much closer to the nominal level than the other tests. In terms of power, some configurations are reported in Table 5.6. The proposed *PT* is more powerful than the *LRT* and *ERLRT* and it is a very good competitor of the *Boot* and *Fitz* methods.

5.6 Concluding Remarks

To test variance components in a linear mixed model with balanced design, the proposed permutation test has a level of significance close to the nominal level and more powerful than the tests based on the 50 : 50 mixture χ^2 distributions and the approximate exact restricted likelihood ratio method given by Crainiceanu and Ruppert (2004). In terms of speed, the proposed permutation test is the fastest method among the resampling-based methods. This is due to the way of obtaining the distribution of the test statistic; the proposed permutation approach requires the fitted model under the null hypothesis only once, while the other algorithms require the fitted model under at least the null hypothesis for every iteration. The proposed permutation test is also fully nonparametric while the other approaches rely on distributional assumptions.

With unbalanced designs, the proposed permutation test still has a level of significance close to the nominal level and it is more powerful than the likelihood ratio test based on the 50 : 50 mixture χ^2 distribution and the approximate exact restricted likelihood ratio method. It is worthwhile to observe that all tests discussed in this chapter are more powerful for the balanced designs than the unbalanced.

When the distributions of the model errors are misspecified all the tests under consideration loose power. Also in this case, the three resampling-based tests, which have similar performances, are clearly preferable to the standard *LRT* and the

ERLRT.

Table 5.1: Times (in seconds) for a single computation of the tests calculated using a PC with a single CPU, considering a design where $n = 100$ and $J = 50$.

Test	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
Time	0.18	0.25	35.00	38.00	0.30

Table 5.2: Empirical level of significance from the simulation study of balanced designs, nominal level $\alpha = 5\%$

(J, n)	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
$(10, 5)$	0.031	0.036	0.050	0.046	0.051
$(10, 25)$	0.026	0.043	0.051	0.046	0.046
$(10, 100)$	0.023	0.047	0.049	0.049	0.051
$(50, 5)$	0.036	0.043	0.049	0.039	0.053
$(50, 25)$	0.038	0.050	0.055	0.055	0.052
$(50, 100)$	0.036	0.048	0.050	0.050	0.051

Table 5.3: Empirical power from the simulation study of unbalanced design, $J = 10$, $n_1 = \cdots = n_5 = 10$ and $n_6 = \cdots = n_{10} = 40$, nominal level $\alpha = 0.05$

σ_ξ^2	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
0.00	0.023	0.046	0.050	0.050	0.055
0.05	0.451	0.560	0.568	0.566	0.524
0.10	0.735	0.810	0.811	0.812	0.804
0.15	0.872	0.922	0.923	0.922	0.924
0.20	0.919	0.948	0.948	0.948	0.953
0.30	0.975	0.981	0.982	0.980	0.987
0.40	0.988	0.991	0.992	0.991	0.992
0.60	0.998	0.999	0.999	0.999	0.999
0.80	0.997	0.999	0.999	0.999	0.999
1.00	0.999	0.999	0.999	0.999	0.999

Table 5.4: Empirical power from the simulation study of balanced designs, nominal level $\alpha = 0.05$

(J, n)	σ_ξ^2	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
$(10, 5)$	0.05	0.086	0.098	0.119	0.124	0.128
	0.10	0.161	0.186	0.216	0.219	0.219
	0.15	0.246	0.278	0.313	0.315	0.308
	0.20	0.318	0.348	0.400	0.396	0.402
	0.30	0.490	0.532	0.575	0.580	0.562
	0.40	0.608	0.648	0.688	0.678	0.680
	0.60	0.779	0.802	0.834	0.835	0.822
	0.80	0.882	0.898	0.910	0.912	0.915
	1.00	0.935	0.945	0.953	0.951	0.953
$(10, 25)$	0.05	0.480	0.588	0.591	0.590	0.591
	0.10	0.776	0.834	0.837	0.833	0.834
	0.15	0.912	0.934	0.937	0.937	0.940
	0.20	0.946	0.962	0.964	0.964	0.964
	0.30	0.989	0.992	0.992	0.992	0.992
	0.40	0.996	0.999	0.999	0.999	0.999
	0.60	0.998	0.999	0.999	0.999	0.999
	0.80	0.999	0.999	0.999	0.999	0.999
	1.00	0.999	0.999	0.999	0.999	0.999
$(50, 5)$	0.05	0.252	0.275	0.289	0.262	0.298
	0.10	0.540	0.570	0.586	0.552	0.584
	0.15	0.780	0.793	0.811	0.785	0.807
	0.20	0.913	0.918	0.927	0.918	0.930
	0.30	0.987	0.988	0.991	0.987	0.991
	0.40	0.998	0.998	0.998	0.998	0.999
	0.60	0.999	0.999	0.999	0.999	0.999
	0.80	0.999	0.999	0.999	0.999	0.999
	1.00	0.999	0.999	0.999	0.999	0.999

Table 5.5: Empirical level of significance from the simulation study when error components follow a gamma distribution, nominal level $\alpha = 0.05$

(J, n)	*LRT*	*ERLRT*	*Boot*	*Fitz*	*PT*
$(10, 5)$	0.026	0.035	0.049	0.047	0.052
$(10, 25)$	0.025	0.048	0.051	0.050	0.051
$(10, 100)$	0.022	0.047	0.048	0.052	0.050
$(50, 5)$	0.039	0.042	0.050	0.041	0.048
$(50, 25)$	0.035	0.045	0.051	0.050	0.047
$(50, 100)$	0.029	0.045	0.045	0.048	0.045

Table 5.6: Empirical power from the simulation study when error components follow a gamma distribution, nominal level $\alpha = 0.05$

(J, n)	σ_ξ^2	LRT	$ERLRT$	$Boot$	$Fitz$	PT
$(10, 5)$	0.05	0.079	0.105	0.127	0.134	0.131
	0.10	0.170	0.204	0.236	0.244	0.232
	0.15	0.235	0.262	0.299	0.306	0.313
	0.20	0.330	0.370	0.406	0.405	0.417
	0.30	0.432	0.465	0.502	0.514	0.510
	0.40	0.541	0.566	0.590	0.594	0.599
	0.60	0.680	0.702	0.728	0.734	0.738
	0.80	0.762	0.785	0.808	0.812	0.810
	1.00	0.830	0.845	0.866	0.865	0.863
$(10, 25)$	0.05	0.412	0.496	0.503	0.507	0.505
	0.10	0.662	0.732	0.739	0.734	0.735
	0.15	0.778	0.827	0.830	0.831	0.832
	0.20	0.871	0.898	0.901	0.903	0.902
	0.30	0.936	0.953	0.955	0.956	0.955
	0.40	0.957	0.972	0.972	0.974	0.972
	0.60	0.982	0.986	0.987	0.988	0.987
	0.80	0.992	0.993	0.993	0.992	0.993
	1.00	0.995	0.996	0.997	0.997	0.997
$(50, 5)$	0.05	0.234	0.252	0.272	0.241	0.269
	0.10	0.498	0.529	0.551	0.514	0.549
	0.15	0.737	0.754	0.768	0.746	0.766
	0.20	0.852	0.866	0.870	0.859	0.869
	0.30	0.964	0.969	0.970	0.967	0.969
	0.40	0.983	0.985	0.988	0.985	0.989
	0.60	0.997	0.998	0.999	0.998	0.998
	0.80	0.999	0.999	0.999	0.999	0.999
	1.00	0.999	0.999	0.999	0.999	0.999

