# NONPARAMETRIC MANOVA APPROACHES FOR NON-NORMAL MULTIVARIATE OUTCOMES

by

**Fanyin He**

B.S., Peking University, China, 2008

Submitted to the Graduate Faculty of

the Department of Biostatistics

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEATH

This dissertation was presented

by

Fanyin He

It was defended on

May 13, 2013

and approved by

**Dissertation Advisor**: Sati Mazumdar, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Stewart Anderson, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Gong Tang, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Robert T. Krafty, PhD, Assistant Professor, Department of Statistics, University of Pittsburgh

Martica Hall, PhD, Associate Professor, Department of Psychiatry, School of Medicine, University of Pittsburgh

Bruce L. Rollman, MD, MPH, Professor, Division of General Internal Medicine, University of Pittsburgh

**NONPARAMETRIC MANOVA APPROACHES**

**FOR NON-NORMAL MULTIVARIATE OUTCOMES**

Fanyin He, PhD

University of Pittsburgh, 2013

## ABSTRACT

Comparisons between groups play a central role in clinical research. As these comparisons often entail many potentially correlated response variables, the classical multivariate general linear model has been accepted as a standard tool. However, parametric methods require distributional assumptions such as multivariate normality while non-normal data often exist in clinical research. For example, a clinical trial investigating a treatment for depression is designed as a longitudinal study and the main outcome is survey scores of subjects on several time points, while the scores are ordinal. Although non-parametric multivariate methods are available in the statistical literature, they are not seen to be commonly used in clinical research. Moreover, automatic deletion of cases with missing values in response variables is a shortcoming of standard software when performing multivariate tests. This dissertation addresses the issues of violation of multivariate normality assumption and missing data, focusing on the non-parametric multivariate Kruskal-Wallis (MKW) test, likelihood-based and permutation-based methods.

First, an R-based program is written to compute the p-value of MKW test for group comparison. Simulation studies show that the permutation-based MKW test provides better coverage and higher power level than likelihood-based MKW test and classical MANOVA. Second, an extension of MKW test is proposed for multivariate data with missingness. The proposed method retrieves information in partially observed cases and is permutation-based. A

sensitivity analysis compares the performance of the proposed extension and the standard test utilizing only complete cases. Results show that the proposed extended method provides higher power level, encompassing a broad spectrum of multivariate effect sizes. An illustrative example using data from a psychiatric clinical trial is provided. The R program is ready to use for applied statistician.

The public health relevance of this work lies in the development of a new powerful methodology with user-friendly computer software for group comparisons in non-normal multivariate data with or without missingness.

## TABLE OF CONTENTS


**ACKNOWLEDGEMENTS ..... XI**

**1.0 INTRODUCTION ..... 1**

**2.0 LITERATURE REVIEW ..... 3**

**2.1 MULTIVARIATE GENERAL LINEAR MODEL (GLM) ..... 3**

**2.1.1 Classical Tests ..... 3**

**2.1.2 Permutation-Based Tests ..... 5**

**2.2 MULTIVARIATE KRUSKAL-WALLIS (MKW) TEST ..... 6**

**2.2.1 Kruskal-Wallis Test ..... 6**

**2.2.2 Likelihood-based Multivariate Kruskal-Wallis Test ..... 7**

**2.2.3 Permutation-based Multivariate Kruskal-Wallis Test ..... 8**

**2.3 MULTIVARIATE EFFECT SIZE ..... 9**

**3.0 PERFORMANCE OF MKW TEST ..... 11**

**3.1 COMPARISON BETWEEN MANOVA AND MKW TEST ..... 11**

**3.2 COVERAGE OF MKW TEST ..... 13**

**3.3 POWER OF MKW TEST ..... 14**

**4.0 MISSING DATA ISSUE ..... 18**

**4.1 MISSINGNESS MECHANISMS ..... 18**

**4.2 FORMULATION OF THE TEST ..... 19**

**4.3 SIMULATION STUDY ..... 21**

**5.0 APPLICATION ..... 30**

**5.1 YOGA DATA ..... 30**

**5.2 APPLICATION ..... 33**

**5.2.1 Univariate Kruskal-Wallis Test ..... 33**

**5.2.2 Multivariate Tests ..... 34**

**6.0 DISCUSSIONS ..... 36**

**6.1 SAMPLE SIZE CALCULATION ..... 36**

**6.2 SINGULARITY ISSUE ..... 37**

**7.0 CONCLUSIONS ..... 38**

**APPENDIX: R CODE FOR MKW TEST ..... 40**

**BIBLIOGRAPHY ..... 45**

## LIST OF TABLES

Table 3.1. Type I errors of MANOVA tests and MKW tests in different scenarios ..... 13

Table 3.2 Type I errors of MKW tests in different scenarios.. ..... 14

Table 3.3 Power simulations of MKW tests in different scenarios (p=4) ..... 16

Table 3.4 Power simulations of MKW tests in different scenarios (p=8) ..... 17

Table 4.1 Data generation summary ..... 22

Table 4.2 Missing patterns of bivariate data ..... 23

Table 4.3 Simulation results of type I errors ..... 25

Table 4.4 Power simulations with normal outcomes, medium rate of missingness and varying effect sizes ..... 26

Table 4.5 Power simulations with normal outcomes and high rate of missingness and varying effect sizes ..... 27

Table 4.6 Power simulations with non-normal outcomes and medium rate of missingness and varying effect sizes ..... 28

Table 4.7 Power simulations with non-normal outcomes and high rate of missingness and varying effect sizes ..... 29

Table 5.1 Univariate Kruskal-Wallis tests on Yoga example ..... 33

Table 5.2 Missing pattern of improvement of cognitive functions in the four domains in Yoga example ..... 34

Table 5.3 p-values of MANOVA and MKW tests on Yoga example ..... 35

Table 6.1 Effect sizes required to detect a difference between two groups with 80% power in pre-specified sample sizes ..... 37

## LIST OF FIGURES

Figure 4.1 Missing patterns of bivaraite data with medium rates of missingness ..... 23

Figure 5.1 Flow chart of Yoga data ..... 32

## ACKNOWLEDGEMENTS

I gratefully acknowledge the suggestions and comments provided by my advisor Dr. Sati Mazumdar and my committee members Dr. Stewart Anderson, Dr. Gong Tang, Dr. Robert T. Krafty, Dr. Martica Hall and Dr. Bruce L. Rollman. I thank Dr. Pranab K. Sen (University of North Carolina) for statistical inputs and Dr. Triptish Bhatia (Dr Ram Manohar Lohia Hospital, New Delhi, India) for letting me use her data (supported by NIH, R01 TW008289). I am grateful for academic support from "Aging Well, Sleeping Efficiently: Intervention Studies" (AgeWise, supported by NIH, P01 AG20677-05), and "Improving Quality of Primary Care for Anxiety Disorders" (supported by NIMH, R01 MH 09421/MH/NIMH NIH HHS/United States).

# 1.0 INTRODUCTION

Comparisons between several treatment groups play a central role in clinical research. As these comparisons often entail many potentially correlated dependent variables, the classical multivariate general linear model has been accepted as a key tool for this endeavor. The widely applied statistical procedures, univariate and multivariate analysis of variance (ANOVA and MANOVA) are subsumed under this model. For practitioners, the use of these statistical procedures does not pose any difficulties under normality assumptions due to the availability of software (SAS, SPSS, and STATA). However, difficulties exist if the assumption of normality is violated. This is especially true for multivariate data. Though practitioners are aware of the benefits of simultaneous inference in parametric and nonparametric methods, lack of readily available computer software for nonparametric MANOVA methods often prevents them from performing appropriate analyses.

This dissertation addresses the issue of nonconformity with the multivariate normality assumption. I consider analytic methods pertaining to outcomes measured at a fixed time point, both continuous and discrete/ordinal variables and use likelihood-based and permutation-based theories for the methods. I focus on the multivariate Kruskal-Wallis (MKW) test (May and Johnson, 1997) for group comparisons. The dissertation concerns the nonparametric hypothesis tests for correlated multivariate outcomes in a MANOVA-like frame. The objective is to provide a guideline to practitioners for analyzing multivariate data for group comparisons.

An R-based program is written to compute the p-value of MKW test for group comparison. Simulation studies are done to compare the coverage and power levels of classical MANOVA, the likelihood-based MKW test and the permutation-based MKW test.

Missing data often exist in clinical trials. However, in the MANOVA-like frame, the standard tests do not utilize information in partially observed cases. In software algorithms such as the SAS MANOVA procedure and the SAS macro written by May and Johnson (1997) for MKW test, cases with missing values in response variables are deleted when performing the tests. I propose a nonparametric method for multivariate non-normal data with missingness, which is an extension of the MKW test. The method retrieves information in partially observed cases in missing data.

I carry out a sensitivity analysis to compare the performance of the proposed extension of MKW test and the standard test utilizing only complete cases under missing completely at random assumption. Results show that the proposed extended method provides higher power level, encompassing a broad spectrum of multivariate effect sizes. An illustrative example using data from a psychiatric clinical trial is provided.

# 2.0 LITERATURE REVIEW

## 2.1 MULTIVARIATE GENERAL LINEAR MODEL (GLM)

### 2.1.1 Classical Tests

The multivariate general linear model (GLM) subsumes MANOVA models utilizing a general statistical framework. Description of GLM can be found in many statistical text books and user guides for most statistical software packages. The term "general" refers to the fact that the GLM implements both regression and ANOVA models, univariate and multivariate in the same framework.

