12-2009

Kendall's Tau and Spearman's Rho for Zero-Inflated Data

Ronald Silva Pimentel
Western Michigan University

KENDALL'S TAU AND SPEARMAN'S RHO FOR ZERO-INFLATED DATA

by

Ronald Silva Pimentel

A Dissertation
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Doctor of Philosophy
Department of Statistics
Advisor: Magdalena Niewiadomska-Bugaj, Ph.D.

Western Michigan University
Kalamazoo, Michigan
December 2009

KENDALL'S TAU AND SPEARMAN'S RHO FOR ZERO-INFLATED DATA

Ronald Silva Pimentel, Ph.D.

Western Michigan University, 2009

Zero-inflated continuous distributions have positive probability mass at zero in addition to a continuous distribution. Such type of data can be encountered, for example, in medical, environmental and financial research. The main focus of this research is to study the association of nonnegative random variables, both having a positive probability mass at zero. New estimators of the classical measures of association, Kendall's tau and Spearman's rho, appropriate for the zero-inflated distributions, are proposed and their asymptotic distributions are derived. Performance of the estimators is assessed by a Monte Carlo simulation study. New ideas are illustrated by a real data example.

ACKNOWLEDGMENTS

First and foremost, I offer my sincerest gratitude to my advisor, Dr. Magdalena Niewiadomska-Bugaj, for the support and guidance throughout the research and dissertation writing. Her willingness to help and encouragement motivated me to finish this dissertation.

Words fail me to express my appreciation to Dr. Joshua Naranjo. His constant oasis of ideas and passion in statistics and many other things, inspire and enrich my growth as a person, a student and a statistician.

My deepest gratitude also to Dr. Jung Chao Wang for agreeing to be a member of my committee, for his tremendous help on my R programs, and his enthusiam about Latex and Beamer, both of which I learned to appreciate.

I gratefully thank Dr. Marlon Rebellato, in the midst of his relocation and new job, he accepted to be a member of the committee.

I convey special acknowledgment to the Department of Statistics for all the years that I received financial aid and for the intership opportunity that started my career; to the staff, for the support and assistance; and to Western Michigan University that has become a home away from home.

I would like to show my gratitude to my co-workers for their understanding and continuous support; to my boss for his encourangements; and to MPI Research for the financial aid through their tuition reimbursement program.

To my invaluable network of supportive and loving friends without whom I could not have survived this process. Thank you all for the fun times we had together and for your company that always kept me sane.

Acknowledgments-Continued

I owe sincere and earnest thankfulness to Chok and Annie, my confidants and voices of reason. Thank you for your patience, understanding and for the friendship.

Finally, I thank *nanay* and *tatay*, my siblings, JV, Jeff and Aika, and my relatives, for their unending love, prayers and support, and for the motivation and encouragement.

Ronald Silva Pimentel

TABLE OF CONTENTS

CHAPTER

Table of Contents – Continued

CHAPTER

Table of Contents – Continued

CHAPTER

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation and Background

Statistical concerns related to analysis of zero-inflated data have been identified as early as in 1955 especially in relation to the estimation of the location parameter (Aitchison 1955). The term "inflation" was used to emphasize that the probability mass at zero exceeds the value coming from a parametric family of distributions. Such data occurrence is common in medical research and also in the fields of finance, insurance, manufacturing, economics and engineering, to name a few. Statistical methodology for such type of data is still being investigated by statisticians in response to the need in these areas.

Some examples of zero-inflated data are as follows:

Example 1. Household expenditure in Aitchison (1955). If a certain commodity is targeted, some households might not be purchasing that commodity. For example, if one is interested in studying the household expenditure on children's clothing, a zero value will be reported for households without any children.

Example 2. Marine surveys in Pennington (1983). Particular species of fish and plankton usually occupies only a part of the total area. In the survey of marine species, zero inflation is brought about by areas unoccupied or maybe unsuitable for some species.

Example 3. Exposure measurements in Taylor, *et. al.* (2001). Depending on work schedules, some workers may be required to spend certain time during the data collection process in control rooms free of contamination. This will give zero exposure measurements for these workers.

Example 4. Antibody response to the measles vaccine in Moulton and Halsey (1995). There are several known factors for the results of these assays to be zero-inflated. One might be due to the passively acquired maternal antibody by the infants that is interfering to respond to the measles vaccine. A Q-Q plot of the partial data is presented in Figure 1.1.

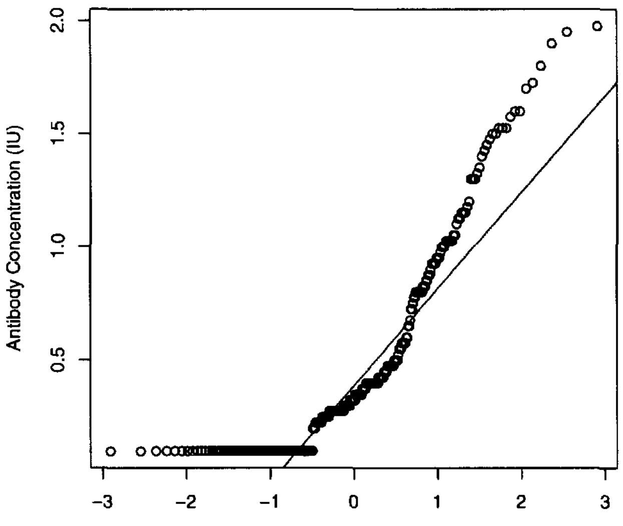

Figure 1.1: Q-Q plot of measles antibody concentration versus the expected distribution.

As indicated in the examples above, the non-ignorable zeroes can be attributed to real zeroes, non-response or non-detects, i.e., falling below some limit of detection. The presence of these zero observations has brought some problems for researchers, statisticians or data analysts. Due to inapplicability of some of the existing statistical methods, common, although not always appropriate, practice in the analysis of zero-inflated data is exclusion or analysis of just the nonzero pairs of observations in a bivariate case or using average ranks in the nonparametric procedures.

Association of two or more variables is a very important research topic. The Pearson's correlation coefficient, while the most commonly used, detects only linear association between two variables, it also needs the normality assumption for each of the random variables. Since real data often violate normality and relationship other than linear is often of interest, Kendall's tau and Spearman's rho are indices that can be used. They are both estimated as rank correlations, so the relations are between the rankings, rather than the actual values of the observations. There have been several adjustments to these rank correlations in the literature that try to take into consideration tied observations but none of them were designed for zero-inflated data. Calculating estimates for these measures of rank correlation using just the nonzero pairs of observations in a zero-inflated data usually leads to inaccurate results.

1.2 Statement of the Problem

This research will focus on studying the well known measures of association, the Kendall's tau and Spearman's rho. Multiple zeroes in the zero-inflated data can be seen as a special case of tied observations. The treatment of these measures with the presence of ties will be studied and compared with a proposed new approach in estimating these measures.

1.3 Organization of the Dissertation

Background information introduced in the remainder of this chapter includes the delta distribution and the classical indices of association not only in the continuous case but also in discrete and categorical cases. A graphical tool will also be presented. Chapter 2 will give a review of the current literature. Chapters 3 and 4 will give the proposed estimators for Kendall's tau and Spearman's rho, respectively. The asymtotic distribution of the proposed estimators will also be defined. Chapter 5 will present the simulation plan and the results. This dissertation will end with the final comments in Chapter 6 which will also outline the future research plan.

1.4 Basic Definitions

We will define the basic distribution, coined by Aitchison as the delta distribution, which incorporates the probability mass at zero while the distribution of the positive values is lognormal. We will also look at the different indices of association for later comparison. A graphical tool called a chi-plot will also be presented which will be used for data evaluation alongside the scatter plot.

1.4.1 Delta Distribution

For the univariate case, assume that a random variable X has continuous distribution for its positive values with density $h_X(x)$ and a positive mass at 0, $P(X = 0) = p > 0$.

Then the distribution function can be written as

$$f(x) = p^{d_X}[(1-p)h_X(x)]^{1-d_X}, \tag{1.1}$$

where $d_X = 0$ if $x > 0$ and $d_X = 1$ if $x = 0$. Consequently,

$$F_X(s) = \begin{cases} 0 & \text{if } s < 0 \\ p & \text{if } s = 0 \\ p + (1-p)\int_0^s h_X(x)dx & \text{if } s > 0. \end{cases}$$

If $h_X(x)$ is a density of a lognormal distribution, X has so-called delta distribution (Aitchison 1955). The mean and variance for this distribution are

$$E(X) = (1-p)\alpha \tag{1.2}$$

and

$$\mathrm{Var}(X) = (1-p)\beta + p(1-p)\alpha^2, \tag{1.3}$$

where α and β are the mean and variance, respectively, of the $h_X(x)$ distribution.

1.4.2 Measures of Association

There are several measures available to study the association of discrete or continuous data. The most common measure for a continuous pair of random variables is the Pearson's correlation coefficient, ρ. Other measures, such as Kendall's tau, τ, and Spearman's rho, ρ_S are also used and will be the focus of this study.

Pearson's Correlation Coefficient, ρ

Pearson's correlation coefficient is a measure of linear relationship between two random variables.

Suppose X and Y are two jointly distributed random variables, the Pearson's correlation coefficient between X and Y is given by

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}, \tag{1.4}$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y and $V(X)$ and $V(Y)$ denote the variances of X and Y, respectively.

From a sample of n paired observations, ρ is estimated by

$$r = \frac{\sum X_i Y_i - (n\overline{X} \times \overline{Y})}{\sqrt{(\sum X_i^2 - n\overline{X}^2)(\sum Y_i^2 - n\overline{Y}^2)}}, \tag{1.5}$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of X_i's and Y_i's, respectively.

Some drawbacks of this measure are: (1) it is not invariant under strictly increasing nonlinear transformations and it is highly affected by extreme outliers, and (2) it is sensitive to the departure from normality. r tends to have large bias and large variance when calculated from a bivariate nonnormal distribution with skewed marginals, $\rho \neq 0$ especially for smaller sample sizes.

Kendall's Tau, τ

Kendall's tau was proposed by Maurice Kendall (1938) as a measure of association of two jointly distributed continuous random variables. It is defined as a difference between the probability of concordance and discordance of two random variables. A pair of observations is said to be concordant if a larger value of X is more likely associated with a larger value of Y. The pair is discordant if a larger value of X is more likely associated with a smaller value of Y. The population Kendall's τ is defined as

$$\tau = \underbrace{P[(X_1 - X_2)(Y_1 - Y_2) > 0]}_{P(concordance)} - \underbrace{P[(X_1 - X_2)(Y_1 - Y_2) < 0]}_{P(discordance)}, \qquad (1.6)$$

where (X_2, Y_2) is an independent replicate of (X_1, Y_1). As a difference of two probabilities, $-1 \leq \tau \leq 1$ with a positive τ indicating positive association between the variables and higher absolute value indicates stronger association.

For (X, Y) following a bivariate normal distribution with correlation coefficient ρ, Kruskall (1958) presented the relationship between Pearson's correlation coefficient and Kendall's tau.

$$\tau = 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)dH(x,y) - 1 = \frac{2}{\pi}\arcsin(\rho). \qquad (1.7)$$

Graphical illustration shown in Figure 1.2 suggests that τ is a nearly linear function of ρ.

To get the estimate of tau, let $(X_1, Y_1), ..., (X_n, Y_n)$ be a random sample from the joint distribution of (X, Y). The Kendall rank correlation statistic K is

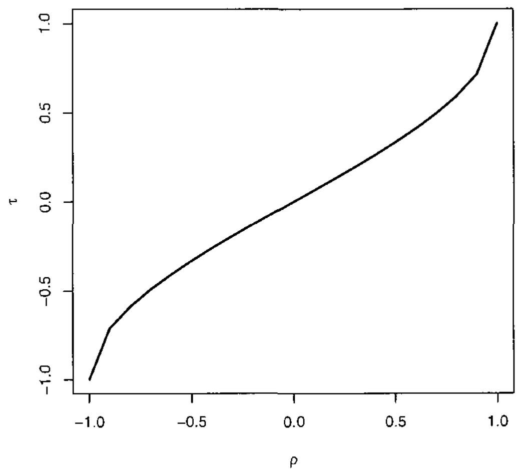

Figure 1.2: Kendall's tau as a function of Pearson's correlation coefficient in the bivariate normal model

calculated as

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((X_i, Y_i), (X_j, Y_j)) \tag{1.8}$$

where

$$Q((a,b),(c,d)) = \begin{cases} 1, & \text{if } (d-b)(c-a) > 0 \\ \text{-1}, & \text{if } (d-b)(c-a) < 0. \end{cases} \tag{1.9}$$

As Kendall proposed, K can be used to obtain a distribution free test of H_0 : X and Y are independent vs. $H_1 : \tau \neq 0$ where τ is defined as in (1.6).

The estimate $\widehat{\tau}$ is based on the statistic K and is defined as

$$\widehat{\tau} \;=\; \frac{K}{\binom{n}{2}} = \frac{2K}{n(n-1)}. \tag{1.10}$$

It can be shown using standard U-statistic theory (see e.g., Randles and Wolfe, 1979) that

$$E(\widehat{\tau}) = \tau \tag{1.11}$$

and

$$\text{Var}(\widehat{\tau}) = \frac{1}{\binom{n}{2}}[2(n-2)\zeta_1 + \zeta_2], \tag{1.12}$$

where $\zeta_1 = \text{Cov}[(Q(X_1,Y_1),(X_2,Y_2)),(Q(X_1,Y_1),(X_3,Y_3))]$, $\zeta_1 > 0$ and $\zeta_2 = \text{Var}[Q(X_1,Y_1),(X_2,Y_2)]$.

If there are ties among the observations $X_1,...,X_n$ and/or separately among the observations $Y_1,...Y_n$, function (1.9) is replaced by

$$Q^*((a,b),(c,d)) = \begin{cases} 1, & \text{if } (d-b)(c-a) > 0 \\ 0, & \text{if } (d-b)(c-a) = 0 \\ \text{-1}, & \text{if } (d-b)(c-a) < 0, \end{cases} \tag{1.13}$$

and K is now defined as

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q^*((X_i,Y_i),(X_j,Y_j)). \tag{1.14}$$

The estimate, $\widehat{\tau}$, of the Kendall population coefficient τ in (1.10) is then redefined as

$$\widehat{\tau} = \frac{2K}{\sqrt{(T_0 - T_x)(T_0 - T_y)}}, \tag{1.15}$$

where $T_0 = n(n-1)$, $T_x = \sum_l s_l(s_l - 1)$ and $T_y = \sum_m t_m(t_m - 1)$. Here, l is the number of tied observations in X and s_l is the size of the l^{th} tied group in X observations and, equivalently, m is the number of tied observations in Y and t_m is the corresponding size of this group. Consequently, the denominator of (1.15) is a geometric average of the number of pairs untied on X and the number of pairs untied on Y. It can easily be seen that (1.15) reduces to (1.10) if there are no tied observations.

Spearman's Rho, ρ_S

Another popular measure of association is the Spearman's rho. Let $(X_1, Y_1), (X_2, Y_2)$ and (X_3, Y_3) be independent random vectors with the same distribution as (X, Y). Then

$$\rho_S = 3P[(X_1 - X_2)(Y_1 - Y_3) > 0] - 3P[(X_1 - X_2)(Y_1 - Y_3) < 0]. \tag{1.16}$$

The coefficient ρ_S is proportional to the difference between the probabilities of concordance and discordance of the random vectors (X_1, Y_1) and (X_2, Y_3), where X_2 and Y_3 are independent variables with the same marginal distributions as X_1 and Y_1, respectively.

For the bivariate normal models with correlation coefficient ρ, Kruskall (1958) similarly has shown that

$$\rho_S = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(x)G(y)dH(x,y) - 3 = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right). \tag{1.17}$$

The rank-based estimator of this correlation parameter was introduced by Spearman in 1904 as

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}, \tag{1.18}$$

where D_i is the difference between the ranks of X_i and Y_i in their separate rankings.

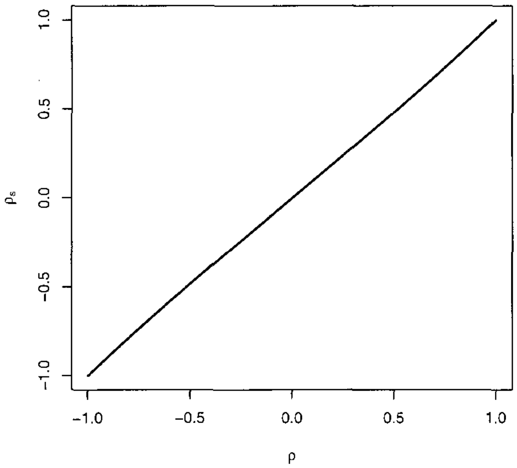

Figure 1.3: Spearman's rho as a function of Pearson's correlation coefficient in the bivariate normal model

With the presence of ties among the n X observations and/or separately among the n Y observations, the estimate in (1.18) can be redefined as

$$r_S = \frac{W_0 - 6\sum_{i=1}^{n} D_i^2 - \frac{1}{2}\{T_x + T_y\}}{\sqrt{(W_0 - T_x)(W_0 - T_y)}}, \tag{1.19}$$

where $W_0 = n(n^2-1)$, $T_x = \sum_l s_l(s_l^2 - 1)$ and $T_y = \sum_m t_m(t_m^2 - 1)$. Similarly to Kendall's τ, l is the number of tied observations in X and s_l is the size of the l^{th} tied group in X observations and, equivalently, m is the number of tied observations in Y and t_m is the corresponding size of this group.

Discrete Case

In a discrete case, ties can be viewed as a combination of three different scenarios (see, e.g., Liebetrau, 1983). Given (X_1, Y_1) and (X_2, Y_2), they can be tied only on X, i.e., $(X_1 = X_2, Y_1 \neq Y_2)$ with probability π_t^X, or tied only on Y, i.e., $(X_1 \neq X_2, Y_1 = Y_2)$ with probability π_t^Y, or tied on both X and Y, i.e., $(X_1 = X_2, Y_1 = Y_2)$ with probability π_t^{XY}. The range of τ depends on the probability of ties, therefore, (1.6) will not be suitable for discrete data. In this case, multinomial sampling is more appropriate. If p_{ij} is defined as $P(X = x_i, Y = y_j)$, then the Kendall's tau, denoted by τ_b for discete case, can be defined as

$$\tau_b = \frac{\pi_C - \pi_D}{\left[\left(1 - \sum_{i=1}^{I} p_{i+}^2\right)\left(1 - \sum_{j=1}^{J} p_{+j}^2\right)\right]^{1/2}}, \tag{1.20}$$

under the multinomial sampling model and π_C is the probability of two randomly selected members of the population that are concordant and π_D is the probability that they are discordant. Also, $1 - \sum_{i=1}^{I} p_{i+}^2 = 1 - \pi_t^X - \pi_t^{XY}$ is the probability that the observations are not tied in Y and equivalently, $1 - \sum_{j=1}^{J} p_{+j}^2 = 1 - \pi_t^Y - \pi_t^{XY}$ is the probability that the observations are not tied in X.

Given that X and Y discrete variables are jointly sampled then (1.20) can be estimated by the formula

$$\widehat{\tau_b} = \frac{2 \times (C - D)}{\left[\left(n^2 - \sum_i n_{i+}^2\right)\left(n^2 - \sum_j n_{+j}^2\right)\right]^{1/2}}, \tag{1.21}$$

where n_{ij}s are the observed frequency. Also, C is the number of concordant pairs and D is the number of discordant pairs.

Similarly, the Spearman's rho can be estimated by the formula

$$\widehat{\rho_S} = \frac{\sum_i \sum_j n_{ij} R(i) C(j)}{\frac{1}{12}\left[\left(n^3 - \sum_i n_{i+}^3\right)\left(n^3 - \sum_j n_{+j}^3\right)\right]^{1/2}}, \tag{1.22}$$

where

$$R(i) = \sum_{k<i} n_{k+} + \frac{n_{i+}}{2} - \frac{N}{2}, \tag{1.23}$$

and

$$C(j) = \sum_{l<j} n_{+l} + \frac{n_{+j}}{2} - \frac{N}{2}. \tag{1.24}$$

Measures of Association for Categorical Variables

For two categorical variables, a contingency table is a tool in understanding their joint distribution. An example is shown in Table 1.1. Here, the two variables are tabulated, one as a row variable and the other as a column variable. The categories (e.g., present or absent) are shown for each variable and the frequency counts for each possible combination of categories are presented. The marginal totals are also shown for each of the variables.

	First Variable		
Second Variable	Present	Absent	Totals
Present	n_{11}	n_{12}	n_{1+}
Absent	n_{21}	n_{22}	n_{2+}
Totals	n_{+1}	n_{+2}	n

Table 1.1: Example of 2×2 contingency table for two categorical variables

Given a contingency table, several measures of association have been proposed like the ϕ coefficient for 2×2 tables; Pearson's C (contingency coefficient) for symmetric contingency tables larger than a 2×2; and Cramer's V for asymmetrical tables.

Given a contingency table for variables measured in ordinal categories such as low/medium/high, with a large number of tied ranks, the gamma coefficient, G is used as the appropriate measure of association, defined as

$$g = \widehat{\gamma} = \frac{C - D}{C + D}. \tag{1.25}$$

The population version of gamma is

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}. \tag{1.26}$$

1.4.3 Chi-plot

In addition to scatter plot of raw data and ranks, association between random variables will be graphically illustrated using chi-plots. These were originally proposed by Fisher and Switzer (1985), and later expanded in Fisher and Switzer (2001), where they showed how a single chi-plot can highlight different forms of dependence.

To generate this plot, given a random sample of n pairs of random samples from a bivariate distribution, one should determine the following quantities.

$$\begin{aligned} H_i &= \frac{1}{n-1}\#(j \neq i : X_j \leq X_i, Y_j \leq Y_i), \qquad (1.27) \\ F_i &= \frac{1}{n-1}\#(j \neq i : X_j \leq X_i), \\ G_i &= \frac{1}{n-1}\#(j \neq i : Y_j \leq Y_i), \text{ and} \\ S_i &= \text{Sign}\left\{(F_i - 0.5)(G_i - 0.5)\right\}. \end{aligned}$$

It can be seen that these quantities depend entirely on the ranks of the distributions. Fisher and Switzer proposed that the chi-plot be a scatter plot of the pairs (λ_i, χ_i), where λ_i is the distance between the observation (x_i, y_i) and the center of the dataset and χ_i is a function of the signed square root of the traditional chi-square test statistic for independence in a two-way table. These are defined as

$$\chi_i = \frac{H_i - F_i G_i}{\{F_i(1-F_i)G_i(1-G_i)\}^{\frac{1}{2}}}, \tag{1.28}$$

$$\lambda_i = 4S_i \max\left\{(F_i - 0.5)^{\frac{1}{2}}, (G_i - 0.5)^{\frac{1}{2}}\right\}, \tag{1.29}$$

where $\chi_i \in$ [-1,1].