Chapter 6

Tests for Random Agreement in Cluster Analysis

Contents

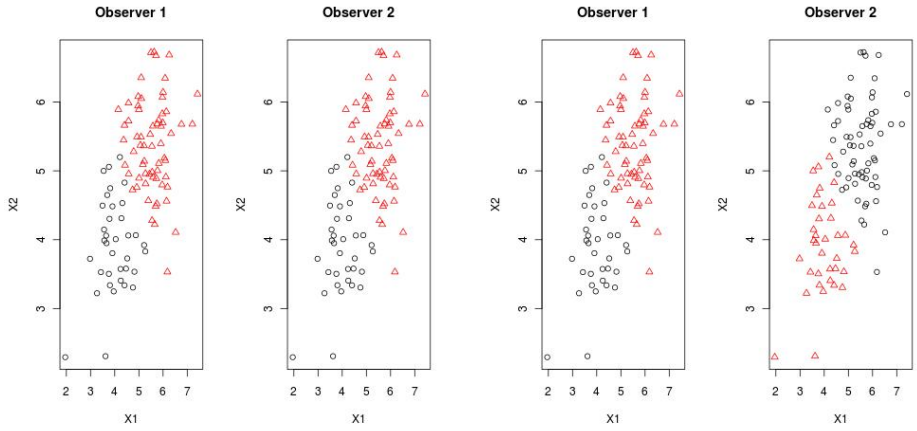
The adjusted Rand index is a measure of similarity or agreement between two clusterings for the same dataset. It is calculated based on counting pairs of points and comparing the *agreement* and the *disagreement* between the two clusterings or two classification rules. In this chapter, the adjusted Rand index is suggested as a test statistic for testing the null hypothesis of random agreement.

6.1 Introduction

Measuring the similarity between two clusterings (two sets of clusters) for the same dataset have received strong interest in the literature. This is due to the existence of many different clustering algorithms (Kaufman and Rousseeuw, 1990; Theodoridis and Koutroumbas, 2006) or different observers may use the same clustering algorithm but different starting points which yield different clusterings (Brennan and Light, 1974). Therefore, measuring the similarity (agreement) is one of the fundamental techniques in the cluster analysis field.

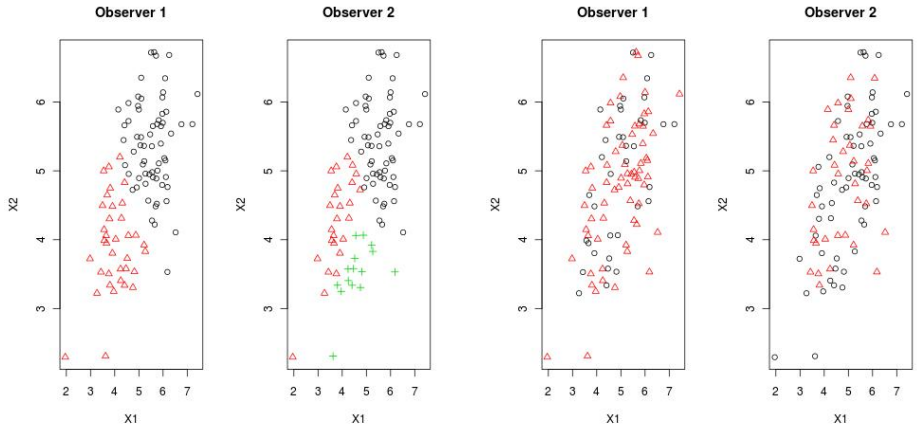
In order to clarify ideas and to avoid misunderstanding of what we mean by the similarity or agreement between two clusterings, it is helpful to refer to an example. Suppose two observers are asked independently to cluster or to partition a dataset into several clusters, so we have two clusterings. The specific criterion for partitioning is left up to each observer. Thus the number of clusters within each clustering could be different. Moreover, each observer may use different labels

for his clusters. An important question to be asked is whether the two observers agree or disagree. For example, consider a two-dimensional dataset of size 100. In Figure 6.1(a) the two observers agree completely. In Figure 6.1(b) they also agree completely although different labels are used. There is a strong agreement in Figure 6.1(c) although different number of clusters are used. Finally, Figure 6.1(d) depicts a random agreement. Note that the random agreement occurred when each of the observers partition the dataset into clusters randomly.



(a) Perfect agreement, $ARI = 1$

(b) Perfect agreement, $ARI = 1$



(c) Strong agreement, $ARI = 0.75$

(d) Random agreement, $ARI \approx 0$

Figure 6.1: The agreement between two clusterings of a dataset obtained independently by two different observers

It is worthwhile to observe that the problem of measuring agreement between two (or more) observers, given that the categories or the cluster labels are predefined and imposed on observers, is investigated in the literature. Cohen (1960) introduced the coefficient kappa to measure the degree of agreement between two observers who cluster the observations among the predefined categories. This measure has been

extended to three or more observers by Light (1971) and Fleiss (1971). See also Cohen (1968), Everitt (1968) and Fleiss et al. (1969).

The problem considered in this chapter is somewhat different. The two observers are asked to cluster the observations into several clusters. The specific criterion for clustering is left up to each observer. Thus the two observers may develop different number of clusters. Moreover, since no precise set of clusters have been labelled in advance, each observer may use different criteria resulting in categories with different labels.

A large number of agreement measures have been proposed in the literature, which can be classified into three groups:

1. Pair counting measures: which are based on counting pairs of points and comparing the *agreement* and the *disagreement* between two clusterings. Jaccard index (Jaccard, 1901), Rand index (Rand, 1971), Folkes and Mallows index (Fowlkes and Mallows, 1983) and adjusted Rand index (Hubert and Arabie, 1985) are examples of this group of measures.