A GLM can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is an $n \times p$ matrix of $n$ observations by $p$ response variables, and each column of $\mathbf{Y}$ corresponds to a specific dependent variable. $\mathbf{X}(n \times k), \boldsymbol{\beta}(k \times p)$ and $\boldsymbol{\varepsilon}(n \times p)$ are design, parameter and error matrices, respectively. The $n$ rows of $\boldsymbol{\varepsilon}$ are assumed to be independent and identically distributed as $\mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where is $\boldsymbol{\Sigma}$ a $p \times p$ positive definite dispersion matrix. The ordinary least square estimate for $\boldsymbol{\beta}$ is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The covariance matrix, $\boldsymbol{\Sigma}$, is estimated by $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})/[n - rank(\mathbf{X})]$.

A linear hypothesis is most commonly written as

$$H_0: \mathbf{L\beta M} = \mathbf{0} \text{ vs. } H_A: \mathbf{L\beta M} \neq \mathbf{0},$$

where **L** and **M** are matrices of specified constants, and **Lβ**$\mathbf{M}$ is estimable.

The common test statistics for testing a linear hypothesis given above are

Wilks' $\lambda = \det(\mathbf{E}) / \det(\mathbf{H} + \mathbf{E})$;

Pillai's trace $V = \text{trace}(\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$;

Hotelling-Lawley trace $U = \text{trace}(\mathbf{E}^{-1}\mathbf{H})$; and

Roy's maximum root $\lambda_1$, the largest eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$;

where **H** and **E** represent, respectively, the sums of squares and cross-product matrices for the hypothesis and error matrices, that is,

$\mathbf{H} = \text{SS}(\mathbf{LbM}) = \mathbf{M}'(\mathbf{Lb})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}']^{-1}(\mathbf{Lb})\mathbf{M}$; and

$\mathbf{E} = \mathbf{M}'(\mathbf{Y}'\mathbf{Y} - \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b})\mathbf{M}$.

Under the multivariate normality assumption, all of the above four test statistics can be approximated by *F* distributions (Anderson 2003).

MANOVA is the most commonly used method to compare groups for a set of continuous dependent variables. MANOVA uses one or more categorical independent variables to form groups, with more than one dependent variable and tests the differences in the centroids of means of several dependent variables, for different groups. MANOVA is subsumed in GLM by specifying **X**, **β**, **L** and **M**. Once the statistics are obtained, they are translated into *F* statistics in order to test the null hypothesis. The four statistics, mentioned in the previous paragraph, may give identical F values. When they differ, Pillai's trace is often used because it is considered by many to be most powerful and robust. Roy's largest root is an upper bound on F, and therefore

gives a lower bound estimate of the probability of $F$. Thus, Roy's largest root is generally disregarded when it is significant but the others are not significant.

### 2.1.2 Permutation-Based Tests

A permutation-based test is a type of statistical test in which the distribution of the test statistics under the null hypothesis is determined by calculating all possible values of the test statistic under rearrangements of the labels of groups on the observed data points. After the results are obtained from the actual experiment, one can determine the results that could have occurred for other rearrangements if the null hypotheses were true. The primary use of permutation-based tests is to obtain p-values. The p-value based on permutation distribution is the proportion of test statistics in the set of all possible re-arrangements that are more extreme than the test statistics value calculated with the original dataset before any rearrangement. Permutation-based tests exist for any statistic regardless of whether or not its distribution is known. The major drawback to permutation-based tests is that they can be computationally intensive and may require "custom" code for difficult-to-calculate statistics (Pesarin, 2001 and Edgington, 2007).

The basic premise in permutation-based tests is the assumption that it is possible that all of the treatment groups are equivalent, and that every member of them is the same before sampling began. This is the notion of exchangeability under the null hypothesis. An important consequence of this assumption is that tests of difference in location (like a permutation-based t-test) require equal variance. In this respect, the permutation-based t-test shares the same weakness as the classical Student's t-test (the Behrens-Fisher problem).

## 2.2 MULTIVARIATE KRUSKAL-WALLIS (MKW) TEST

### 2.2.1 Kruskal-Wallis Test

The Kruskal-Wallis test (Kruskal and Wallis, 1952) is a univariate nonparametric method to test whether the variable of interest is differently distributed in two or more independent groups. Kruskal-Wallis test is analogous to one-way analysis of variance (ANOVA), while ANOVA is a parametric test. The Kruskal-Wallis test is an extension of Wilcoxon rank sum test, while Wilcoxon rank sum test is for two-group test. The Kruskal-Wallis test is widely used in analyses involving non-normal data.

Let $Y_{ij}$ be the original observation of $j$th subject from $i$th group, where $j = 1, \dots, n_i; i = 1, \dots, g$. $n = \sum_{i=1}^{g} n_i$ is the total sample size. We first rank the $n$ observations among all the groups to get ranks $R_{ij}$ corresponding to $Y_{ij}$. Tied values are assigned to average ranks. The null hypothesis of Kruskal-Wallis test is that the medians of $Y_{ij}$ are the same in the $g$ groups.

The mean rank of group $i$ is denoted as

$$\bar{R}_{i\cdot} = \frac{\sum_{j=1}^{n_i} R_{ij}}{n_i}.$$

Let

$$\bar{R} = \frac{n+1}{2}.$$

The test statistic is

$$K = (n-1)\frac{\sum_{i=1}^{g} n_i(\bar{R}_{i\cdot} - \bar{R})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(R_{ij} - \bar{R}\right)^2}.$$

In large samples, $K$ is approximately $\chi^2$ distributed with $(g-1)$ degrees of freedom.

### 2.2.2 Likelihood-based Multivariate Kruskal-Wallis Test

The multivariate extension of the univariate Kruskal-Wallis test is a rank-order procedure in which the $n$ scores of each of the $p$ variables are ranked separately. If certain observations are tied, each of these observations is assigned the mean of the ranks for which the observations are tied. It should be noted that this procedure of assigning ranks poses no difficulty if the number of scores for the different variables are not equal. The null hypothesis is that for each variable, the expected values of the mean ranks are equal for the different groups. Large sample theory suggests that the MKW statistic is approximately $\chi^2$ distributed. However, in small samples permutation theory is needed to get the exact distribution.

Katz and McSweeney (1980) provided an explicit description of this MKW test. They also provided computational formulas and post-hoc techniques which could be used to isolate sources of differences if the null hypothesis is rejected. However, the testing procedure discussed in their paper was based on large sample properties of the statistic which was approximately $\chi^2$ distributed. May and Johnson (1997) have written a SAS macro that computes the probability values and tabulates the exact distribution for the univariate and MKW test.

Multivariate Kruskal-Wallis test transforms original data to its ranking, and therefore it is distribution-free. The ranking is performed separately for each dependent variable, and is across groups. Let $Y_{ijk}$ be the original observation of *k*th variate for *j*th subject from *i*th group, where $k = 1, \dots, p; j = 1, \dots, n_i; i = 1, \dots, g$. Denote $R_{ijk}$ as the rank corresponding to $Y_{ijk}$. In case of ties, mid-ranks are used. Let

$$\bar{R}_{i.k} = \sum_{j=1}^{n_i} \frac{R_{ijk}}{n_i},$$

then $E(\bar{R}_{i.k}) = m = (n+1)/2$. The vector $\mathbf{U_i} = (\bar{R}_{i.1} - m, \ldots, \bar{R}_{i.p} - m)'$ denotes the average ranks for the *i*th group corrected for *m* for each variate. $\mathbf{U_i}$ is a measure of directed distance from the mean vector of ranks for the *i*th group. An estimate of the pooled within-group covariance matrix is

$$\mathbf{V} = \frac{1}{n-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(\mathbf{R_{ij}} - m\mathbf{1_p}\right)\left(\mathbf{R_{ij}} - m\mathbf{1_p}\right)'.$$

Under the null hypothesis that there is no difference in group means for the *p* variables,

$$E(\mathbf{U_i}) = \mathbf{0_p}.$$

The MKW test is expressed as

$$W^2 = \sum_{i=1}^{g} n_i \mathbf{U_i'} \mathbf{V^{-1}} \mathbf{U_i}.$$

In large samples, $W^2$ is approximately centrally $\chi^2$ distributed with $p(g-1)$ degrees of freedom.

### 2.2.3 Permutation-based Multivariate Kruskal-Wallis Test

When there are too many possible orderings of the data to allow complete enumeration in a reasonably time efficient manner, an asymptotically equivalent permutation-based test can be created by generating the exact distribution by Monte Carlo sampling, which takes a small (relative to the total number of required permutations) random sample of the possible replicates. This type of permutation-based test is known as Monte Carlo permutation-based test. The necessary size of the Monte Carlo sampling depends on the need for accuracy of the test (Edgington and Onghena, 2007).

Monte Carlo permutation-based test can be used to get a more accurate $p$-value of MKW test in small samples. The procedure is as follows:

(a) Calculate the statistic for the data, and denote it as $W^{2*}$.

(b) Randomly assign subjects to groups, and calculate the new $W^2$ for this permuted data.

(c) Repeat (b) $M$ times to get the permutation distribution of $W^2$ under null hypothesis.

(d) Calculate the $p$-value $= \frac{\text{number of } W^2 \geq W^{2*}}{M}$.

## 2.3 MULTIVARIATE EFFECT SIZE

Effect sizes are commonly used for power analysis and experiment designs. In hypothesis testing, ES is an index reflecting the degree to which the null hypothesis is false, or the discrepancy between the null hypothesis and the alternative hypothesis (Cohen 1992), without the influence of sample sizes. One of the widely used effect sizes index in one-way ANOVA setting is Cohen's $f^2$, the ratio of the variance of the group means to the variance of the values within groups (Cohen 1988). Cohen's $f^2$ is defined as

$$f^2 = \frac{R^2}{1 - R^2},$$

where $R^2$ is the squared multiple correlation.

