# EVALUATION OF THE NORMAL APPROXIMATION FOR THE PAIRED TWO SAMPLE PROBLEM WITH MISSING DATA

By

Shang-Lin Yang

B.S., National Taiwan University, 1996

M.S., University of Pittsburgh, 2005

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Shang-Lin Yang

It was defended on

June 07, 2005

and approved by

Thesis Advisor:
John W. Wilson, PhD
Associate Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Vincent C. Arena, PhD
Associate Professor,
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Maria M. Brooks, PhD
Assistant Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

John W. Wilson, Ph.D.

# EVALUATION OF THE NORMAL APPROXIMATION FOR THE PAIRED TWO SAMPLE PROBLEM WITH MISSING DATA

Shang-Lin Yang, M.S.

University of Pittsburgh, 2005

## Abstract

Previous authors have combined tests for pairs and unpaired data so that population means can be compared using a paired study design with incomplete data. The primary object of my thesis is to determine the appropriate sample size and the appropriate proportion and configuration of complete data and incomplete data so that a normal approximation can be used to calculate p-values. The test statistic studied is one due to Wilson (1992) in which the sign test and rank sum test are combined to form of composite test statistic.

To fulfill these objectives, the following approach is adopted:

(1) Choose different data scenarios in terms of different sample sizes of paired data and different proportions of complete data.

(2) Obtain the exact sampling distribution of the test statistic under each data scenario we study.

(3) Obtain the normal approximation distribution under each data scenario we study.

(4) Compare the exact and approximate cumulate distribution by their difference on each possible test statistic value.

The results show that when the study groups are approximately balanced with respect to incomplete data, and have at least 9 observations in each group, the normal approximation appears to be useful when the number of complete pairs is as low as 5. However, when the groups are highly unbalanced with respect to incomplete data, using the normal approximation seems not to be appropriate, at least when the total sample size is 70 or less. These results may make public health studies easier to carry out when the data include both complete and incomplete pairs.

# ACKNOWLEDGEMENT

First, I would like to thank my advisor Dr. John Wilson for his invaluable guidance at each stage of my thesis. I really appreciate my advisor that he offered me his great idea and always patiently assisted me throughout the thesis including software guiding and entire process. Without his patience, without his help and support, this paper would not be completed.

Second, I would like to thank other members of my thesis committees, Dr. Vincent Arena and Dr. Brooks, giving me valuable suggestions and comments to fulfill my work.

Then I would like to express my gratitude to Miss Xiaoqing Wang for her encouragement and other technical support.

Lastly, but most importantly, I want to thank my greatest parents, especially my mother, for their constant support and encouragement for my life and my study, without whom this thesis work would have been impossible, and my life would not have been successful and smooth.

## TABLE OF CONTENTS

PAGE

1. INTRODUCTION……………………………………………………………………..........1

2. THE PAIRED TWO-SAMPLE WITH MISSING DATA MODEL AND ITS NOTATION….4

3. DESCRIPTION OF NON-PARAMETRIC TEST STATISTIC T1

(WILSON 1992) ………………………………………………………..……………..…...8

3.1. Test Statistic T1…………………….……………………….…………………………..8

3.2. Expected Value and Variance of T1 under the Null Hypothesis…….……………………11

4. LARGE SAMPLE APPROXIMATION AND DIFFERENT DATA SCENARIOS…………12

4.1. Central Limit Theorem……………………………………………………………...........12

4.2. Different Data Scenarios…………………………………………………………………13

5. ANALYSIS PROCEDURE AND CRITERION CHOOSING………………………………15

5.1. Procedure of Analysis…………………………………………………………….....15

5.2. Choice of Criterion for Deciding Whether Normal Approximation is Appropriate……..19

6. RESULTS AND DISCUSSION…………………………………………………………21

7. CONCLUSIONS…………………………………………………………...………25

BIBLIOGRAPHY……………………………………………………………………………..26

# LIST OF TABLES

Table No. Page

1. Basic data configuration based on the experiment carried by Weidmann et al……….……..7

2. The ranks based on T1 test for the T lymphocytes (3-2-3) data……………………….........10

3. Listing of studied data configurations……………………………………………….……14

4. Calculation procedure for the (3-2-3) basic data configuration……………………………..17

5. “Differences” from studied data configurations under the given α level……………….......18

6. “Differences” for the Wilcoxon Rank-Sum and Sign-Rank tests………………………..….20

7. “Differences” from J-8-8, J-7-8, and J-7-7 data configurations…………………………….23

8. “Differences” for data configuration J-10-10 and J-9-9 under the given α level…………...24

# LIST OF FIGURES

Figure No. Page

1. Model of paired two-sample problem with missing data

   (adapted from Im 2002)…………………………………………………...……..6

# 1. INTRODUCTION

It is common to apply normal approximations in statistics problems because they can simplify the calculation of p-values and confidence limits. However, these approximations typically work well only when sample sizes are moderately large. For example, the Wilcoxon Signed Rank test requires at least 16 pairs for the normal approximation to apply and the Wilcoxon Rank Sum test requires at least 10 observations in each group for the normal approximation to be useful (Rosner 2000). Of course, the exact distribution of the test statistic can often be obtained as in a rank-based test, but the distribution is also different for each different configuration of ranks. This means that it is not practical to find the exact distribution of a test statistic for a large sample size because it may become tedious or impractical to compute. Furthermore, to derive its p-value and to test the hypothesis of interest of may be computationally difficult. However, the normal approximation often provides a simpler way to handle the problem. In my thesis, I will discuss when it is appropriate to apply a normal approximation for a particular nonparametric test for comparing group medians when there are both complete and incomplete pairs of data.