2. Set matching measures: which are based on measuring the shared set cardinality between two clusterings. F-measures (Rijsbergen, 1979) and misclassification rate (Meilă, 2005) are examples of this group of measures.

3. Information theoretic measures: which are based on the conditional probabilities resulting from the number of points shared between clusters of the two clusterings. Mutual information (Strehl and Ghosh, 2003) and variation of information (Meilă, 2005) are examples of this group of measures.

For more details see Hubalek (1982), Albatineh et al. (2006), Milligan and Cooper (1986) and Warrens (2008a,b).

Few publications are found in the literature concerning distributional properties of agreement measures. Janson and Vegelius (1981) derived the mean and the variance of Jaccard index. McCormick et al. (1992) derived the exact distribution of Jaccard index assuming an underlying multinomial distribution with all categories equally likely except one. Hubert and Arabie (1985) derived the mean of the Rand index under the hypergeometric distribution assumption. Fowlkes and Mallows (1983) derived the mean and variance for Rand index. Albatineh (2010) generalized the derivation of Fowlkes and Mallows (1983) for the mean and the variance to a large number of similarity measures. Finally, Shuweihdi and Taylor (2007) showed that the Rand index is linearly related to the Pearson statistic given that the cluster sizes (i.e. the number of observations within each cluster) within each clustering are equal.

In this chapter, the ARI is used as a test statistic for testing the null hypothesis of random agreement. The concept of the ARI and its properties are reviewed in Section 6.2. Tests for the null hypothesis of random agreement using χ^2 distribution and permutation approaches are investigated in Section 6.3. Simulation study to investigate the empirical level of significance is carried out in Section 6.4. Finally, concluding remarks are contained in Section 6.5.

6.2 Adjusted Rand Index

6.2.1 Definition and notation

Consider a dataset with n items denoted by $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Let $\mathcal{U}$ with r clusters and $\mathcal{V}$ with c clusters are two clusterings to be compared. $\mathcal{U}$ and $\mathcal{V}$ are obtained independently by two observers, same observer but in different occasions or different starting points, or by applying two different clustering algorithms. The information on the overlap between $\mathcal{U}$ and $\mathcal{V}$ can be summarized by considering one of the following representations.

- **Representation 1** Each clustering is represented by a string of symbols containing the cluster labels of the corresponding data points. For example, $\mathcal{U} = \{u_1, u_1, u_3, u_4, u_4, \ldots\}$ and $\mathcal{V} = \{v_3, v_3, v_1, v_2, v_4, \ldots\}$ means the first data point $\mathbf{X}_1$ is labeled by u_1 in clustering $\mathcal{U}$ whereas it is labeled by v_3 in clustering $\mathcal{V}$, and so on.

- **Representation 2** Let $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_c\}$, where $\mathbf{u}_i$ is the set of all data points clustered into the i^{th} cluster, $i = 1, \ldots, r$, by $\mathcal{U}$, and $\mathbf{v}_j$ is the set of all data points clustered into the j^{th} cluster, $j = 1, \ldots, c$, by $\mathcal{V}$. Then the information on cluster overlap between $\mathcal{U}$ and $\mathcal{V}$ can be summarized in the form of a $r \times c$ contingency table as illustrated in Table 6.1, where n_{ij} is the number of items classified into cluster $\mathbf{u}_i$ according to $\mathcal{U}$ and into cluster $\mathbf{v}_j$ according to $\mathcal{V}$. The cluster sizes in the two clusterings are the row and column totals of the contingency table given by $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$.

Table 6.1: Two-way contingency table

		$\mathcal{V}$				
		$\mathbf{v}_1$	$\mathbf{v}_2$	$\ldots$	$\mathbf{v}_c$	n_{i+}
$\mathcal{U}$	$\mathbf{u}_1$	n_{11}	n_{12}	$\ldots$	n_{1c}	n_{1+}
	$\mathbf{u}_2$	n_{21}	n_{22}	$\ldots$	n_{2c}	n_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	$\mathbf{u}_r$	n_{r1}	n_{r2}	$\ldots$	n_{rc}	n_{r+}
	n_{+j}	n_{+1}	n_{+2}	$\ldots$	n_{+c}	n

- **Representation 3** Any pair of data points from the total of $N = \binom{n}{2}$ different pairs in the dataset $\mathbf{X}$ falls into one of the following four types of pairs:

  1. N_{11}: the number of pairs that are in the same cluster in both $\mathcal{U}$ and $\mathcal{V}$;
  2. N_{00}: the number of pairs that are in different clusters in both $\mathcal{U}$ and $\mathcal{V}$;
  3. N_{01}: the number of pairs that are in the same cluster in $\mathcal{U}$ but in different clusters in $\mathcal{V}$;
  4. N_{10}: the number of pairs that are in different clusters in $\mathcal{U}$ but in the same cluster in $\mathcal{V}$.

These quantities can be calculated using the n_{ij}'s (Hubert and Arabie, 1985). Intuitively, N_{00} and N_{11} are typically interpreted as agreements in the classification of the items whereas N_{01} and N_{10} represent disagreements. The information on cluster overlap between $\mathcal{U}$ and $\mathcal{V}$ can be summarized in the form of a 2×2 contingency table as illustrated in Table 6.2.

Table 6.2: 2×2 contingency table

$\mathcal{U} \downarrow \quad \mathcal{V} \rightarrow$	Pairs in same cluster	Pairs in different clusters
Pairs in same cluster	N_{11}	N_{01}
Pairs in different clusters	N_{10}	N_{00}

The Rand index (Rand, 1971) is simply defined as the probability of agreement:

$$RI = \frac{N_{00} + N_{11}}{N}.$$

The Rand index lies between 0 and 1. It takes the value of 1 when the two clusterings are identical and 0 when the two clusterings have no agreement. In fact, the latter happens if and only if one clustering consists of a single cluster and the other only of clusters containing single points. However as can be seen, the unique case where $RI = 0$ is quite extreme and has little practical value. In most situations the Rand index often lies within the narrower range of $[0.5, 1]$. Therefore, the Rand index possibly gives high values to pairs of randomly generated clusterings, e.g. 0.5, and this baseline value does not take on the same value in different scenarios. In fact, it is desirable for the similarity measure between two random clusterings to take values close to zero, or at least a constant value. A further problem with the Rand index is that its expected value between two random clusterings does not even take a constant value. Hubert and Arabie (1985), by taking the generalized hypergeometric distribution as the model of randomness, i.e. the two clusterings are picked at random subject to having the original number of classes and objects in each, found the expected value for $N_{00} + N_{11}$. They suggested using a corrected version of the Rand index of the form:

$$Adjusted\_Index = \frac{Index - \mathbb{E}(Index)}{Max(Index) - \mathbb{E}(Index)}$$

thus giving rise to the adjusted Rand index given by:

$$ARI(\mathcal{U}, \mathcal{V}) = \frac{\sum_i \sum_j \binom{n_{ij}}{2} - \sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}{0.5 \left(\sum_i \binom{n_{i+}}{2} + \sum_j \binom{n_{+j}}{2} \right) - \sum_i \binom{n_{i+}}{2} \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}. \tag{6.1}$$

The ARI is bounded above by 1 and takes on the value 0 when the index equals its expected value (under the generalized hypergeometric distribution assumption for randomness). For more details see Hubert and Arabie (1985); Yeung and Ruzzo (2001).