Cohen (1988) suggested a generalization of $f^2$ based on Wilks' λ, which can be used in multivariate settings:

$$f^2 = \lambda^{-1/r} - 1 = \frac{\sqrt[r]{\det(\mathbf{H} + \mathbf{E})} - \sqrt[r]{\det(\mathbf{E})}}{\sqrt[r]{\det(\mathbf{E})}},$$

where

$$r = \sqrt{\frac{p^2(g-1)^2 - 4}{p^2 + (g-1)^2 - 5}},$$

$p$ is the number of response variables, $g$ is the number of groups, and **E** and **H** refer to the population error and hypothesis matrices.

From the latter form of $f^2$ we see that it is a signal to noise ratio: the ratio of variance of the model to the variance of errors. $f^2$ is a non-increasing function of $p$ and $g$, which means that large data sizes (more groups and/or more dependent variables) have a negative effect on effect sizes. For two-group cases, $r = 1$ and $f^2$ reduces to $\lambda^{-1} - 1$. For 3-group cases, $r = 2$ and $f^2$ reduces to $\lambda^{-1/2} - 1$. If these two cases have the same Wilks' $\lambda$, the latter case will have a smaller effect size. Cohen (1988) also suggested "small", "medium" and "large" $f^2$ values to be 0.02, 0.15 and 0.35, respectively.

# 3.0 PERFORMANCE OF MKW TEST

## 3.1 COMPARISON BETWEEN MANOVA AND MKW TEST

An R-based program is written to compute the approximate $\chi^2$ distribution and the exact null distribution under null hypothesis for the MKW test for multi-group comparisons. Simulation studies are done to compare the performance of MKW tests to the MANOVA tests (classical and permutation-based).

Zeng *et al.* (2011) proposed a permutation-based test with classical MANOVA test statistics. They claimed it to be better than classical MANOVA tests, and concluded that the SAS GLM procedure with the 'exact' option provided best approximation among commonly available software. I simulate data for several different scenarios in their paper, and compare their type I errors to the type I errors of MKW tests. Zeng provided type I errors of four common statistics, and I only compare with Wilks' $\lambda$ since none of the statistics performed superior.

Simulations were done in the same way as in Zeng *et al.* (2011). I use Clayton's family of Archimedean copulas with compounding approach (Frees and Valdez, 1998) to simulate correlated outcomes. Compound symmetry is assumed for the simulations and used Kendall's $\tau$ = 0.8 and 0.5 to indicate highly-correlated and medium-correlated outcomes. I consider $n$ = 5 and 10 to indicate small and medium sample sizes in each group. I use exponential distributions to be the marginal distributions of the outcomes. Both likelihood-based p-values and

permutation-based p-values are considered. $g$ = 5 groups and $p$ = 4 outcomes are generated for each scenario. The marginals are set as exponential distributions with means = 100, 200, 300 and 400. Significance level $\alpha$=0.05 is used. I use `mvdc()` function from the `copula` package (Yan 2007, Kojadinovic 2010 and Hofert 2011) in R for the data generation.

Getting permutation distribution from all possible rearrangements is very time-consuming. For example, in one scenario where $g = 5, n = 10$ and $p = 4$, the number of all possible rearrangements of group labels is $50!/(10!)^5 = 4.8 \times 10^{31}$. It took roughly a half of a minute to get one $p$-value based on permutation distribution on a personal computer, and several days to get the simulated type I errors. To balance the efficiency and the accuracy, the permutation distribution is estimated from 300 Monte-Carlo samples in each simulated data set. The type I error is estimated from 5000 simulations in each scenario. The simulations are supported in part by Computational Resources on PittGrid (www.pittgrid.pitt.edu).

Table 3.1 represents results for skewed continuous cases. Exponential marginal distributions are used. May and Johnson (1997) mentioned that the $\chi^2$ approximation performed well in MKW tests with as few as 10 subjects per group, which is not consistent with my simulation results. When the sample sizes per group are not large ($\leq$10), the type I errors of MKW tests based on $\chi^2$ approximation are far less than the nominal significance level 0.05, and the same problem occurs in classical MANOVA, even with the 'exact' option in SAS. These results suggest that tests based on large sample approximation are not doing well in medium sized samples ($\leq$10). On the other hand, MANOVA and MKW tests performed well when they were based on permutation distributions. MKW performed slightly better when sample sizes are medium sized.

**Table 3.1.** Type I errors of MANOVA tests and MKW tests in different scenarios

| Kendall's $\tau$ | N/group | Wilks' $\lambda$* | | MKW | |
|---|---|---|---|---|---|
| | | large sample approximation | permutation | large sample approximation | permutation |
| 0.5 | 5 | 0.038 | 0.052 | 0.018 | 0.045 |
| | 10 | 0.034 | 0.043 | 0.036 | 0.051 |
| 0.8 | 5 | 0.027 | 0.052 | 0.021 | 0.053 |
| | 10 | 0.032 | 0.055 | 0.034 | 0.051 |

* Type I errors for Wilks' $\lambda$ were from Zeng (2011).

The simulated data have 5 groups and 4 outcomes, with exponential marginal distributions of means 100, 200, 300 and 400.

## 3.2 COVERAGE OF MKW TEST

MKW test can be used when the response variables are ordinal. To examine the coverage of MKW test, simulation studies are done with correlated data with Poisson marginal distributions.

The same setting as section 3.1 is used to generate data and $g = 2, 3$ and 4 are examined. Marginal distributions of the 4 outcomes are set as Poisson distribution with means = 10, 20, 30 and 40.

Table 3.2 shows the results of type I errors of the simulations. The p-values of MKW tests from $\chi^2$ approximation are far from 0.05, while the permutation-based p-values are very close to 0.05. However, the MKW tests are not always applicable when the total sample size is very small ($< 10$).

**Table 3.2** Type I errors of MKW tests in different scenarios

| # of groups | Kendall's $\tau$ | N/group | MKW | |
|---|---|---|---|---|
| | | | approximation | permutation |
| 2 | 0.5 | 5 | NA | NA |
| | | 10 | 0.029 | 0.051 |
| | 0.8 | 5 | NA | NA |
| | | 10 | 0.022 | 0.048 |
| 3 | 0.5 | 5 | 0.015 | 0.051 |
| | | 10 | 0.026 | 0.045 |
| | 0.8 | 5 | 0.011 | 0.050 |
| | | 10 | 0.026 | 0.046 |
| 4 | 0.5 | 5 | 0.017 | 0.050 |
| | | 10 | 0.038 | 0.053 |
| | 0.8 | 5 | 0.017 | 0.052 |
| | | 10 | 0.032 | 0.050 |

The simulated data have 4 outcomes, with Poisson marginal distributions with means 10, 20, 30 and 40.

## 3.3 POWER OF MKW TEST

Simulations are performed to examine the power of MKW test under different scenarios. Data are simulated similarly as in section 3.1. I use Clayton's family of Archimedean copulas with compounding approach (Frees and Valdez, 1998) to simulate $p = 4$ and 8 correlated outcomes for $g = 2$ groups. I assume compound symmetry for the simulations and use Kendall's $\tau = 0.8$

and 0.5 to indicate highly-correlated and medium-correlated outcomes. I consider $n$ = 10 and 15 subjects in each group. To simulate data with 4 outcomes, the marginal distributions are set as Poisson distributions with means = 10, 20, 30 and 40 for the first group, and with means=10, 20, 30 and $40 - \Delta$ for the second group, where different $\Delta$'s are set to get varying effect sizes. Similarly, to simulate data with 8 outcomes, the marginal distributions are set as Poisson distributions with means = 10, 20, 30, 40, 50, 60, 70 and 80 for the first group, and with means=10, 20, 30, 40, 50, 60, 70 and $80 - \Delta$ for the second group.

Five thousand data sets are generated for each scenario. The power is estimated by the proportion of rejections of null hypothesis out of the 5000 simulations with significance level 0.05.

Tables 3.3 and 3.4 represent the results of the simulations. As expected, power is higher with larger sample size per group, lager $\Delta$ and higher Kendall's $\tau$. When there are four outcomes, medium samples (10 subjects per group with 2 groups) have very high power (>99%) when $\Delta$ is large (10 or 15). Medium sample sizes also result in high power (>90%) when Kendall's $\tau$ is high (0.8). When there are eight outcomes, medium sized samples have <80% power when $\Delta$ is medium (5 or 7). It suggests that with more response variables, we need larger sample size to obtain a specified power level. Since there is a positive relationship between effect size and power, this also agrees with the statement that large data sizes (more groups and/or more dependent variables) have a negative effect on effect sizes.

**Table 3.3** Power simulations of MKW tests in different scenarios (p=4)

| n/group | Δ | Kendall's $\tau$ | power |
|---|---|---|---|
| 10 | 5 | 0.5 | 0.45 |
| | | 0.8 | 0.92 |
| | 7 | 0.5 | 0.75 |
| | | 0.8 | 0.99 |
| | 10 | 0.5 | 0.97 |
| | | 0.8 | 0.9995 |
| | 15 | 0.5 | 0.9998 |
| | | 0.8 | 1 |
| 15 | 10 | 0.5 | 0.999 |
| | | 0.8 | 1 |
| | 15 | 0.5 | 1 |
| | | 0.8 | 1 |

The simulated data have 2 groups and 4 outcomes, with Poisson marginal distributions with means 10, 20, 30 and 40 for the first group, and with means 10, 20, 30 and 40-Δ for the second group.

**Table 3.4** Power simulations of MKW tests in different scenarios (p=8)

| n/group | Δ | Kendall's $\tau$ | power |
|---|---|---|---|
| 10 | 5 | 0.5 | 0.15 |
| | | 0.8 | 0.53 |
| | 7 | 0.5 | 0.28 |
| | | 0.8 | 0.77 |

The simulated data have 2 groups and 8 outcomes, with Poisson distributions with means = 10, 20, 30, 40, 50, 60, 70 and 80 for the first group, and with means=10, 20, 30, 40, 50, 60, 70 and 80-Δ for the second group.