In order to help with the interpretation of the chi-plot, Fisher and Switzer recommended that a pair of horizontal lines be displayed showing $\pm c_p/\sqrt{n}$, where c_p is selected such that approximately $(100 \times p)\%$ of pairs (λ_i, χ_i) lie between these lines. They reported c_p values 1.54, 1.78, and 2.18 that correspond to $p = 0.90$, 0.95 and 0.99, respectively, obtained through simulations.

Figures 1.4 and 1.5 are shown to illustrate the expected behavior of the chi-plots with two independent random variables and with the presence of increasing monotone association. Data were randomly generated from a bivariate standard normal distribution with $n = 100$ and correlation $\rho = 0.0$, 0.20, 0.50, 0.95. The left portion of each figure shows the scatter plot for each case while the corresponding graph on its right is the chi-plot. The horizontal lines represents the 95% control limit, which suggests that 95% of the χ_i values should fall within these lines if there is no association between the variables. The points depart from this band as the association becomes more prominent. In Figure 1.4(b), majority of the points are within the 95% band which indicates the lack of association between the variables as depicted in its corresponding scatter plot in Figure 1.4(a). As the correlation coefficient is increased, the points depart from the band which leads to a picture similar to the one shown in Figure 1.5(h). In this figure, there is evidence of monotone dependence between the two variables.

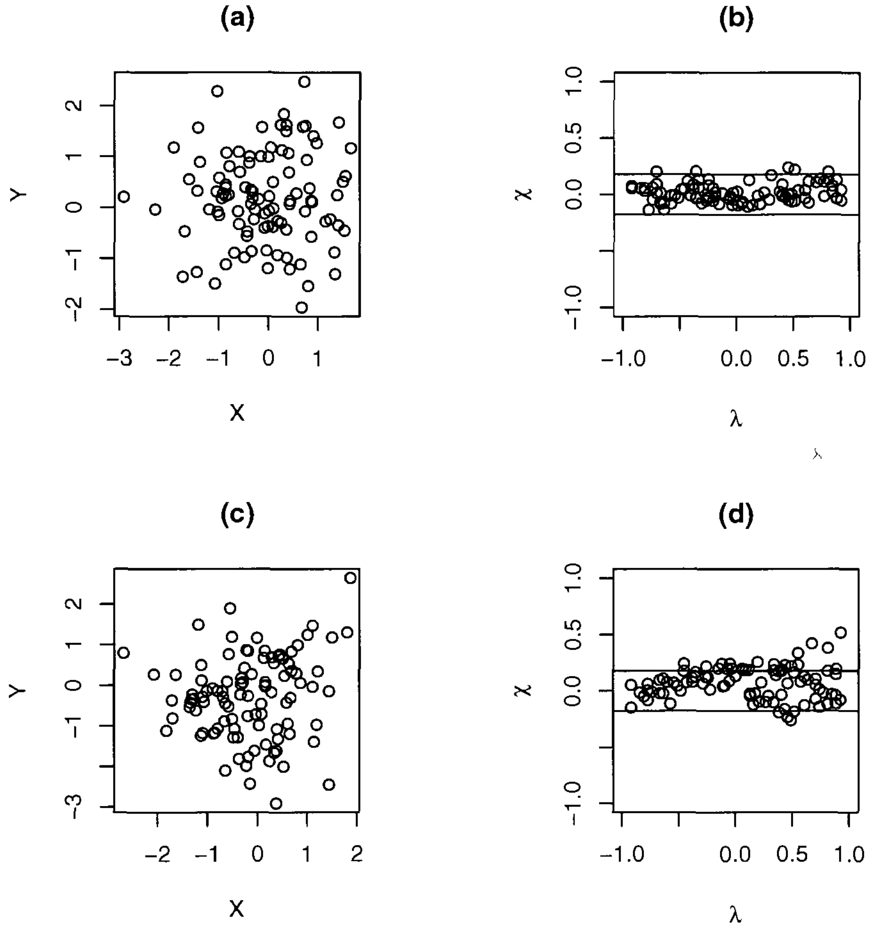

Figure 1.4: Sample chi-plot. Left column shows the scatter plots and the right column their corresponding chi-plots, for simulated samples of size 100 from the bivariate normal distribution with correlation coefficients, 0.0 and 0.20, respectively.

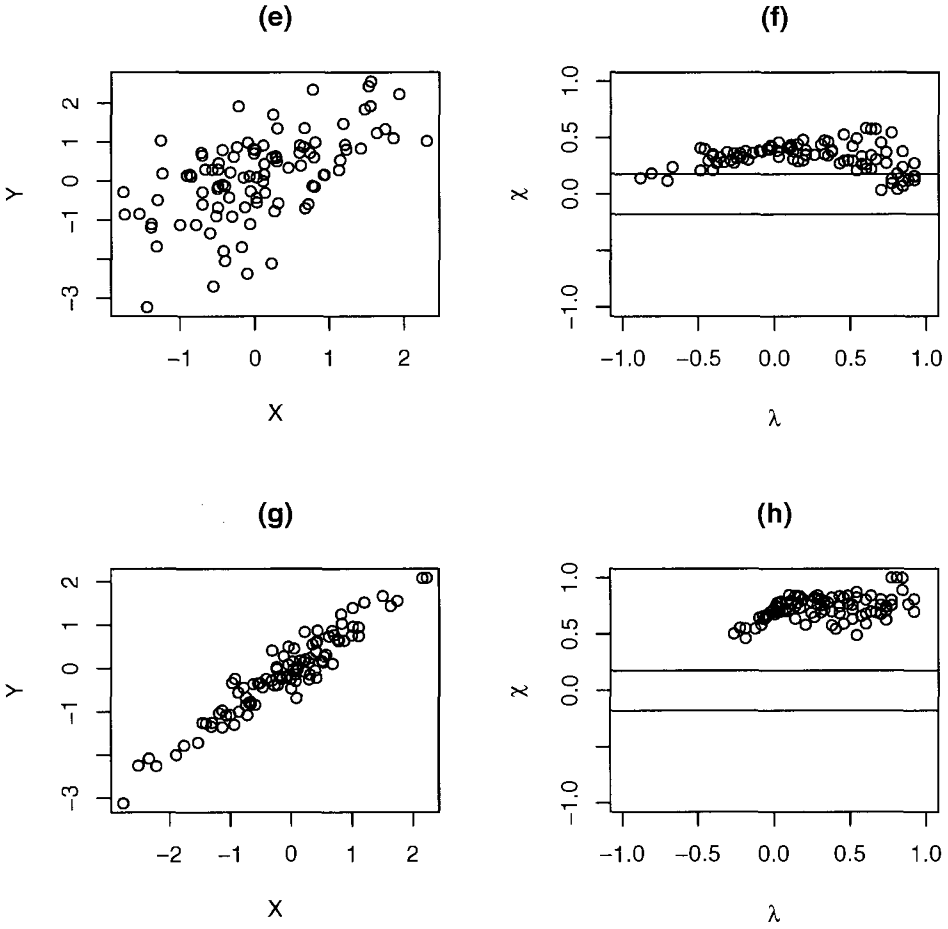

Figure 1.5: Additional sample chi-plot. Left column shows the scatter plots and the right column their corresponding chi-plots, for simulated samples of size 100 from the bivariate normal distribution with correlation coefficients, 0.50 and 0.95, respectively.

Chapter 2

LITERATURE REVIEW

Several studies have been published regarding the location parameters for single, paired or independent samples having a mass at zero, the earliest was Aitchison (1955). Examples have been provided to illustrate the problem at hand, one of which is the analysis of household expenditure on a certain commodity. Some households may not use or buy the product which results to a zero observation. The presence of these cases skews the distribution which can then be approximated by a lognormal curve. Aitchison proposed efficient estimates of the mean and variance. He further applied his results using several distributions and then used real data as examples.

The concepts presented by Aitchison were used by Pennington (1983) in finding efficient estimators of abundance for fish and plankton surveys. He pointed out that inflation at zero can also be observed in marine survey, which is brought about by having areas that are not occupied or unsuitable for some species. He applied Aitchison's estimators on ichthyoplankton survey and concluded the efficiency of the mean estimator based on the delta-distribution due to the large variability of the log of the nonzero values. He was also able to extend Aitchison's work and presented an estimate of the variance for the estimator of the mean.

Owen and DeRouen (1980) also studied the mean estimation with zero-inflated data. In addition to just having zero observations, they also looked into having a left-censoring and a combination of both and used the mean square error approach. They reported that the maximum variance unbiased estimator of the zero inflated mean has lower MSE than the MLE with just the nonzero censored data.

Several other papers were published that dealt with zero-inflated data. One of the main motivations for this research was the study by Moulton and Halsey (1995). They presented a measles vaccine data from an immunogenicity study on sera collected from children 12 months of age. The zero values in this data arise from values falling below a limit of detection. A mixture model approach using lognormal distribution for the nonzero values was used.

An interesting point to further illustrate when zero-inflated data can occur was made by Taylor, *et. al.* (2001). In their paper, they presented the study of exposure measurement falling below a fixed limit of detection. In this type of data, at least 20% of the data are expected to fall below the set limit of detection, which give rise to the zero-inflation problem. However, they pointed out that it is false to assume that all zero values are due to the fact that the observed value is below a limit of detection. Some of those are real zeroes which were observed from personnel assigned to work in a controlled environment for a certain period of time.

Bascoul-Mollevi, *et. al.* (2005) presented several two-part statistics that can be used to analyze paired data from a mixed distribution. These statistics are a sum of a test of proportions (for the count of zero values) and a parametric or non-parametric statistic comparing the means from two paired samples. The test of proportions is based on a χ^2 distribution with 1 degree of freedom (d.f.) and the test for the nonzero value is based on a statistic that also tends to a χ^2 distribution with 1 d.f. The resulting statistics tend to a χ^2 distribution with 2 d.f. These tests were proposed by Lachenbruch (2001) who considered two independent groups. Both papers compared the two-part statistics with the usual tests used in testing difference in proportions and tests in difference in means. Lachenbruch concluded that the two-part statistic performed better if the larger proportion of zero values corresponded to the population with the larger mean. Bascoul-Mollevi, *et. al.* concluded that all tests were efficient for the case when small number of zero values corresponded to the population with the larger mean. On all other cases, the two-part statistic performed better, thus, showing consistency between the independent and matched-pair scenario.

Lachenbruch (2002) revisited and summarized the studies he had presented regarding the analysis of data with excess zeroes. The two-part models that he presented considered the nonzero part having continuous distribution rather than Poison or negative binomial. From his paper in 2001, he only considered the t-test and the Wilcoxon test, and the two-part tests using these tests. He further studied the size and power of all the tests he presented and later concluded that the two-part models are useful alternatives to the usual t-test and Wilcoxon test.

Zhou and Tu (1999) also compared means of independent populations having zero observations. They looked into the analysis of medical cost data having significant zero values from different patient groups. The problem was recognized when in the first intervention group with 142 patients, 108 of them were not hospitalized, and therefore, no charges were incurred from them. From the second group with 113 patients, 85 were not hospitalized. And from the control group with 119 patients, 98 were not hospitalized. Due to the inappropriateness of standard tests like the analysis of variance, they proposed to use the Wald test and the likelihood ratio test in which they found that both have reasonable power to detect true difference in the means. They argued that both tests performed satisfactory based on their simulation results. The power of both tests are equally comparable and the type I error rate of the Wald test is relatively close to that of the likelihood ratio test, especially when the sample sizes are large. Overall, they concluded that due to the ease of implementation and computationally being more efficient, the Wald test was preferred over the likelihood ratio test.

Daoud (2007) extended the two-part tests to comparison of means in k independent populations with zero inflated distributions.

Kendall's tau was proposed to measure the strength of dependence between two continuously distributed variables and was first introduced by Kendall in 1938, applied in solving psychology related problems, while Spearman introduced his measure of rank correlation in 1904. Kendall, *et. al.* (1939) determined the theoretical sampling distribution of Spearman's rank correlation coefficient. Then Kendall (1942) proposed a coefficient of partial correlation. The motivation was that one can naturally inquire that a significant rank correlation between two ranked observations maybe due to the correlation of both qualities with some more fundamental quality. In 1945, Kendall studied the effects of tied ranks on the coefficients of rank correlations. If observations between the i^{th} and k^{th} observations are ties, the midrank method, $(i + k)/2$, is used to calculate the rank of these observations. He presented the adjustments for the calculation of the Kendall's and Spearman's rank correlation coefficients and further discussed them in more detail in his book together with Gibbons (1948). Hollander and Wolfe (1999) also discussed the concepts of the rank correlation coefficients. They noted though as a comment that the modified formula taking into consideration the presence of ties works best if the size of the tied observation in either variables or both do not represent a big percentage of the data. Noether (1967) proposed a consistent estimator of the variance of τ based on the test statistic proposed by Kendall. Flinger and Rust (1983) also proposed a consistent variance estimator that corrects the problem of obtaining negative values by Noether. They further indicated that even with discontinuous distribution, the function $F(x, y)$, $n^{1/2}(\widehat{\tau} - \tau)/\widehat{\sigma}_j$, where $\widehat{\tau}$ is the proposed estimator of τ with corresponding standard deviation $\widehat{\sigma}_j$, still maintains a limiting standard normal distribution. Samara and Randall (1988) further studied the subject and also proposed their consistent estimator for the variance of Kendall's tau. A corresponding modified Kendall's test statistic was also defined. Cliff and Charlin (1991) mentioned that only when there are no tied observations is it possible to attain the limits of Kendall's τ which is [-1, 1]. In their paper, they also generalized the formula for estimating the variance of the sample tau.

An extreme case of tied observations can sometimes lead to having a dichotomy in both rankings of two variables, i.e., the values can just be classified as either present or not present. This gives rise to the 2×2 contingency table (see, e.g., Kendall and Gibbons, 1948). The Kendall's rank correlation coefficient in this case is calculated using the observed frequencies and the marginal distributions. Contingency tables were further extended to 2×c, r×2, and r×c tables and they were used to study the relationship between the categorical variables of interest as presented by Agresti (1990, 1996) and Stokes *et. al.* (1995).

In modeling multivariate distributions, one has to take into account the effects of the marginal distributions as well as of the dependence between them.

In order to achieve this, Sklar (1959) first introduced the concept of a copula. A copula is a function which couples a joint distribution function with its univariate, uniformly distributed margins $[U(0,1)]$. It also aids in understanding the concept of monotone dependence between continuous variables. Literature in different areas of research, especially in the field of finance and banking have been published using this concept. Nelsen (1999) published a comprehensive introduction and background on copulas. In the simplest way, Sklar's theorem can be summarized such that, if H is a bivariate cdf with marginals F and G, there exists a bivariate copula C wherein for all $(x, y) \in \mathbb{R}$, $H(x, y)$ can be written as $C\{F(x), G(y)\}$. If F and G are continuous, then C is unique, otherwise, if they are discrete, there is no unique way to express the joint distribution as a function of the marginal distributions unless on Range$(F)\times$ Range(G). Given the copula C, the Kendall's tau can be defined as $\tau = 4\int_0^1 \int_0^1 C(u, v)dC(u, v) - 1$. And also, the Spearman's rho can be defined as $\rho = 12\int_0^1 \int_0^1 uvdC(u, v) - 3$.

Herath and Kumar (1991) applied the use of copulas in the field of engineering economy, specifically in the area of project risk and regression analysis for forecasting. They attempted to look for an alternative for the Pearson's product moment correlation due to its limitations.

Wang (2007) studied the relationship between semen and plasma viral loads, both zero-inflated variables, using the Clayton copula model and proposed a modified estimate for τ for bivariate truncated data. A goodness of fit test was first introduced to check if the Clayton copula model assumptions were met. The nonzero part of the data was expected to retain the Clayton copula distribution after truncation. The modified τ uses only the nonzero pairs of observations. From the example given, 85 pairs of plasma and semen viral loads were collected but only 19 pairs were used for the modified tau. The scatter plot of the data is shown in Figure 2.1. The graph illustrates that the pairs of observations are grouped into four different sections namely, Section I with probability p_{00}; Section II with probability p_{10}; Section III with probability p_{01}; and Section IV with probability $1-(p_{00}+p_{10}+p_{01})$.

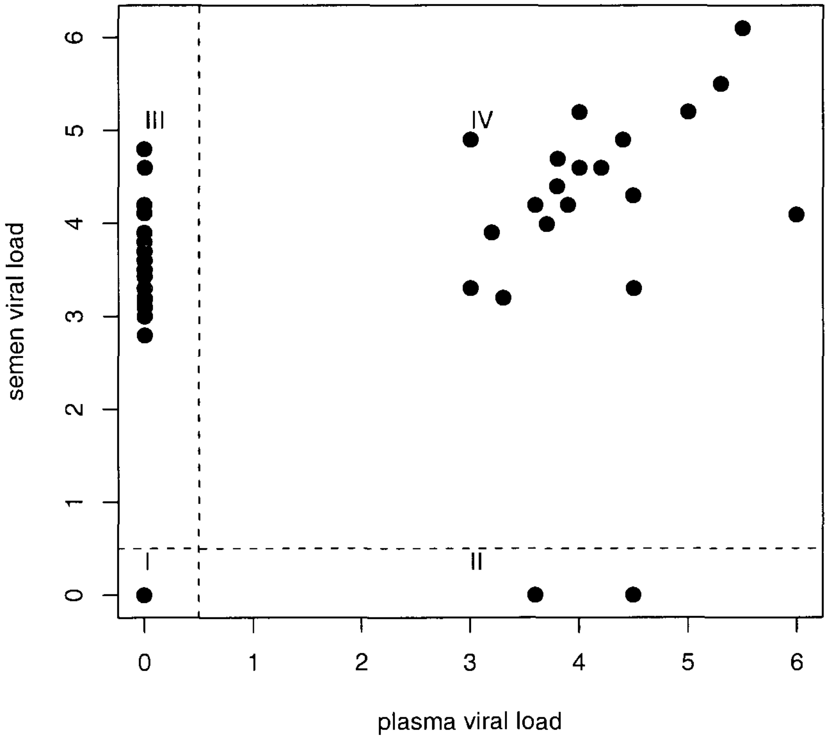

Figure 2.1: Scatter plot of plasma and semen viral loads from Wang (2007).

In terms of tied observations, this sample data shows a big part of data tied at zero (0) with $\widehat{p_{00}}, \widehat{p_{01}}, \widehat{p_{10}} > 0$.

Most of the coefficients of dependence or association have been defined for continuous random variables. If applied to discrete data, some properties of these dependence measures are lost. Nešlehová (2007) generalized the rank correlation measures for non-continuous random variables. Since copulas are defined to be unique in the continuous case, a technique was proposed that will allow the application of the copulas to the non-continuous random variables. An important finding was indicated on the role of the standard extension copula that was first introduced by Schweizer and Sklar (1974). This standard extension copula, compared with the role of the unique copula with the continuous variables, allowed for the generalization of the rank correlation measures.

With discrete data, the limits [-1,1] are also not attainable. Denuit and Lambert (2005) studied the constraints of the dependence measures in bivariate discrete data. They presented a continuous extension of a discrete variable and focused on the Kendall's tau. They indicated that a discrete variable can be associated with a continuous random variable X^* defined as $X^* = X + (U - 1)$, where U is a continuous random variable on (0,1) and is independent of X. They showed that the extension preserves the concordance order, that is, $(X_1, Y_1) \prec_c (X_2, Y_2) \Rightarrow (X_1^*, Y_1^*) \prec_c (X_2^*, Y_2^*)$. In general, $(X_1, Y_1) \prec_c (X_2, Y_2)$ denotes that (X_2, Y_2) is more concordant than (X_1, Y_1) if $P(X_1 < s, Y_1 < t) \leq P(X_2 < s, Y_2 < t)$ for all $s, t \in \mathbb{R}^2$ given that (X_1, Y_1) and (X_2, Y_2) are independent and identically distributed. In preserving the concordance order, the continuous extension also preserves the Kendall's tau, that is, $\tau(X, Y) = \tau(X^*, Y^*)$. They also presented the boundaries of the Kendall's tau using continuous extension of discrete data.

Mesfioui and Tajar (2005) established monotonicity of τ and ρ with respect to concordance ordering described above. They also studied the dependence measures for discrete data and also proposed the use of continuous extension. The continuousation is done such that $X^* = X + U$ where U was chosen to be uniformly distributed on [0,1]. They also established that ρ is larger than τ for positively dependent discrete random variables and derived the maximum limits of the estimate for τ for discrete data.

Chapter 3

PROPOSED ESTIMATOR OF KENDALL'S TAU

Kendall's tau, τ, a widely used and accepted measure of association, is defined as

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0], \qquad (3.1)$$

where (X_1, Y_1), (X_2, Y_2) are independent replicates of jointly distributed variables X and Y.

3.1 Adjustment of Kendall's Tau with Ties

The formula (3.1) was proposed under the assumption that both X and Y have continuous distribution. However, as it was defined in (1.13), if two pairs of observations are tied in either X, or Y, or both, a score of zero will be given to be accounted for in the calculation of K in (1.14). Similarly, the denominator has to be adjusted and the estimator is defined in (1.15).

If there are no paired observations, there is a total of $n(n-1)/2$ pairs, which is also the sum of number of concordant and discordant pairs. In a simple case, if a single value in X is tied s times, there will be $s(s-1)/2$ pairs with those observations only. Then, if there are l of these values, each tied in varying s times, then there is a total of

$$T_x = \sum_l s_l(s_l - 1)/2 \tag{3.2}$$

ties in X.

Similarly for ties in Y, we can define

$$T_y = \sum_m t_m(t_m - 1)/2. \tag{3.3}$$

Therefore, these will lead to the redefined formula in (1.15). This estimator will always be greater than that without the adjustment on the denominator, i.e., still use $n(n-1)/2$ even with ties.

Zero-inflated type data is a special case of tied observations, under the assumption that there is a very small or nearly zero chance for the continuous part to have ties. In this case, there will only be one tied value observed n_{01} in X only, n_{10} in Y only, and n_{00} in both X and Y. Incorporating these in (3.2) and (3.3), will lead to

$$\widehat{\tau} = \frac{K}{\sqrt{\binom{n}{2} - \binom{n_{00}+n_{01}}{2}}\sqrt{\binom{n}{2} - \binom{n_{00}+n_{10}}{2}}}, \tag{3.4}$$

where n_{00} = number of zero pairs of observations; n_{01} = number of cases when $x = 0$ and $y > 0$; and n_{10} = number of cases when $x > 0$ and $y = 0$.