Using **Representation 3**, Warrens (2008b) showed that the *ARI* can be rewritten as follows:

$$ARI(\mathcal{U},\mathcal{V}) = \frac{2(N_{11}N_{00} - N_{01}N_{10})}{(N_{11}+N_{01})(N_{00}+N_{01}) + (N_{00}+N_{10})(N_{10}+N_{11})}.$$

Albatineh et al. (2006) introduced a family of similarity measures which can be written in the form $\beta_0 + \beta_1 \sum_i \sum_j n_{ij}^2$, where β_0 and β_1 are unique for each measure. The *ARI* can be written by the same way. By the use of Equation 6.1, after simple algebra, the *ARI* is written in the following form:

$$ARI(\mathcal{U},\mathcal{V}) = \beta_0 + \beta_1 \sum_i \sum_j n_{ij}^2, \tag{6.2}$$

where

$$\beta_0 = \frac{-n - \frac{PQ}{n(n-1)}}{0.5(P+Q) - \frac{PQ}{n(n-1)}}$$

and

$$\beta_1 = \frac{1}{0.5(P+Q) - \frac{PQ}{n(n-1)}}$$

with $P = \sum_i n_{i+}^2 - n$ and $Q = \sum_j n_{+j}^2 - n$.

6.2.2 *ARI* and Pearson statistic

Let the totals within each marginal are equal, that is,

$$n_{i+} = \frac{n}{r}, \forall i = 1, \ldots, r \tag{6.3}$$

and

$$n_{+j} = \frac{n}{c}, \forall j = 1, \ldots, c. \tag{6.4}$$

Shuweihdi and Taylor (2007) showed that the Rand index is linearly related with the Pearson statistic. By the same way, the relationship between *ARI* and Pearson statistic can be derived. The Pearson statistic is given by

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}.$$

Under restrictions 6.3 and 6.4, the Pearson statistic becomes

$$X^2 = \frac{rc}{n} \sum_i \sum_j n_{ij}^2 - n.$$

Therefore, after simple algebra,

$$ARI = \gamma_0 + \gamma_1 X^2, \tag{6.5}$$

where $\gamma_0 = \frac{c+r-rc-1}{d}$ and $\gamma_1 = \frac{n-1}{nd}$ with $d = 0.5nc - rc + 0.5c + 0.5nr - n + 0.5r$.

6.3 Tests for Random Agreement

Consider two independent clusterings $\mathcal{U}$ and $\mathcal{V}$. The hypotheses of interest are given by

$$H_0 : \{\text{There is a random agreement between } \mathcal{U} \text{ and } \mathcal{V}\}$$

and

$$H_1 : \{\mathcal{U} \text{ and } \mathcal{V} \text{ are not random}\}.$$

Performing the test based on the statistic ARI requires the knowledge of its probability distribution under the null hypothesis which is tedious to find in closed form. To overcome this problem, two approaches are proposed; χ^2 distribution approach (Section 6.3.1) and permutation approach (Section 6.3.2).

6.3.1 χ^2 distribution approach

When the clusterings $\mathcal{U}$ and $\mathcal{V}$ have equal cluster sizes, it is shown in Section 6.2.2 that the ARI can be written as a linear function with Pearson statistic (see Equation 6.5).

Since X^2 has an asymptotic χ^2 distribution with $\nu = (r-1)(c-1)$ degrees of freedom, then the probability distribution of ARI is given by

$$f_{ARI}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)\gamma_1}\left(\frac{x-\gamma_0}{\gamma_1}\right)^{\nu/2-1}\exp\left\{\frac{-(x-\gamma_0)}{2\gamma_1}\right\}, \text{ where } x \geq \gamma_0.$$

with mean

$$\mathbb{E}\left(ARI(\mathcal{U},\mathcal{V})\right) = \gamma_0 + \gamma_1\nu,$$

and variance

$$\mathbb{V}\left(ARI(\mathcal{U},\mathcal{V})\right) = 2\nu\gamma_1^2.$$

To test the null hypothesis of random agreement, the following test statistic is used.

$$X^2_{ARI}(\mathcal{U},\mathcal{V}) = \frac{ARI - \gamma_0}{\gamma_1},$$

which has an asymptotic χ^2 distribution with $\nu = (r-1)(c-1)$ degrees of freedom. Therefore, the p-value is given by

$$\lambda_1 = 1 - F_{X^2}(X^{2o}_{ARI}) = \int_{X^{2o}_{ARI}}^{\infty} f_{ARI}(x)\,dx,$$

where X^{2o}_{ARI} is the observed test statistic and $F_{X^2}(\cdot)$ is the cdf of χ^2 distribution.

The size of the test has the correct nominal level α in the sense that $\int_{X^2_\alpha}^{\infty} f_{ARI}(x)\,dx = \alpha$.

6.3.2 Permutation approach

χ^2 distribution approach, discussed in Section 6.3.1, is valid when the cluster sizes within each clustering are equal and the expected number of cells is greater than 5. In practice, these restrictions are hard to attain. Therefore, an alternative approach is required. In this section, a permutation test is proposed.

The goal of using permutation method is the computation of the conditional probability distribution of the ARI. For the purpose of finding the permutation sample space, **Representation 1** of the two clusterings (discussed in Section 6.2.1) is considered. The cluster labels within each clustering are permuted then ARI is calculated using $\mathcal{U}^*$ and $\mathcal{V}^*$. Algorithm 6.1 is used to obtain the permutation (conditional) p-value for testing the null hypothesis of random agreement.

Algorithm 6.1 Conditional p-value of the ARI

1. For the given two clusterings $\mathcal{U}$ and $\mathcal{V}$, calculate the observed test statistic $ARI(\mathcal{U}, \mathcal{V})$, denoted by ARI^o.
2. Take a random permutation $\mathcal{U}^*$ of $\mathcal{U}$ and $\mathcal{V}^*$ of $\mathcal{V}$.
3. Calculate the test statistic $ARI^* = ARI(\mathcal{U}^*, \mathcal{V}^*)$.
4. Independently repeat Steps 2 and 3 many times, say B times, giving B test statistics, say $\{ARI_b^*, b = 1, \ldots, B\}$.
5. The permutation mid p-value is estimated as

$$\lambda_2 = \frac{\sum_{b=1}^{B} \mathbb{I}(ARI_b^* > ARI^o)}{B} + \frac{\sum_{b=1}^{B} \mathbb{I}(ARI_b^* = ARI^o)}{2B}.$$

Note that the permutation mid p-value (Lancaster, 1961) is calculated due to the discreteness of the permutation distribution of the test statistic.