This thesis is motivated by the work of J. Wilson (Wilson 1992). Wilson proposed a nonparametric rank-based test to compare the proportions of Vβ T-cell receptors of tumor-infiltrating lymphocytes (TILs) and peripheral blood lymphocytes (PBLs) in patients with hepatocellular carcinoma (Weidmann et al. 1992). Because Weidmann et al.'s experiment involved data from a small sample of eight patients, and data collected on two main measures were complete for only three of these patients, standard statistical methods for paired data were not appropriate. Wilson (1992), Brunner and Neumann (1984), P.K, Sen (personal communication to J. Wilson), and Im (2002) proposed different test statistics to handle this kind of problem.

Wilson (1992) proposed combining two different nonparametric rank-based tests for this type of data, using a sign test for the complete pairs and a Wilcoxon rank sum test for the incomplete pairs. That is, the observations within each complete pair were ranked 1 or 2 and then the sum was taken for group 1 across all pairs. The contribution to the test statistic from the incomplete pairs was a rank sum statistic (summing the ranks for group 1) obtained in the standard manner (Rosner 2000). The sum of the test statistics from the complete and incomplete pairs was denoted T1 (Wilson 1992).

The test statistic we will denote T2 (Brunner and Neumann 1984) is the sum of two rank sum statistics from both complete and incomplete pairs. The contribution to the T2 test statistic made by the complete pairs is the sum of the ranks in one group (group 1) that combine all complete pairs from the two groups being compared (group 1 and group 2). For the incomplete pairs in T2, all the values in incomplete pairs are combined together and ranked, and the ranks in the same group as are chosen as in the complete pairs were summed. The test statistic T3 (P. K, Sen) is also the sum of two test statistics from the complete and incomplete pairs. For the complete pairs in T3, each observation for the complete pairs is ranked after the mean of each pair is subtracted from the raw observation values.  In other words, the contribution to the T3 test statistic from the complete pairs is the aligned rank statistic (Lehmann 1975). The statistic for the incomplete pairs in T3 is the same as the one for the incomplete pairs in T2. The T4 statistic proposed by Im (2002) is again the sum of two statistics from the complete and incomplete pairs as in T2 and T3. The test statistic of complete pairs in T4 is the sum of the ranks from the absolute differences with a negative sign subtracted from the sum of the ranks from the absolute differences with a positive sign. For the incomplete pairs in T4, first rank the combined incomplete observations from group 1 and group 2, and sum the ranks in group 1. The test statistic for the incomplete pairs in T4 is the value that subtracts the expected sum of ranks from the observed sum of ranks for the incomplete pairs in group 1. Im (2002) compared the power of these four test statistics by using Monte Carlo studies. These four test statistics are very conservative at the 0.05 nominal $\alpha$ level. In general, T1 was not as powerful as the other three statistics for most scenarios examined. T2, T3, and T4 performed similarly to one another. However, the power of T1 approached the power of the other 3 statistics when its natural alpha was close to 0.05.

Wilson's proposed method did not specify when or whether a normal approximation can be used to approximate this statistic when the sample size becomes large and exact distributions are inconvenient to obtain. It is the purpose of this thesis to extend Wilson's work and determine when to properly apply a normal approximation for large samples.

The contents of this thesis are as follows: Section 2 introduces the paired two-sample with missing data model and its notation. Section 3 reviews the nonparametric test proposed previously by Wilson. Large sample size and different data scenarios will be presented in Section 4. Analysis methods and criterion choosing will be derived in Section 5. The results from the calculation and a discussion will be presented in Section 6, followed by conclusions in Section 7.

## 2. THE PAIRED TWO-SAMPLE WITH MISSING DATA MODEL AND ITS NOTATION

The model of paired two-sample problem with missing data is presented in Figure 1. As shown in the figure, this model involves J pairs of completed data, K observations with data missing from one group, and L observations with missing data in the another group. In such a model, we will assume that the probability that an observation is missing is independent of the outcome and that the underlying populations are continuous and have the same shape under the null hypothesis. Table 1 shows the basic data configuration based on the experiment carried by Weidmann et al. This experiment compares the proportions of certain T cell receptor gene families, the Vβ gene families, between tumor infiltrating lymphocytes (TILs) and peripheral blood lymphocytes (PBLs). It was hypothesized that the Vβ gene families would show detectable changes in the presence of tumor. To test the hypotheses, the relative proportions of Vβ family usage for several patients' TILs and PBLs were estimated and compared. However, due to some non-measurement factors, some data were not collected. As shown in Table 1, in the original experiment dataset, data from both TILs and PBLs were available for only 3 of the 8 patients studied. For 2 of the 8 patients, data were available from TILs only. Data were only available from PBLs for the remaining 3 patients. So the test comparing the data from TILs and PBLs must take into account not only the small sample size but also the incomplete nature of the dataset.