The denominator of (3.4) is a geometric mean of the untied pairs of X observations and untied pairs of Y observations.

3.2 Proposed Estimator of Kendall's Tau, τ^*

Since the estimator (3.4) is not a MLE of Kendall's tau defined in (3.1), and it was pointed out in Hollander and Wolfe (1999) that the adjustment for ties is only satisfactory as long as the number of pairs of observations that are tied in X and/or in Y does not represent a sizable proportion of the total number of pairs, we propose a new estimator τ^*, being an estimator of (3.1) for the case of pairs being tied at 0 on at least one variable.

Proposition 1 *Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed random vectors, each with joint distribution function H. Then the population Kendall's τ given by (3.1) is*

$$\tau^* = p_{11}^2 \tau_{11} + 2(p_{00}p_{11} - p_{01}p_{10}), \tag{3.5}$$

where τ_{11} is the population Kendall's τ defined in (3.1) for the pairs of positive observations, $(X > 0, Y > 0)$, and $p_{00} = P(X = 0, Y = 0); p_{10} = P(X > 0, Y = 0); p_{01} = P(X = 0, Y > 0)$; and $p_{11} = P(X > 0, Y > 0)$.

Proof. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed. In the same manner as (1.6), we will define

$$\tau^* = \underbrace{P(\text{concordance})}_{P(C)} - \underbrace{P(\text{discordance})}_{P(D)} \tag{3.6}$$

Using the total probability formula, we can first derive $P(C)$.

$$\begin{aligned} P(C) &= \sum_{\forall i} P(C|X_{i1}, Y_{i1}, X_{i2}, Y_{i2}) P(X_{i1}, Y_{i1}, X_{i2}, Y_{i2}) \\ &= P(C|X_1 = 0, Y_1 = 0, X_2 = 0, Y_2 = 0) \end{aligned}$$

$$
\begin{aligned}
& \times P(X_1 = 0, Y_1 = 0, X_2 = 0, Y_2 = 0) \\
& + P(C|X_1 = 0, Y_1 > 0, X_2 = 0, Y_2 = 0) \\
& \times P(X_1 = 0, Y_1 = 0, X_2 > 0, Y_2 = 0) \\
& + P(C|X_1 = 0, Y_1 = 0, X_2 = 0, Y_2 > 0) \\
& \times P(X_1 = 0, Y_1 = 0, X_2 = 0, Y_2 > 0) \\
& + P(C|X_1 = 0, Y_1 = 0, X_2 > 0, Y_2 > 0) \\
& \times P(X_1 = 0, Y_1 = 0, X_2 > 0, Y_2 > 0) \\
& + \vdots \\
& + P(C|X_1 > 0, Y_1 > 0, X_2 = 0, Y_2 = 0) \\
& \times P(X_1 > 0, Y_1 >= 0, X_2 = 0, Y_2 = 0) \\
& + P(C|X_1 > 0, Y_1 > 0, X_2 = 0, Y_2 = 0) \\
& \times P(X_1 > 0, Y_1 > 0, X_2 > 0, Y_2 = 0) \\
& + P(C|X_1 > 0, Y_1 > 0, X_2 = 0, Y_2 > 0) \\
& \times P(X_1 > 0, Y_1 > 0, X_2 = 0, Y_2 > 0) \\
& + P(C|X_1 > 0, Y_1 > 0, X_2 > 0, Y_2 > 0) \\
& \times P(X_1 > 0, Y_1 > 0, X_2 > 0, Y_2 > 0) \\
P(C) &= p_{00}^2(0) + p_{00}p_{10}(0) + p_{00}p_{01}(0) + p_{00}p_{11} \\
& + \cdots + p_{11}p_{00} + p_{11}p_{10}(0) + p_{11}p_{01}(0) + p_{11}^2\tau_{11} \\
&= p_{00}p_{11} + p_{11}p_{00} + p_{11}^2\tau_{11} \\
P(C) &= p_{11}^2\tau_{11} + 2p_{00}p_{11}. \qquad (3.7)
\end{aligned}
$$

Similarly,

$$P(D) = p_{10}p_{01} + p_{01}p_{10}. \qquad (3.8)$$

Substituting, (3.7) and (3.8) in (3.6), we will get (3.5), which completes the proof. ∎

It can easily be seen that $\tau^* = \tau_{11}$ when there are no tied observations in both variables, i.e., $p_{00} = p_{10} = p_{01} = 0$ and $p_{11} = 1$.

From Proposition 1, an unbiased estimator of τ^* is defined as

$$\widehat{\tau^*} = \widehat{p_{11}}^2 \widehat{\tau_{11}} + 2(\widehat{p_{00}}\widehat{p_{11}} - \widehat{p_{01}}\widehat{p_{10}}), \tag{3.9}$$

where $\widehat{p_{ij}} = n_{ij}/n$ for $i = 0, 1$, $j = 0, 1$ and $\widehat{\tau_{11}}$ is given by (3.4) calculated from the nonzero pairs of observations.

A simulation study will be used to investigate the properties of the proposed estimate.

3.3 Asymptotic Distribution of $\widehat{\tau^*}$

The asymptotic distribution of the estimator $\widehat{\tau^*}$ will be determined partly using the delta method (see, e.g., Agresti, 2002). Given that $\mathbf{p}$ is the vector of cell probabilities in a multinomial distribution and $\widehat{\mathbf{p}}$ is the vector of sample proportions.

Theorem 1 *Let $g(\boldsymbol{p})$ denote a differentiable function of $\{p_{ij}\}$, with sample value $g(\widehat{\boldsymbol{p}})$ for a multinomial sample. Let*

$$\phi_{ij} = \frac{\partial g(\boldsymbol{p})}{\partial p_{ij}}, \text{ where } i, j = 0, 1 \tag{3.10}$$

Then,

$$\sqrt{n}[g(\widehat{\boldsymbol{p}}) - g(\boldsymbol{p})] \xrightarrow{D} N(0, \sigma^2), \tag{3.11}$$

where the asymptotic variance is defined as

$$\sigma^2 = \sum p_{ij}\phi_{ij}^2 - \left(\sum p_{ij}\phi_{ij}\right)^2 . \tag{3.12}$$

Using Theorem 1, the following proposition states the asymptotic distribution of the proposed estimator, τ^*.

Proposition 2 *Suppose* $(n_{00}, n_{01}, n_{10}, n_{11})$ *have a multinomial distribution with cell probabilities* $\boldsymbol{p} = (p_{00}, p_{01}, p_{10}, p_{11})'$. *Let* $n = n_{00} + n_{01} + n_{10} + n_{11}$, *and let* $\widehat{\boldsymbol{p}} = (\widehat{p_{00}}, \widehat{p_{01}}, \widehat{p_{10}}, \widehat{p_{11}})'$ *denote the sample proportions, where* $\widehat{p_{ij}} = n_{ij}/n$. *Then* $\sqrt{n}(\widehat{\tau^*} - \tau^*) \xrightarrow{D} N(0, \sigma^2_{\widehat{\tau^*}})$ *where*

$$\sigma^2_{\widehat{\tau^*}} = 2\tau^*(p_{00} + p_{11}) - p_{11}^2\tau_{11}(2p_{11} - 6p_{00} - 4p_{11}\tau_{11}) + 4p_{01}p_{10} - 4\tau^{*2}. \tag{3.13}$$

Proof. The proof of $E(\widehat{\tau^*})$ is straightforward and since $\widehat{\tau_{11}}$ is an MLE of τ, and $\widehat{\tau^*}$ is a function of this MLE, then $\widehat{\tau^*}$ is an MLE of τ^*.

To derive the asymptotic variance, first define τ^* as a function of $\mathbf{p}$. We have $g(\mathbf{p}) = p_{11}^2\tau_{11} + 2(p_{00}p_{11} - p_{01}p_{10})$. The elements of ϕ_{ij} of Φ, given by (3.10) are: $\phi_{00} = \partial g(\mathbf{p})/\partial p_{00} = 2p_{11}$, $\phi_{11} = 2p_{11}\tau_{11} + 2p_{00}$, $\phi_{01} = -2p_{10}$, and $\phi_{10} = -2p_{01}$. And consequently,

$$\begin{aligned}
\sum p_{ij}\phi_{ij}^2 &= p_{00}(2p_{11})^2 + p_{11}(2p_{11}\tau_{11} + 2p_{00})^2 + p_{01}(-2p_{10})^2 + p_{10}(-2p_{01})^2 \\
&= 4p_{00}p_{11}^2 + 4p_{11}^3\tau_{11}^2 + 8p_{11}^2p_{00}\tau_{11} + 4p_{11}p_{00}^2 + 4p_{01}p_{10}^2 + 4p_{10}p_{01}^2 \\
\sum p_{ij}\phi_{ij}^2 &= 2\tau^*(p_{00} + p_{11}) - p_{11}^2\tau_{11}(2p_{11} - 6p_{00} - 4p_{11}\tau_{11}) + 4p_{01}p_{10}, \qquad (3.14)
\end{aligned}$$

where $(p_{00} + p_{11}) = 1 - (p_{01} + p_{10})$ and $(p_{01} + p_{10}) = 1 - (p_{00} + p_{11})$, from $1 = p_{00} + p_{01} + p_{10} + p_{11}$.

Next,

$$\begin{aligned}\left(\sum p_{ij}\phi_{ij}\right)^2 &= [p_{00}(2p_{11}) + p_{11}(2p_{11}\tau_{11} + 2p_{00}) + p_{01}(-2p_{10}) + p_{10}(-2p_{01})]^2 \\ &= [4p_{00}p_{11} + 2p_{11}^2\tau_{11} - 4p_{01}p_{10}]^2 \\ \left(\sum p_{ij}\phi_{ij}\right)^2 &= 4\tau^{*2}. \end{aligned} \qquad (3.15)$$

Substituting (3.14) and (3.15) in (3.10) will give the desired result in (3.13). ■

As a consequence of Proposition 2, an estimate of the standard error of $g(\widehat{\mathbf{p}})$ is given by,

$$S_{\widehat{\tau^*}} = \frac{\sqrt{2\widehat{\tau^*}(\widehat{p_{00}} + \widehat{p_{11}}) - \widehat{p_{11}}^2\widehat{\tau_{11}}(2\widehat{p_{11}} - 6\widehat{p_{00}} - 4\widehat{p_{11}}\widehat{\tau_{11}}) + 4\widehat{p_{01}}\widehat{p_{10}} - 4\widehat{\tau^*}^2}}{\sqrt{n}}. \qquad (3.16)$$

Chapter 4

PROPOSED ESTIMATOR OF SPEARMAN'S RHO

Another commonly used measure of association is the Spearman's rho. Given (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) are independent replicates of (X, Y), the population Spearman's ρ is defined as

$$\rho_S = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]). \tag{4.1}$$

where X_2 and Y_3 are independent variables with the same marginal distributions as X_1 and Y_1, respectively.

4.1 Adjustment of Spearman's Rho with Ties

If there are tied ranks in X, or Y, or both, the estimator in (1.18) has to be adjusted. If there are l distinct tied observations in X, each having varying size s then we can define

$$T_x = \sum_l s_l(s_l^2 - 1) \tag{4.2}$$

ties in X.

Similarly in Y, we can define

$$T_y = \sum_m t_m(t_m^2 - 1). \tag{4.3}$$

Therefore, these will lead to the adjusted formula in (1.19), which we are stating here again.

$$r_S = \frac{W_0 - 6\sum_{i=1}^{n} D_i^2 - \frac{1}{2}\{T_x + T_y\}}{\sqrt{(W_0 - T_x)(W_0 - T_y)}}. \tag{4.4}$$

In the presence of zero values in either or both variables, the sample size is $n_{00} + n_{01}$ in X and $n_{00} + n_{10}$ in Y. Hence, (4.2) will reduce to $(n_{00} + n_{01})[(n_{00} + n_{01})^2 - 1]$ and similarly, (4.3) will reduce to $(n_{00} + n_{10})[(n_{00} + n_{10})^2 - 1]$.

4.2 Proposed Estimator of Spearman's Rho, ρ_S^*

Since the estimator (4.4) is not a MLE of Spearman's rho defined in (4.1), we propose a new estimator, ρ_S^*, being an estimator of (4.1) for the case of pairs being tied at 0 on at least one variable.

Proposition 3 *Let (X_1, Y_1) and (X_2, Y_3) be identically distributed random vectors, and X_2 and Y_3 are independent. Then the population Spearman's ρ given by (4.4) has a form*

$$\rho_S{}^* \;=\; p_{11}p_{+1}p_{1+}\rho_{S11} + 3(p_{00}p_{11} - p_{10}p_{01}) \tag{4.5}$$

where ρ_{S11} is the population Spearman's ρ defined in (4.1) for the non-zero pairs of observations.

Proof. Given a zero-inflated data in a bivariate setting, the scope of the analysis will be divided into four different quadrants as shown in Figure 2.1. Again, these areas will each contain the pairs $(0,0), (x>0,0), (0,y>0)$ and $(x>0,y>0)$. Using the total probability theorem, the adjusted formula for Spearman's ρ can then be derived as

$$
\begin{aligned}
{\rho_S}^* &= \sum_{\forall i} P(\rho_S | X_{i1}, Y_{i1}, X_{i2}, Y_{i3}) P(X_{i1}, Y_{i1}, X_{i2}, Y_{i3}) \\
&= 3(p_{00}p_{1+}p_{+1} - p_{10}p_{0+}p_{+1} - p_{01}p_{1+}p_{+0} + p_{11}p_{0+}p_{+0}) + p_{11}p_{+1}p_{1+}\rho_{S11} \\
&= 3[p_{00}(p_{10}+p_{11})(p_{01}+p_{11}) - p_{10}(p_{00}+p_{01})(p_{01}+p_{11}) \\
&\quad -p_{01}(p_{10}+p_{11})(p_{00}+p_{10}) + p_{11}(p_{00}+p_{01})(p_{00}+p_{10})] \\
&\quad +p_{11}p_{+1}p_{1+}\rho_{S11} \\
&= 3[p_{00}p_{10}p_{01} + p_{00}p_{10}p_{11} + p_{00}p_{11}p_{01} + p_{00}p_{11}p_{11} - p_{10}p_{00}p_{01} - p_{10}p_{00}p_{11} \\
&\quad -p_{10}p_{01}p_{01} - p_{10}p_{01}p_{11} - p_{01}p_{10}p_{00} - p_{01}p_{10}p_{10} - p_{01}p_{11}p_{00} - p_{01}p_{11}p_{10} \\
&\quad +p_{11}p_{00}p_{00} + p_{11}p_{00}p_{10} + p_{11}p_{01}p_{00} + p_{11}p_{01}p_{10}] + p_{11}p_{+1}p_{1+}\rho_{S11} \\
&= 3[p_{00}p_{01}p_{11} + p_{00}p_{11}p_{11} - p_{01}p_{01}p_{10} - p_{10}p_{01}p_{11} - p_{00}p_{10}p_{01} - p_{01}p_{10}p_{10} \\
&\quad +p_{00}p_{10}p_{11} + p_{00}p_{00}p_{11}] + p_{11}p_{+1}p_{1+}\rho_S \\
&= 3[p_{00}p_{11}(p_{01}+p_{10}+p_{00}+p_{11}) - p_{01}p_{10}(p_{01}+p_{10}+p_{00}+p_{11})] + p_{11}p_{+1}p_{1+}\rho_{S11}
\end{aligned}
$$

$$
{\rho_S}^* = p_{11}p_{+1}p_{1+}\rho_{S11} + 3(p_{00}p_{11} - p_{10}p_{01}) \ \blacksquare \tag{4.6}
$$

Equation (4.6) can be easily reduced to (4.1) when there are no tied observations in both variables.

4.3 Asymptotic Distribution of $\widehat{\rho_S^*}$

Proposition 4 *Suppose* $(n_{00}, n_{01}, n_{10}, n_{11})$ *have a multinomial distribution with cell probabilities* $\boldsymbol{p} = (p_{00}, p_{01}, p_{10}, p_{11})'$. *Let* $n = n_{00} + n_{01} + n_{10} + n_{11}$, *and let* $\widehat{\boldsymbol{p}} = (\widehat{p}_{00}, \widehat{p}_{01}, \widehat{p}_{10}, \widehat{p}_{11})'$ *denote the sample proportions, where* $\widehat{p}_{ij} = n_{ij}/n$. *Then* $\sqrt{n}(\widehat{\rho_S^*} - \rho_S^*) \overset{D}{\rightarrow} N(0, \sigma^2_{\widehat{\rho_S^*}})$ *where*

$$\begin{aligned}
\sigma^2_{\widehat{\rho_S^*}} &= \rho_S^*[3(p_{00} + p_{11}) + 2p_{11}\rho_{S11}(p_{+1} + p_{1+})] \\
&\quad +9p_{01}p_{10} + p_{11}p_{1+}p_{+1}\rho_{S11}(p_{1+}p_{+1}\rho_{S11} + 3(p_{00} - p_{11}) + 2p_{11}^2\rho_{S11}) \\
&\quad +p_{11}^2\rho_{S11}^2(p_{01}p_{1+}^2 + p_{10}p_{+1}^2 + p_{11}p_{1+}^2 + p_{11}^3) \\
&\quad -\left\{3\rho_S^* - 3(p_{00}p_{11} - p_{01}p_{10})\right\}^2. \qquad (4.7)
\end{aligned}$$

Proof. The proof of $E(\widehat{\rho_S^*})$ is straightforward and since $\widehat{\rho_{S11}}$ is an MLE of ρ_S, and $\widehat{\rho_S^*}$ is a function of this MLE, then $\widehat{\rho_S^*}$ is an MLE of ρ_S^*.

To derive the asymptotic variance of the proposed estimator $\widehat{\rho_S^*}$ of Spearman's rho, first define ρ_S^* as a function of $\mathbf{p}$, we have

$$g(\mathbf{p}) = p_{11}p_{+1}p_{1+}\rho_{S11} + 3(p_{00}p_{11} - p_{01}p_{10}). \qquad (4.8)$$

The elements of ϕ_{ij} of Φ, given by (3.10) are: $\phi_{00} = \partial g(\mathbf{p})/\partial p_{00} = 3p_{11}$, $\phi_{11} = p_{1+}p_{+1}\rho_{S11} + p_{11}\rho_{S11}(p_{1+} + p_{+1}) + 3p_{00}$, $\phi_{01} = p_{1+}p_{11}\rho_{S11} - 3p_{10}$, and $\phi_{10} = p_{+1}p_{11}\rho_{S11} - 3p_{01}$. And consequently,

$$\begin{aligned}
\sum p_{ij}\phi_{ij}^2 &= p_{00}(3p_{11})^2 \\
&\quad +p_{11}\left\{p_{1+}p_{+1}\rho_{S11} + 3p_{00} + p_{11}\rho_{S11}[p_{1+} + p_{+1}]\right\}^2 \\
&\quad +p_{01}[p_{11}p_{1+}\rho_{S11} - 3p_{10}]^2 + p_{10}[p_{+1}p_{11}\rho_{S11} - 3p_{01}]^2 \\
&= 9p_{00}p_{11}^2 + p_{11}(p_{1+}p_{+1}\rho_{S11} + 3p_{00})^2
\end{aligned}$$

$$
\begin{aligned}
& +2p_{11}(p_{1+}p_{+1}\rho_{S11} + 3p_{00})(p_{11}\rho_{S11}[p_{1+} + p_{+1}]) \\
& +p_{11}^3\rho_{S11}^2[p_{1+} + p_{+1}]^2 \\
& +p_{01}[p_{11}^2p_{1+}^2\rho_{S11}^2 - 6p_{11}p_{1+}\rho_{S11}p_{10} + 9p_{10}^2] \\
& +p_{10}[p_{+1}^2p_{11}^2\rho_{S11}^2 - 6p_{+1}p_{11}\rho_{S11}p_{01} + 9p_{01}^2] \\
= \; & 9p_{00}p_{11}(p_{00} + p_{11}) + 9p_{01}p_{10}(p_{01} + p_{10}) \\
& +p_{11}p_{1+}p_{+1}\rho_{S11}(2p_{1+}p_{11}\rho_{S11} + 2p_{+1}p_{11}\rho_{S11}) \\
& +p_{11}p_{1+}p_{+1}\rho_{S11}(p_{1+}p_{+1}\rho_{S11} + 6p_{00} + 2p_{11}^2\rho_{S11}) \\
& +6p_{11}\rho_{S11}(p_{00}p_{11}p_{+1} + p_{00}p_{11}p_{1+} - p_{01}p_{10}p_{+1} - p_{01}p_{10}p_{1+}) \\
& +p_{11}^2\rho_{S11}^2(p_{01}p_{1+}^2 + p_{10}p_{+1}^2 + p_{11}p_{1+}^2 + p_{11}^3) \\
= \; & 9(p_{00} + p_{11})(p_{00}p_{11} - p_{01}p_{10}) + 9p_{01}p_{10} \\
& +p_{11}p_{1+}p_{+1}\rho_{S11}[2p_{11}\rho_{S11}(p_{+1} + p_{1+})] \\
& +3(p_{00} + p_{11})p_{11}p_{1+}p_{+1}\rho_{S11} \\
& +2p_{11}\rho_{S11}(p_{+1} + p_{1+})[3(p_{00}p_{11} - p_{01}p_{10})] \\
& +p_{11}p_{1+}p_{+1}\rho_{S11}(p_{1+}p_{+1}\rho_{S11} + 3(p_{00} - p_{11}) + 2p_{11}^2\rho_{S11}) \\
& +p_{11}^2\rho_{S11}^2(p_{01}p_{1+}^2 + p_{10}p_{+1}^2 + p_{11}p_{1+}^2 + p_{11}^3) \\
= \; & 3(p_{00} + p_{11})[p_{11}p_{1+}p_{+1}\rho_{S11} + 3(p_{00}p_{11} - p_{01}p_{10})] \\
& 2p_{11}\rho_{S11}(p_{+1} + p_{1+})[p_{11}p_{1+}p_{+1}\rho_{S11} + 3(p_{00}p_{11} - p_{01}p_{10})] \\
& +9p_{01}p_{10} + p_{11}p_{1+}p_{+1}\rho_{S11}(p_{1+}p_{+1}\rho_{S11} + 3(p_{00} - p_{11}) + 2p_{11}^2\rho_{S11}) \\
& +p_{11}^2\rho_{S11}^2(p_{01}p_{1+}^2 + p_{10}p_{+1}^2 + p_{11}p_{1+}^2 + p_{11}^3) \\
\sum p_{ij}\phi_{ij}^2 \; = \; & \rho_S^*[3(p_{00} + p_{11}) + 2p_{11}\rho_{S11}(p_{+1} + p_{1+})] \\
& +9p_{01}p_{10} + p_{11}p_{1+}p_{+1}\rho_{S11}(p_{1+}p_{+1}\rho_{S11} + 3(p_{00} - p_{11}) + 2p_{11}^2\rho_{S11}) \\
& +p_{11}^2\rho_{S11}^2(p_{01}p_{1+}^2 + p_{10}p_{+1}^2 + p_{11}p_{1+}^2 + p_{11}^3) \qquad (4.9)
\end{aligned}
$$

where $(p_{00}+p_{11}) = 1-(p_{01}+p_{10})$, $(p_{01}+p_{10}) = 1-(p_{00}+p_{11})$, $p_{+1} = p_{01}+p_{11}$ and $p_{1+} = p_{10}+p_{11}$.