6.4 Simulation Study

In this section, the empirical level of significance of the proposed tests is investigated.

To assess the empirical level of significance, the tests are performed on a two random clusterings. A random clustering can be created by assigning data points to clusters randomly. As an example, two clusterings each with three categories ($r = c = 3$) are created under the null hypothesis and three different configurations are considered: (a) $n_{i+} = 50, \forall i = 1, 2, 3$ and $n_{+j} = 50, \forall j = 1, 2, 3$; (b) $n_{1+} = n_{+1} = 5$, $n_{i+} = 50$, $i = 2, 3$ and $n_{+j} = 50$, $j = 2, 3$; (c) $n_{1+} = 5$, $n_{2+} = 3$, $n_{3+} = 7$ and $n_{+1} = 1$, $n_{+2} = 10$, $n_{+3} = 4$. Steps for assessing the empirical significance level are summarized in Algorithm 6.2. A simulation study based on $R = 5000$ datasets are performed. The considered permutations on each dataset are $B = 1000$.

Algorithm 6.2 Empirical level of significance

1. For the given dataset, randomly create two clusterings $\mathcal{U}$ and $\mathcal{V}$.
2. Use the aforementioned approaches to obtain the p-values, λ_1 and λ_2.
3. Independently repeat Steps 1 and 2 many times, say R times, giving R p-values for each approach, say $\{\lambda_{ir}, r = 1, \ldots, R\}$, $i = 1, 2$.
4. For a preassigned nominal level of significance α, the empirical level of significance is given by
$$\hat{\alpha}_i = \frac{\sum_{r=1}^{R} \mathbb{I}(\lambda_{ir} \leq \alpha)}{R}, \; i = 1, 2.$$

The simulation results are reported in Tables 6.3-6.5 for each configuration. It is clear that the empirical level of significance for the proposed tests in configuration (a) is closed to the nominal one; that is, the p-values under the null hypothesis are uniformly distributed over its support, $[0, 1]$. While in configurations (b) and (c) the proposed permutation test is still valid but not the χ^2 distribution.

Table 6.3: The empirical level of significance, $n_{i+} = 50$, $\forall i = 1, 2, 3$ and $n_{+j} = 50$, $\forall j = 1, 2, 3$

	Nominal level α						
Method	0.05	0.10	0.20	0.40	0.60	0.80	0.90
χ^2 distribution	0.049	0.104	0.215	0.427	0.604	0.813	0.906
permutation	0.051	0.105	0.208	0.410	0.600	0.805	0.905

Table 6.4: The empirical level of significance, $n_{1+} = n_{+1} = 5$, $n_{i+} = 50$, $i = 2, 3$ and $n_{+j} = 50$, $j = 2, 3$

	Nominal level α						
Method	0.05	0.10	0.20	0.40	0.60	0.80	0.90
χ^2 distribution	0.049	0.103	0.184	0.409	0.550	0.804	0.999
permutation	0.049	0.098	0.200	0.408	0.596	0.800	0.898

Table 6.5: The empirical level of significance, $n_{1+} = 5$, $n_{2+} = 3$, $n_{3+} = 7$ and $n_{+1} = 1$, $n_{+2} = 10$, $n_{+3} = 4$

	Nominal level α						
Method	0.05	0.10	0.20	0.40	0.60	0.80	0.90
χ^2 distribution	0.040	0.049	0.182	0.4100	0.828	0.999	0.999
permutation	0.048	0.103	0.190	0.4100	0.575	0.828	0.871

6.5 Concluding Remarks

Testing for random agreement for two clusterings of a dataset is investigated in this chapter. The adjusted Rand index is proposed as a test statistic. Two proposed methods are discussed; the first one is based on the χ^2 distribution by the use of the relationship between Pearson statistic and the adjusted Rand index; the second one is based on the permutation approach. Comparison between these proposed methods is carried out in terms of empirical level of significance.

Appendix A

Perspectives of Future Work

I would be most interested in continuing to work and to extend some approaches discussed in this thesis.

In Chapter 2, the power functions of permutation tests (conditional and unconditional) are defined for two-sample design for one-sided alternatives. It is of interest to extend these definitions to two-sided alternatives, one-sample, and $k > 2$-sample designs also with categorical variables and in multidimensional settings. Moreover, the power functions are defined for fixed effects and extension to random effects can be provided.

In Chapter 3, some applications of empirical conditional power function are investigated. It is of interest to extend these applications for bioequivalence and non-inferiority testing problems (see, for example Wellek, 2010).

In Chapter 4, two-sample permutation design is studied with ranked set sampling for perfect ranking. It is of interest to study different permutation designs (such as paired and ANOVA designs) with ranked set sampling and imperfect ranking may also be considered. Moreover, it is of interest to study the use of permutation tests with multistage ranked set sampling (Al-Saleh and Al-Omari, 2002) and to check the effectiveness of the number of stages on the power of the test.

In Chapter 5, permutation tests in linear mixed models is proposed for one variance component and the random intercept model is considered as a guide. It is of interest to study the use of permutation tests for more than one variance component.

In Chapter 6, tests for random agreement are investigated for a two different clusterings created for the same dataset. It is of interest to study these tests when the two clusterings are created for two different datasets. It is found in the literature a measure of similarity called ADCO proposed by Bae et al. (2010) which could be considered as a test statistic for the null hypothesis of random agreement.

Appendix B

Curriculum Vitae – MONJED SAMUH

Personal Details

Date of Birth: January 16, 1980
Place of Birth: Hebron, Palestine
Nationality: Palestinian

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.
Tel. +39 049 827 4174
e-mail: monjed@stat.unipd.it

Current Position

Since January 2009; (expected completion: December 2011)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Some Advances in Permutation Testing
Supervisor: Prof. FORTUNATO PESARIN
Co-supervisor: Prof. FRIEDRICH LEISCH.