# 4.0 MISSING DATA ISSUE

Missing data often exist in clinical trials. However, MKW test assumes that data is fully observed. All incomplete cases are deleted before the MKW test is performed, which means that the information in partially observed cases is lost. To retrieve this part of information, we propose a method that extends MKW test to data with missingness. The proposed method is more powerful than the standard MKW test in simulated data.

## 4.1 MISSINGNESS MECHANISMS

There are three mechanisms of missing data. If missingness does not depend on the observed values or the missing values, the missing data are called missing completely at random (MCAR). If missingness depends only on the observed data and not on the missing values of the data, the missing data are called missing at random (MAR). The missing data are called not missing at random (NMAR) if missingness depends on the missing values of the data. MCAR is the most restrictive assumption among these three. In this section, we always assume MCAR.

## 4.2 FORMULATION OF THE TEST

Let $\mathbf{Y}_{n\times p}$ denote the data. Let $Y_{ijk}$ be the original observation of *k*th variate for *j*th subject from *i*th group, where $= 1, \dots, p; j = 1, \dots, n_i; i = 1, \dots, g$. $n = \Sigma n_i$. The data has $n$ subjects in total with $g$ groups and $p$ outcomes.

For the *j*th subject in *i*th group, its observation vector $\mathbf{Y}_{ij} = (Y_{ij1}, \dots, Y_{ijp})$ has a corresponding missing indicator vector $\mathbf{r}_{ij} = (R_{ij1}, \dots, R_{ijp})$, where $R_{ijk} = 1$ if the *k*th variate is missing, and 0 if it is observed. $\mathbf{r}_{ij}$ is called the missing pattern of the *j*th subject in *i*th group. $\mathbf{r}_{ij}$ is a vector of length $p$, with each element valued at 0 or 1. For example, if $p = 4$, and if a subject is observed on the first and second outcomes, and is missing on the third and fourth outcomes, its missing pattern is (0, 0, 1, 1). In a dataset with $p$ variables, there are in total $2^p$ possible distinct missing patterns. MCAR is assumed at this stage.

Suppose there are $L$ distinct missing patterns in $\mathbf{Y}$ ($L \leq 2^p$). Let $S_l$ denote the set of cases with missing pattern $l$,$l = 1, \dots, L$, and let $m_l$ denote the number of observations in $S_l$, and let $n = \sum_{l=1}^{L} m_l$. Let $p_l$ denote the number of observed variables in missing pattern $l, l = 1, \dots, L$. Let $m_{il}$ denote the number of observations in group $i$, pattern $l$, $i = 1, \dots, g; l = 1, \dots, L$.

I assume that $m_l > p_l, l = 1, \dots, L$. If the number of observations with missing pattern $l$ is too small ($m_l \leq p_l$), we delete the cases in $S_l$ from the total sample before performing the method. The assumption is to avoid the situation that the estimated variance-covariance matrix within missing pattern is singular and the statistic cannot be calculated.

The statistic $W_l^2$ in each $S_l$ with regard to observed variables can be calculated from standard MKW test. Then the proposed test statistic would be

$$W^2 = \sum_{l=1}^{L} t_l W_l^2,$$

where the $t_l$s are weights and $\sum t_l = 1$.

The standard MKW test is a special case of the proposed test, when $t_l$ is assigned to 1 if $S_l$ is the set of complete cases, and 0 otherwise. Two weighting schemes are proposed:

(1) Unweighted: $t_l = 1/L, l = 1, \ldots, L$. Then $W^2$ is the arithmetic mean of $W_l^2$s.

(2) Weighted: $t_l = m_l/n, l = 1, \ldots, L$. Then each $W_l^2$ would contribute to $W^2$ proportional to the number of cases in its missing pattern.

In large samples, $W_l^2$ is approximately $\chi^2$ distributed with degrees of freedom $\nu_l = p_l(g-1), l = 1, \ldots, L$. And $W^2$ is a linear combination of the $L$ independent $\chi^2$ distributed statistics. Two ways are considered to get the approximate distribution of $W^2$ in large samples. One is that based on Welch-Satterthwaite equation (Satterthwaite, 1946 and Welch, 1947), $W^2$ is also approximately $\chi^2$ distributed, with degrees of freedom

$$\nu \approx \frac{(W^2)^2}{\sum_{l=1}^{L} \frac{(t_l W_l^2)^2}{p_l(g-1)}}.$$

The other is based on empirical distribution. Generate $W_{l1}^2, \ldots, W_{lM}^2$ as random samples from $\chi^2$ distribution with $\nu_l$ degrees of freedom, where $l = 1, \ldots, L$ and $M$ is a large integer. Set

$$X_m = \sum_{l=1}^{L} t_l W_{lm}^2, m = 1, \ldots, M.$$

An empirical distribution of $W^2$ can be calculated from $X_1, \ldots, X_M$. And the p-value can be obtained based on the empirical distribution. In small samples, permutation distribution of $W^2$ can be obtained by permuting the group labels among the whole data set under MCAR.

## 4.3 SIMULATION STUDY

To examine the coverage and power level of the proposed method, simulations in different scenarios are performed. Data with $g = 2$ groups and $p = 2$ outcome variables are simulated. To generate correlated outcomes, I use a latent variable $X$. Two scenarios are examined. One is based on normal distributed $X$, and the other is based on binomial distributed $X$.

For the first scenario, set $X \sim N(0,1)$.

For group 1, generate $X_1, \dots, X_{n_1}$ as a random sample of $X$.

Set $Y_{11}|X \sim N(1+X, 2), Y_{12}|X \sim N(X, 1)$.

Generate $\left[\left(Y_{1j1}, Y_{1j2}\right)|X_j\right]$ as a random sample of $\left[(Y_{11}, Y_{12})|X_j\right], j = 1, \dots, n_1$.

For group 2, generate $X_1, \dots, X_{n_2}$ as another random sample of $X$.

Set $Y_{21}|X \sim N(1+X, 2), Y_{22}|X \sim N(\Delta + X, 1)$.

Generate $\left[\left(Y_{2j1}, Y_{2j2}\right)|X_j\right]$ as a random sample of $\left[(Y_{21}, Y_{22})|X_j\right], j = 1, \dots, n_2$.

For the second scenario, set $X \sim BIN(5,0.5)$.

For group 1, generate $X_1, \dots, X_{n_1}$ as a random sample of $X$.

Set $Y_{11}|X \sim POI(1+X),\ Y_{12}|X \sim POI(2+X)$.

Generate $\left[\left(Y_{1j1}, Y_{1j2}\right)|X_j\right]$ as a random sample of $\left[(Y_{11}, Y_{12})|X_j\right], j = 1, \dots, n_1$.

For group 2, generate $X_1, \dots, X_{n_2}$ as another random sample of $X$.

Set $Y_{21}|X \sim POI(1+X),\ Y_{22}|X \sim POI(2+\Delta+X)$.

Generate $\left[\left(Y_{2j1}, Y_{2j2}\right)|X_j\right]$ as a random sample of $\left[(Y_{21}, Y_{22})|X_j\right], j = 1, \dots, n_2$.

**Table 4.1** Data generation summary

| Scenario | Group | Variable 1 | Variable 2 |
|---|---|---|---|
| $X \sim N(0,1)$ | Group 1 | $Y_{11}\|X \sim N(1+X,2)$ | $Y_{12}\|X \sim N(X,1)$ |
| | Group 2 | $Y_{21}\|X \sim N(1+X,2)$ | $Y_{22}\|X \sim N(\Delta+X,1)$ |
| $\mathrm{X} \sim \mathrm{BIN}(5,0.5)$ | Group 1 | $Y_{11}\|X \sim POI(1+X)$ | $Y_{12}\|X \sim POI(2+X)$ |
| | Group 2 | $Y_{21}\|X \sim POI(1+X)$ | $Y_{22}\|X \sim POI(2+X+\Delta)$ |

The summary of data generation is shown in Table 4.1. Letting $n_1 = n_2 = 50$, the simulated data is given as

$$\mathbf{Y} = \begin{bmatrix} Y_{111} & Y_{112} \\ \vdots & \vdots \\ Y_{1,50,1} & Y_{1,50,2} \\ Y_{211} & Y_{212} \\ \vdots & \vdots \\ Y_{2,50,1} & Y_{2,50,2} \end{bmatrix} = (\mathbf{Y_1}, \mathbf{Y_2}).$$

When $\Delta$ is assigned to zero, the underlying distributions of the two outcomes are the same in the two groups. Hence, type I error rates can be examined.

$\Delta$ can be assigned a spectrum of non-zero numbers to get different effect sizes. The underlying distributions of the first outcome variable are the same in the two groups, and the underlying distributions of the second outcome variable are differing across the two groups. Power levels can be examined.

There are $L = 4$ possible missing patterns in bivariate data. They are listed in Table 4.2. Since observations with missing pattern M4 do not contain any information, I only consider the first three missing patterns in our simulations. Two rates of missingness, also shown in Table 4.2, are randomly assigned to data to examine the coverage and power of the proposed method. Figure 4.1 shows the structure of the simulated bivariate data with missingness.