Next,

$$\begin{aligned}
\sum p_{ij}\phi_{ij} &= 3p_{00}p_{11} + p_{11}\left[p_{1+}p_{+1}\rho_{S11} + p_{11}\rho_{S11}(p_{1+}+p_{+1}) + 3p_{00}\right] \\
&\quad + p_{01}(p_{1+}p_{11}\rho_{S11} - 3p_{10}) + p_{10}(p_{+1}p_{11}\rho_{S11} - 3p_{01}) \\
&= p_{+1}p_{1+}p_{11}\rho_{S11} + 6(p_{00}p_{11} - p_{01}p_{10}) \\
&\quad + p_{11}\rho_{S11}(p_{1+}p_{11} + p_{11}p_{+1} + p_{1+}p_{01} + p_{+1}p_{10}) \\
&= 3p_{1+}p_{+1}p_{11}\rho_{S11} + 9(p_{00}p_{11} - p_{01}p_{10}) - 3(p_{00}p_{11} - p_{01}p_{10}) \\
\left(\sum p_{ij}\phi_{ij}\right)^2 &= \{3\rho_S^* - 3(p_{00}p_{11} - p_{01}p_{10})\}^2 \qquad (4.10)
\end{aligned}$$

Substituting (4.9) and (4.10) in (3.10) will give the desired result in (4.7). ■

Chapter 5

SIMULATION STUDY AND RESULTS

5.1 Simulation and Results: Kendall's Tau

5.1.1 Simulation Plan

A Monte Carlo simulation procedure was employed to study the proposed estimator for Kendall's τ defined in Proposition 1. Samples of $n = 30, 50, 100$ pairs of data were simulated from a bivariate lognormal distribution with $\mu_X = 0$ and $\mu_Y = 0$. Using the relationship between Kendall's τ and Pearson's ρ defined as $\tau = \frac{2}{\pi} \arcsin(\rho)$, $\tau = 0.1$ to 0.9 by 0.1 was used to calculate $\rho = \sin\left(\frac{\tau\pi}{2}\right)$ for generating the data. The proportion of zeroes used were $p_{00} = p_{01} = 0.1$ and $p_{10} = (0.1, 0.2, 0.3)$.

For each case, a multinomial distribution was used to randomly determine the n_{00} pairs of observations that will be $(0,0)$ with probability p_{00}; n_{01} pairs of $(0,y)$ with probability p_{01}; and n_{10} pair of $(x,0)$ with probability p_{10}, where $p_{11} = 1 - (p_{00} + p_{01} + p_{10})$. From the nonzero pairs of observations, Kendall's coefficient of correlation, $\widehat{\tau_{11}}$, was calculated. In addition, the following estimates were determined from the simulated data, $\widehat{p_{00}} = n_{00}/n$; $\widehat{p_{10}} = n_{10}/n$; $\widehat{p_{01}} = n_{01}/n$;

and $\widehat{p_{11}} = n_{11}/n = 1 - (\widehat{p_{00}} + \widehat{p_{10}} + \widehat{p_{01}})$. Then the proposed estimator of Kendall's τ defined in Proposition 1 was calculated as $\widehat{\tau^*} = \widehat{p_{11}}^2\widehat{\tau_{11}} + 2\widehat{p_{00}}\widehat{p_{11}} - 2\widehat{p_{01}}\widehat{p_{10}}$. The estimate of (3.4) was also calculated for comparison. For each case, the process was repeated 1000 times.

The same plan was utilized to study the asymptotic variance of the estimator proposed in Proposition 1. Sample sizes up to 200 were considered. Although not all combinations of the cell probabilities listed above were used, some additional combinations not considered before were presented. Only a low, mid, and a high level of the population τ were considered and 2000 replicates were performed.

5.1.2 Results

Table 5.1 shows the percentile intervals for each of the cases mentioned above. For all cases, the intervals based on $\widehat{\tau^*}$ contain the value of τ^*. The intervals also are narrower as the sample size is increased. On the other hand, the intervals based on $\widehat{\tau}$ also contain the value for τ^* but the intervals are consistently wider.

In order to check normality of the estimates, a Shapiro-Wilk test was employed and the value of test statistic and the corresponding p-value from the 1000 estimates of $\widehat{\tau}^*$ are displayed in Table 5.2. There is no evident pattern for non-normality which occurred by chance. To further look into the normality of the estimate by lowering the number of samples, a random sample of 100 out of the 1000 estimates was used and the results are tabulated in Table 5.3. Again, the non-normality of the estimate is not apparently present.

The mean, $\overline{\widehat{\tau^*}}$, and corresponding standard deviation, $S(\widehat{\tau^*})$, from the 1000 estimates for each case are displayed in Tables 5.4, 5.5, and 5.6 for sample sizes 30, 50 and 100, respectively. These measures were used to look at the bias of the proposed estimator $\widehat{\tau^*}$ in Proposition 1. As shown on the tables, when compared to τ^*, the bias tends to be very small as the sample size increases for $\widehat{\tau^*}$ than for $\widehat{\tau}$. The corresponding adjusted variance also drops with the increase in the sample size. Also presented is the MSE of the estimator and these values will be compared later with the results of the asymptotic variance. The MSE for the $\widehat{\tau}$ estimates are consistently larger than that for $\widehat{\tau^*}$.

The variance of the estimator proposed in (3.13) for each of the known population values are calculated and these values approach 0 as the sample size is increased from 50 to 200, regardless of the size of the pairs with zeros on either variable or on both. The sample variance of the 2000 calculated estimates was determined for each case. These are presented in Tables 5.7 and 5.9 under the column $S^2_{\widehat{\tau^*}}$. The sample variance tends to be larger than the asymptotic variance for smaller proportions of zero but stabilizes as the sample size is increased and there are more zeroes in the data. The estimate of the asymptotic variance was also calculated from each of the cases and the mean of these 2000 estimates are presented in Tables 5.8 and 5.10. These values are consistent with the population values regardless of the size of zeroes in the data.

					Interval for τ^* based on $\widehat{\tau^*}$			Interval for τ^* based on $\widehat{\tau}$		
p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	**n=30**	**n=50**	**n=100**	**n=30**	**n=50**	**n=100**
0.1	0.1	0.1	0.1	0.169	(-0.069,0.396)	(0.014,0.331)	(0.052,0.285)	(-0.124,0.454)	(-0.020,0.386)	(0.027,0.323)
			0.2	0.218	(0.002,0.432)	(0.057,0.379)	(0.108,0.335)	(-0.062,0.478)	(0.013,0.425)	(0.082,0.369)
			0.3	0.267	(0.061,0.466)	(0.112,0.427)	(0.153,0.384)	(-0.008,0.535)	(0.075,0.475)	(0.130,0.418)
			0.4	0.316	(0.100,0.509)	(0.159,0.483)	(0.206,0.423)	(0.035,0.583)	(0.121,0.538)	(0.181,0.463)
			0.5	0.365	(0.153,0.569)	(0.214,0.521)	(0.256,0.481)	(0.105,0.639)	(0.188,0.572)	(0.234,0.522)
			0.6	0.414	(0.210,0.624)	(0.251,0.577)	(0.302,0.525)	(0.149,0.681)	(0.222,0.623)	(0.277,0.564)
			0.7	0.463	(0.248,0.671)	(0.296,0.629)	(0.355,0.578)	(0.188,0.730)	(0.257,0.673)	(0.337,0.627)
			0.8	0.512	(0.287,0.722)	(0.338,0.692)	(0.384,0.634)	(0.233,0.782)	(0.316,0.740)	(0.372,0.680)
			0.9	0.561	(0.341,0.776)	(0.376,0.746)	(0.428,0.676)	(0.279,0.824)	(0.340,0.788)	(0.421,0.727)
		0.2	0.1	0.116	(-0.076,0.314)	(-0.025,0.261)	(0.013,0.217)	(-0.163,0.402)	(-0.072,0.325)	(-0.020,0.275)
			0.2	0.152	(-0.037,0.332)	(0.011,0.298)	(0.040,0.256)	(-0.110,0.422)	(-0.031,0.355)	(0.007,0.307)
			0.3	0.188	(-0.016,0.397)	(0.036,0.336)	(0.080,0.304)	(-0.088,0.472)	(-0.014,0.400)	(0.052,0.355)
			0.4	0.224	(0.014,0.425)	(0.072,0.401)	(0.116,0.332)	(-0.086,0.516)	(0.024,0.458)	(0.085,0.391)
			0.5	0.260	(0.055,0.479)	(0.091,0.428)	(0.158,0.387)	(-0.033,0.572)	(0.037,0.488)	(0.126,0.438)
			0.6	0.296	(0.075,0.506)	(0.129,0.478)	(0.176,0.418)	(-0.014,0.582)	(0.078,0.531)	(0.148,0.471)
			0.7	0.332	(0.105,0.555)	(0.162,0.521)	(0.214,0.455)	(0.029,0.629)	(0.097,0.571)	(0.182,0.511)
			0.8	0.368	(0.128,0.621)	(0.174,0.548)	(0.243,0.502)	(0.068,0.692)	(0.128,0.612)	(0.218,0.563)
			0.9	0.404	(0.146,0.679)	(0.225,0.606)	(0.271,0.547)	(0.098,0.746)	(0.181,0.656)	(0.247,0.616)
		0.3	0.1	0.065	(-0.114,0.252)	(-0.068,0.217)	(-0.025,0.166)	(-0.214,0.367)	(-0.153,0.273)	(-0.078,0.218)
			0.2	0.090	(-0.085,0.295)	(-0.048,0.236)	(-0.005,0.190)	(-0.168,0.388)	(-0.127,0.307)	(-0.053,0.246)
			0.3	0.115	(-0.062,0.318)	(-0.020,0.261)	(0.020,0.211)	(-0.152,0.408)	(-0.088,0.338)	(-0.024,0.271)
			0.4	0.140	(-0.056,0.346)	(-0.016,0.301)	(0.039,0.244)	(-0.146,0.454)	(-0.097,0.384)	(0.016,0.306)
			0.5	0.165	(-0.029,0.379)	(0.020,0.322)	(0.058,0.286)	(-0.135,0.485)	(-0.036,0.414)	(0.021,0.353)
			0.6	0.190	(-0.009,0.406)	(0.038,0.353)	(0.075,0.310)	(-0.120,0.487)	(-0.048,0.453)	(0.052,0.385)
			0.7	0.215	(0.003,0.451)	(0.040,0.396)	(0.098,0.335)	(-0.100,0.539)	(-0.025,0.487)	(0.068,0.387)
			0.8	0.240	(0.016,0.485)	(0.061,0.424)	(0.112,0.372)	(-0.083,0.593)	(0.021,0.512)	(0.082,0.436)
			0.9	0.265	(0.038,0.523)	(0.077,0.464)	(0.129,0.403)	(-0.024,0.588)	(0.035,0.539)	(0.103,0.468)

Table 5.1: $(2.5^{th}, 97.6^{th})$ percentile intervals for τ^* based from the 1000 estimates

					Shapiro-Wilk Test Statistic (p-value)		
p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	**n=30**	**n=50**	**n=100**
0.1	0.1	0.1	0.1	0.169	0.9977(0.1783)	0.9988(0.7761)	0.9988(0.7664)
			0.2	0.218	0.9983(0.4266)	0.9979(0.2595)	0.9986(0.6436)
			0.3	0.267	0.9985(0.5730)	0.9986(0.6352)	0.9988(0.7577)
			0.4	0.316	0.9979(0.2461)	0.9979(0.2326)	0.9985(0.5763)
			0.5	0.365	0.9983(0.4237)	0.9989(0.8462)	0.9978(0.2177)
			0.6	0.414	0.9987(0.6707)	0.9988(0.7774)	0.9994(0.9920)
			0.7	0.463	0.9976(0.1411)	0.9982(0.3899)	0.9976(0.1466)
			0.8	0.512	0.9952(**0.0033**)	0.9976(0.1639)	0.9970(0.0618)
			0.9	0.561	0.9946(**0.0013**)	0.9978(0.2127)	0.9970(0.0538)
		0.2	0.1	0.116	0.9967(**0.0347**)	0.9975(0.1278)	0.9985(0.5501)
			0.2	0.152	0.9984(0.4656)	0.9983(0.4414)	0.9987(0.7083)
			0.3	0.188	0.9989(0.8221)	0.9970(0.0602)	0.9978(0.2003)
			0.4	0.224	0.9971(0.0643)	0.9972(0.0840)	0.9991(0.9360)
			0.5	0.260	0.9979(0.2317)	0.9988(0.7848)	0.9972(0.0856)
			0.6	0.296	0.9986(0.6466)	0.9982(0.3963)	0.9982(0.3683)
			0.7	0.332	0.9989(0.7974)	0.9982(0.4017)	0.9986(0.6372)
			0.8	0.368	0.9990(0.8827)	0.9981(0.3094)	0.9992(0.9657)
			0.9	0.404	0.9976(0.1459)	0.9986(0.6125)	0.9982(0.3542)
		0.3	0.1	0.065	0.9971(0.0664)	0.9984(0.5052)	0.9979(0.2254)
			0.2	0.090	0.9961(**0.0123**)	0.9987(0.6915)	0.9978(0.2140)
			0.3	0.115	0.9957(**0.0065**)	0.9983(0.4459)	0.9980(0.2954)
			0.4	0.140	0.9978(0.2010)	0.9990(0.8950)	0.9991(0.9148)
			0.5	0.165	0.9979(0.2579)	0.9981(0.3157)	0.9982(0.3616)
			0.6	0.190	0.9985(0.5638)	0.9982(0.3616)	0.9990(0.8574)
			0.7	0.215	0.9989(0.8094)	0.9982(0.3991)	0.9979(0.2588)
			0.8	0.240	0.9984(0.4980)	0.9990(0.8888)	0.9990(0.8852)
			0.9	0.265	0.9989(0.8046)	0.9988(0.7484)	0.9989(0.8194)

Table 5.2: Normality test from the 1000 $\widehat{\tau^*}$ estimates

					Shapiro-Wilk Test Statistic (p-value)		
p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	**n=30**	**n=50**	**n=100**
0.1	0.1	0.1	0.1	0.169	0.9951(0.9772)	0.9939(0.9374)	0.9901(0.6695)
			0.2	0.218	0.9917(0.7969)	0.9909(0.7380)	0.9894(0.6141)
			0.3	0.267	0.9935(0.9164)	0.9904(0.6977)	0.9855(0.3454)
			0.4	0.316	0.9903(0.6905)	0.9907(0.7175)	0.9918(0.8035)
			0.5	0.365	0.9960(0.9934)	0.9927(0.8696)	0.9851(0.3231)
			0.6	0.414	0.9799(0.1316)	0.9886(0.5556)	0.9851(0.3239)
			0.7	0.463	0.9890(0.5818)	0.9882(0.5202)	0.9732(**0.0390**)
			0.8	0.512	0.9870(0.4389)	0.9925(0.8534)	0.9805(0.1453)
			0.9	0.561	0.9882(0.5208)	0.9913(0.7707)	0.9883(0.5299)
		0.2	0.1	0.116	0.9903(0.6909)	0.9843(0.2845)	0.9904(0.6946)
			0.2	0.152	0.9890(0.5867)	0.9862(0.3842)	0.9912(0.7642)
			0.3	0.188	0.9887(0.5577)	0.9817(0.1800)	0.9912(0.7625)
			0.4	0.224	0.9784(0.1003)	0.9930(0.8901)	0.9783(0.0973)
			0.5	0.260	0.9959(0.9915)	0.9911(0.7560)	0.9859(0.3690)
			0.6	0.296	0.9819(0.1873)	0.9893(0.6054)	0.9830(0.2256)
			0.7	0.332	0.9963(0.9956)	0.9890(0.5835)	0.9768(0.0748)
			0.8	0.368	0.9921(0.8300)	0.9783(0.0978)	0.9845(0.2898)
			0.9	0.404	0.9883(0.5289)	0.9952(0.9802)	0.9882(0.5239)
		0.3	0.1	0.065	0.9821(0.1914)	0.9817(0.1798)	0.9886(0.5539)
			0.2	0.090	0.9907(0.7199)	0.9920(0.8222)	0.9868(0.4269)
			0.3	0.115	0.9892(0.5984)	0.9805(0.1460)	0.9825(0.2067)
			0.4	0.140	0.9832(0.2324)	0.9915(0.7830)	0.9921(0.8310)
			0.5	0.165	0.9929(0.8842)	0.9737(**0.0429**)	0.9880(0.5054)
			0.6	0.190	0.9824(0.2050)	0.9798(0.1281)	0.9902(0.6831)
			0.7	0.215	0.9892(0.6037)	0.9864(0.3970)	0.9914(0.7766)
			0.8	0.240	0.9918(0.8102)	0.9777(0.0876)	0.9928(0.8755)
			0.9	0.265	0.9829(0.2211)	0.9934(0.9090)	0.9855(0.3423)

Table 5.3: Normality test from the 100 randomly selected $\widehat{\tau^*}$ estimates

					$\widehat{\tau^*}$				$\widehat{\tau}$			
p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	**Mean**	**SD**	**Bias**	**MSE**	**Mean**	**SD**	**Bias**	**MSE**
0.1	0.1	0.1	0.1	0.169	0.1626	0.1111	0.0064	0.0124	0.1734	0.1398	-0.0044	0.0196
			0.2	0.218	0.2166	0.1080	0.0014	0.0117	0.2250	0.1394	-0.0070	0.0195
			0.3	0.267	0.2694	0.1055	-0.0024	0.0111	0.2809	0.1354	-0.0139	0.0185
			0.4	0.316	0.3119	0.1013	0.0041	0.0103	0.3277	0.1367	-0.0117	0.0188
			0.5	0.365	0.3695	0.1044	-0.0045	0.0109	0.3879	0.1375	-0.0229	0.0194
			0.6	0.414	0.4233	0.1041	-0.0093	0.0109	0.4425	0.1374	-0.0285	0.0197
			0.7	0.463	0.4633	0.1096	-0.0003	0.0120	0.4815	0.1390	-0.0185	0.0197
			0.8	0.512	0.5167	0.1136	-0.0047	0.0129	0.5352	0.1402	-0.0232	0.0202
			0.9	0.561	0.5626	0.1152	-0.0016	0.0133	0.5810	0.1408	-0.0200	0.0202
		0.2	0.1	0.116	0.1096	0.1009	0.0064	0.0102	0.1234	0.1376	-0.0074	0.0190
			0.2	0.152	0.1506	0.0963	0.0014	0.0093	0.1589	0.1386	-0.0069	0.0193
			0.3	0.188	0.1875	0.1036	0.0005	0.0107	0.2052	0.1421	-0.0172	0.0205
			0.4	0.224	0.2284	0.1047	-0.0044	0.0110	0.2429	0.1472	-0.0189	0.0220
			0.5	0.260	0.2675	0.1069	-0.0075	0.0115	0.2860	0.1518	-0.0260	0.0237
			0.6	0.296	0.2998	0.1110	-0.0038	0.0123	0.3142	0.1488	-0.0182	0.0225
			0.7	0.332	0.3297	0.1145	0.0023	0.0131	0.3472	0.1570	-0.0152	0.0249
			0.8	0.368	0.3744	0.1199	-0.0064	0.0144	0.4030	0.1545	-0.0350	0.0251
			0.9	0.404	0.4050	0.1301	-0.0010	0.0169	0.4246	0.1677	-0.0206	0.0286
		0.3	0.1	0.065	0.0634	0.0921	0.0016	0.0085	0.0727	0.1474	-0.0077	0.0218
			0.2	0.090	0.0951	0.0939	-0.0051	0.0088	0.1043	0.1430	-0.0143	0.0207
			0.3	0.115	0.1184	0.0962	-0.0034	0.0093	0.1341	0.1431	-0.0191	0.0208
			0.4	0.140	0.1406	0.1017	-0.0006	0.0103	0.1472	0.1512	-0.0072	0.0229
			0.5	0.165	0.1677	0.1045	-0.0027	0.0109	0.1839	0.1567	-0.0189	0.0249
			0.6	0.190	0.1935	0.1064	-0.0035	0.0113	0.2134	0.1575	-0.0234	0.0253
			0.7	0.215	0.2267	0.1122	-0.0117	0.0127	0.2495	0.1609	-0.0345	0.0271
			0.8	0.240	0.2450	0.1195	-0.0050	0.0143	0.2622	0.1687	-0.0222	0.0289
			0.9	0.265	0.2778	0.1235	-0.0128	0.0154	0.2953	0.1578	-0.0303	0.0258

Table 5.4: Summary statistics including the bias and MSE for $\widehat{\tau^*}$ and $\widehat{\tau}$ based on the 1000 estimates with n=30 sample size