Permanent Address

Since February 2006
Full-time lecturer in Statistics
College of Applied Sciences - Palestine Polytechnic University
Wadi Al-Hariyah Building No. A
Hebron - Palestine
P. O. Box 198
e-mail: mhstat@ppu.edu
Homepage: staff.ppu.edu/mhstat

Research interests

- Permutation Tests
- Ranked Set Sampling

Education

September 2003 – January 2006
Master degree (*laurea specialistica/magistrale*) **in Statistics**.
Yarmouk University, Irbid - Jordan
Title of dissertation: "On Multistage Ranked Set Sampling with Application to Distribution and Median Estimation"
Supervisor: Prof. Mohammad Fraiwan AlSaleh
Final mark: 91.8%

September 1998 – June 2002
Bachelor degree (*four years*) **in Applied Mathematics**.
Palestine Polytechnic University, Hebron - Palestine
Title of dissertation: "Hilbert Spaces"
Supervisor: Prof. Ibrahim Al-Masri
Final mark: 81.4%.

Visiting periods

March 2011 – June 2011
University of Natural Resources and Applied Life Sciences, Vienna - Austria .
Supervisor: Prof. Friedrich Leisch

Work experience

February 2006 – December 2008
Full-time Lecturer.
Palestine Polytechnic University.

September 2006 – January 2007
Part-time Lecturer.
Al-Quds Open University.

September 2005 – January 2006
Teaching Assistant.
Yarmouk University.

September 2002 – January 2003
Teaching Assistant.
Palestine Polytechnic University.

Awards and Scholarship

January 2009 - December 2011
PhD Scholarship: Fondazione Cassa di Risparmio di Padova e Rovigo (CARIPARO).

September 2003 - August 2005
Master Scholarship: Saudi Committee for the Relief of the Palestinian People Under the Cooperation of the Palestinian Ministry of Higher Education.

September 2000, February 2001 and February 2002
Dean Honors List.

Computer skills

- Programming Languages and Statistical Packages: SPSS, Minitab, R, C++.
- Operating Systems: Windows, Linux, DOS.
- Other Packages: LaTeX, MATLAB, Mathematica, Scientific Workplace

Language skills

Arabic: native; English: Good; Italian: Slight; French: Slight.

Publications

Articles in journals
Samuh, M., Al-Saleh, M. F. (2011). The effectiveness of multistage ranked set sampling in stratifying the population. *Communications in Statistics - Theory and Methods* **40**, 1063–1080.

Al-Saleh, M. F., Samuh, M. (2008). On multistage ranked set sampling for distribution and median estimation. *Computational Statistics & Data Analysis* **52**, 2066–2078.

Grilli, L., Rampichini, C., Salmaso, L., Lunardon, N., Samuh, M. (2011). The use of permutation tests for variance components in linear mixed models. *Communications in Statistics - Theory and Methods* **to be appear**.

Conference presentations

Samuh, M. (2011). Tests for random agreement in cluster analysis (poster) *The European Researchers Night in Veneto (Venetonight 2011)*, Padova, Italy, September 23, 2011.

Samuh, M. (2011). Permutation tests with ranked set sampling (accepted talk) *7th Conference on Statistical Computation and Complex Systems (SCo 2011)*, Padova, Italy, September 19-21, 2011.

Samuh, M. (2010). Empirical post hoc conditional power function (accepted talk) *Palestinian Conference on Modern Trends in Mathematics and Physics II (PCMTMP II)*, An-Najah National University, Palestine, August 2-4, 2010.

Samuh, M. (2010). Conditional power function: background, planning and use (accepted talk) *International Symposium on Business and Industrial Statistics (ISBIS 2010)*, Portoroz, Slovenia, July 5-9, 2010.

Samuh, M. (2010). A review of diagnostic tests in multilevel models (poster) *Statistics for complex problems: the multivariate permutation approach and related topics in honor of the 70th birthday of Fortunato Pesarin*, Padova, Italy, June 14-15, 2010.

Samuh, M. (2008). The effectiveness of multistage ranked set sampling in stratifying the population (accepted talk) *Palestinian Conference on Modern Trends in Mathematics and Physics I (PCMTMP I)*, Birzeit, Palestine, July 28-30, 2008.

Samuh, M. (2008). (presence) *The 8th German Open Conference in Probability and Statistics (GOCPS 2008)*, Aachen, Germany, March 4-7, 2008.

Samuh, M. (2008). The effectiveness of multistage ranked set sampling in stratifying the population (accepted talk) *8th International Conference on Ordered Statistical Data and Its Applications (OSDA 2008)*, Aachen, Germany, March 7-8, 2008.

Samuh, M. (2007). Multistage ranked set sampling as a tool of data reduction for huge datasets (accepted talk) *7th International Conference on Ordered Statistical Data and Inequalities (OSDI 2007)*, Amman, Jordan, June 12-14, 2007.

Samuh, M. (2000). (presence) *The 3rd International Palestinian Conference on Mathematics and Mathematics Education (IPCM 2000)*, Bethlehem, Palestine, August 2000.

Teaching experience

February 2006 – December 2008
Introduction to Statistics, Probability Theory, Regression Analysis, Variance Analysis, Sampling Theory, Probability and Statistics for Engineers, Statistical Lab 1, Statistical Lab 2, Statistical Lab 3
Full-time Lecturer
Palestine Polytechnic University

September 2005 – January 2006

Statistical Lab 1, Statistical Lab 2
Graduate Teaching Assistant
Yarmouk University

September 2002 – January 2003
Statistical Lab 1
Undergraduate Teaching Assistant
Palestine Polytechnic University

References

Prof. Ibrahim Al-Masri
Palestine Polytechnic University
P. O. Box 198, Hebron - Palestine
Phone: 00972 2 2233050
e-mail: imasri@ppu.edu

Prof. Mohammad Fraiwan Al-Saleh
Yarmouk University
P. O. Box 566, 21163 Irbid - Jordan
Phone: 00962 2 7211111
e-mail: m-saleh@yu.edu.jo

Prof. Alessandra Salvan
Padova University
Via C. Battisti, 241, 35121 Padova - Italy
Phone: 0039 049 8274139
e-mail: salvan@stat.unipd.it

Prof. Fortunato Pesarin
Padova University
Via C. Battisti, 241, 35121 Padova - Italy
Phone: 0039 049 8274143
e-mail: pesarin@stat.unipd.it

Prof. Amjad D. Al-Nasser
University of Dubai
Maktoom Road, Al Masaood Building, P.O.Box 14143, Dubai
Phone: 00971 4 2072656
e-mail: amjadyu@yahoo.com

Bibliography

Afshartous, D. and de Leeuw, J. (2004). An application of multilevel model prediction to NELS:88. *Behaviormetrika*, 31:43–66. 43

Al-Saleh, M. F. and Al-Omari, A. (2002). Multistage ranked set sampling. *Journal of Statistical Planning and Inference*, 102:273–286. 32, 63