**Table 4.2** Missing patterns of bivariate data

| missing pattern | description | percentages out of all data (%) | |
|---|---|---|---|
| | | medium rates of missingness | high rates of missingness |
| M1 | $\mathbf{Y_1}$ and $\mathbf{Y_2}$ both observed | 40% | 20% |
| M2 | $\mathbf{Y_1}$observed, $\mathbf{Y_2}$ missing | 30% | 40% |
| M3 | $\mathbf{Y_1}$ missing, $\mathbf{Y_2}$ observed | 30% | 40% |
| M4 | $\mathbf{Y_1}$ and $\mathbf{Y_2}$both missing | 0% | 0% |

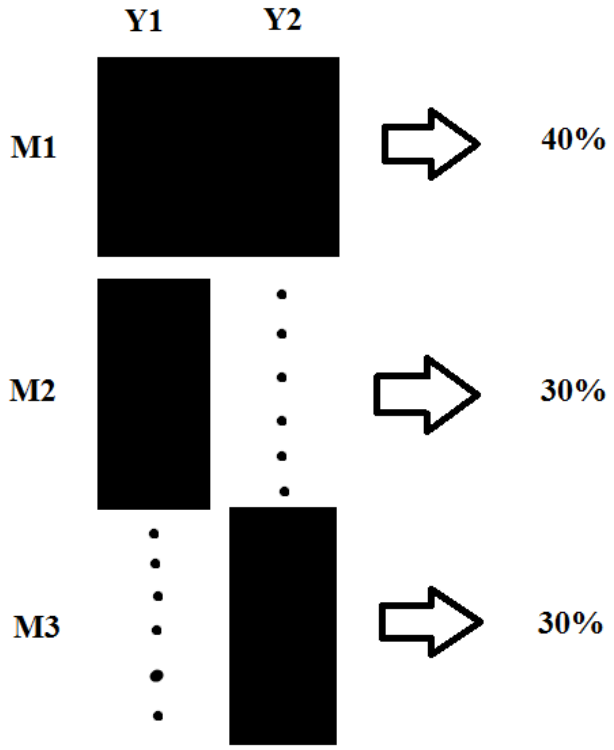


**Figure 4.1** Missing patterns of bivaraite data with medium rates of missingness

For each scenario, I perform the proposed method in $n_{sim} = 1000$ simulated incomplete data, and get the p-values. I estimate the power level of the method as $\frac{\text{number of } p-\text{values}<0.05}{n_{sim}}$ when $\Delta > 0$, and estimate the type I error of the method as $\frac{\text{number of } p-\text{values}<0.05}{n_{sim}}$ when $\Delta = 0$.

The simulation results of type I errors are shown in Table 4.3. Permutation-based p-values are close to the nominal significance level 0.05, and are slightly more accurate compared with p-values based on large sample approximation. Higher missing rates imply less information. Therefore it can be seen that type I errors are closer to 0.05 in "medium" missing rates scenarios compared with "high" missing rates ones, either in normal data or in non-normal data.

The simulation results of power levels are shown in Table 4.4 through Table 4.7. As expected, the power levels of the proposed method are always higher than the power levels of standard MKW test applied only on complete cases, and the difference is larger with higher missing rates. Comparing adjacent columns, the permutation-based tests provide higher power levels than tests based on large sample approximation. Neither of the two weighting schemes shows great superiority to the other. The weighted test statistic and the unweighted one provide very similar powers. And both increase when effect size increases. In three of the four simulation sets, the power levels of extended MKW test reach 80% when effect size is "medium" (<0.3). When percentage of missing values increases, the power level decreases. The performance of the extended test in non-normal data is as powerful as in normal data.

**Table 4.3** Simulation results of type I errors

| distribution | missing rates | unweighted † statistic | | weighted ǂ statistic | |
|---|---|---|---|---|---|
| | | permutation-based | large sample approximation** | permutation-based | large sample approximation** |
| normal | medium* | 0.056 | 0.058 | 0.05 | 0.056 |
| | high* | 0.062 | 0.066 | 0.054 | 0.054 |
| Poisson | medium* | 0.044 | 0.038 | 0.046 | 0.044 |
| | high* | 0.062 | 0.070 | 0.062 | 0.066 |

* medium: M1=40%, M2=M3=30%. high: M1=20%, M2=M3=40%. $n_1 = n_2 = 50$.

† $t_l = 1/3\,, l = 1,2,3$.

ǂ $t_1 = 0.4, t_2 = t_3 = 0.3$.

** Approximated from empirical distribution.

**Table 4.4** Power simulations with normal outcomes, medium rate of missingness[*] and varying effect sizes

| Effect size | standard MKW test (Deleting all missing data) | | extended MKW test (Partially observed data) | | | | standard MKW test in original data (Assuming no missing) | |
|---|---|---|---|---|---|---|---|---|
| | | | unweighted[†] | | weighted[‡] | | | |
| | permutation-based | large sample approximation | permutation-based | large sample approximation[**] | permutation-based | large sample approximation[**] | permutation-based | large sample approximation |
| 0.08 | 0.21 | 0.19 | 0.24 | 0.24 | 0.24 | 0.23 | 0.49 | 0.48 |
| 0.12 | 0.34 | 0.34 | 0.40 | 0.39 | 0.43 | 0.42 | 0.73 | 0.73 |
| 0.18 | 0.51 | 0.50 | 0.61 | 0.62 | 0.64 | 0.62 | 0.93 | 0.93 |
| 0.24 | 0.67 | 0.66 | 0.81 | 0.80 | 0.82 | 0.81 | 0.99 | 0.99 |

[*] M1=40%, M2=M3=30%. $n_1 = n_2 = 50$.

[†] $t_l = 1/3\,, l = 1,2,3.$

[‡] $t_1 = 0.4, t_2 = t_3 = 0.3.$

[**] Approximated from empirical distribution.

**Table 4.5** Power simulations with normal outcomes and high rate of missingness[*] and varying effect sizes

| Effect size | standard MKW test (Deleting all missing data) | | extended MKW test (Partially observed data) | | | | standard MKW test in original data (Assuming no missing) | |
|---|---|---|---|---|---|---|---|---|
| | | | unweighted[†] | | weighted[‡] | | | |
| | permutation-based | large sample approximation | permutation-based | large sample approximation[**] | permutation-based | large sample approximation[**] | permutation-based | large sample approximation |
| 0.12 | 0.16 | 0.15 | 0.34 | 0.32 | 0.32 | 0.31 | 0.73 | 0.74 |
| 0.18 | 0.26 | 0.24 | 0.54 | 0.51 | 0.55 | 0.54 | 0.93 | 0.92 |
| 0.25 | 0.33 | 0.31 | 0.72 | 0.67 | 0.71 | 0.67 | 0.98 | 0.98 |
| 0.33 | 0.40 | 0.38 | 0.84 | 0.83 | 0.82 | 0.83 | 0.998 | 0.998 |

[*] M1=20%, M2=M3=40%. $n_1 = n_2 = 50$.

[†] $t_l = 1/3\,, l = 1,2,3.$

[‡] $t_1 = 0.2, t_2 = t_3 = 0.4.$

[**] Approximated from empirical distribution.

**Table 4.6** Power simulations with non-normal outcomes and medium rate of missingness[*] and varying effect sizes

| Effect size | standard MKW test (Deleting all missing data) | | extended MKW test (Partially observed data) | | | | standard MKW test in original data (Assuming no missing) | |
|---|---|---|---|---|---|---|---|---|
| | | | unweighted[†] | | weighted[‡] | | | |
| | permutation-based | large sample approximation | permutation-based | large sample approximation[**] | permutation-based | large sample approximation[**] | permutation-based | large sample approximation |
| 0.08 | 0.26 | 0.25 | 0.28 | 0.28 | 0.28 | 0.30 | 0.57 | 0.57 |
| 0.10 | 0.32 | 0.31 | 0.41 | 0.39 | 0.40 | 0.41 | 0.67 | 0.69 |
| 0.12 | 0.38 | 0.36 | 0.45 | 0.42 | 0.47 | 0.46 | 0.78 | 0.79 |
| 0.16 | 0.43 | 0.44 | 0.58 | 0.58 | 0.58 | 0.56 | 0.89 | 0.88 |
| 0.19 | 0.55 | 0.56 | 0.73 | 0.70 | 0.71 | 0.71 | 0.96 | 0.96 |
| 0.26 | 0.73 | 0.71 | 0.87 | 0.85 | 0.89 | 0.88 | 0.99 | 0.99 |
| 0.36 | 0.85 | 0.84 | 0.96 | 0.96 | 0.97 | 0.96 | 0.998 | 0.998 |

[*] M1=40%, M2=M3=30%. $n_1 = n_2 = 50$.

[†] $t_l = 1/3\,, l = 1,2,3$.

[‡] $t_1 = 0.4, t_2 = t_3 = 0.3$.

[**] Approximated from empirical distribution.

**Table 4.7** Power simulations with non-normal outcomes and high rate of missingness[*] and varying effect sizes

| Effect size | standard MKW test (Deleting all missing data) | | extended MKW test (Partially observed data) | | | | standard MKW test in original data (Assuming no missing) | |
|---|---|---|---|---|---|---|---|---|
| | | | unweighted[†] | | weighted[‡] | | | |
| | permutation-based | large sample approximation | permutation-based | large sample approximation[**] | permutation-based | large sample approximation[**] | permutation-based | large sample approximation |
| 0.08 | 0.15 | 0.13 | 0.23 | 0.22 | 0.23 | 0.22 | 0.55 | 0.54 |
| 0.11 | 0.16 | 0.15 | 0.33 | 0.31 | 0.34 | 0.32 | 0.72 | 0.71 |
| 0.13 | 0.19 | 0.17 | 0.40 | 0.38 | 0.41 | 0.38 | 0.76 | 0.77 |
| 0.16 | 0.24 | 0.22 | 0.52 | 0.50 | 0.54 | 0.52 | 0.89 | 0.89 |
| 0.19 | 0.26 | 0.24 | 0.60 | 0.58 | 0.61 | 0.61 | 0.95 | 0.95 |
| 0.27 | 0.38 | 0.35 | 0.81 | 0.77 | 0.79 | 0.78 | 0.99 | 0.99 |
| 0.37 | 0.52 | 0.50 | 0.93 | 0.93 | 0.93 | 0.92 | 1 | 1 |

[*] M1=20%, M2=M3=40%. $n_1 = n_2 = 50$.