					$\widehat{\tau^*}$				$\widehat{\tau}$			
p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	**Mean**	**SD**	**Bias**	**MSE**	**Mean**	**SD**	**Bias**	**MSE**
0.1	0.1	0.1	0.1	0.169	0.1706	0.0791	-0.0016	0.0063	0.1821	0.1030	-0.0131	0.0108
			0.2	0.218	0.2154	0.0818	0.0026	0.0067	0.2252	0.1032	-0.0072	0.0107
			0.3	0.267	0.2708	0.0805	-0.0038	0.0065	0.2837	0.1007	-0.0167	0.0104
			0.4	0.316	0.3138	0.0807	0.0022	0.0065	0.3282	0.1047	-0.0122	0.0111
			0.5	0.365	0.3675	0.0794	-0.0025	0.0063	0.3862	0.1003	-0.0212	0.0105
			0.6	0.414	0.4166	0.0812	-0.0026	0.0066	0.4337	0.1006	-0.0197	0.0105
			0.7	0.463	0.4644	0.0837	-0.0014	0.0070	0.4809	0.1063	-0.0179	0.0116
			0.8	0.512	0.5141	0.0895	-0.0021	0.0080	0.5376	0.1070	-0.0256	0.0121
			0.9	0.561	0.5619	0.0943	-0.0009	0.0089	0.5823	0.1158	-0.0213	0.0139
		0.2	0.1	0.116	0.1135	0.0721	0.0025	0.0052	0.1262	0.1023	-0.0102	0.0106
			0.2	0.152	0.1500	0.0747	0.0020	0.0056	0.1658	0.0994	-0.0138	0.0101
			0.3	0.188	0.1875	0.0795	0.0005	0.0063	0.2031	0.1069	-0.0151	0.0117
			0.4	0.224	0.2288	0.0805	-0.0048	0.0065	0.2446	0.1096	-0.0206	0.0124
			0.5	0.260	0.2605	0.0870	-0.0005	0.0076	0.2731	0.1165	-0.0131	0.0137
			0.6	0.296	0.2958	0.0862	0.0002	0.0074	0.3123	0.1136	-0.0163	0.0132
			0.7	0.332	0.3329	0.0891	-0.0009	0.0079	0.3517	0.1187	-0.0197	0.0145
			0.8	0.368	0.3651	0.0950	0.0029	0.0090	0.3879	0.1237	-0.0199	0.0157
			0.9	0.404	0.4119	0.0958	-0.0079	0.0092	0.4329	0.1198	-0.0289	0.0152
		0.3	0.1	0.065	0.0652	0.0716	-0.0002	0.0051	0.0735	0.1093	-0.0085	0.0120
			0.2	0.090	0.0900	0.0709	0.0000	0.0050	0.0996	0.1110	-0.0096	0.0124
			0.3	0.115	0.1155	0.0722	-0.0005	0.0052	0.1287	0.1102	-0.0137	0.0123
			0.4	0.140	0.1382	0.0808	0.0018	0.0065	0.1539	0.1209	-0.0139	0.0148
			0.5	0.165	0.1683	0.0783	-0.0033	0.0061	0.1890	0.1138	-0.0240	0.0135
			0.6	0.190	0.1935	0.0812	-0.0035	0.0066	0.2138	0.1197	-0.0238	0.0149
			0.7	0.215	0.2165	0.0905	-0.0015	0.0082	0.2427	0.1284	-0.0277	0.0173
			0.8	0.240	0.2455	0.0936	-0.0055	0.0088	0.2692	0.1276	-0.0292	0.0171
			0.9	0.265	0.2699	0.1034	-0.0049	0.0107	0.2941	0.1350	-0.0291	0.0191

Table 5.5: Summary statistics including the bias and MSE for $\widehat{\tau^*}$ and $\widehat{\tau}$ based on the 1000 estimates with n=50 sample size

					$\widehat{\tau^*}$				$\widehat{\tau}$			
p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	**Mean**	**SD**	**Bias**	**MSE**	**Mean**	**SD**	**Bias**	**MSE**
0.1	0.1	0.1	0.1	0.169	0.1667	0.0581	0.0023	0.0034	0.1753	0.0732	-0.0063	0.0054
			0.2	0.218	0.2198	0.0571	-0.0018	0.0033	0.2298	0.0730	-0.0118	0.0055
			0.3	0.267	0.2666	0.0573	0.0004	0.0033	0.2791	0.0713	-0.0121	0.0052
			0.4	0.316	0.3161	0.0569	-0.0001	0.0032	0.3300	0.0718	-0.0140	0.0054
			0.5	0.365	0.3656	0.0568	-0.0006	0.0032	0.3826	0.0733	-0.0176	0.0057
			0.6	0.414	0.4131	0.0566	0.0009	0.0032	0.4274	0.0729	-0.0134	0.0055
			0.7	0.463	0.4682	0.0587	-0.0052	0.0035	0.4890	0.0736	-0.0260	0.0061
			0.8	0.512	0.5161	0.0643	-0.0041	0.0042	0.5364	0.0787	-0.0244	0.0068
			0.9	0.561	0.5612	0.0643	-0.0002	0.0041	0.5825	0.0765	-0.0215	0.0063
		0.2	0.1	0.116	0.1140	0.0521	0.0020	0.0027	0.1225	0.0726	-0.0065	0.0053
			0.2	0.152	0.1508	0.0537	0.0012	0.0029	0.1609	0.0749	-0.0089	0.0057
			0.3	0.188	0.1880	0.0558	0.0000	0.0031	0.2001	0.0779	-0.0121	0.0062
			0.4	0.224	0.2254	0.0549	-0.0014	0.0030	0.2402	0.0763	-0.0162	0.0061
			0.5	0.260	0.2631	0.0591	-0.0031	0.0035	0.2841	0.0806	-0.0241	0.0071
			0.6	0.296	0.2986	0.0616	-0.0026	0.0038	0.3194	0.0795	-0.0234	0.0069
			0.7	0.332	0.3331	0.0627	-0.0011	0.0039	0.3568	0.0845	-0.0248	0.0077
			0.8	0.368	0.3701	0.0657	-0.0021	0.0043	0.3960	0.0853	-0.0280	0.0081
			0.9	0.404	0.4083	0.0696	-0.0043	0.0049	0.4352	0.0919	-0.0312	0.0094
		0.3	0.1	0.065	0.0656	0.0480	-0.0006	0.0023	0.0733	0.0751	-0.0083	0.0057
			0.2	0.090	0.0919	0.0502	-0.0019	0.0025	0.1018	0.0768	-0.0118	0.0060
			0.3	0.115	0.1158	0.0493	-0.0008	0.0024	0.1308	0.0768	-0.0158	0.0062
			0.4	0.140	0.1438	0.0519	-0.0038	0.0027	0.1596	0.0748	-0.0196	0.0060
			0.5	0.165	0.1707	0.0573	-0.0057	0.0033	0.1922	0.0858	-0.0272	0.0081
			0.6	0.190	0.1936	0.0601	-0.0036	0.0036	0.2179	0.0855	-0.0279	0.0081
			0.7	0.215	0.2150	0.0605	0.0000	0.0037	0.2371	0.0842	-0.0221	0.0076
			0.8	0.240	0.2421	0.0643	-0.0021	0.0041	0.2661	0.0881	-0.0261	0.0085
			0.9	0.265	0.2654	0.0693	-0.0004	0.0048	0.2924	0.0923	-0.0274	0.0093

Table 5.6: Summary statistics including the bias and MSE for $\widehat{\tau^*}$ and $\widehat{\tau}$ based on the 1000 estimates with n=100 sample size

p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	Var(τ^*)	n	$\widehat{\tau_{11}}$	$\widehat{\tau^*}$	$S^2_{\widehat{\tau^*}}$	**Shapiro-Wilk Test**
										Statistic (p-value)
0.1	0.1	0.1	0.2	0.218	0.00350	50	0.196	0.2154	0.00652	0.9731 (0.0386)
					0.00175	100	0.200	0.2163	0.00320	0.9775 (0.0847)
					0.00117	150	0.198	0.2171	0.00214	0.9265 (0.0000)
					0.00087	200	0.201	0.2174	0.00157	0.9892 (0.6032)
0.1	0.1	0.1	0.5	0.365	0.00476	50	0.503	0.3642	0.00678	0.9864 (0.3998)
					0.00238	100	0.499	0.3629	0.00315	0.9909 (0.7373)
					0.00159	150	0.498	0.3641	0.00214	0.9895 (0.6215)
					0.00119	200	0.500	0.3667	0.00155	0.9891 (0.5932)
0.1	0.1	0.1	0.8	0.512	0.00750	50	0.799	0.5140	0.00804	0.9927 (0.8701)
					0.00375	100	0.801	0.5123	0.00399	0.9877 (0.4884)
					0.00250	150	0.800	0.5131	0.00251	0.9666 (0.0122)
					0.00188	200	0.799	0.5133	0.00183	0.9867 (0.4148)
0.1	0.1	0.2	0.2	0.152	0.00383	50	0.200	0.1530	0.00604	0.9790 (0.1110)
					0.00192	100	0.201	0.1532	0.00279	0.9874 (0.4658)
					0.00128	150	0.200	0.1521	0.00183	0.9822 (0.1958)
					0.00096	200	0.201	0.1530	0.00140	0.9898 (0.6501)
0.1	0.1	0.2	0.5	0.260	0.00563	50	0.497	0.2610	0.00651	0.9873 (0.4575)
					0.00282	100	0.501	0.2596	0.00331	0.9690 (0.0185)
					0.00188	150	0.499	0.2590	0.00216	0.9852 (0.3281)
					0.00141	200	0.501	0.2607	0.00170	0.9919 (0.8159)
0.1	0.1	0.3	0.2	0.090	0.00391	50	0.197	0.0862	0.00505	0.9888 (0.5678)
					0.00196	100	0.200	0.0911	0.00257	0.9899 (0.6552)
					0.00130	150	0.201	0.0909	0.00166	0.9633 (0.0070)
					0.00098	200	0.199	0.0897	0.00123	0.9913 (0.7667)
0.1	0.1	0.3	0.5	0.165	0.00568	50	0.501	0.1677	0.00633	0.9885 (0.5471)
					0.00284	100	0.497	0.1646	0.00318	0.9853 (0.3343)
					0.00189	150	0.500	0.1674	0.00207	0.9927 (0.8693)
					0.00142	200	0.500	0.1656	0.00156	0.9742 (0.0466)
0.1	0.1	0.3	0.8	0.240	0.00835	50	0.797	0.2423	0.00851	0.9613 (0.0050)
					0.00418	100	0.800	0.2408	0.00401	0.9851 (0.3245)
					0.00278	150	0.801	0.2397	0.00280	0.9800 (0.1341)
					0.00209	200	0.800	0.2393	0.00196	0.9786 (0.1036)

Table 5.7: Sample variance from the 2000 $\widehat{\tau^*}$ estimates. The estimates are calculated from 2000 simulations and the Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	Var(τ^*)	n	$\widehat{\tau_{11}}$	$\widehat{\tau^*}$	$\widehat{\text{Var}(\tau^*)}$	**Shapiro-Wilk Test Statistic (p-value)**
0.1	0.1	0.1	0.2	0.218	0.00350	50	0.198	0.2135	0.00333	0.9935 (0.9169)
					0.00175	100	0.201	0.2163	0.00171	0.9675 (0.0143)
					0.00117	150	0.201	0.2178	0.00115	0.9711 (0.0269)
					0.00088	200	0.199	0.2161	0.00086	0.9929 (0.8792)
0.1	0.1	0.1	0.5	0.365	0.00476	50	0.499	0.3674	0.00448	0.9826 (0.2121)
					0.00238	100	0.500	0.3641	0.00232	0.9920 (0.8191)
					0.00159	150	0.499	0.3643	0.00155	0.9794 (0.1192)
					0.00119	200	0.499	0.3630	0.00117	0.9915 (0.7843)
0.1	0.1	0.1	0.8	0.512	0.00750	50	0.802	0.5148	0.00709	0.9645 (0.0085)
					0.00375	100	0.800	0.5108	0.00365	0.9830 (0.2272)
					0.00250	150	0.799	0.5110	0.00245	0.9527 (0.0013)
					0.00188	200	0.800	0.5129	0.00185	0.9825 (0.2061)
0.1	0.1	0.2	0.2	0.152	0.00383	50	0.200	0.1521	0.00365	0.9953 (0.9811)
					0.00192	100	0.200	0.1522	0.00188	0.9951 (0.9782)
					0.00128	150	0.202	0.1527	0.00126	0.9918 (0.8098)
					0.00096	200	0.202	0.1532	0.00095	0.9591 (0.0035)
0.1	0.1	0.2	0.5	0.260	0.00563	50	0.500	0.2609	0.00534	0.9861 (0.3830)
					0.00282	100	0.500	0.2618	0.00274	0.9779 (0.0903)
					0.00188	150	0.499	0.2599	0.00184	0.9938 (0.9313)
					0.00141	200	0.502	0.2606	0.00139	0.9775 (0.0841)
0.1	0.1	0.2	0.8	0.368	0.00867	50	0.801	0.3753	0.00821	0.9661 (0.0112)
					0.00434	100	0.800	0.3693	0.00422	0.9916 (0.7880)
					0.00289	150	0.800	0.3682	0.00284	0.9668 (0.0127)
					0.00217	200	0.800	0.3682	0.00214	0.9813 (0.1684)
0.1	0.1	0.3	0.2	0.090	0.00391	50	0.197	0.0886	0.00376	0.9862 (0.3888)
					0.00196	100	0.196	0.0899	0.00190	0.9847 (0.3023)
					0.00130	150	0.202	0.0912	0.00129	0.9873 (0.4572)
					0.00098	200	0.200	0.0892	0.00097	0.9670 (0.0131)
0.1	0.1	0.3	0.5	0.165	0.00568	50	0.501	0.1665	0.00542	0.9841 (0.2720)
					0.00284	100	0.499	0.1651	0.00276	0.9794 (0.1193)
					0.00189	150	0.500	0.1673	0.00186	0.9841 (0.2750)
					0.00142	200	0.499	0.1655	0.00140	0.9833 (0.2380)
0.1	0.1	0.3	0.8	0.240	0.00835	50	0.799	0.2456	0.00796	0.9673 (0.0137)
					0.00418	100	0.800	0.2442	0.00409	0.9797 (0.1255)
					0.00278	150	0.800	0.2418	0.00274	0.9908 (0.7281)
					0.00209	200	0.800	0.2417	0.00206	0.9946 (0.9634)

Table 5.8: Asymptotic variance of τ^*. The estimates are calculated from 2000 simulations and the Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	Var(τ^*)	n	$\widehat{\tau_{11}}$	$\widehat{\tau^*}$	$S^2_{\widehat{\tau^*}}$	**Shapiro-Wilk Test**
										Statistic (p-value)
0.2	0.1	0.1	0.2	0.292	0.00401	50	0.200	0.291	0.00603	0.9403 (0.0002)
					0.00201	100	0.202	0.290	0.00294	0.9830 (0.2271)
					0.00134	150	0.199	0.292	0.00205	0.9683 (0.0164)
					0.00100	200	0.199	0.291	0.00149	0.9770 (0.0772)
0.2	0.1	0.1	0.5	0.400	0.00512	50	0.501	0.396	0.00630	0.9851 (0.3251)
					0.00256	100	0.500	0.399	0.00305	0.9835 (0.2460)
					0.00171	150	0.502	0.400	0.00204	0.9842 (0.2781)
					0.00128	200	0.499	0.398	0.00154	0.9897 (0.6427)
0.2	0.1	0.1	0.8	0.508	0.00747	50	0.798	0.505	0.00733	0.9670 (0.0132)
					0.00374	100	0.799	0.508	0.00393	0.9682 (0.0160)
					0.00249	150	0.799	0.509	0.00258	0.9876 (0.4776)
					0.00187	200	0.799	0.508	0.00178	0.9832 (0.2340)
0.3	0.1	0.1	0.2	0.330	0.00385	50	0.197	0.325	0.00486	0.9889 (0.5789)
					0.00192	100	0.202	0.328	0.00242	0.9887 (0.5577)
					0.00128	150	0.199	0.328	0.00152	0.9870 (0.4399)
					0.00096	200	0.203	0.329	0.00118	0.9711 (0.0268)
0.3	0.1	0.1	0.5	0.405	0.00514	50	0.501	0.402	0.00561	0.9811 (0.1631)
					0.00257	100	0.500	0.403	0.00297	0.9852 (0.3307)
					0.00171	150	0.499	0.406	0.00183	0.9749 (0.0526)
					0.00128	200	0.501	0.405	0.00145	0.9892 (0.5990)
0.3	0.1	0.1	0.8	0.480	0.00733	50	0.800	0.479	0.00733	0.9805 (0.1446)
					0.00366	100	0.800	0.479	0.00366	0.9758 (0.0619)
					0.00244	150	0.799	0.482	0.00247	0.9755 (0.0589)
					0.00183	200	0.801	0.481	0.00178	0.9872 (0.4525)
0.4	0.1	0.1	0.2	0.332	0.00383	50	0.202	0.327	0.00450	0.9874 (0.4647)
					0.00192	100	0.195	0.329	0.00216	0.9887 (0.5628)
					0.00128	150	0.202	0.330	0.00150	0.9900 (0.6645)
					0.00096	200	0.199	0.329	0.00109	0.9913 (0.7702)
0.4	0.1	0.1	0.5	0.380	0.00525	50	0.502	0.377	0.00597	0.9840 (0.2684)
					0.00262	100	0.499	0.379	0.00274	0.9879 (0.4990)
					0.00175	150	0.500	0.377	0.00194	0.9647 (0.0088)
					0.00131	200	0.499	0.379	0.00148	0.9888 (0.5653)
0.4	0.1	0.1	0.8	0.428	0.00721	50	0.800	0.426	0.00698	0.9850 (0.3186)
					0.00361	100	0.799	0.427	0.00363	0.9921 (0.8283)
					0.00240	150	0.799	0.427	0.00234	0.9829 (0.2202)
					0.00180	200	0.800	0.427	0.00176	0.9918 (0.8049)

Table 5.9: Additional results for the sample variance from the 2000 $\widehat{\tau^*}$ estimates. The estimates are calculated from 2000 simulations and the Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

p_{00}	p_{01}	p_{10}	τ_{11}	τ^*	$\text{Var}(\tau^*)$	n	$\widehat{\tau_{11}}$	$\widehat{\tau^*}$	$\widehat{\text{Var}(\tau^*)}$	**Shapiro-Wilk Test Statistic (p-value)**
0.2	0.1	0.1	0.2	0.292	0.00401	50	0.200	0.2901	0.00387	0.9608 (0.0046)
					0.00201	100	0.202	0.2900	0.00197	0.9902 (0.6818)
					0.00134	150	0.198	0.2910	0.00132	0.9890 (0.5866)
					0.00100	200	0.199	0.2912	0.00099	0.9633 (0.0070)
0.2	0.1	0.1	0.5	0.400	0.00512	50	0.499	0.3989	0.00486	0.9872 (0.4503)
					0.00256	100	0.501	0.3975	0.00251	0.9861 (0.3779)
					0.00171	150	0.500	0.4003	0.00167	0.9932 (0.9010)
					0.00128	200	0.501	0.3997	0.00127	0.9756 (0.0597)
0.2	0.1	0.1	0.8	0.508	0.00747	50	0.799	0.5084	0.00706	0.9878 (0.4965)
					0.00374	100	0.800	0.5055	0.00365	0.9758 (0.0627)
					0.00249	150	0.801	0.5088	0.00244	0.9765 (0.0702)
					0.00187	200	0.799	0.5073	0.00184	0.9923 (0.8426)
0.3	0.1	0.1	0.2	0.330	0.00385	50	0.199	0.3246	0.00379	0.9915 (0.7850)
					0.00192	100	0.201	0.3283	0.00190	0.9802 (0.1368)
					0.00128	150	0.199	0.3288	0.00127	0.9894 (0.6136)
					0.00096	200	0.199	0.3282	0.00096	0.9890 (0.5881)
0.3	0.1	0.1	0.5	0.405	0.00514	50	0.495	0.3994	0.00493	0.9806 (0.1479)
					0.00257	100	0.498	0.4025	0.00251	0.9679 (0.0154)
					0.00171	150	0.500	0.4048	0.00169	0.9851 (0.3251)
					0.00128	200	0.500	0.4042	0.00127	0.9866 (0.4099)
0.3	0.1	0.1	0.8	0.480	0.00733	50	0.799	0.4780	0.00700	0.9690 (0.0186)
					0.00366	100	0.802	0.4818	0.00358	0.9767 (0.0737)
					0.00244	150	0.799	0.4788	0.00240	0.9714 (0.0282)
					0.00183	200	0.800	0.4793	0.00181	0.9664 (0.0118)
0.4	0.1	0.1	0.2	0.332	0.00383	50	0.198	0.3256	0.00379	0.9861 (0.3802)
					0.00192	100	0.203	0.3317	0.00190	0.9910 (0.7479)
					0.00128	150	0.202	0.3296	0.00128	0.9800 (0.1332)
					0.00096	200	0.205	0.3328	0.00096	0.9950 (0.9763)
0.4	0.1	0.1	0.5	0.380	0.00525	50	0.504	0.3760	0.00510	0.9814 (0.1703)
					0.00262	100	0.501	0.3802	0.00257	0.9887 (0.5617)
					0.00175	150	0.502	0.3789	0.00173	0.9796 (0.1239)
					0.00131	200	0.499	0.3794	0.00130	0.9764 (0.0699)
0.4	0.1	0.1	0.8	0.428	0.00721	50	0.799	0.4267	0.00693	0.9879 (0.5015)
					0.00361	100	0.800	0.4247	0.00355	0.9633 (0.0069)
					0.00240	150	0.800	0.4274	0.00237	0.9897 (0.6370)
					0.00180	200	0.800	0.4277	0.00178	0.9791 (0.1137)

Table 5.10: Additional results for the asymptotic variance of τ^*. The estimates are calculated from 2000 simulations. The Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

5.2 Graphical Illustration

As introduced in the previous chapters, chi-plots will be used as a tool, alongside the scatter plot, to further study the behavior of the data with the presence of zero observation. These plots will be presented here first for better understanding and to give a clearer picture of its behavior when the zero observations are introduced. Only the case $(X = 0, Y = 0)$ will be presented for simplicity. Also, this case already gives a good picture of the behavior of the data. Figure 5.1 shows the scatter plots on the top and their corresponding chi-plots on the bottom from data simulated from bivariate lognormal distribution with correlation 0.0 and p_{00} 0%, 30%, 60% and 80%, respectively. The chi-plots also show the 95% control lines. The chi-plot in (b) shows the baseline plot where p_{00} is 0%. Here, the calculated χ_is are still falling within the band 95% of the time. Comparing this with the chi-plots in (d), (f), and (h), it is apparent how the points depart from the control band when p_{00} was increased. This just illustrates how the classical estimate for τ can be misleading whenever a proportion of data clusters into a single value.

Figures 5.2, 5.3 and 5.4 show other levels of correlation.