Al-Saleh, M. F. and Samuh, M. H. (2008). On multistage ranked set sampling for distribution and median estimation. *Computational Statistics & Data Analysis*, 52:2066–2078. 32

Albatineh, A. N. (2010). Means and variances for a family of similarity indices used in cluster analysis. *Journal of Statistical Planning and Inference*, 140:2828–2838. 55

Albatineh, A. N., Niewiadomska-Bugaj, M., and Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, 23:301–313. 55, 58

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43:75–88. 2

Bae, E., Bailey, J., and Dong, G. (2010). A clustering comparison measure using density profiles and its application to the discovery of alternate clusterings. *Data Mining and Knowledge Discovery*, 21:427–471. 63

Barnett, V. and Moore, K. (1997). Best linear unbiased estimates in ranked-set sampling with particular reference to imperfect ordering. *Journal of Applied Statistics*, 24:697–710. 30

Barton, D. E. (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. 18

Basso, D., Pesarin, F., Salmaso, L., and Solari, A. (2009). *Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications in R*. Springer, New York. 6

Bates, D. M. (2010). *lme4: Mixed-Effects Modeling with R*. Springer, New York. 45

Bell, C. B., Moser, J. M., and Thompson, R. (1966). Goodness criteria for two-sample distribution-free tests. *The Annals of Mathematical Statistics*, 37:133–142. 18

Bohn, L. L. and Wolfe, D. A. (1992). Nonparametric two-sample procedures for ranked-set samples data. *Journal of the American Statistical Association*, 87:552–561. 32

Bohn, L. L. and Wolfe, D. A. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association*, 89:168–176. 32

Box, G. E. P. and Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumptions. *Journal of the Royal Statistical Society*, 17:1–34. 10

Box, G. E. P. and Tiao, G. C. (1964). A note on criterion robustness and inference robustness. *Biometrika*, 51:169–173. 1

Brennan, R. L. and Light, R. J. (1974). Measuring agreement when two observers classify people into categories not defined in advance. *British Journal of Mathematical and Statistical Psychology*, 27:154–163. 53

Brewer, J. K. and Sindelar, P. T. (1988). Adequate sample size: A priori and post hoc considerations. *The Journal of Special Education*, 21:74–84. 21

Chen, Z., Bai, Z., and Sinha, B. K. (2004). *Ranked Set Sampling: Theory and Applications*. Springer-Verlag, New York. 30

Chow, S.-C. and Liu, J.-P. (2004). *Design and Analysis of Clinical Trials: Concepts and Methodologies, 2nd Edition*. Wiley-Blackwell, New York. 19

Chow, S.-C., Shao, J., and 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. 19

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46. 54

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220. 55

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey. 11, 18

Collier, R. O. and Baker, F. B. (1966). Some Monte Carlo results on the power of the F-test under permutation in the simple randomized block design. *Biometrika*, 53:199–203. 10

Collings, B. J. and Hamilton, M. A. (1988). Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics*, 44:847–860. 18, 19

Cooper, H. and Hedges, L. V. (1997). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York. 11

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society*, 66:165–185. 45, 46, 49

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics, 3rd Edition*. Wiley, New York. 31

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, UK. 46

De Martini, D. (2002). Pointwise estimate of the power and sample size determination for permutation tests. *Statistica*, 62:779–790. 20

De Martini, D. (2008). Reproducibility probability estimation for testing statistical hypotheses. *Statistics & Probability Letters*, 78:1056–1061. 2

Dell, J. R. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28:545–553. 29

Dixon, W. J. (1954). Power under normality of several nonparametric tests. *The Annals of Mathematical Statistics*, 25:610–614. 18

Drikvandi, R., Modarres, R., and Jalilian, A. H. (2011). A bootstrap test for symmetry based on ranked set samples. *Computational Statistics & Data Analysis*, 55:1807–1814. 32

Edgington, E. S. (1995). *Randomization Tests, 3rd Edition*. Marcel Dekker, New York. 6

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York. 46

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. 18

Everitt, B. S. (1968). Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 21:97–103. 55

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall, New York. 46

Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh. 1, 5

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh. 1, 5

Fitzmaurice, G. M., Lipsitz, S. R., and Ibrahim, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics*, 63:942–946. 2, 46, 47, 48

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382. 55

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72:323–327. 55

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78:553–569. 55

Garner, C. L. and Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education*, 64:251–262. 43

Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses, 3rd Edition*. Springer-Verlag, New York. 6

Goodman, S. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, 11:875–879. 2, 21

Hallahan, M. and Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy*, 34:489–499. 17

Hamilton, M. A. and Collings, B. J. (1991). Determining the appropriate sample size for nonparametric tests for location shift. *Technometrics*, 33:327–337. 19

Haynam, G. E. and Govindarajulu, Z. (1966). Exact power of the Mann-Whitney test for exponential and rectangular alternatives. *The Annals of Mathematical Statistics*, 37:945–953. 18

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York. 11

Hemelrijk, J. (1961). Experimental comparison of Student's and Wilcoxon's two sample test. *Quantitative Methods in Pharmacology*, pages 118–133. 18

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23:169–192. 6

Hubalek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews*, 57:669–689. 55

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218. 55, 57

Jaccard, P. (1901). Étude comparative de la distribution orale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579. 3, 55

Janson, S. and Vegelius, J. (1981). Measures of ecological association. *Oecologia*, 49:371–376. 55

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York. 53

Kempthorne, O., Zyskind, G., Addleman, S., Throckmorton, T., and White, R. (1961). Analysis of variance procedures. Technical report, Aeronautical Research Laboratory 149, Wright-Patterson Air Force Base, Ohio. 10

Kotia, K. M. and Babua, G. J. (1996). Sign test for ranked-set sampling. *Communications in Statistics - Theory and Methods*, 25:1617–1630. 32

Kraemer, H. C. and Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Sage Publications, Newbury Park, CA. 17

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56:223–234. 60

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses, 3rd Edition*. Springer, New York. 5

Lehmann, E. L. and Stein, C. (1949). On the theory of some non-parametric hypotheses. *Annals of Mathematical Statistics*, 20:28–45. 6

Lenth, R. V. (2007). Post hoc power: Tables and commentary. Technical Report 378, The University of Iowa - Department of Statistics and Actuarial Science. 21

Levine, M. and Ensom, M. H. H. (2001). Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy*, 21:405–409. 21

Liangyong, Z. and Xiaofang, X. (2010). Optimal ranked set sampling design for the sign test. *Chinese Journal of Applied Probability and Statistics*, 26:225–233. 32

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76:365–377. 55

Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Sage Publications, Newbury Park, CA. 17

Markowski, E. P. and Markowski, C. A. (1999). Practical uses of statistical power in business research studies. *Journal of Education for Business*, 75:122–125. 18