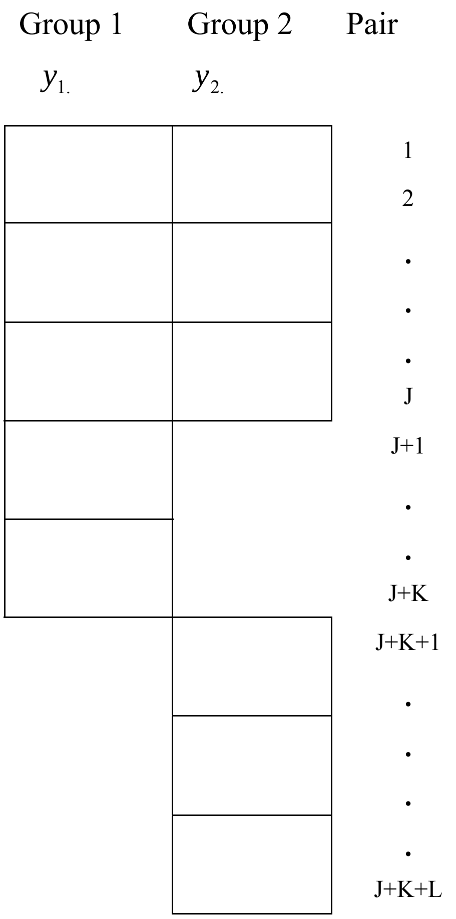
It is hypothesized that there will be detectable changes in the surface receptors of T lymphocytes in the presence of tumor. The following one-sided test was used throughout the analysis:

$H_0$ : $m_{TIL} = m_{PBL}$

$H_A$: $m_{TIL} > m_{PBL}$

Where $m_{TIL}$ = Median percentages of Vβ per cent usage in TIL

$m_{PBL}$ = Median percentages of Vβ per cent usage in PBL



$y_{ij}$ = Response from $i^{th}$ group and $j^{th}$ pair

**Figure 1. Model of paired two-sample problem with missing data (adapted from Im 2002)**

Thus we have:

Group 1: $Y_{1,1}$, $Y_{1,2}$,…, $Y_{1,J}$, $Y_{1,J+1}$,…, $Y_{1,J+K}$
Group 2: $Y_{2,1}$, $Y_{2,2}$,…, $Y_{2,J}$, $Y_{2,J+K+1}$, $Y_{2,J+K+2}$, …, $Y_{2,J+K+L}$

Note: For the convenience, all the data configurations mentioned later in this thesis will be referred to the notation by J-K-L. "J" indicates the number of complete pairs, "K" the number of incomplete pairs with missing data in Group 2, and "L" the number of incomplete pairs with missing data in Group 1.

**Table 1.Basic data configuration based on the experiment carried by Weidmann et al.**

| | Data Vβ8 ( %) | |
|---|---|---|
| Patient | TIL | PBL |
| 1 | 6.7 | 2.8 |
| 2 | 3.7 | 3.5 |
| 3 | 4.4 | 4.1 |
| 4 | 2.3 | • |
| 5 | 4.5 | • |
| 6 | • | 4.0 |
| 7 | • | 14.7 |
| 8 | • | 3.2 |

## 3. DESCRIPTION OF NON-PARAMETRIC TEST STATISTIC T1 (WILSON 1992)

### 3.1. Test Statistic T1

This section describes the test statistic that is the focus of this thesis. The notation follows that given in Im (2002). The test statistic T1 proposed by Wilson (1992) combines two different nonparametric rank-based tests, the sign test and the Wilcoxon rank sum test, to handle the paired design with missing data. For the observations with complete pairs of data, each complete pairs of values are compared and assigned a rank of 1 or 2, depending on which value is larger (the larger value is assigned 2 and the smaller one assigned 1). Then ranks for only one group are summed. For the incomplete pairs, all the values (including data from group 1 and group 2) are combined together and ranked, and the ranks in the same group as are chosen as in the complete parts were summed. The T1 test statistic is the sum of the two statistics computed separately from the complete and incomplete parts. The equation and its notation are stated as follows:

For the Complete Pairs:

Let $R_W(y_{1i}) = \begin{bmatrix} 1 & if & y_{1i} < y_{2i}, & i = 1 \ldots J \\ 2 & if & y_{1i} > y_{2i}, & i = 1 \ldots J \end{bmatrix}$ ["w" stands for "within each pair"]

Rank the observations within each pair. Then sum one group (in our case, we chose group 1), and let $T_c$ stand for the test statistic of the sum of ranks from group1 for the complete pairs.

$$T_c = \sum_{i=1}^{J} R_W(y_{1i})$$

For the incomplete pairs:

Rank all incomplete data in two groups, and sum the ranks from group1.

$$T_i = \sum_{i=J+1}^{J+k} r(y_{1i}),$$

where $r(y_{1i})$ is the rank in group 1, and $T_i$ is the Wilcoxon rank sum test statistic when handling the incomplete pairs.