[†] $t_l = 1/3\,, l = 1,2,3$.

[‡] $t_1 = 0.2, t_2 = t_3 = 0.4$.

[**] Approximated from empirical distribution.

# 5.0 APPLICATION

## 5.1 YOGA DATA

Data from an open, non-randomized clinical trial for the treatment of persons with schizophrenia (SZ) for improvement in cognitive functions are used in this dissertation as an example (Bhatia *et al*., 2012) in order to illustrate empirically the statistical approaches that are described in earlier sections. The original study was supported partly by grants from the Central Council for Research in Yoga and Naturopathy, AYUSH, MoHFW, India (12-1/CCRYN/2005-2006/Res.P-III) and NIH (MH56242, MH66263, MH 63480 and Indo-US Project Agreement # N-443-645). The objective of this clinical trial was to evaluate adjunctive Yoga Therapy (YT) for cognitive domains impaired in patients with SZ. The participants were outpatients at Dr Ram Manohar Lohia Hospital (RMLH), Delhi, India. All participants were older than 18 years and resided in Delhi. Persons dependent on alcohol/illicit substances or individuals with neurological disorders that interfered with diagnosis or cognitive evaluations were excluded. At the psychiatric outpatient clinics of RMLH, all patients clinically diagnosed with psychoses who fulfilled these criteria were invited to participate in the Yoga therapy.

A total of 396 patients with SZ had suitable clinical diagnoses assigned by their physicians, on the basis of unstructured interviews. 207 of them were included in the trial after some exclusion criteria. This was a non-randomized clinical trial. Some SZ patients refused to

receive yoga therapy and they were assigned to treatment as usual group (TAU) and received conventional pharmacological treatment (without changing for yoga) from their psychiatrists throughout the study. And SZ patients agreed to receive yoga therapy were assigned to yoga therapy group (YT) and received conventional treatment plus yoga therapy as an adjunctive treatment prescribed protocol daily for approximately one hour for 21 consecutive days (excluding Sundays).

Cognitive functions were assessed with a Hindi version of the Penn computerized neuropsychological battery (CNB) (Gur *et al.* 2001 a, b). The CNB included neurocognitive domains known to be impaired among individuals with SZ. The verbal domains were available only in English. As many Indian participants did not speak English, the verbal domains were excluded. Accuracy which reflects the number of correct responses and speed which reflects the median reaction time for eight cognitive domains were assessed. The domains are: abstraction and mental flexibility, attention, working memory, face memory, spatial memory, spatial ability, sensorimotor dexterity and emotion processing. The CNB was assessed at baseline, 21 days post treatment and 2 months post treatment. 63 subjects in YT group and 24 subjects in TAU group completed the intervention period. The trial primarily compares YT (N=63) patients and TAU (N=24) patients to evaluate the adjunctive YT for cognitive domains impaired in SZ. Changes of the CNB in cognitive domains at 2-month assessment point in the TAU group were compared with the YT group. The assessments are illustrated in the flow chart (Figure 5.1).

Except for those with significantly more education and significantly poorer global assessment of worst point functioning scores during recent SZ episode, SZ patients who participated in Yoga and those who refused Yoga were found to be similar in standard

demographic and clinical characteristics with regard to age, sex, marital status and occupation (Table 1 of Bhatia 2012).

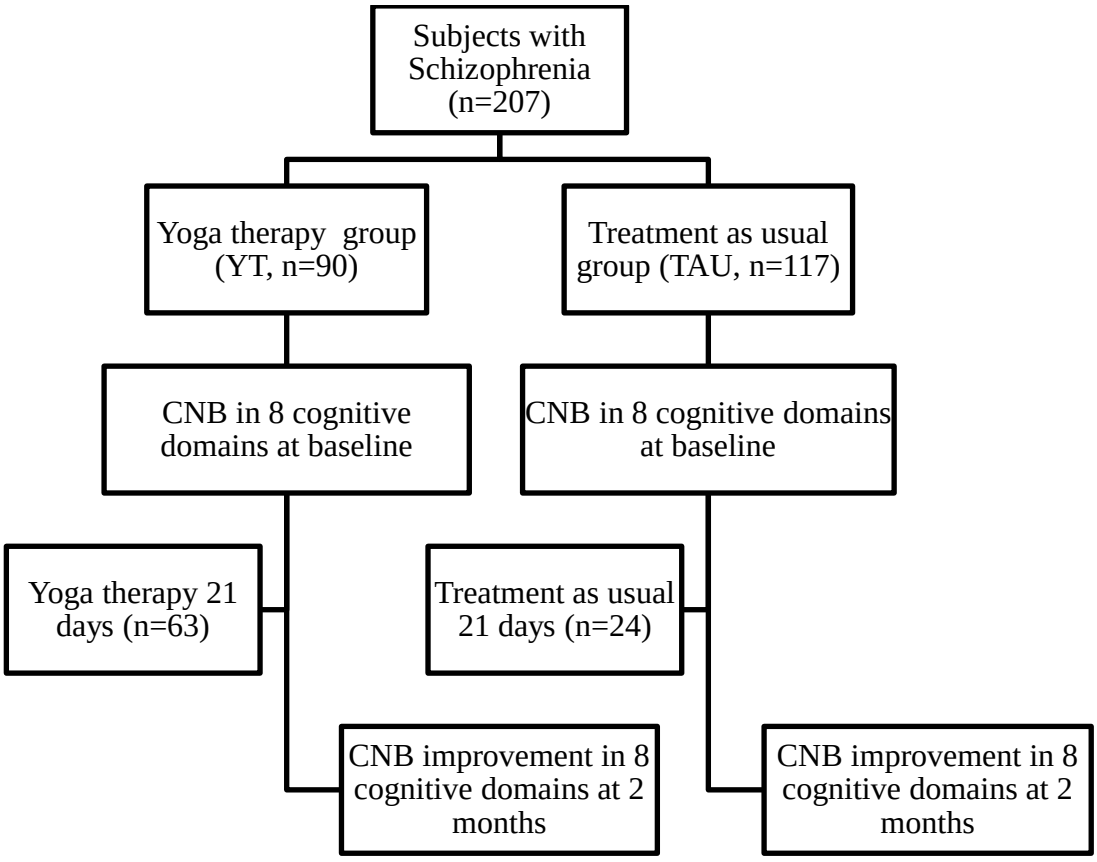


**Figure 5.1** Flow chart of Yoga data

## 5.2 APPLICATION

### 5.2.1 Univariate Kruskal-Wallis Test

In the original study, the changes in CNB in the cognitive domains were compared between YT and TAU groups. A large amount of missing values existed in the data. Only 10 subjects in the YT group and 9 subjects in the TAU group completed CNB in all domains at all assessment points. Moreover, the CNB was skewed distributed. Therefore, the researchers selected a univariate and non-parametric test — Kruskal-Wallis test — to do the comparisons, followed by corrections for multiple comparisons. Univariate Kruskal-Wallis tests use complete cases only.

**Table 5.1** Univariate Kruskal-Wallis tests on Yoga example

| Domains | Variables | Number Of Complete Cases | | | p-value | Adjusted p-value |
|---|---|---|---|---|---|---|
| | | YT (N=63) | TAU (N=24) | All | | |
| Abstraction And Mental Flexibility | $Y_1$ | 23 | 21 | 44 | 0.028 | 0.084 |
| Attention | $Y_2$ | 18 | 16 | 34 | 0.014 | 0.056 |
| Face Memory | $Y_3$ | 26 | 22 | 48 | 0.069 | 0.138 |
| Spatial Memory | $Y_4$ | 24 | 19 | 43 | 0.66 | 0.66 |

I reproduce the tests in the speed summary functions in four domains: abstraction and mental flexibility, attention, face memory and spatial memory. The results are shown in Table 5.1. The speed functions in abstraction and mental flexibility and in attention are shown to

improve more in SZ/YT group than in SZ/TAU group (p-values 0.028 and 0.014, respectively). However, after Hochberg adjustment for multiple comparisons, neither of them remains significantly different between the two groups.

### 5.2.2 Multivariate Tests

MANOVA, standard MKW test and extended MKW tests (likelihood-based and permutation-based) were applied on the improvement of cognitive functions in the four domains to test group differences. The missing pattern of the data is shown in Table 5.2, and the results are shown in Table 5.3.

**Table 5.2** Missing pattern of improvement of cognitive functions in the four domains in Yoga example
O=Observed, M=Missing

| Missing pattern | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $m_l$ | Used in MANOVA and standard MKW test | Used in extended MKW test |
|---|---|---|---|---|---|---|---|
| 1 | O | O | O | O | 32 | Yes | Yes |
| 2 | O | M | O | O | 9 | No | Yes |
| 3 | M | M | O | O | 2 | No | No |
| 4 | O | O | O | M | 1 | No | No |
| 5 | M | O | O | M | 1 | No | No |
| 6 | M | M | O | M | 3 | No | Yes |
| 7 | O | M | M | M | 2 | No | Yes |
| 8 | M | M | M | M | 37 | No | No |

The classical MANOVA did not detect any significant difference between the two groups. The p-value was borderline (p=0.054). And the permutation-based MANOVA proposed by Zeng (2011) did not detect any difference either (p=0.081). On the other hand, the standard MKW test (p=0.0295) and extended MKW tests (p=0.0311 and 0.0343 for two weighting schemes, respectively) gave significant p-values based on permutation distribution, implying that the improvement functions in at least one of the four domains were different between the two groups. The standard MKW test with $\chi^2$ approximation also provides significant p-values (p=0.0381). However, when performing extended MKW test and calculating the p-values based on linear combination of $\chi^2$ distributions, the results are not significant (p=0.5942 and 0.3204 for two weighting schemes, respectively).

