2008

CONFIDENCE INTERVAL ESTIMATION FOR A DIFFERENCE BETWEEN CORRELATED INTRACLASS CORRELATION COEFFICIENTS

Chinthanie Ramasundarahettige,
Western University

CONFIDENCE INTERVAL ESTIMATION FOR A DIFFERENCE BETWEEN CORRELATED INTRACLASS CORRELATION COEFFICIENTS

(Spine title: Interval estimation for intraclass correlation)

(Thesis format: Monograph)

by

Chinthanie Ramasundarahettige, M.Sc

Graduate Program in Epidemiology & Biostatistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science

School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario
July 2008

THE UNIVERSITY OF WESTERN ONTARIO

SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisor

Dr. Guangyong Zou

Co-Supervisor

Dr. Allan Donner

Examiners

Dr. John Koval

Dr. Duncan Murdoch

Dr. Yves Bureau

The thesis by

Chinthanie Ramasundarahettige

entitled:

Confidence interval estimation for a difference between correlated intraclass correlation coefficients

is accepted in partial fulfillment of the
requirements for the degree of
Master of Science

Date: _____

Chair of the Thesis Examination Board

Abstract

The intraclass correlation coefficient (ICC), an index of similarity, plays an important role in a wide range of disciplines, for example in the assessment of instrument reliability. In this case, the study design may involve recruiting a sample of subjects each of whom are assessed several times with a new device and the standard. The ICC estimates for the two devices may then be compared using a test of hypothesis. However it is well known that conclusions drawn from hypothesis testing are confounded by sample size, i.e., a significant p-value can result from a sufficiently large sample size. In such cases, a confidence interval for a difference between two ICCs is more informative since it combines point estimation and hypothesis testing into a single inference statement.

The sampling distribution for the ICC is well known to be left-skewed and thus confidence limits are usually constructed using Fisher's Z-transformation or the F-distribution. Unfortunately, such an approach is not applicable to a difference between two ICCs. The remaining alternative is to apply a simple asymptotic approach, i.e., point estimate plus/minus normal quantile multiplied by the estimate of standard error. However this method is known to perform poorly because it ignores the features of the underlying sampling distribution. In this thesis I develop a confidence interval procedure using the method of variance estimate recovery (MOVER). Specifically, the variance estimates required for the upper and lower limits of a difference are

recovered from those obtained for separate ICCs. An advantage of this approach is that it provides a confidence interval that reflects the underlying sampling distribution. Simulation results show that the MOVER method performs very well in terms of overall coverage percentage and tail errors. Two data sets are used to illustrate this procedure.

Key Words: Intraclass correlation, confidence interval, reliability

ACKNOWLEDGMENTS

I am indebted to many people who have contributed to my success during my course of study in the University of Western Ontario. I gratefully recognize my supervisors Dr. Guangyong Zou and Dr. Allan Donner for their guidance, kind support and encouragement throughout my research study. Their direction and assistance made it possible for me to finish this thesis in a reasonable time.

I am grateful to the Department of Epidemiology and Biostatistics at the University of Western Ontario, for providing financial support in terms of graduate scholarships throughout my studies. The research assistantship from my supervisor is also appreciated.

My very personal thanks go to my beloved husband, Ajith and my two sons, Aloka and Arjun without whose encouragement and patience, this work would have not been a success. I express my gratitude to my parents and entire family for their love and support during this course of study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

This thesis concerns setting approximate confidence intervals for a difference between two correlated intraclass correlation coefficients in the context of reliability studies. Comparison of two ICCs has been a major focus in many measurement reliability studies in education, psychology and in biomedical research. The problem could arise in such situations as when a new instrument is being developed to replace an existing instrument or when two competing measurements are being evaluated for reliability. For example, reliability of a new automatic blood pressure monitor compared with the reliability of an existing sphygmomanometer which uses the ausculatatory method with the purpose of replacing the existing device or using it interchangeably.

In this introductory chapter, I start with providing some background in section 1.1 which follows by inference procedures for single intraclass correlation coefficient in section 1.2. Section 1.3 and 1.4 are allocated to present motivation and the objective of the thesis. The chapter finishes with the structure of this thesis.