The final test statistic, T1, is defined as: T1 = $T_C$ + $T_i$

The ranks based on T1 test for the T lymphocytes (3-2-3) data are shown in Table 2. In this case, T1 statistic for 3-2-3 data is:

T1=$T_C$ + $T_i$ = 6+5 = 11

**Table 2. The ranks based on T1 test for the T lymphocytes (3-2-3) data**

|  | Data Vβ8 ( %) |  | Ranks for Calculation of T1 |  |
|---|---|---|---|---|
| Patient | TIL | PBL | TIL | PBL |
| 1 | 6.7 | 2.8 | 2 | 1 |
| 2 | 3.7 | 3.5 | 2 | 1 |
| 3 | 4.4 | 4.1 | 2 | 1 |
| 4 | 2.3 | • | 1 | • |
| 5 | 4.5 | • | 4 | • |
| 6 | • | 4.0 | • | 3 |
| 7 | • | 14.7 | • | 5 |
| 8 | • | 3.2 | • | 2 |

### 3.2. Expected Value and Variance of T1 under the Null Hypothesis

For each of the complete pairs: (sign test)

Suppose that X is a random variable for which the p.d.f. f(x) is as follows:

$$f(x) = \begin{cases} \frac{1}{2} & for \quad x = 1 \\ \frac{1}{2} & for \quad x = 2 \end{cases}$$

So we have E(x) = $1 * \frac{1}{2} + 2 * \frac{1}{2} = \frac{3}{2}$ , and

V(x) = E( $x^2$ ) – $(E(x))^2$ = [ $\frac{1}{2}$ * ( $1^2$ ) + $\frac{1}{2}$ * ( $2^2$ ) ] - $\left(\frac{3}{2}\right)^2$ = $\frac{1}{4}$

Because these are J independent pairs in the complete part of our dataset, the expected value and variance of $T_C$ are as follows:

E ( $T_C$ ) = J * $\frac{3}{2}$ and V ( $T_C$ ) = $\frac{J}{4}$

For the incomplete parts: (Wilcoxon rank sum test)

E ( $T_i$ ) = $\frac{K(K+L+1)}{2}$ (Rosner 2000)

V ( $T_i$ ) = $\frac{KL(K+L+1)}{12}$ (Rosner 2000)

Because the pairs are independent of each other, the test statistics from complete and incomplete parts are also independent of each other. So from the equation T1 = $T_C$ + $T_i$, we get:

$$E(T1) = E(T_C) + E(T_i) = \frac{3J}{2} + \frac{K(K+L+1)}{2}, \qquad (1)$$

$$V(T1) = V(T_C) + V(T_i) = \frac{J}{4} + \frac{KL(K+L+1)}{12} \qquad (2)$$

# 4. LARGE SAMPLE APPROXIMATION AND DIFFERENT DATA SCENARIOS

## 4.1. Central Limit Theorem

The central limit theorem is a common and important theorem in statistical inference because it is the basis for applying normal approximations. In many cases, experimental data do not appear to follow a normal distribution, and calculation of exact probabilities for tests and confidence intervals can be difficult. In such cases, the normal approximation may offer a convenient way to carry out confidence intervals and tests.

When the underlying distribution is normal, it can be shown that the sample mean will be normally distributed itself with a mean $\mu$ and variance ${\sigma^2}/{n}$. When the underlying distribution is not normal and the sample size is large, the central limit theorem will allow us to use a normal approximation to carry out statistical inference irrespective of the population distribution.

The Theorem is given as follows,

Suppose $Y_i$ is the sum of the i random variables $X_1$, $X_2$, …, $X_i$, $\mu_i$ is the mean of $Y_i$, and $\sigma_i^2$ is the variance of $Y_i$. When the number of random variables goes to infinity, the distribution function of the random variable $\frac{Y_i - \mu_i}{\sigma_i}$ will approximate the standard normal distribution function. (Conover 1999)

The version of the CLT used here is

$$\mathrm{P}\left[\frac{\sum Y_i - \sum \mu_i}{\sum \sigma_i} \le c\right] \xrightarrow{\ell} \Phi(c)$$

Although the number of random variables would never reach infinity, the central limit theorem often holds when sample size is moderately large. How large a sample size is considered to be “reasonably good” when applying normal approximation will be different case by case. For example, for the Wilcoxon signed rank test, a sample size larger than 16 is considered to be large enough for the normal approximation to apply (Rosner 2000). For the Wilcoxon rank-sum test, the normal approximation requires at least 10 for each group (Rosner 2000). What we would like to discuss in this thesis is when to apply the normal approximation applying the T1 test statistic for the incomplete paired data.

## 4.2. Different Data Scenarios

The T lymphocyte data include both complete and incomplete pairs. Under these circumstances, the applicability of the normal approximation will depend on more than just the total number of pairs. We will consider:

(1) Different proportions of the total sample size that comprise complete pairs. In this paper, we discuss 3 different proportions (25%, 50%, and 75%).

(2) Different proportions of missing data in the two groups for the incomplete pairs. We discuss 2 different degrees of balance between groups 1 and 2: Even (or nearly even balance) and highly uneven balance.

The different data scenarios studied in this thesis are presented in Table 3.