**Table 5.3** p-values of MANOVA and MKW tests on Yoga example

| method | | large sample approximation | permutation-based |
|---|---|---|---|
| MANOVA, Wilks' $\lambda$ | | 0.0541 | 0.081 |
| standard MKW tests | | 0.0381* | 0.0295 |
| extended MKW tests | unweighted | 0.5942† | 0.0311 |
| | weighted | 0.3204† | 0.0343 |

* Approximated by $\chi^2_4$ distribution

† Approximated from empirical distribution

# 6.0 DISCUSSIONS

## 6.1 SAMPLE SIZE CALCULATION

Simulation studies of the MKW test are performed under different scenarios to find out the effect sizes required to detect a difference between groups with 80% power in pre-specified sample sizes.

As in section 4, data with $g = 2$ groups and $p = 2$ outcome variables are simulated. To generate correlated outcomes, I use a latent variable $X$.

Set $X \sim BIN(5,0.5)$.

For group 1, generate $X_1, \dots, X_{n_1}$ as a random sample of $X$.

Set $Y_{11}|X \sim POI(1 + X)$, $Y_{12}|X \sim POI(2 + X)$.

Generate $\left[\left(Y_{1j1}, Y_{1j2}\right)|X_j\right]$ as a random sample of $\left[(Y_{11}, Y_{12})|X_j\right], j = 1, \dots, n_1$.

For group 2, generate $X_1, \dots, X_{n_2}$ as another random sample of $X$.

Set $Y_{21}|X \sim POI(1 + X)$, $Y_{22}|X \sim POI(2 + \Delta + X)$.

Generate $\left[\left(Y_{2j1}, Y_{2j2}\right)|X_j\right]$ as a random sample of $\left[(Y_{21}, Y_{22})|X_j\right], j = 1, \dots, n_2$.

For each pre-specified sample size, various Δ's are selected for differing effect sizes. The cut-off effect sizes for corresponding sample sizes are summarized in Table 6.1. Practitioners can use this type of table as a guide when designing the experiment. Additional tables can be generated with the help of the computer program developed in this thesis.

**Table 6.1** Effect sizes required to detect a difference between two groups with 80% power in pre-specified sample sizes

| Sample Size / Group | Effect Size |
|---|---|
| 20 | 0.67 |
| 30 | 0.25 |
| 50 | 0.14 |

## 6.2 SINGULARITY ISSUE

One drawback of MKW test is the singularity issue of the estimate of covariance matrix **V**. For example, in a 2-group, 2-outcome case, suppose two observations are (0.5, 2) and (1, 1.5). The rank variables are (1,2) and (2,1).

The estimate of the covariance matrix is

$$\mathbf{V} = 2\begin{pmatrix} 1-\frac{3}{2} \\ 2-\frac{3}{2} \end{pmatrix}\left(1-\frac{3}{2}, 2-\frac{3}{2}\right)' = \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix},$$

which is singular.

When singularity issue arises, the inverse of **V** cannot be obtained and thus MKW test cannot be performed.

When sample sizes increase, the possibility of singularity issue will decrease. In real data when there are more than 10 subjects per group, singularity issue will rarely happen. Future work with generalized inverses is warranted.

# 7.0 CONCLUSIONS

Comparisons between groups play a central role in clinical research. In trials with multivariate outcomes, the classical parametric methods such as MANOVA model have two major drawbacks. They require distributional assumptions such as multivariate normality. When sample size is small, or response variables are ordinal outcomes, the violation of normality prevents clinical researchers from group comparisons using parametric multivariate methods. And when performing the tests by standard software, incomplete cases are deleted. Non-parametric multivariate methods are available in the statistical literature. They circumvent the former issue. However, they do not resolve the latter one, the missing data issue. The non-parametric multivariate methods are not commonly used in clinical research.

In this dissertation, I examine the performance of the non-parametric multivariate Kruskal-Wallis (MKW) test with a simulation study in a variety of scenarios, and conclude that this test can be applied in multivariate non-normal data, with good coverage and power.

Further, I propose an extension of the MKW test for multivariate data with missingness. The proposed method retrieves information in partially observed cases. A simulation study shows that the proposed extended method provides higher power level than the standard MKW test, encompassing a broad spectrum of multivariate effect sizes. MKW test with the proposed extension is a powerful alternative to test group difference on non-normal multivariate data with missing values.

An R-based program is written to implement the standard MKW test and the extended MKW test. The program is user-friendly and ready to use for researchers. It is provided in Appendix A. The parameters need to be reassigned before performing the test.

The public health relevance of this work lies in the development of a new powerful methodology with user-friendly computer software for group comparisons in non-normal multivariate data with or without missingness.

# APPENDIX

## R CODE FOR MKW TEST

```
library(lattice)
library(Matrix)

################################
###### 1. mult-KW funtion ######
################################

multkw<- function(group,y,simplify=FALSE){

### sort data by group ###
      o<-order(group)
      group<-group[o]
      y<-as.matrix(y[o,])

      n<-length(group)
      p<-dim(y)[2]
      if (dim(y)[1] != n)
            return("number of oberservations not equal to length of group")
      groupls<-unique(group)
      g<-length(groupls)     #number of groups#
      groupind<-sapply(groupls,"==",group)  #group indicator#
      ni<-colSums(groupind)   #num of subj of each group#
      r<-apply(y,2,rank) #corresponding rank variable#

### calculation of statistic   ###
      r.ik<-t(groupind)%*%r*(1/ni)  #gxp,  mean  rank  of  kth  variate  in  ith
group#
      m<- (n+1)/2          #expected value of rik#
      u.ik<-t(r.ik-m)
      U<-as.vector(u.ik)
      V<-1/(n-1)*t(r-m)%*%(r-m) #pooled within-group cov matrix
      Vstar<-bdiag(lapply(1/ni,"*",V))
      W2<-as.numeric(t(U)%*%solve(Vstar)%*%U)

### return stat and p-value ###
      returnlist<-list(statistic=W2,d.f.=p*(g-1),
            p.value=pchisq(W2,p*(g-1),lower.tail=F))
```

```
      if (simplify==TRUE) return (W2)
      else return (returnlist)
}


########################################
######## 2 MKW with missing values #######
########################################

mkw.m<-function(group,y,r,weight){
### count missng patterns ###
      p<-dim(y)[2]
      r.order<-r
      y.order<-y
      g.order<-group
      for (i in 1:p){
            oo<-order(r.order[,i])
            y.order<-y.order[oo,]
            g.order<-g.order[oo]
            r.order<-r.order[oo,]
      }
      J<-nrow(unique(r.order,MARGIN=1)) #number of missing patterns
      D<-data.frame(r.order)
      n<-length(group)
      ones<-rep(1,n)
      mc<-aggregate(ones,by=as.list(D),FUN=sum)   #counts   of   each   missing
pattern
      mi<-mc$x
      pi<-p-rowSums(mc[,1:p])

### get W^2_j ###
      W2<-rep(0,J)
      W2.c<-0
      i.st<-1
      for (j in 1:J){
            i.end<-i.st+mi[j]-1
            gg<-g.order[i.st:i.end]
            yy<-y.order[i.st:i.end,]
            ii<-mc[j,1:p]==F
            if (sum(as.numeric(ii))>0){
                  yy1<-as.matrix(yy[,ii])
                  if (mi[j]>pi[j]) W2[j]<-multkw(gg,yy1,simplify=T)  ##### if
mi[j]>p needs to dig more
            }
            if (prod(as.numeric(ii))==1) W2.c<-W2[j]
            i.st<-i.end+1
      }
      if (weight=="prop") tj<-mi/sum(mi)  else tj<-1/J
      W2<-sum(tj*W2)
      nu<-(W2)^2/sum((tj*W2)^2/pi/(g-1))
      return(list(W2.m=W2,nu=nu,W2.c=W2.c))
}


########################################
###### 3. monte carlo permutation ######
########################################
```

```
multkw.perm<-function(nmc,group,y,r,weight){
### count missng patterns ###
      p<-dim(y)[2]
      r.order<-r
      y.order<-y
      g.order<-group
      for (i in 1:p){
            oo<-order(r.order[,i])
            y.order<-y.order[oo,]
            g.order<-g.order[oo]
            r.order<-r.order[oo,]
      }
      J<-nrow(unique(r.order,MARGIN=1)) #number of missing patterns
      D<-data.frame(r.order)
      n<-length(group)
      ones<-rep(1,n)
      mc<-aggregate(ones,by=as.list(D),FUN=sum)   #counts   of   each   missing
pattern
      mi<-mc$x

      W2.m.perm<-rep(0,nmc)
      W2.c.perm<-rep(0,nmc)
      stats0<-mkw.m(group,y,r,weight)
      W2.m<-stats0$W2.m
      W2.c<-stats0$W2.c
      nu<-stats0$nu
      for (i in 1:nmc){
            i.st<-1
            group.perm<-rep(0,n)
            group.perm<-sample(group,size=n)
            stats<-mkw.m(group.perm,y,r,weight)
            W2.m.perm[i]<-stats$W2.m
            W2.c.perm[i]<-stats$W2.c
      }
      p.mkw.m.perm<-sum(W2.m<W2.m.perm)/nmc
      p.mkw.m.chi2<-pchisq(W2.m,nu,lower.tail=FALSE)
      p.mkw.c.perm<-sum(W2.c<W2.c.perm)/nmc
      p.mkw.c.chi2<-pchisq(W2.c,p*(g-1),lower.tail=FALSE)
      return(list(W2.m=W2.m,p.mkw.m.perm=p.mkw.m.perm,p.mkw.m.chi2=p.mkw.m.ch
i2,
                  p.mkw.c.perm=p.mkw.c.perm,p.mkw.c.chi2=p.mkw.c.chi2))
}


########################################
######### 4. data generation ############
########################################

data.gen<-function(p,g=2,ni,delta,mpcnt){
      n<-ni*g
      X1<-rnorm(ni)
      Y11<-sapply(X1+1,rnorm,n=1,sd=sqrt(2))
      Y12<-sapply(X1,rnorm,n=1,sd=1)
      X2<-rnorm(ni)
      Y21<-sapply(X2+1,rnorm,n=1,sd=sqrt(2))
      Y22<-sapply(X2+delta,rnorm,n=1,sd=1)
```

```
      Y1<-matrix(c(Y11,Y12),nrow=ni,ncol=p)
      Y2<-matrix(c(Y21,Y22),nrow=ni,ncol=p)
      y<-rbind(Y1,Y2)
      group<-rep(1:g,each=ni)
      m1<-matrix(rep(c(0,0),times=n*mpcnt[1]),ncol=p,byrow=T)
      m2<-matrix(rep(c(0,1),times=n*mpcnt[2]),ncol=p,byrow=T)
      m3<-matrix(rep(c(1,0),times=n*mpcnt[3]),ncol=p,byrow=T)
      m<-rbind(m1,m2,m3)
      perm<-sample(n)
      r<-m[perm,]
      return(list(group=group,y=y,r=r))
}

data.gen2<-function(p,g=2,ni,delta,mpcnt){
      n<-ni*g
      X1<-rbinom(ni,5,0.5)
      W1<-rbinom(ni,2,0.5)
      Y11<-sapply(X1+1,rpois,n=1)
      Y12<-sapply(X1+2,rpois,n=1)
      X2<-rbinom(ni,5,0.5)
      Y21<-sapply(X2+1,rpois,n=1)
      Y22<-sapply(X2+2+delta,rpois,n=1)
      Y1<-matrix(c(Y11,Y12),nrow=ni,ncol=p)
      Y2<-matrix(c(Y21,Y22),nrow=ni,ncol=p)
      y<-rbind(Y1,Y2)
      group<-rep(1:g,each=ni)
      m1<-matrix(rep(c(0,0),times=n*mpcnt[1]),ncol=p,byrow=T)
      m2<-matrix(rep(c(0,1),times=n*mpcnt[2]),ncol=p,byrow=T)
      m3<-matrix(rep(c(1,0),times=n*mpcnt[3]),ncol=p,byrow=T)
      m<-rbind(m1,m2,m3)
      perm<-sample(n)
      r<-m[perm,]
      return(list(group=group,y=y,r=r))
}


########################################
############ 5. simulation #############
########################################
p<-2
g<-2
ni<-50
delta<-2.5
mpcnt<-c(0.2,0.4,0.4)

nsim<-1000
psim<-matrix(0,nrow=nsim,ncol=8)
wilks<-rep(0,nsim)
tau<-rep(0,nsim)

nmc<-1000
r.comp<-matrix(0,nrow=ni*g,ncol=p)

chi2<-matrix(0,nrow=10000,ncol=3)
chi2[,1]<-rchisq(10000,df=2)
chi2[,2]<-rchisq(10000,df=1)
chi2[,3]<-rchisq(10000,df=1)
```

```
ptable1<-rowMeans(chi2)
ptable2<-chi2%*%as.matrix(mpcnt)

Sys.time()
for(i in 1:nsim){
      data1<-data.gen2(p,g,ni,delta,mpcnt)
      group<-data1$group
      y<-data1$y
      r<-data1$r
      fit<-manova(y ~ group)
      wilks[i]<-(summary(fit,test="Wilks"))$stats[1,2]   #value   of   wilks'
lambda
      ps1<-multkw.perm(nmc,group,y,r,weight="plain")
      ps2<-multkw.perm(nmc,group,y,r,weight="prop")
      ps.comp<-multkw.perm(nmc,group,y,r.comp,weight="prop")
      psim[i,1]<-ps1$p.mkw.m.perm
      psim[i,2]<-mean(ps1$W2.m<ptable1)
      psim[i,3]<-ps1$p.mkw.c.perm
      psim[i,4]<-ps1$p.mkw.c.chi2
      psim[i,5]<-ps2$p.mkw.m.perm
      psim[i,6]<-mean(ps2$W2.m<ptable2)
      psim[i,7]<-ps.comp$p.mkw.m.perm
      psim[i,8]<-ps.comp$p.mkw.m.chi2
}
Sys.time()

pw<-colMeans(psim<0.05)
f2<-mean(1/wilks-1)

fname<-paste("power est n",ni, "delta",delta,"-poi-h",".txt")
write.table(c(f2,pw),file=fname,row.names=T,append=T)
```