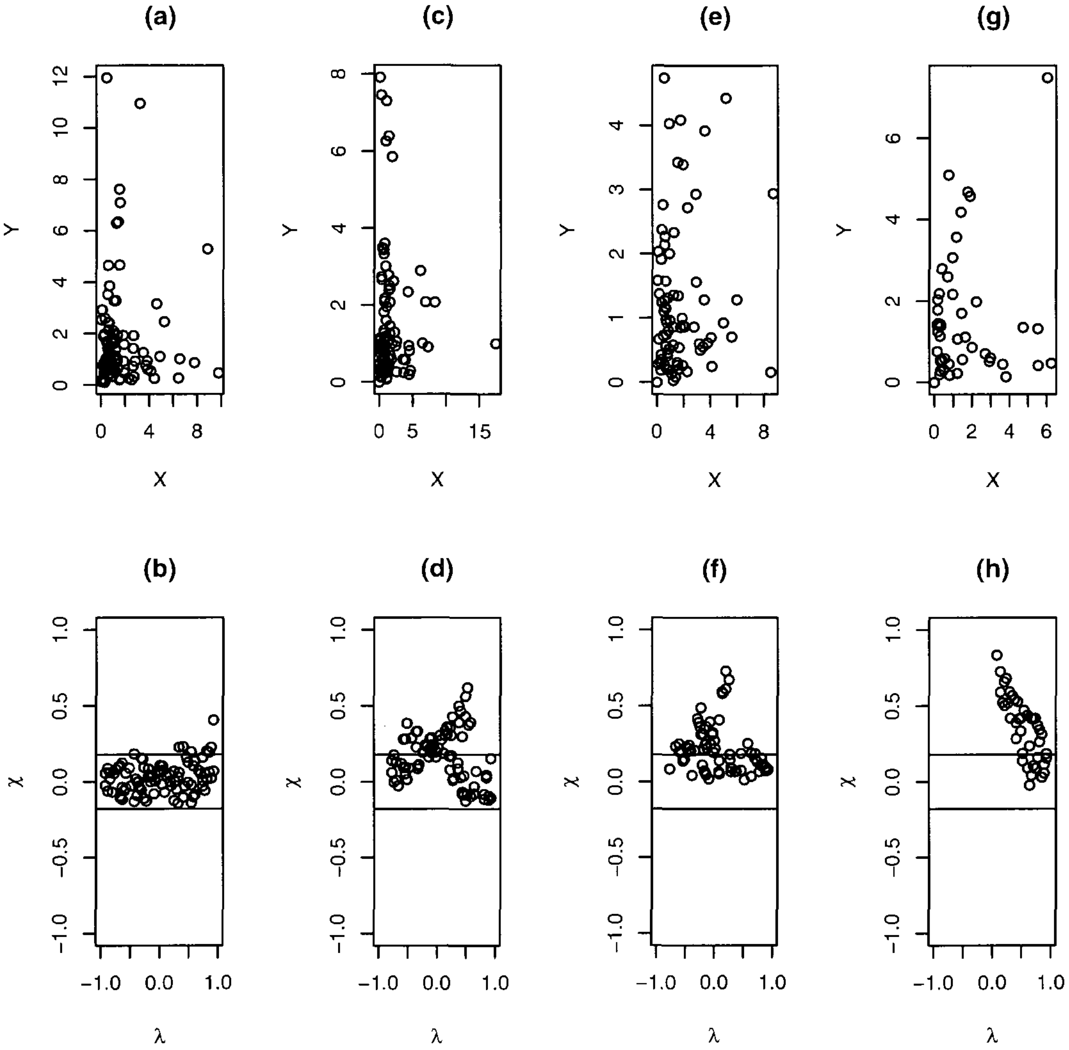

Figure 5.1: Behavior of the chi-plot on varying proportions of zero, p_{00} = 0%, 30%, 60%, 80%; $\rho = 0.0$. The top row shows the scatter plots and the bottom row their corresponding chi-plots, for simulated samples of size 100 from the bivariate lognormal distribution.

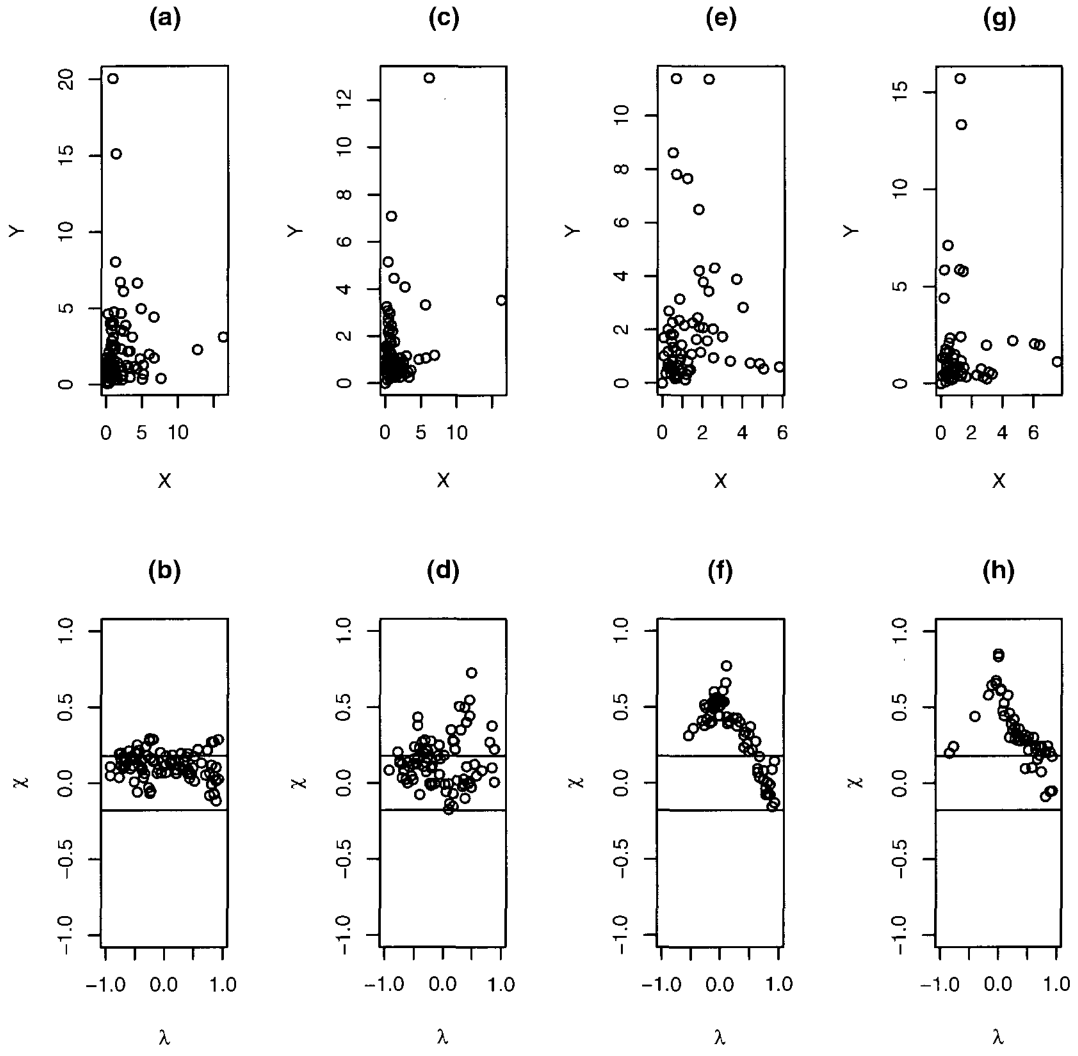

Figure 5.2: Behavior of the chi-plot on varying proportions of zero, p_{00} = 0%, 30%, 60%, 80%; $\rho = 0.20$. The top row shows the scatter plots and the bottom row their corresponding chi-plots, for simulated samples of size 100 from the bivariate lognormal distribution.

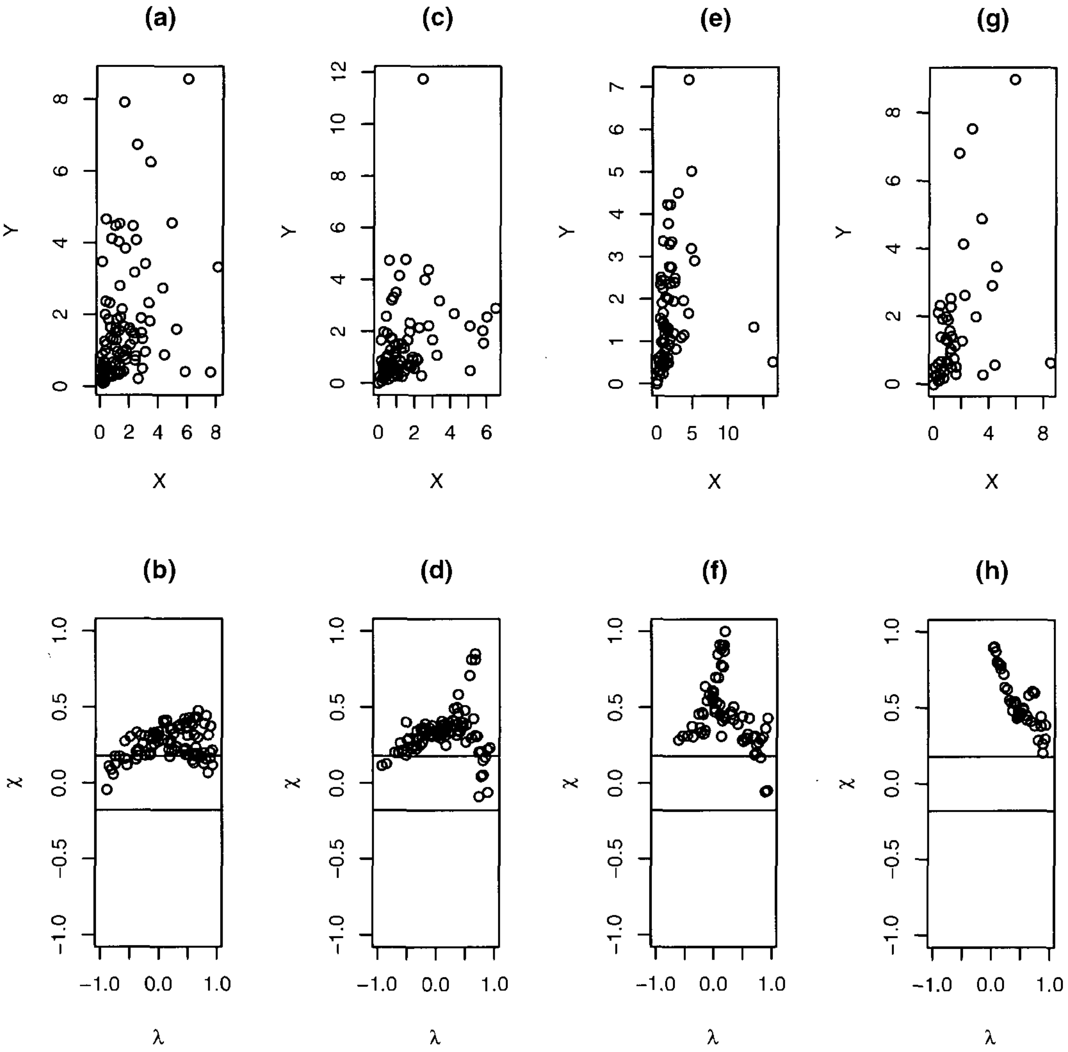

Figure 5.3: Behavior of the chi-plot on varying proportions of zero, $p_{00} = 0\%$, 30%, 60%, 80%; $\rho = 0.50$. The top row shows the scatter plots and the bottom row their corresponding chi-plots, for simulated samples of size 100 from the bivariate lognormal distribution.

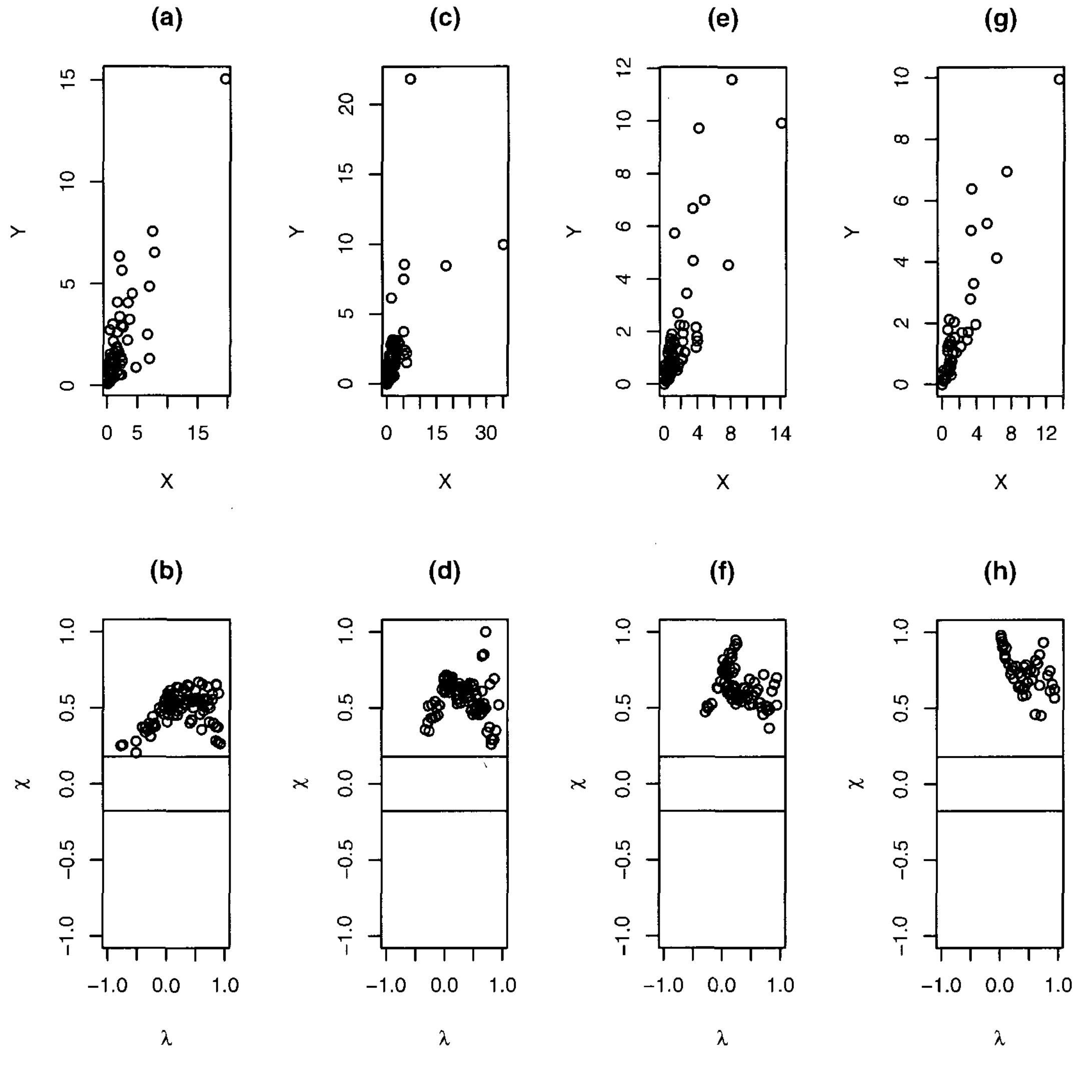

Figure 5.4: Behavior of the chi-plot on varying proportions of zero, $p_{00} = 0\%$, 30%, 60%, 80%; $\rho = 0.80$. The top row shows the scatter plots and the bottom row their corresponding chi-plots, for simulated samples of size 100 from the bivariate lognormal distribution.

5.3 Simulation and Results: Spearman's Rho

5.3.1 Simulation Plan

A similar Monte Carlo simulation procedure used in the previous section was employed to study the proposed estimator for Spearman's ρ defined in Proposition 3. Samples of $n = 30$, 50, 100 pairs of data were simulated from a bivariate lognormal distribution with $\mu_X = 0$ and $\mu_Y = 0$. Using the relationship between Spearman's ρ and Pearson's ρ defined as $\rho_S = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right)$, $\rho_S = 0.1$ to 0.9 by 0.1 was used to get $\rho = 2\sin\left(\frac{\rho_S \pi}{6}\right)$. The proportion of zeroes used were $p_{00} = p_{01} = 0.1$ and $p_{10} = (0.1, 0.2, 0.3)$.

For each case, a multinomial distribution was used to randomly determine the n_{00} pairs of observations that will be $(0, 0)$ with probability p_{00}; n_{01} pairs of $(0, y)$ with probability p_{01}; and n_{10} pair of $(x, 0)$ with probability p_{10}, where $p_{11} = 1 - (p_{00} + p_{01} + p_{10})$. From the nonzero pairs of observations, calculate the Spearman's coefficient of correlation, $\widehat{\rho_{S11}}$. In addition, the following estimates were determined from the simulated data, $\widehat{p_{00}} = n_{00}/n$; $\widehat{p_{10}} = n_{10}/n$; $\widehat{p_{01}} = n_{01}/n$; and $\widehat{p_{11}} = n_{11}/n = 1 - (\widehat{p_{00}} + \widehat{p_{10}} + \widehat{p_{01}})$. Then the estimate of the proposed Spearman's ρ defined in Proposition 3 was calculated as $\widehat{\rho_S^*} = \widehat{p_{11}}\widehat{p_{+1}}\widehat{p_{1+}}\widehat{\rho_{S11}} + 3(\widehat{p_{00}}\widehat{p_{11}} - \widehat{p_{01}}\widehat{p_{10}})$. The estimate of (4.1) was also calculated for comparison. For each case, the process was repeated for 1000 times.

The same plan was utilized to study the asymptotic variance of the estimator proposed in Proposition 3. Sample sizes up to 200 were considered. Although not all combinations of the cell probabilities listed above were used, some additional combinations not considered before were presented. Only a low, mid, and a high level of the population ρ_S were considered and 2000 replicates were performed.

5.3.2 Results

From the 1000 estimates of $\rho_S{}^*$ for each case, the 95% percentile intervals were determined and reported in Table 5.11. For all cases, the intervals based on $\widehat{\rho_S^*}$ contains the population value and it gets narrower as the sample size increases.

A normality test on the ρ_S^* was performed for each case and Table 5.12 shows values of the Shapiro-Wilk test statistic and corresponding p-values. There is no evidence of the lack of normality of the estimator and to show a more consistent result, a sample of 100 estimates was selected and normality test was performed on those estimates. The results are reported in Table 5.13.

Tables 5.14, 5.15, and 5.16 shows the mean and standard deviation of the 1000 estimates for ρ_S^*. The bias of the estimate was also shown and the corresponding estimate of the variance. When compared to the population ρ_S, $\widehat{\rho_S^*}$ tend to be less bias and have smaller MSE than $\widehat{\rho_S}$.

The value of the variance calculated using the known population values approaches 0 as the sample size gets larger. As presented in Tables 5.17 and 5.19 using 2000 replications in calculating the proposed estimator, the variance from these estimates were determined and found to be consistent with the population value (Var$[\rho_S^*]$ vs. $S^2_{\rho_S^*}$) especially for cases with larger proportion of zeroes in the data and higher association between the variables. This is also supplemented by a normality test with the Shapiro-Wilk test statistic and p-value also reported. The asymptotic variance was also calculated from each case and the mean of the 2000 estimates are presented in Tables 5.18 and 5.20. These values are consistent with its corresponding population value.

					Interval for ρ_S^* based on $\widehat{\rho_S^*}$			Interval for ρ_S based on $\widehat{\rho_S}$		
p_{00}	p_{01}	p_{10}	ρ_S	${\rho_S}^*$	n=30	n=50	n=100	n=30	n=50	n=100
0.1	0.1	0.1	0.1	0.2248	(-0.080,0.508)	(-0.012,0.444)	(0.057,0.389)	(-0.137,0.577)	(-0.065,0.484)	(0.016,0.439)
			0.2	0.2696	(-0.043,0.538)	(0.043,0.494)	(0.109,0.412)	(-0.107,0.595)	(-0.012,0.541)	(0.063,0.454)
			0.3	0.3144	(0.022,0.562)	(0.083,0.507)	(0.164,0.454)	(-0.063,0.622)	(0.003,0.573)	(0.111,0.492)
			0.4	0.3592	(0.054,0.603)	(0.135,0.553)	(0.206,0.495)	(-0.016,0.654)	(0.086,0.613)	(0.162,0.528)
			0.5	0.4040	(0.122,0.668)	(0.194,0.592)	(0.250,0.533)	(0.039,0.739)	(0.128,0.638)	(0.198,0.583)
			0.6	0.4488	(0.164,0.675)	(0.253,0.652)	(0.315,0.593)	(0.072,0.748)	(0.170,0.702)	(0.263,0.627)
			0.7	0.4936	(0.209,0.728)	(0.287,0.672)	(0.359,0.621)	(0.103,0.793)	(0.208,0.741)	(0.313,0.657)
			0.8	0.5384	(0.281,0.767)	(0.330,0.722)	(0.394,0.664)	(0.178,0.830)	(0.241,0.775)	(0.356,0.698)
			0.9	0.5832	(0.326,0.802)	(0.387,0.748)	(0.440,0.707)	(0.214,0.867)	(0.302,0.817)	(0.383,0.744)
		0.2	0.1	0.1536	(-0.120,0.407)	(-0.030,0.353)	(0.017,0.313)	(-0.234,0.477)	(-0.111,0.429)	(-0.026,0.348)
			0.2	0.1872	(-0.112,0.448)	(-0.016,0.400)	(0.039,0.333)	(-0.195,0.534)	(-0.082,0.456)	(-0.011,0.369)
			0.3	0.2208	(-0.056,0.482)	(-0.003,0.428)	(0.070,0.367)	(-0.173,0.576)	(-0.078,0.499)	(0.028,0.426)
			0.4	0.2544	(-0.060,0.509)	(0.057,0.447)	(0.107,0.395)	(-0.165,0.581)	(-0.017,0.518)	(0.045,0.457)
			0.5	0.2880	(0.017,0.531)	(0.071,0.484)	(0.133,0.432)	(-0.117,0.607)	(-0.002,0.549)	(0.076,0.480)
			0.6	0.3216	(0.053,0.573)	(0.111,0.521)	(0.182,0.462)	(-0.043,0.655)	(0.040,0.604)	(0.124,0.510)
			0.7	0.3552	(0.088,0.607)	(0.149,0.555)	(0.208,0.497)	(-0.027,0.700)	(0.075,0.621)	(0.146,0.534)
			0.8	0.3888	(0.105,0.643)	(0.169,0.578)	(0.235,0.535)	(-0.019,0.714)	(0.095,0.662)	(0.171,0.602)
			0.9	0.4224	(0.153,0.673)	(0.204,0.623)	(0.271,0.562)	(0.023,0.793)	(0.121,0.704)	(0.212,0.628)
		0.3	0.1	0.0840	(-0.166,0.343)	(-0.120,0.271)	(-0.049,0.220)	(-0.287,0.460)	(-0.185,0.345)	(-0.105,0.285)
			0.2	0.1080	(-0.156,0.358)	(-0.089,0.313)	(-0.022,0.246)	(-0.272,0.475)	(-0.183,0.395)	(-0.088,0.305)
			0.3	0.1320	(-0.119,0.377)	(-0.073,0.331)	(-0.017,0.276)	(-0.252,0.495)	(-0.165,0.422)	(-0.068,0.338)
			0.4	0.1560	(-0.097,0.413)	(-0.046,0.351)	(0.010,0.304)	(-0.230,0.521)	(-0.128,0.429)	(-0.066,0.352)
			0.5	0.1800	(-0.087,0.410)	(-0.015,0.367)	(0.030,0.326)	(-0.189,0.539)	(-0.093,0.438)	(-0.034,0.390)
			0.6	0.2040	(-0.056,0.460)	(-0.011,0.395)	(0.058,0.340)	(-0.171,0.583)	(-0.080,0.512)	(-0.006,0.416)
			0.7	0.2280	(-0.037,0.483)	(0.016,0.436)	(0.095,0.373)	(-0.182,0.592)	(-0.053,0.529)	(0.019,0.427)
			0.8	0.2520	(-0.030,0.515)	(0.034,0.455)	(0.103,0.397)	(-0.138,0.608)	(-0.065,0.539)	(0.056,0.468)
			0.9	0.2760	(0.010,0.548)	(0.047,0.496)	(0.125,0.422)	(-0.123,0.671)	(-0.041,0.590)	(0.064,0.488)