**Table 3 Listing of studied data configurations**

| %Total Sample Completed<br>Total Sample Size (J+K+L) | 25% | 50% | 75% |
|---|---|---|---|
| | J-K-L | J-K-L | J-K-L |
| 20 | 5-7-8 | 10-5-5 | 15-2-3 |
| | 5-1-14 | 10-1-9 | 15-1-4 |
| 30 | 8-11-11 | 15-7-8 | 22-4-4 |
| | 8-2-20 | 15-1-14 | 22-1-7 |
| 40 | 10-15-15 | 20-10-10 | 30-5-5 |
| | 10-2-28 | 20-2-18 | 30-1-9 |
| 50 | 12-19-19 | 25-12-13 | 38-6-6 |
| | 12-2-36 | 25-2-23 | 38-1-11 |
| 60 | 15-22-23 | 30-15-15 | 45-7-8 |
| | 15-2-43 | 30-2-28 | 45-1-14 |
| 70 | 18-26-26 | 35-17-18 | 52-9-9 |
| | 18-2-50 | 35-2-33 | 52-1-17 |

Note: “J” indicates the number of complete pairs, “K” the number of incomplete pairs with missing data in Group 2, and “L” the number of incomplete pairs with missing data in Group 1.

## 5. ANALYSIS PROCEDURE AND CRITERION CHOOSING

### 5.1. Procedure of Analysis

Judging whether the normal approximation is appropriate involves comparing exact and approximate sampling distributions to see whether they are close enough for the approximation to be useful in practice. The following steps were carried out in making this assessment.

For each data configuration in Table 3, we repeated the following steps:

(1) Obtain the exact sampling distribution (probability distribution) of test statistic T1, assuming the null hypothesis is true.

(2) Obtain the cumulative exact sampling distribution of test statistic T1.

(3) Obtain the normal approximation distribution of T1 using the central limit theorem.

(4) Obtain the cumulative normal approximation distribution of test statistic T1.

(5) Compare the exact and approximate cumulate distributions by their difference on each possible T1 value.

(6) Identify the values of T1 that would lead to rejection of the null hypothesis at 1-sided $\alpha$ values of 0.01, 0.05, and 0.10. For each of these T1 values, assess whether the absolute difference between the exact and approximate cumulative distributions is smaller than the criterion we choose. The choice of this criterion is presented in Section 5.2 below.

(7) The normal approximation will be considered adequate when this criterion is met for all 3 nominal $\alpha$ levels.

The null exact distribution of T1 for each data configurations in Table 3 is generated by permuting the observations within each stratum of complete pairs and within the stratum of incomplete pairs. For step (1) in the above algorithm, we use the statistical package StatXact to obtain the null exact distribution by permuting the observations within each stratum and evaluating T1 for all possible permutations. We used StatXact's stratified permutation function to permute the ranks within each stratum of complete pairs, and within the stratum of incomplete pairs.

After computing the exact sample distribution from StatXact, we used the statistical package STATA to obtain the cumulative exact sample distribution of T1 and the cumulative normal approximation distribution function. For the normal approximation, the expectation and variance of T1 were calculated as equations (1) and (2).

Table 4 presents the null distributions of the test statistics and the results from steps (2)-(5) based on the (3-2-3) data configuration, observed in the V$\beta$ study.

The results for step (6) are presented in Table 5, which show the values of differences between cumulative exact distribution and cumulative approximation normal distribution under different significance levels ($\alpha$=0.01, 0.05, and 0.1 respectively) of cumulative exact sampling distribution.

**Table 4. Calculation procedure for the (3-2-3) basic data configuration**

| | Step (1) | Step (2) | Step (3) | Step (4) | Step (5) |
|---|---|---|---|---|---|
| Possible value of T1 | Exact Probability under H0 | Exact Cumulative Probability | Normal Z-Variate | Cumulative Normal Approximate Probability | Difference: Exact-Approximate |
| 6 | 0.0125 | 0.0125 | -2.323790 | 0.010068 | 0.002432 |
| 7 | 0.0500 | 0.0625 | -1.807392 | 0.035351 | 0.027149 |
| 8 | 0.1000 | 0.1625 | -1.290994 | 0.098353 | 0.064147 |
| 9 | 0.1500 | 0.3125 | -0.774597 | 0.219289 | 0.093211 |
| 10 | 0.1875 | 0.5000 | -0.258199 | 0.398127 | 0.101873 |
| 11 | 0.1875 | 0.6875 | 0.258199 | 0.601873 | 0.085627 |
| 12 | 0.1500 | 0.8375 | 0.774597 | 0.780711 | 0.056789 |
| 13 | 0.1000 | 0.9375 | 1.290994 | 0.901647 | 0.035853 |
| 14 | 0.0500 | 0.9875 | 1.807392 | 0.964649 | 0.022851 |
| 15 | 0.0125 | 1.0000 | 2.323790 | 0.989932 | 0.010068 |