McCormick, W. P., Lyons, N. I., and Hutcheson, K. (1992). Distributional properties of Jaccard's index of similarity. *Communication in Statistics - Theory and Methods*, 21:51–68. 55

McHugh, R. B. (1961). Confidence interval inference and sample size determination. *The American Statistician*, 15:14–17. 18

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, 3:385–390. 29

McIntyre, G. (2005). A method for unbiased selective sampling, using ranked sets. *The American Statistician*, 59:230–232. 29

Mehta, C. R. and Patel, N. R. (1997). Exact inference for categorical data. *Biometrics*, 53:112–117. 2

Meilă, M. (2005). Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 577–584, New York, NY, USA. ACM. 55

Milligan, G. W. and Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21:441–458. 55

Milton, R. C. (1970). *Rank Order Probabilities: Two-Sample Normal Shift Alternatives.* John Wiley & Sons Inc, New York. 18

Moher, D., Dulberg, C. S., and Wells, G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, 272:122–124. 18

Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, 54:1560–1568. 45

Neyman, J. and Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of Cambridge Philosophical Society*, 20:492–510. 9

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*, 82:645–647. 19

Onwuegbuzie, A. J. and Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, 3:201–230. 21

Owen, D. B. (1965). The power of Student's t-test. *Journal of the American Statistical Association*, 60:320–333. 18

Ozturk, O. (1999). Two-sample inference based on one-sample ranked set sample sign statistics. *Journal of Nonparametric Statistics*, 10:197–212. 32

Ozturk, O. and Wolfe, D. A. (2000). Optimal allocation procedure in ranked set two-sample median test. *Journal of Nonparametric Statistics*, 13:57–76. 32

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective.* World Scientific Publishing Company, Singapore. 45

Patil, G. P. (1995). Editorial: ranked set sampling. *Environmental and Ecological Statistics*, 2:271–285. 30

Pesarin, F. (2001). *Multivariate Permutation Tests: With Application in Biostatistics*. John Wiley & Sons, Ltd., Chichester. 6

Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Theory, Application and Software*. John Wiley & Sons, Ltd., Chichester. 5, 6, 10, 21, 47

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society*, Series B, 4:119–130. 2, 5

Pitman, E. J. G. (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. 2, 5

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any population. III. the analysis of variance test. *Biometrika*, 29:322–335. 2, 5

Posten, H. O. (1978). The robustness of the two-sample t-test over the Pearson system. *Journal of Statistical Computation and Simulation*, 6:295–311. 1

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850. 3, 55, 57

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. John Wiley & Sons, New York. 18

Rasch, D. and Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, 46:175–208. 1

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edition*. Sage Publications, Newbury Park, California. 43

Rey, D. S. (2004). *The Informational Order in Ranked Set Sampling Experiments*. PhD thesis, Georg-August-Universitat zu Gottingen. 30

Rijsbergen, C. J. V. (1979). *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, London, England. 55

Salmaso, L. (2003). Synchronized permutation tests in 2^k factorial designs. *Communication in Statistics - Theory and Methods*, 32:1419–1437. 6

Samawi, H. M. (1999). On quantiles estimation with application to normal ranges and hodges-lehmann estimate using a variety of ranked set sample. Technical report, Department of Statistics, Yarmouk University, Irbid, Jordan. 29, 30

Samawi, H. M. and Al-Sagheer, O. A. M. (2001). On the estimation of the distribution function using extreme and median ranked set sampling. *Biometrical Journal*, 43:357–373. 29

Samuh, M. H. and Al-Saleh, M. F. (2011). The effectiveness of multistage ranked set sampling in stratifying the population. *Communications in Statistics - Theory and Methods*, 40:1063–1080. 32, 35

Scheipl, F. (2010). RLRsim: Exact (restricted) likelihood ratio tests for mixed and additive models. *R package version 2.0-5*. 45

Schmitt, M. C. (1987). *The Effects on an Elaborated Directed Reading Activity on the Metacomprehension Skills of Third Graders*. PhD thesis, Purdue University. 13

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610. 44, 45

Shao, J. and Chow, S.-C. (2002). Reproducibility probability in clinical trials. *Statistics in Medicine*, 21:1727–1742. 2, 21, 22

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley Series in Probability & Mathematical Statistics, New York. 9

Shuweihdi, F. and Taylor, C. (2007). Inference for similarity indices. *In S. Barber, P.D. Baxter, & K.V.Mardia (eds), Systems Biology & Statistical Bioinformatics. Leeds, Leeds University Press*, pages 139–142. 55, 58

Silvapulle, M. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, 90:342–349. 46

Simonoff, J. S., Hochberg, Y., and Reiser, B. (1986). Alternative estimation procedures for $P_r(X < Y)$ in categorized data. *Biometrics*, 42:895–907. 19

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, 37:219–234. 46

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE Publications, London. 43

Stokes, S. L. and Sager, T. W. (1988). Characterization of a ranked set sampling with application to estimating distribution functions. *Journal of the American Statistical Association*, 83:374–381. 32

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50:1171–1177. 44, 45

Strehl, A. and Ghosh, J. (2003). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617. 55

Sun, Y. and Sherman, M. (1996). Some permutation tests for survival data. *Biometrics*, 52:87–97. 2

Swamy, P. A. V. B. (1970). Efficient inference in a random coefficient regression model. *Econometrica*, 38:311–323. 43

Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20:1–31. 31

Teichroew, D. (1955). Empirical power functions for nonparametric two-sample tests for small samples. *The Annals of Mathematical Statistics*, 26:340–344. 18

Theodoridis, S. and Koutroumbas, K. (2006). *Pattern Recognition, 3rd Edition*. Academic Press, Inc., Orlando, FL, USA. 53

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology*, 11:276–280. 21

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York. 43, 45

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, 59:254–262. 44

Wang, H., Chow, S. C., and Chen, M. (2005). A Bayesian approach on sample size calculation for comparing means. *Journal of Biopharmaceutical Statistics*, 15:799–807. 18

Warrens, M. J. (2008a). On similarity coefficients for 2×2 tables and correction for chance. *Psychometrika*, 73:487–502. 55

Warrens, M. J. (2008b). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification*, 25:177–183. 55, 58

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority, 2nd Edition*. Chapman & Hall/CRC. 63

Wolfe, D. A. (2004). Ranked set sampling: An approach to more efficient data collection. *Statistical Science*, 19:636–643. 2

Yeung, K. and Ruzzo, W. (2001). Details of the adjusted Rand index and clustering algorithms. Supplement to the paper (an experimental study on principal component analysis for clustering gene expression data). *Bioinformatics*, 17:763–774. 57