Table 5.11: $(2.5^{th}, 97.6^{th})$ percentile intervals for ρ_S^* based from the 1000 estimates

					Shapiro-Wilk Test Statistic (p-value)		
p_{00}	p_{01}	p_{10}	ρ_S	$\rho_S{}^*$	**n=30**	**n=50**	**n=100**
0.1	0.1	0.1	0.1	0.2248	0.9984(0.4930)	0.9990(0.8831)	0.9986(0.6427)
			0.2	0.2696	0.9967(**0.0375**)	0.9992(0.9460)	0.9983(0.4432)
			0.3	0.3144	0.9966(**0.0288**)	0.9977(0.1757)	0.9980(0.3010)
			0.4	0.3592	0.9937(**0.0003**)	0.9979(0.2440)	0.9988(0.7508)
			0.5	0.4040	0.9985(0.5633)	0.9980(0.2680)	0.9963(**0.0189**)
			0.6	0.4488	0.9971(0.0657)	0.9970(0.0546)	0.9983(0.4557)
			0.7	0.4936	0.9979(0.2603)	0.9976(0.1611)	0.9983(0.4361)
			0.8	0.5384	0.9972(0.0849)	0.9979(0.2444)	0.9984(0.4707)
			0.9	0.5832	0.9963(**0.0167**)	0.9970(0.0536)	0.9984(0.4872)
		0.2	0.1	0.1536	0.9988(0.7722)	0.9967(**0.0335**)	0.9976(0.1618)
			0.2	0.1872	0.9980(0.3004)	0.9979(0.2412)	0.9991(0.9352)
			0.3	0.2208	0.9984(0.5171)	0.9986(0.6037)	0.9989(0.8205)
			0.4	0.2544	0.9974(0.1096)	0.9983(0.4242)	0.9983(0.4373)
			0.5	0.2880	0.9969(**0.0454**)	0.9980(0.2714)	0.9983(0.4110)
			0.6	0.3216	0.9987(0.7201)	0.9986(0.6261)	0.9981(0.3363)
			0.7	0.3552	0.9964(**0.0207**)	0.9989(0.8060)	0.9986(0.6459)
			0.8	0.3888	0.9959(**0.0089**)	0.9975(0.1388)	0.9977(0.1840)
			0.9	0.4224	0.9981(0.3227)	0.9976(0.1578)	0.9970(0.0574)
		0.3	0.1	0.0840	0.9992(0.9591)	0.9975(0.1198)	0.9984(0.5051)
			0.2	0.1080	0.9986(0.6073)	0.9987(0.6925)	0.9987(0.7054)
			0.3	0.1320	0.9991(0.9014)	0.9984(0.4619)	0.9984(0.4760)
			0.4	0.1560	0.9973(0.1006)	0.9976(0.1598)	0.9969(**0.0473**)
			0.5	0.1800	0.9984(0.4895)	0.9976(0.1551)	0.9987(0.6573)
			0.6	0.2040	0.9973(0.0886)	0.9985(0.5816)	0.9984(0.5125)
			0.7	0.2280	0.9983(0.4442)	0.9991(0.9313)	0.9991(0.9095)
			0.8	0.2520	0.9990(0.8650)	0.9986(0.5944)	0.9984(0.4933)
			0.9	0.2760	0.9988(0.7261)	0.9975(0.1346)	0.9979(0.2522)

Table 5.12: Normality test from the 1000 $\widehat{\rho_S^*}$ estimates

					Shapiro-Wilk Test Statistic (p-value)		
p_{00}	p_{01}	p_{10}	ρ_S	$\rho_S{}^*$	**n=30**	**n=50**	**n=100**
0.1	0.1	0.1	0.1	0.2248	0.9844(0.2864)	0.9925(0.8571)	0.9929(0.8820)
			0.2	0.2696	0.9875(0.4699)	0.9914(0.7770)	0.9875(0.4696)
			0.3	0.3144	0.9898(0.6454)	0.9897(0.6428)	0.9861(0.3784)
			0.4	0.3592	0.9761(0.0656)	0.9894(0.6195)	0.9895(0.6268)
			0.5	0.4040	0.9870(0.4391)	0.9930(0.8852)	0.9786(0.1034)
			0.6	0.4488	0.9839(0.2649)	0.9863(0.3916)	0.9924(0.8522)
			0.7	0.4936	0.9914(0.7764)	0.9933(0.9037)	0.9870(0.4365)
			0.8	0.5384	0.9877(0.4842)	0.9894(0.6190)	0.9900(0.6670)
			0.9	0.5832	0.9797(0.1254)	0.9954(0.9846)	0.9900(0.6664)
		0.2	0.1	0.1536	0.9930(0.8881)	0.9806(0.1469)	0.9880(0.5059)
			0.2	0.1872	0.9860(0.3740)	0.9936(0.9233)	0.9674(**0.0141**)
			0.3	0.2208	0.9885(0.5482)	0.9910(0.7434)	0.9900(0.6652)
			0.4	0.2544	0.9920(0.8223)	0.9943(0.9516)	0.9948(0.9705)
			0.5	0.2880	0.9892(0.5984)	0.9939(0.9374)	0.9875(0.4747)
			0.6	0.3216	0.9919(0.8156)	0.9907(0.7216)	0.9819(0.1860)
			0.7	0.3552	0.9789(0.1087)	0.9854(0.3395)	0.9921(0.8302)
			0.8	0.3888	0.9669(**0.0130**)	0.9929(0.8849)	0.9912(0.7604)
			0.9	0.4224	0.9889(0.5794)	0.9890(0.5848)	0.9937(0.9276)
		0.3	0.1	0.0840	0.9905(0.7030)	0.9846(0.2995)	0.9783(0.0982)
			0.2	0.1080	0.9741(**0.0460**)	0.9896(0.6357)	0.9919(0.8127)
			0.3	0.1320	0.9844(0.2850)	0.9947(0.9674)	0.9844(0.2889)
			0.4	0.1560	0.9912(0.7573)	0.9939(0.9364)	0.9927(0.8696)
			0.5	0.1800	0.9919(0.8135)	0.9921(0.8272)	0.9908(0.7288)
			0.6	0.2040	0.9838(0.2588)	0.9940(0.9410)	0.9870(0.4385)
			0.7	0.2280	0.9850(0.3201)	0.9915(0.7851)	0.9770(0.0778)
			0.8	0.2520	0.9896(0.6295)	0.9847(0.3003)	0.9923(0.8426)
			0.9	0.2760	0.9742(**0.0471**)	0.9856(0.3508)	0.9961(0.9943)

Table 5.13: Normality test from the 100 randomly selected $\widehat{\rho_S^*}$ estimates

					$\widehat{\rho_S^*}$				$\widehat{\rho_S}$			
p_{00}	p_{01}	p_{10}	ρ_S	${\rho_S}^*$	**Mean**	**SD**	**Bias**	**MSE**	**Mean**	**SD**	**Bias**	**MSE**
0.1	0.1	0.1	0.1	0.2248	0.2201	0.1489	0.0047	0.0222	0.2243	0.1829	0.0005	0.0335
			0.2	0.2696	0.2551	0.1507	0.0145	0.0229	0.2610	0.1818	0.0086	0.0331
			0.3	0.3144	0.2959	0.1388	0.0185	0.0196	0.3018	0.1766	0.0126	0.0314
			0.4	0.3592	0.3479	0.1386	0.0113	0.0193	0.3536	0.1708	0.0056	0.0292
			0.5	0.4040	0.3942	0.1364	0.0098	0.0187	0.4000	0.1791	0.0040	0.0321
			0.6	0.4488	0.4342	0.1292	0.0146	0.0169	0.4434	0.1728	0.0054	0.0299
			0.7	0.4936	0.4807	0.1278	0.0129	0.0165	0.4832	0.1751	0.0104	0.0308
			0.8	0.5384	0.5279	0.1291	0.0105	0.0168	0.5291	0.1701	0.0093	0.0290
			0.9	0.5832	0.5770	0.1207	0.0062	0.0146	0.5818	0.1665	0.0014	0.0277
		0.2	0.1	0.1536	0.1444	0.1358	0.0092	0.0185	0.1448	0.1833	0.0088	0.0337
			0.2	0.1872	0.1743	0.1394	0.0129	0.0196	0.1805	0.1848	0.0067	0.0342
			0.3	0.2208	0.2108	0.1354	0.0100	0.0184	0.2247	0.1868	-0.0039	0.0349
			0.4	0.2544	0.2449	0.1398	0.0095	0.0196	0.2514	0.1908	0.0030	0.0364
			0.5	0.2880	0.2774	0.1325	0.0106	0.0177	0.2770	0.1878	0.0110	0.0354
			0.6	0.3216	0.3135	0.1312	0.0081	0.0173	0.3222	0.1847	-0.0006	0.0341
			0.7	0.3552	0.3576	0.1290	-0.0024	0.0167	0.3622	0.1839	-0.0070	0.0339
			0.8	0.3888	0.3837	0.1397	0.0051	0.0195	0.3900	0.1874	-0.0012	0.0351
			0.9	0.4224	0.4234	0.1346	-0.0010	0.0181	0.4268	0.1959	-0.0044	0.0384
		0.3	0.1	0.0840	0.0819	0.1270	0.0021	0.0161	0.0892	0.1902	-0.0052	0.0362
			0.2	0.1080	0.1055	0.1252	0.0025	0.0157	0.1166	0.1896	-0.0086	0.0360
			0.3	0.1320	0.1254	0.1244	0.0066	0.0155	0.1283	0.1893	0.0037	0.0358
			0.4	0.1560	0.1516	0.1309	0.0044	0.0172	0.1503	0.1955	0.0057	0.0383
			0.5	0.1800	0.1714	0.1295	0.0086	0.0168	0.1786	0.1877	0.0014	0.0352
			0.6	0.2040	0.1961	0.1385	0.0079	0.0192	0.2098	0.1965	-0.0058	0.0386
			0.7	0.2280	0.2206	0.1309	0.0074	0.0172	0.2283	0.1952	-0.0003	0.0381
			0.8	0.2520	0.2504	0.1355	0.0016	0.0184	0.2586	0.1967	-0.0066	0.0387
			0.9	0.2760	0.2720	0.1387	0.0040	0.0192	0.2776	0.2013	-0.0016	0.0405

Table 5.14: Summary statistics including the bias and MSE for $\widehat{\rho_S^*}$ and $\widehat{\rho_S}$ based on the 1000 estimates with n=30 sample size

					$\widehat{\rho_S^*}$				$\widehat{\rho_S}$			
p_{00}	p_{01}	p_{10}	ρ_S	${\rho_S}^*$	**Mean**	**SD**	**Bias**	**MSE**	**Mean**	**SD**	**Bias**	**MSE**
0.1	0.1	0.1	0.1	0.2248	0.2178	0.1153	0.0070	0.0133	0.2197	0.1429	0.0051	0.0204
			0.2	0.2696	0.2664	0.1138	0.0032	0.0130	0.2667	0.1428	0.0029	0.0204
			0.3	0.3144	0.3060	0.1077	0.0084	0.0117	0.3121	0.1427	0.0023	0.0204
			0.4	0.3592	0.3488	0.1049	0.0104	0.0111	0.3543	0.1352	0.0049	0.0183
			0.5	0.4040	0.3940	0.1032	0.0100	0.0108	0.3973	0.1322	0.0067	0.0175
			0.6	0.4488	0.4513	0.1034	-0.0025	0.0107	0.4553	0.1359	-0.0065	0.0185
			0.7	0.4936	0.4864	0.1006	0.0072	0.0102	0.4892	0.1330	0.0044	0.0177
			0.8	0.5384	0.5294	0.0994	0.0090	0.0100	0.5324	0.1361	0.0060	0.0186
			0.9	0.5832	0.5752	0.0965	0.0080	0.0094	0.5777	0.1299	0.0055	0.0169
		0.2	0.1	0.1536	0.1547	0.1038	-0.0011	0.0108	0.1576	0.1388	-0.0040	0.0193
			0.2	0.1872	0.1868	0.1050	0.0004	0.0110	0.1925	0.1375	-0.0053	0.0189
			0.3	0.2208	0.2201	0.1046	0.0007	0.0109	0.2280	0.1460	-0.0072	0.0214
			0.4	0.2544	0.2492	0.0995	0.0052	0.0099	0.2551	0.1388	-0.0007	0.0193
			0.5	0.2880	0.2869	0.1037	0.0011	0.0107	0.2964	0.1415	-0.0084	0.0201
			0.6	0.3216	0.3142	0.1055	0.0074	0.0112	0.3227	0.1455	-0.0011	0.0212
			0.7	0.3552	0.3469	0.1026	0.0083	0.0106	0.3569	0.1409	-0.0017	0.0198
			0.8	0.3888	0.3799	0.1058	0.0089	0.0113	0.3924	0.1488	-0.0036	0.0222
			0.9	0.4224	0.4213	0.1089	0.0011	0.0119	0.4329	0.1498	-0.0105	0.0226
		0.3	0.1	0.0840	0.0809	0.0997	0.0031	0.0099	0.0839	0.1409	0.0001	0.0199
			0.2	0.1080	0.1069	0.1054	0.0011	0.0111	0.1116	0.1504	-0.0036	0.0226
			0.3	0.1320	0.1303	0.1046	0.0017	0.0109	0.1346	0.1509	-0.0026	0.0228
			0.4	0.1560	0.1485	0.1005	0.0075	0.0102	0.1525	0.1421	0.0035	0.0202
			0.5	0.1800	0.1787	0.0978	0.0013	0.0096	0.1828	0.1440	-0.0028	0.0208
			0.6	0.2040	0.1991	0.1029	0.0049	0.0106	0.2104	0.1499	-0.0064	0.0225
			0.7	0.2280	0.2234	0.1028	0.0046	0.0106	0.2333	0.1502	-0.0053	0.0226
			0.8	0.2520	0.2500	0.1051	0.0020	0.0110	0.2568	0.1575	-0.0048	0.0248
			0.9	0.2760	0.2778	0.1132	-0.0018	0.0128	0.2853	0.1571	-0.0093	0.0248

Table 5.15: Summary statistics including the bias and MSE for $\widehat{\rho_S^*}$ and $\widehat{\rho_S}$ based on the 1000 estimates with n=50 sample size

					$\widehat{\rho_S^*}$				$\widehat{\rho_S}$			
p_{00}	p_{01}	p_{10}	ρ_S	${\rho_S}^*$	**Mean**	**SD**	**Bias**	**MSE**	**Mean**	**SD**	**Bias**	**MSE**
0.1	0.1	0.1	0.1	0.2248	0.2215	0.0812	0.0033	0.0066	0.2248	0.1048	0.0000	0.0110
			0.2	0.2696	0.2650	0.0785	0.0046	0.0062	0.2676	0.0999	0.0020	0.0100
			0.3	0.3144	0.3095	0.0756	0.0049	0.0057	0.3150	0.0959	-0.0006	0.0092
			0.4	0.3592	0.3513	0.0734	0.0079	0.0054	0.3530	0.0926	0.0062	0.0086
			0.5	0.4040	0.3959	0.0753	0.0081	0.0057	0.4006	0.0970	0.0034	0.0094
			0.6	0.4488	0.4478	0.0710	0.0010	0.0050	0.4546	0.0937	-0.0058	0.0088
			0.7	0.4936	0.4910	0.0668	0.0026	0.0045	0.4975	0.0887	-0.0039	0.0079
			0.8	0.5384	0.5322	0.0680	0.0062	0.0047	0.5352	0.0914	0.0032	0.0084
			0.9	0.5832	0.5791	0.0695	0.0041	0.0048	0.5799	0.0926	0.0033	0.0086
		0.2	0.1	0.1536	0.1543	0.0753	-0.0007	0.0057	0.1584	0.0971	-0.0048	0.0094
			0.2	0.1872	0.1846	0.0749	0.0026	0.0056	0.1859	0.0998	0.0013	0.0100
			0.3	0.2208	0.2215	0.0751	-0.0007	0.0056	0.2237	0.1025	-0.0029	0.0105
			0.4	0.2544	0.2548	0.0727	-0.0004	0.0053	0.2581	0.1023	-0.0037	0.0105
			0.5	0.2880	0.2879	0.0742	0.0001	0.0055	0.2938	0.1011	-0.0058	0.0103
			0.6	0.3216	0.3179	0.0709	0.0037	0.0050	0.3253	0.1003	-0.0037	0.0101
			0.7	0.3552	0.3535	0.0731	0.0017	0.0053	0.3577	0.1012	-0.0025	0.0103
			0.8	0.3888	0.3892	0.0762	-0.0004	0.0058	0.3977	0.1096	-0.0089	0.0121
			0.9	0.4224	0.4210	0.0745	0.0014	0.0056	0.4291	0.1044	-0.0067	0.0109
		0.3	0.1	0.0840	0.0872	0.0695	-0.0032	0.0048	0.0919	0.1014	-0.0079	0.0103
			0.2	0.1080	0.1072	0.0680	0.0008	0.0046	0.1104	0.1020	-0.0024	0.0104
			0.3	0.1320	0.1280	0.0718	0.0040	0.0052	0.1333	0.1026	-0.0013	0.0105
			0.4	0.1560	0.1522	0.0707	0.0038	0.0050	0.1560	0.1045	0.0000	0.0109
			0.5	0.1800	0.1782	0.0722	0.0018	0.0052	0.1843	0.1077	-0.0043	0.0116
			0.6	0.2040	0.2026	0.0728	0.0014	0.0053	0.2133	0.1060	-0.0093	0.0113
			0.7	0.2280	0.2302	0.0720	-0.0022	0.0052	0.2357	0.1078	-0.0077	0.0117
			0.8	0.2520	0.2530	0.0736	-0.0010	0.0054	0.2624	0.1091	-0.0104	0.0120
			0.9	0.2760	0.2751	0.0788	0.0009	0.0062	0.2860	0.1114	-0.0100	0.0125

Table 5.16: Summary statistics including the bias and MSE for $\widehat{\rho_S^*}$ and $\widehat{\rho_S}$ based on the 1000 estimates with n=100 sample size

p_{00}	p_{01}	p_{10}	ρ_{S11}	ρ_S^*	Var(ρ_S^*)	n	$\widehat{\rho_{S11}}$	$\widehat{\rho_S^*}$	$S^2_{\widehat{\rho_S^*}}$	**Shapiro-Wilk Test**
										Statistic (p-value)
0.1	0.1	0.1	0.2	0.270	0.00696	50	0.199	0.2685	0.01247	0.9771 (0.0787)
					0.00348	100	0.195	0.2648	0.00616	0.9703 (0.0235)
					0.00232	150	0.198	0.2668	0.00405	0.9862 (0.3891)
					0.00174	200	0.199	0.2689	0.00306	0.9883 (0.5333)
0.1	0.1	0.1	0.5	0.404	0.00676	50	0.487	0.4005	0.01064	0.9782 (0.0964)
					0.00338	100	0.494	0.4005	0.00570	0.9929 (0.8814)
					0.00225	150	0.498	0.4043	0.00356	0.9910 (0.7446)
					0.00169	200	0.496	0.4017	0.00271	0.9896 (0.6356)
0.1	0.1	0.1	0.8	0.538	0.00789	50	0.783	0.5309	0.00961	0.9762 (0.0675)
					0.00395	100	0.791	0.5364	0.00487	0.9908 (0.7298)
					0.00263	150	0.794	0.5380	0.00327	0.9924 (0.8518)
					0.00197	200	0.793	0.5351	0.00253	0.9827 (0.2144)
0.1	0.1	0.2	0.2	0.187	0.00746	50	0.193	0.1805	0.01138	0.9602 (0.0042)
					0.00373	100	0.197	0.1848	0.00535	0.9687 (0.0176)
					0.00249	150	0.197	0.1852	0.00374	0.9832 (0.2342)
					0.00186	200	0.199	0.1857	0.00265	0.9909 (0.7349)
0.1	0.1	0.2	0.5	0.288	0.00829	50	0.484	0.2859	0.01018	0.9901 (0.6687)
					0.00415	100	0.494	0.2824	0.00530	0.9905 (0.7066)
					0.00276	150	0.492	0.2853	0.00358	0.9885 (0.5449)
					0.00207	200	0.496	0.2848	0.00268	0.9836 (0.2512)
0.1	0.1	0.2	0.8	0.389	0.01018	50	0.781	0.3873	0.01093	0.9888 (0.5694)
					0.00509	100	0.793	0.3862	0.00542	0.9898 (0.6520)
					0.00339	150	0.793	0.3865	0.00366	0.9912 (0.7579)
					0.00255	200	0.796	0.3888	0.00286	0.9889 (0.5731)
0.1	0.1	0.3	0.2	0.108	0.00766	50	0.188	0.1068	0.01011	0.9806 (0.1483)
					0.00383	100	0.193	0.1055	0.00501	0.9852 (0.3297)
					0.00255	150	0.197	0.1090	0.00339	0.9820 (0.1895)
					0.00191	200	0.197	0.1066	0.00252	0.9820 (0.1895)
0.1	0.1	0.3	0.5	0.180	0.00886	50	0.483	0.1775	0.01003	0.9940 (0.9411)
					0.00443	100	0.490	0.1781	0.00512	0.9943 (0.9510)
					0.00295	150	0.494	0.1780	0.00358	0.9886 (0.5534)
					0.00221	200	0.493	0.1792	0.00269	0.9827 (0.2160)
0.1	0.1	0.3	0.8	0.252	0.01081	50	0.777	0.2471	0.01134	0.9817 (0.1791)
					0.00540	100	0.789	0.2496	0.00565	0.9659 (0.0109)
					0.00360	150	0.795	0.2527	0.00366	0.9828 (0.2193)
					0.00270	200	0.795	0.2509	0.00279	0.9759 (0.0634)