**Table 5. “Differences” from studied data configurations under the given $\alpha$ level**

| %Total Sample Completed | 25% | | | | 50% | | | | 75% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total Sample Size (J+K+L) | J-K-L | $\alpha$ =0.01 | $\alpha$ =0.05 | $\alpha$ =0.10 | J-K-L | $\alpha$ =0.01 | $\alpha$ =0.05 | $\alpha$ =0.10 | J-K-L | $\alpha$ =0.01 | $\alpha$ =0.05 | $\alpha$ =0.10 |
| 20 | 5-7-8 | **-0.000221** | **0.004956** | 0.011018 | 10-5-5 | **-0.000356** | 0.006638 | 0.017110 | 15-2-3 | 0.002346 | 0.008234 | 0.019971 |
| | 5-1-14 | -0.014556 | **0.001492** | **0.001492** | 10-1-9 | -0.004443 | **0.000115** | 0.0147995 | 15-1-4 | 0.001001 | 0.016090 | 0.016090 |
| 30 | 8-11-11 | **-0.000219** | **0.003105** | **0.007472** | 15-7-8 | **-0.000387** | 0.005263 | 0.011133 | 22-4-4 | 0.001214 | 0.007754 | 0.021017 |
| | 8-2-20 | -0.006133 | **0.004965** | 0.017922 | 15-1-14 | -0.007886 | **-0.000603** | 0.012447 | 22-1-7 | **-0.000095** | 0.007212 | 0.016916 |
| 40 | 10-15-15 | **-0.000262** | **0.002287** | **0.004815** | 20-10-10 | **-0.000316** | **0.003747** | **0.008381** | 30-5-5 | **0.000328** | 0.005742 | 0.013323 |
| | 10-2-28 | -0.006699 | 0.005982 | 0.017720 | 20-2-18 | -0.004672 | **0.004353** | 0.017327 | 30-1-9 | **-0.000104** | 0.006947 | 0.014844 |
| 50 | 12-19-19 | **-0.000226** | **0.001607** | **0.003593** | 25-12-13 | **-0.000257** | **0.002808** | **0.006394** | 38-6-6 | **0.000069** | 0.007151 | 0.012817 |
| | 12-2-36 | 0.003849 | 0.024159 | 0.045204 | 25-2-23 | -0.004718 | **0.003877** | 0.017237 | 38-1-11 | -0.001156 | 0.006233 | 0.012940 |
| 60 | 15-22-23 | **-0.000219** | **0.001288** | **0.002968** | 30-15-15 | **-0.000237** | **0.002320** | **0.004822** | 45-7-8 | **-0.000155** | **0.004565** | **0.009483** |
| | 15-2-43 | -0.006802 | **0.004590** | 0.015564 | 30-2-28 | -0.005329 | **0.003548** | 0.014526 | 45-1-14 | -0.002159 | 0.006298 | 0.012949 |
| 70 | 18-26-26 | **-0.000203** | **0.001051** | **0.002468** | 35-17-18 | **-0.000256** | **0.001727** | **0.003972** | 52-9-9 | **-0.000200** | **0.004090** | **0.008473** |
| | 18-2-50 | -0.006576 | **0.004860** | 0.014242 | 35-2-33 | -0.005777 | 0.005590 | 0.014831 | 52-1-17 | -0.003546 | **0.001618** | 0.013192 |

Note: “Differences” are the values of differences between cumulative exact distribution and cumulative approximation normal distribution under different significance levels (α=0.01, 0.05, and 0.1 respectively) of the cumulative exact sampling distribution. “J” indicates the number of complete pairs, “K” the number of incomplete pairs with missing data in Group 2, and “L” the number of incomplete pairs with missing data in Group 1. The numbers bolded and underlined identify the data configurations for which our criterion for applying the normal approximation are met.

## 5.2. Choice of Criterion for Deciding Whether Normal Approximation is Appropriate

Although statistical references provide guidelines for applying the normal approximation for standard nonparametric tests, they do not supply the criterion for choosing those particular “rules of thumb”. According the rules from statistical references, we will try to find out a criterion that can apply to our studies, and furthermore to find out the general rules of when to apply the normal approximation to the paired two sample problem with missing data.

In Rosner (2000), for the Wilcoxon signed rank test a sample size of at least 16 is considered acceptable for the normal approximation to apply; for the Wilcoxon rank-sum test (Rosner 2000), at least 10 in each group would be sufficient to use the normal approximation.

Based on the above reference, we carried out steps (1)-(7) in section 5.1 using the Wilcoxon signed rank test with N=16, and the Wilcoxon rank-sum test with ($N_1$=10, $N_2$=10). The results from step (6) are listing in Table 6.

From Table 6, we can see that the “Differences” for these two tests, Wilcoxon rank-sum and signed-rank tests, are below α/10 under the chosen α levels. We therefore chose this criterion for our studies: If the difference from step (7) in section 5.1 under certain α level is lower than α/10 for $\alpha$ = 0.01, 0.05, and 0.10, then the normal approximation will be considered suitable for that data configuration.

We realize that this criterion is arbitrary and that the normal approximation may perform acceptably well when this criterion is not strictly met. However, we will adhere to the more stringent criterion in this thesis.

**Table 6. “Differences” for the Wilcoxon Rank-Sum and Sign-Rank tests**

| | | $\alpha$=0.01 | $\alpha$=0.05 | $\alpha$=0.1 |
|---|---|---|---|---|
| $N_1$=10, $N_2$=10 | Wilcoxon Rank-Sum | -0.000283 | 0.003555 | 0.008349 |
| N=16 | Wilcoxon Sign-Rank | -0.000861 | 0.002755 | 0.007000 |

Note: “Differences” are the values of differences between cumulative exact distribution and cumulative approximation normal distribution under different significance levels (α=0.01, 0.05, and 0.1 respectively) of cumulative exact sampling distribution.