## BIBLIOGRAPHY


Anderson, T.W. (2003), *An Introduction to Multivariate Statistical Analysis* (Third Edition). Hoboken, NJ: Wiley-Interscience.

Bhatia, T., Agarwal, A.S., Wood, J., Richard, J., Gur, R.E., Gur, R.C., Nimgaonkar, V.L., Mazumdar, S., & Deshpande, S.N. (2012), Adjunctive cognitive remediation for schizophrenia using yoga: an open non-randomised trial. *Acta Neuropsychiatrica.* 24(2), 91-100.

Bilker, W.B., Brensinger, C., & Gur, R.C. (2004), A Two Factor ANOVA-Like Test For Correlated Correlations: CORANOVA. *Multivariate Behavioral Research*. 39 (4), 565-594.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992), A Power Primer. *Psychological Bulletin*. 112(1), 155-159.

Edgington, E.S. & Onghena, P. (2007), *Randomization Tests* (Fourth Edition). Chapman and Hall/CRC.

Ferguson, C.J. (2009), An Effect Size Primer: A Guide For Clinicians and Researchers. *Professional Psychology: Research and Practice*. 40(5), 532-538.

Filiz, Z. (2003), Multivariate Repeated Measures Experiment And An Application. *Hacettepe Journal of Mathematics and Statistics*. 2003. 32, 75-90.

Fleishman, A.I. (1978), A Method For Simulating Non-Normal Distributions. *Psychometrika*. 43(4), 521-531.

Frees, E.W., and Valdez, E. A. (1998), Understanding Relationships Using Copulas. *North American Actuarial Journal*, 2, 1–25.

Ghosh, M., Grizzle, J.E., & Sen, P.K. (1973), Nonparametric Methods in Longitudinal Studies. *Journal of the American Statistical Association*. 68(341), 29-36.

Gur R.C., Ragland J.D., Moberg P.J., Turner T.H., Bilker W.B., Kohler C., Siegel S.J., and Gur R.E. (2001), Computerized neurocognitive scanning: I. Methodology and validation in healthy people. *Neuropsychopharmacology*. 25, 766-776.

Gur R.C., Ragland J.D., Moberg P.J., Bilker W.B., Kohler C., Siegel S.J., and Gur R.E. (2001), Computerized neurocognitive scanning: II. The profile of schizophrenia. *Neuropsychopharmacology*. 25, 777-788.

Harwell, M.R. and Serlin, R.C. (1989), A Nonparametric Test Statistic for the General Linear Model. *Journal Of Educational And Behavioral Statistics*. 14(4), 351-371.

Hofert, M. & Maechler, M. (2011), Nested Archimedean Copulas Meet R: The nacopula Package. *Journal of Statistical Software*. 39(9), 1-20.

Katz, B.M. & McSweeney, M. (1980), A Multivariate Kruskal-Wallis Test With Post Hoc Procedures. *Multivariate Behavioral Research*. 15(3), 281-297.

Kawaguchi, A. & Koch, G. (2010), Multivariate Mann–Whitney Estimators for the Comparison of Two Treatments in a Three-Period Crossover Study with Randomly Missing Data. *Journal of Biopharmaceutical Statistics*, 20, 720–744,

Kim, K. & Timm, N. (2006), *Univariate and Multivariate General Linear Models: Theory and Applications with SAS* (Second Edition), Chapman and Hall/CRC.

Kojadinovic, J. & Yan J. (2010), Modeling Multivariate Distributions with Continuous Margins Using the copula R Package. *Journal of Statistical Software*. 34(9), 1-20.

Kruskal, W.H., and Wallis W.A. (1952), Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260), 583-621.

Pesarin, F. (2001), *Multivariate Permutation Tests: With Applications to Biostatistics*, John Wiley & Sons.

May, W.L. and Johnson, W.D. (1997), A SAS Macro For The Multivariate Extension Of The Kruskal-Wallis Test Including Multiple Comparisons: Randomization And $\chi^2$ Criteria. *The Statistical Software Newsletter*. 26(2), 239-250.

Satterthwaite, F.E. (1946), An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*. 2, 110–114.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Sen, P. K. (1998), Multivariate Median And Rank Sum Tests. *Encyclopedia of Biostatistics*. Vol. 4, 2887-2900. Chichester: J. Wiley.

Steyn, H.S. & Ellis, S.M. (2009), Estimating an Effect Size in One-Way Multivariate Analysis of Variance (MANOVA). *Multivariate Behavioral Research.* 44(1), 106-129.

Welch, B.L. (1947), The Generalization Of "Student's" Problem When Several Different Population Variances Are Involved. *Biometrika* 34, 28–35.

Yan, J. (2007), Enjoy the Joy of Copulas: With a Package copula. *Journal of Statistical Software*. 21(4), 1-21.

Zeng, C., Pan, Z., MaWhinney, S., Baron, A. E., and Zerbe, G. O. (2011), Permutation and F Distribution of Tests in the Multivariate General Linear Model. *The American Statistician*. 65(1), 31-36.