Table 5.17: Sample variance from the 2000 $\widehat{\rho_S^*}$ estimates. The estimates are calculated from 2000 simulations and the Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

p_{00}	p_{01}	p_{10}	ρ_{S11}	ρ_S^*	Var(ρ_S^*)	n	$\widehat{\rho_{S11}}$	$\widehat{\rho_S^*}$	$\widehat{\text{Var}(\rho_S^*)}$	**Shapiro-Wilk Test Statistic (p-value)**
0.1	0.1	0.1	0.2	0.270	0.00696	50	0.203	0.2680	0.00662	0.9858 (0.3619)
					0.00348	100	0.196	0.2663	0.00339	0.9676 (0.0145)
					0.00232	150	0.196	0.2678	0.00229	0.9683 (0.0165)
					0.00174	200	0.199	0.2693	0.00172	0.9766 (0.0722)
0.1	0.1	0.1	0.5	0.404	0.00676	50	0.485	0.3973	0.00633	0.9921 (0.8311)
					0.00338	100	0.493	0.4011	0.00329	0.9819 (0.1865)
					0.00225	150	0.498	0.4032	0.00222	0.9839 (0.2647)
					0.00169	200	0.498	0.4022	0.00167	0.9920 (0.8182)
0.1	0.1	0.1	0.8	0.538	0.00789	50	0.781	0.5278	0.00739	0.9883 (0.5326)
					0.00395	100	0.790	0.5351	0.00381	0.9853 (0.3343)
					0.00263	150	0.795	0.5377	0.00256	0.9848 (0.3050)
					0.00197	200	0.796	0.5378	0.00194	0.9828 (0.2169)
0.1	0.1	0.2	0.2	0.187	0.00746	50	0.188	0.1822	0.00712	0.9666 (0.0121)
					0.00373	100	0.198	0.1855	0.00362	0.9842 (0.2755)
					0.00249	150	0.197	0.1875	0.00246	0.9844 (0.2884)
					0.00186	200	0.198	0.1860	0.00185	0.9826 (0.2098)
0.1	0.1	0.2	0.5	0.288	0.00829	50	0.489	0.2806	0.00787	0.9871 (0.4460)
					0.00415	100	0.494	0.2875	0.00403	0.9696 (0.0205)
					0.00276	150	0.495	0.2862	0.00271	0.9677 (0.0147)
					0.00207	200	0.497	0.2859	0.00205	0.9921 (0.8268)
0.1	0.1	0.2	0.8	0.389	0.01018	50	0.779	0.3827	0.00953	0.9817 (0.1800)
					0.00509	100	0.791	0.3873	0.00492	0.9902 (0.6838)
					0.00339	150	0.796	0.3860	0.00333	0.9928 (0.8772)
					0.00255	200	0.795	0.3875	0.00250	0.9928 (0.8729)
0.1	0.1	0.3	0.2	0.108	0.00766	50	0.194	0.1065	0.00731	0.9807 (0.1514)
					0.00383	100	0.196	0.1083	0.00375	0.9710 (0.0262)
					0.00255	150	0.201	0.1089	0.00252	0.9923 (0.8411)
					0.00191	200	0.196	0.1078	0.00189	0.9922 (0.8333)
0.1	0.1	0.3	0.5	0.180	0.00886	50	0.482	0.1768	0.00836	0.9796 (0.1234)
					0.00443	100	0.492	0.1769	0.00430	0.9863 (0.3923)
					0.00295	150	0.495	0.1797	0.00290	0.9805 (0.1448)
					0.00221	200	0.495	0.1814	0.00218	0.9889 (0.5744)
0.1	0.1	0.3	0.8	0.252	0.01081	50	0.776	0.2479	0.01013	0.9883 (0.5323)
					0.00540	100	0.788	0.2501	0.00524	0.9919 (0.8103)
					0.00360	150	0.791	0.2508	0.00352	0.9822 (0.1951)
					0.00270	200	0.793	0.2491	0.00266	0.9937 (0.9278)

Table 5.18: Asymptotic variance of ρ_S^*. The estimates are calculated from 2000 simulations and the Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

p_{00}	p_{01}	p_{10}	ρ_{S11}	ρ_S^*	Var(ρ_S^*)	n	$\widehat{\rho_{S11}}$	$\widehat{\rho_S^*}$	$S^2_{\widehat{\rho_S^*}}$	**Shapiro-Wilk Test Statistic (p-value)**
0.2	0.1	0.1	0.2	0.389	0.00813	50	0.193	0.3795	0.01104	0.9448 (0.0004)
					0.00407	100	0.201	0.3834	0.00560	0.9943 (0.9521)
					0.00271	150	0.200	0.3882	0.00360	0.9902 (0.6805)
					0.00203	200	0.197	0.3859	0.00285	0.9924 (0.8526)
0.2	0.1	0.1	0.5	0.477	0.00781	50	0.487	0.4725	0.01044	0.9680 (0.0155)
					0.00391	100	0.493	0.4730	0.00488	0.9875 (0.4731)
					0.00260	150	0.490	0.4723	0.00340	0.9905 (0.7062)
					0.00195	200	0.494	0.4724	0.00240	0.9883 (0.5341)
0.2	0.1	0.1	0.8	0.565	0.00854	50	0.778	0.5580	0.00941	0.9917 (0.7962)
					0.00427	100	0.790	0.5625	0.00471	0.9832 (0.2358)
					0.00285	150	0.792	0.5632	0.00318	0.9949 (0.9727)
					0.00213	200	0.796	0.5644	0.00232	0.9805 (0.1449)
0.3	0.1	0.1	0.2	0.456	0.00767	50	0.187	0.4463	0.00914	0.9873 (0.4598)
					0.00383	100	0.196	0.4539	0.00437	0.9708 (0.0255)
					0.00256	150	0.192	0.4526	0.00298	0.9640 (0.0079)
					0.00192	200	0.201	0.4540	0.00225	0.9910 (0.7467)
0.3	0.1	0.1	0.5	0.510	0.00792	50	0.480	0.5022	0.00838	0.9857 (0.3574)
					0.00396	100	0.489	0.5077	0.00412	0.9934 (0.9127)
					0.00264	150	0.492	0.5060	0.00294	0.9938 (0.9305)
					0.00198	200	0.495	0.5062	0.00216	0.9910 (0.7488)
0.3	0.1	0.1	0.8	0.564	0.00880	50	0.780	0.5578	0.00876	0.9788 (0.1069)
					0.00440	100	0.789	0.5605	0.00472	0.9848 (0.3086)
					0.00293	150	0.793	0.5635	0.00300	0.9819 (0.1870)
					0.00220	200	0.794	0.5620	0.00238	0.9894 (0.6200)
0.4	0.1	0.1	0.2	0.470	0.00750	50	0.196	0.4622	0.00798	0.9851 (0.3220)
					0.00375	100	0.195	0.4641	0.00412	0.9893 (0.6073)
					0.00250	150	0.203	0.4680	0.00260	0.9855 (0.3466)
					0.00188	200	0.195	0.4668	0.00194	0.9918 (0.8056)
0.4	0.1	0.1	0.5	0.500	0.00821	50	0.478	0.4911	0.00825	0.9843 (0.2833)
					0.00410	100	0.486	0.4967	0.00414	0.9827 (0.2153)
					0.00274	150	0.490	0.4958	0.00291	0.9718 (0.0306)
					0.00205	200	0.497	0.4974	0.00218	0.9824 (0.2023)
0.4	0.1	0.1	0.8	0.530	0.00920	50	0.771	0.5208	0.00940	0.9686 (0.0174)
					0.00460	100	0.786	0.5267	0.00489	0.9783 (0.0986)
					0.00307	150	0.790	0.5257	0.00305	0.9808 (0.1546)
					0.00230	200	0.790	0.5273	0.00239	0.9942 (0.9470)

Table 5.19: Additional results for the sample variance from the 2000 $\widehat{\rho_S^*}$ estimates. The estimates are calculated from 2000 simulations and the Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

p_{00}	p_{01}	p_{10}	ρ_{S11}	ρ_S^*	Var(ρ_S^*)	n	$\widehat{\rho_{S11}}$	$\widehat{\rho_S^*}$	$\widehat{\text{Var}(\rho_S^*)}$	**Shapiro-Wilk Test Statistic (p-value)**
0.2	0.1	0.1	0.2	0.389	0.00813	50	0.198	0.3824	0.00785	0.9688 (0.0181)
					0.00407	100	0.198	0.3869	0.00399	0.9442 (0.0003)
					0.00271	150	0.195	0.3846	0.00268	0.9505 (0.0009)
					0.00203	200	0.199	0.3865	0.00202	0.9906 (0.7166)
0.2	0.1	0.1	0.5	0.477	0.00781	50	0.486	0.4689	0.00746	0.9637 (0.0075)
					0.00391	100	0.493	0.4723	0.00382	0.9933 (0.9066)
					0.00260	150	0.493	0.4725	0.00257	0.9791 (0.1141)
					0.00195	200	0.495	0.4743	0.00193	0.9873 (0.4585)
0.2	0.1	0.1	0.8	0.565	0.00854	50	0.781	0.5602	0.00800	0.9871 (0.4424)
					0.00427	100	0.790	0.5595	0.00415	0.9778 (0.0892)
					0.00285	150	0.793	0.5629	0.00279	0.9848 (0.3076)
					0.00213	200	0.796	0.5633	0.00210	0.9889 (0.5774)
0.3	0.1	0.1	0.2	0.456	0.00767	50	0.191	0.4445	0.00756	0.9683 (0.0165)
					0.00383	100	0.194	0.4531	0.00378	0.9866 (0.4100)
					0.00256	150	0.195	0.4532	0.00254	0.9843 (0.2804)
					0.00192	200	0.197	0.4548	0.00190	0.9796 (0.1227)
0.3	0.1	0.1	0.5	0.510	0.00792	50	0.481	0.5028	0.00758	0.9925 (0.8584)
					0.00396	100	0.494	0.5060	0.00389	0.9820 (0.1896)
					0.00264	150	0.489	0.5074	0.00260	0.9892 (0.6034)
					0.00198	200	0.494	0.5093	0.00195	0.9907 (0.7219)
0.3	0.1	0.1	0.8	0.564	0.00880	50	0.776	0.5584	0.00829	0.9937 (0.9244)
					0.00440	100	0.789	0.5611	0.00428	0.9769 (0.0758)
					0.00293	150	0.792	0.5606	0.00288	0.9883 (0.5291)
					0.00220	200	0.795	0.5635	0.00217	0.9820 (0.1884)
0.4	0.1	0.1	0.2	0.470	0.00750	50	0.187	0.4623	0.00734	0.9840 (0.2689)
					0.00375	100	0.202	0.4663	0.00372	0.9826 (0.2117)
					0.00250	150	0.193	0.4672	0.00248	0.9589 (0.0034)
					0.00188	200	0.198	0.4681	0.00187	0.9913 (0.7696)
0.4	0.1	0.1	0.5	0.500	0.00821	50	0.478	0.4940	0.00790	0.9707 (0.0248)
					0.00410	100	0.494	0.4953	0.00404	0.9742 (0.0465)
					0.00274	150	0.489	0.4959	0.00271	0.9899 (0.6558)
					0.00205	200	0.496	0.4965	0.00204	0.9910 (0.7410)
0.4	0.1	0.1	0.8	0.530	0.00920	50	0.774	0.5218	0.00880	0.9797 (0.1253)
					0.00460	100	0.786	0.5278	0.00448	0.9818 (0.1843)
					0.00307	150	0.789	0.5278	0.00301	0.9859 (0.3699)
					0.00230	200	0.792	0.528923	0.00227	0.9885 (0.5440)

Table 5.20: Additional results for the asymptotic variance of ρ_S^*. The estimates are calculated from 2000 simulations and the Shapiro-Wilk statistic was calculated using a random sample of 100 estimates

5.4 An Example

To apply the estimators proposed in Sections 3 and 4 for Kendall's tau and Spearman's rho, respectively, the dataset from Wang (2007) will be used. The data was from a cohort study of HIV-infected men conducted at the Hospital Universitrio Clementino Farga Filho on Rio de Janeiro, Brazil. One of the objectives of the study was to assess the association between plasma and semen viral loads. A summary of the data is shown in Table 5.22.

	All Positive Values		Positive Paired Values	
	Plasma Viral Loads	Semen Viral Loads	Plamsa Viral Loads	Semen Viral Loads
Mean	4.13	4.04	4.14	4.45
SD	0.811	0.772	0.841	0.762
N	21	38	19	19

Table 5.21: Summary of HIV data

Reported data were from 85 men wherein 75% (n=64) of the semen samples and 55% (n=47) of the blood samples have undetectable viral loads (falling below the limit of detection, 2.60). For the purposes of our current study, these will be treated as zero values. The proportion of data for each of the multinomial probabilities are $p_{00} = 0.52941$, $p_{01} = 0.22353$, $p_{10} = 0.02353$, and $p_{11} = 0.22353$. With these values, Kendall's tau and Spearman's rho will be estimated using the proposed estimators with their corresponding asymptotic variances. However, we will first examine the data graphically with the aid of the scatter plot and the corresponding chi-plot in Figure 5.5. The scatter plot from the 85 pairs of observations shows the ”‘L”’ shaped curve on the left corner where the zero values are presented. The continuous pairs, however, are on the upper-right corner of the plot exhibiting some kind of positive linear relationship between plasma viral load and semen viral load. In the corresponding chi-plot, all points are about and outside the 95% band which is indicative of a definite dependence between plasma and semen viral loads. The underlying question now is how to quantify that dependence.

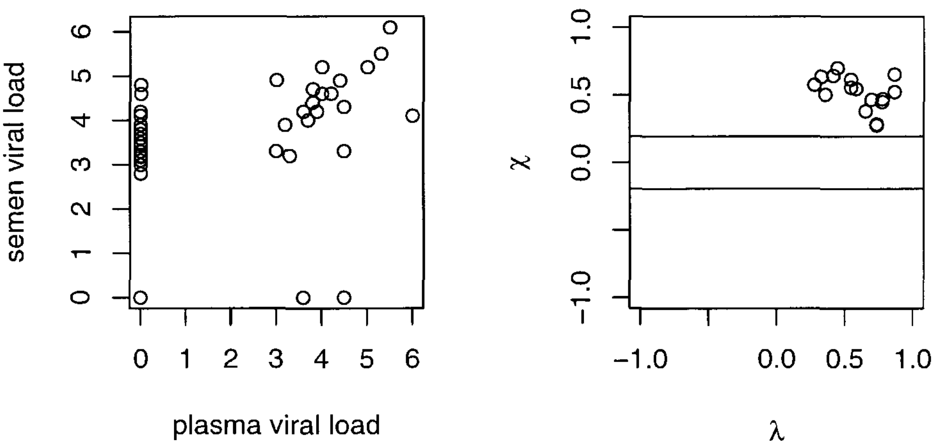

Figure 5.5: Scatter plot and corresponding chi-plot of plasma and semen viral loads from Wang (2007)

Table 5.22 shows the calculated estimates of the population values for Kendall's tau and Spearman's rho. For Kendall's tau, the value of $\widehat{\tau^*}$ is lower than $\widehat{\tau_{11}}$ after considering the proportion of zero in either plasma viral load or semen viral load and in both. The variance of this estimate is 0.002456. In the same manner, the value of $\widehat{\rho_S^*}$ is also lower than $\widehat{\rho_{S11}}$, where only the positive pairs of observations are considered. The corresponding variance of this estimate is 0.004632.

Kendall's tau	Spearman's Rho
$\widehat{\tau_{11}} = 0.3844$	$\widehat{\rho_{S11}} = 0.4619$
$\widehat{\tau^*} = 0.2454$	$\widehat{\rho_S^*} = 0.3506$
$S^2_{\widehat{\tau^*}} = 0.002456$	$S^2_{\widehat{\rho_S^*}} = 0.004632$

Table 5.22: Calculated value of the estimators and the corresponding variances of the HIV data

Chapter 6

FINAL COMMENTS AND FUTURE RESEARCH

The estimation of dependence measures is an important problem in many fields of research. Although there have been several adjustments proposed because of violation of continuity assumption, none of them really focused on having a probability mass at zero. In this research, we introduced the problem of having zero-inflated data. With the presence of a probability mass at zero in a bivariate model, we proposed an adjusted Kendall's tau and Spearman's rho estimators. These were compared to their counterparts and it was shown that the intervals are narrower for the proposed estimators and are less bias than their counterparts. Their corresponding asymptotic variances were also determined and were found to be consistent with the population value. A real data from Wang (2007) was used to illustrate and apply the proposed estimators.

As been discussed in the previous chapters of this research, the widely used and accepted classical dependence measures might not be appropriate for cases when the underlying distribution of the data being analyzed is zero-inflated. Considering only the nonzero pairs of observations usually leads to misleading results.

A next step for the researcher is to define a procedure for a confidence interval estimation. It will also be of interest to further look at the asymptotic variance by considering the Var(τ_{11}). Also, an α level test of association using the proposed estimators and their asymptotic variances. The power of the test will also be determined. This research can be further extended to left truncated data. Also, other measures of association such as the Gini's index can be studied with zero-inflated data. It is also of interest for the researcher to look into the small sample application and further apply the concepts proposed in this research and other methods for handling zero-inflated data in the pre-clinical field. There are several areas in this field where zero-inflated data can be observed either by recording real zeroes or values falling below a limit of detection. Some of these areas are immunotoxicology and developmental and reproductive toxicology. Pharmacokinetic data is also a good example where zero-inflation can occur. Drug concentration in the blood or metabolites is usually not detectable, thus reported as falling below the limit of quantification (BLOQ).

Bibliography

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.

Aitchison, J. (1955). On The Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin, *Journal of the American Statistical Association*, **50**, 901-908.

Bascoul-Mollevi, C., Gourgou-Bourgade, S., and Kramar, A. (2005). Two-part statistics with paired data, *Statistics in Medicine*, **24**, 1435-1448.

Cliff, N. and Charlin, V. (1991). Variances and Covariances of Kendall's Tau and Their Estimation, *Multivariate Behavioral Research*, **26**, 4, 693-707.

Denuit, M. and Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data, *Journal of Multivariate Analysis*, **93**, 40-57.

Fisher, N.I. and Switzer, P. (1985). Chi-plots for Assessing Dependence, *Biometrika*, **72**, 253-265.

Fisher, N.I. and Switzer, P. (2001). Graphical Assessment of Dependence: Is Picture Worth 100 Tests?, *The American Statistician*, **55**, 3, 233-239.

Fligner, M.A. and Rust, S.W. (1983). On the independence problem and Kendall's tau, *Communications in Statistics - Theory and Methods*, **12**, 14, 1597 - 1607.

Herath, H.S.B. and Kumar, P. (2007). New Research Directions in Engineering Economics - Modeling Dependencies with Copulas, *The Engineering Economist*, **52**, 305-331.

Hollander, M. and Wolfe, D.A. (1999). *Nonparametric Statistical Methods.* 2nd ed., Wiley, New York.

Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika*, **30**, 81-93.

Kendall, M.G., Kendall, S.F.H., and Smith, B.B. (1939). The Distribution of Spearman's Coefficient of Rank Correlation in a Universe in which all Ranking Occur an Equal Number of Times. *Biometrika*, **30**, 251-273.

Kendall, M.G. (1942). Partial Rank Correlation. *Biometrika*, **32**, 277-283.

Kendall, M.G. (1945). The Treatment of Ties in Ranking Problems. *Biometrika*, **3**, 239-251.

Kendall, M.G. and Gibbons, J.D. (1948). *Rank Correlation Methods*, Oxford University Press, New York.

Lachenbruch, P.A. (2001). Comparisons of two-part models with competitors, *Statistics in Medicine*, **20**, 1215-1234.

Liebetrau, A. M. (1983). *Measures of Association*, Sage University Paper Series on Quantitative Application in the Social Sciences, 07-032. Berverly Hills and London: Sage Pubns.

Mesfioui, M. and Tajar, A. (2005). On the properties of some nonparametric concordance measures in the discrete case, *Nonparametric Statistics*, **17**, 5, 541-554.

Moulton, L.H. and Halsey, N.A. (1995). A Mixture Model With Detection Limits for Regression Analyses of Antibody Response to Vaccine, *Biometrics*, **51**, 1570-1578.

Nelsen, R.B. (1999). *An Introduction to Copulas*, Springer, New York.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables, *Journal of Multivariate Analysis*, **98**, 544 567.

Noether, G.E. (1967) *Elements of Nonparametric Statistics*, John Wiley, New York.

Pennington, M. (1983) Efficient Estimators of Abundance, for Fish and Plankton Surveys, *Biometrics*, **39**, 281-286.

Owen, W.J. and DeRouen, T.A. (1980). Estimation of the Mean for Lognormal Data Containing Zeroes and Left-Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants, *Biometrics*, **36**, 707-719.

Samara, B. and Randles, R.H. (1988). A Test for Correlation Based on Kendall's Tau, *Communications in Statistics - Theory and Methods*, **17**, 9, 3191-3205.

Sklar, A. (1959). *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statist. Univ. Paris 8, 229-231.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, **15**, 72-101.

Stokes, M.E., Davis, C.S., and Koch, G.G. (1995). *Categorical Data Analysis Using the SAS System*, SAS Institute Inc., Cary, NC.

Taylor, D.J., Kupper, L.L., Rappaport, S.M. and Lyles, R.H. (2001). A mixture model for occupational exposure mean testing with a limit of detection, *Biometrics*, **57**, 3, 681-688.

Wang, A. (2007). The Analysis of Bivariate Truncated Data Using the Clayton Copula Model, *The International Journal of Biostatistics*, **3**, 1.

Zhou, X. and Tu, W. (1999). Comparison of Several Independent Population Means When Their Samples Contain Log-Normal and Possibly Zero Observations, *Biometrics*, **55**, 645-651.