## 6. RESULTS AND DISCUSSION

From Table 5, we can see the following patterns for the different data configurations (J-K-L):

(1) For each highly uneven proportion of missing data between the two groups, our criterion is not met for any sample size or proportion of complete data examined.

(2) Under the listed data configurations in Table 5, for each even proportion of missing data between two groups of incomplete pairs, if K and L are large enough, that is, $K \geq 9$ and $L \geq 9$, the normal approximation appears to be appropriate, regardless of the different sample sizes of complete pairs.

(3) The total sample size (J + K + L) and proportion of complete data do not appear to influence the appropriateness of using the normal approximation beyond the considerations in points (1) and (2).

Besides the results listed above, there are still two main issues to address. First, from Table 6 we can see that the data configuration 45-7-8 meets our criterion but that the (5-7-8) and (15-7-8) data configurations do not. It is possible that the larger sample size of complete pairs increases the likelihood that the data would meet our criterion. To address this issue, we examined other data configurations, with 7 or 8 in each incomplete pair group. The data configurations and calculation results are presented in Table 7. We did not see a clear pattern emerging as the number of complete pair increased. This suggests that the number of complete pairs is not a significant factor in inducing the applicability of the normal approximation.

Second, according to point (2) above, we would like to confirm that "K$\geq$9 and L$\geq$9" is the point to properly apply the normal approximation for the even-weighted missing data between two groups of incomplete pairs under the $\alpha/10$ criterion. Thus we analyzed the data configuration J-10-10 (J =5, 6... 15), and J-9-9 (J=5, 6… 15), and the results are presented in Table 8. We found that all data configurations in Table 8 meet our criterion. We therefore believe that it is reasonable to use the normal approximation if "K$\geq$9 and L$\geq$9" for data configurations where the groups are balanced in the number of incomplete pairs. Since only the distributions of incomplete pairs in the two groups seem to matter, this result also implies that the rank-sum test for the incomplete pairs plays an important role in deciding when to apply the normal approximation for the paired two sample data problem with missing data.

Looking at the expected value and the variance of T1 [equations (1) and (2)] may explain the importance of the number of incomplete pairs in determining whether the normal approximation performs well. When the number of complete pairs (J) increases, the expected value and variance of T1 increase by the same magnitude. However, when the number of incomplete pairs (K or L) increases, the expected value and variance of T1 increase according to the squared increase. Therefore, the number of incomplete pairs (K and L) has a larger impact on the expected value and variance and on the applicability of the normal approximation than does the number of complete pairs.

**Table 7. “Differences” from J-8-8, J-7-8, and J-7-7 data configurations**

| $\alpha$ level<br>(J-K-L) | $\alpha$ =0.01 | $\alpha$ =0.05 | $\alpha$ =0.1 |
|---|---|---|---|
| 15-8-8 | -0.000369 | 0.004104 | 0.010563 |
| 20-8-8 | -0.000140 | 0.004888 | 0.009701 |
| 25-8-8 | -0.000220 | 0.004330 | 0.010630 |
| 30-8-8 | -0.000283 | 0.005098 | 0.009791 |
| 35-8-8 | -0.000071 | 0.004542 | 0.010684 |
| 40-8-8 | -0.000147 | 0.004032 | 0.009869 |
| 45-8-8 | -0.000209 | 0.004740 | 0.009091 |
| 50-8-8 | -0.000012 | 0.004231 | 0.009937 |
| 55-8-8 | -0.000086 | 0.004925 | 0.009179 |
| 60-8-8 | 0.000122 | 0.004417 | 0.009995 |
| 65-8-8 | 0.000037 | 0.003950 | 0.009256 |
| 70-8-8 | -0.000034 | 0.004592 | 0.010046 |
| 15-7-8 | -0.000387 | 0.005263 | 0.011133 |
| 20-7-8 | -0.000103 | 0.004612 | 0.010130 |
| 25-7-8 | -0.000191 | 0.005545 | 0.011231 |
| 30-7-8 | -0.000260 | 0.004895 | 0.010260 |
| 35-7-8 | 0.000002 | 0.004305 | 0.011313 |
| 40-7-8 | -0.000086 | 0.005155 | 0.010372 |
| 45-7-8 | -0.000155 | 0.004565 | 0.009483 |
| 50-7-8 | 0.000087 | 0.005394 | 0.010469 |
| 55-7-8 | 0.000001 | 0.004805 | 0.009605 |
| 60-7-8 | -0.000068 | 0.004267 | 0.010552 |
| 65-7-8 | 0.000156 | 0.005028 | 0.009711 |
| 70-7-8 | 0.000072 | 0.004490 | 0.010622 |
| 15-7-7 | -0.000101 | 0.006039 | 0.013163 |
| 20-7-7 | -0.000195 | 0.005267 | 0.011937 |
| 25-7-7 | -0.000265 | 0.004574 | 0.010792 |
| 30-7-7 | 0.000065 | 0.005624 | 0.012062 |
| 35-7-7 | -0.000034 | 0.004927 | 0.010959 |
| 40-7-7 | -0.000111 | 0.005947 | 0.012164 |
| 45-7-7 | 0.000194 | 0.005250 | 0.011100 |
| 50-7-7 | 0.000093 | 0.004620 | 0.010100 |
| 55-7-7 | 0.000012 | 0.005545 | 0.011218 |
| 60-7-7 | 0.000295 | 0.004914 | 0.010249 |
| 65-7-7 | 0.000193 | 0.004342 | 0.011317 |
| 70-7-7 | 0.000110 | 0.005184 | 0.010378 |

Note: J=15, 20, 25 …70

**Table 8. “Differences” for data configuration J-10-10 and J-9-9 under the given α level**

| $\alpha$ level<br>(J-K-L) | $\alpha$ =0.01 | $\alpha$ =0.05 | $\alpha$ =0.1 |
|---|---:|---:|---:|
| 5-10-10 | -0.000334 | 0.004018 | 0.007829 |
| 6-10-10 | -0.000236 | 0.003619 | 0.008354 |
| 7-10-10 | -0.000320 | 0.004036 | 0.007836 |
| 8-10-10 | -0.000222 | 0.003638 | 0.008359 |
| 9-10-10 | -0.000307 | 0.004054 | 0.007842 |
| 10-10-10 | -0.000208 | 0.003657 | 0.008363 |
| 11-10-10 | -0.000293 | 0.004071 | 0.007848 |
| 12-10-10 | -0.000194 | 0.003675 | 0.008367 |
| 13-10-10 | -0.000280 | 0.003298 | 0.007854 |
| 14-10-10 | -0.000179 | 0.003693 | 0.008371 |
| 15-10-10 | -0.000266 | 0.003317 | 0.007860 |
| 5-9-9 | -0.000332 | 0.003951 | 0.008899 |
| 6-9-9 | -0.000194 | 0.004498 | 0.009598 |
| 7-9-9 | -0.000310 | 0.003980 | 0.008909 |
| 8-9-9 | -0.000171 | 0.004526 | 0.009605 |
| 9-9-9 | -0.000288 | 0.004010 | 0.008918 |
| 10-9-9 | -0.000386 | 0.004554 | 0.009611 |
| 11-9-9 | -0.000267 | 0.004039 | 0.008928 |
| 12-9-9 | -0.000366 | 0.004582 | 0.009617 |
| 13-9-9 | -0.000246 | 0.004068 | 0.008938 |
| 14-9-9 | -0.000346 | 0.004609 | 0.009622 |
| 15-9-9 | -0.000224 | 0.004096 | 0.008947 |

Note: J= 5, 6, 7…15. “Differences” are the values of differences between cumulative exact distribution and cumulative approximation normal distribution under different significance levels (α=0.01, 0.05, and 0.1 respectively) of cumulative exact sampling distribution. “J” indicates the number of complete pairs, “K” the number of incomplete pairs with missing data in Group 2, and “L” the number of incomplete pairs with missing data in Group 1.

## 7. CONCLUSIONS

The purpose of this thesis is to evaluate the normal approximation for the paired two sample problem with missing data. This thesis is based on Wilson's work (Wilson 1992) and the data we studied are the extension of experimental data from Weidmann et al (1992). The focus of this thesis is to detect when it is appropriate to apply the normal approximation to the T1 test statistic for incomplete paired data. After comparing the exact and approximate cumulative distributions according to the difference for each possible T1 value for each data configuration we studied, we have drawn the following conclusions under the criterion we chose:

(1) For the highly uneven proportion of missing data between the two groups, it seems inappropriate to apply the normal approximation under our evaluation.

(2) It is reasonable to use the normal approximation if "$K \geq 9$ and $L \geq 9$" for data configurations where the groups are balanced in the number of incomplete pairs.

(3) The total sample size ($J + K + L$) and proportion of complete data do not appear to influence the appropriateness of using the normal approximation beyond the considerations in conclusions (1) and (2).

(4) The number of complete pairs is not a significant factor affecting the applicability of the normal approximation.

# BIBLIOGRAPHY

Conover W. J. (1999) Practical Nonparametric Statistics ($3^{rd}$ ed.) Wiley, New York.

Hajek J., Sidak Z., Sen P.K. (1999) Theory of Rank Tests. Academic Press, San Diego, California.

Im KyungAh (2002) A Modified Signed Rank Test to Account for Missing Data in Small Samples with Paired Data. Master Thesis, University of Pittsburgh.

Lehmann E.L (1975) Nonparametrics: Statistical methods based on ranks. Holden-Day, Inc., San Francisco, California.

Rosner B. (2000) Fundamentals of Biostatistics ($5^{th}$ ed.) Duxbruy, Pacific Grove, California.

Weidmann E., Whiteside TL., Giorda R., Herberman RB., Trucco M. (1992) The T-cell receptor V beta gene usage in tumor-infiltrating lymphocytes and blood of patients with hypatocellular carcinoma. Cancer Research. 52(21): p5913-5920

Wilson J. W. Nonparametric tests for the paired two-sample problem with missing data. [unpublished manuscript, 1992]