1.1 Reliability and intraclass correlation coefficient

In health science and in biomedical research, it is very rare to obtain perfectly reliable measurements, rather data are often measured with error. Whether it is taking blood pressure from a patient or assessing the results of diagnostic procedure or obtaining the effects of a therapy, a repetition of the measurement under the same condition may not yield identical values. This phenomenon is known as measurement error with its extent usually quantified by a reliability coefficient.

A basic requirement of a measurement is its reliability. Reliability refers to the consistency of measurements or reproducibility of the same values when the procedure is applied to the same subject under the same condition repeatedly. It can also be interpreted as the degree to which the measurement is influenced by the measurement errors. Low reliability of the measurements can have severe consequences on the validity of the research results (Lachin, 2004). Data with large amount of measurement error will fail to reflect the criterion of interest and therefore reliability is a prerequisite for validity.

Low reliability of the measurements can seriously affect the statistical analysis and subsequent interpretation. In correlation analysis, unreliability of measurements attenuates the correlation between two variables and hence reduces the power to detect the relationship between the two variables. In regression analysis, unreliable measurements shrink the estimate of the slope towards zero. In randomization trials,

responses measured with unreliable devices reduce the overall power of the test by increasing the variance of the outcome measure.

According to classical measurement theory an observed score Y_i for subject i ($i = 1, 2 \ldots, n$) is the sum of two components, true score μ_i and random measurement error e_i such that (Shrout, 1998),

$$Y_i = \mu_i + e_i.$$

It is assumed that e_i is independent of μ_i and normally distributed with mean 0 and variance σ_e^2. The true score μ_i for a subject i is fixed but varies among subjects in a population of subjects with mean μ and variance σ_μ^2. In classical reliability theory it is customary to assume that the variance of the error component σ_e^2 is the same for different subjects. Under these assumptions, the observed variability of a randomly chosen score Y_i is,

$$\sigma_Y^2 = \sigma_\mu^2 + \sigma_e^2.$$

The reliability coefficient ρ is defined as the ratio of the true score variance to the total score variance of the observed measure Y, i.e.,

$$\begin{aligned} \rho &= \frac{\sigma_\mu^2}{\sigma_Y^2} \\ &= \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_e^2} \end{aligned} \tag{1.1}$$

In other words the reliability coefficient can be interpreted as the proportion of the total variance σ_Y^2 that can be attributed to the variation among the values of the true score σ_μ^2.

The reliability coefficient may be estimated by a one-way random effects ANOVA model. Specifically, let Y_{ij} denotes the jth observation on the ith subject where $i = 1, 2 \ldots, n; j = 1, 2 \ldots, k$. Then the one-way random effects model may be written as

$$Y_{ij} = \mu + a_i + e_{ij}, \tag{1.2}$$

where μ is the grand mean of all the observations in the population, the subject effect $\{a_i\}$ are normally distributed with mean 0 and variance σ_a^2, the measurement error $\{e_{ij}\}$ are normally distributed with mean 0 and variance σ_e^2, and $\{a_i\}, \{e_{ij}\}$ are completely independent. Here n refers to the number of subjects and k refers to the number of measurements per subject. Throughout out this thesis I use this notation.

From the model (1.2) it is easy to show that the correlation (ρ) between any two observations Y_{ij} and $Y_{ij'}$ for $j \neq j'$ is given by

$$\begin{aligned} \rho = \text{corr}(Y_{ij}, Y_{ij'}) &= \frac{\text{cov}(Y_{ij}, Y_{ij'})}{\sqrt{\text{var}(Y_{ij})}\sqrt{\text{var}(Y_{ij'})}} \\ &= \frac{\text{E}[(Y_{ij} - \text{E}(Y_{ij}))(Y_{ij'} - \text{E}(Y_{ij'}))]}{\sqrt{\sigma_a^2 + \sigma_e^2}\sqrt{\sigma_a^2 + \sigma_e^2}} \end{aligned}$$

$$= \frac{\mathrm{E}[(a_i + e_{ij})(a_i + e_{ij'})]}{\sigma_a^2 + \sigma_e^2}$$
$$= \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.$$

Thus, if one refers to a subject as a class, the reliability coefficient ρ is identical to the well known ICC (Fisher, 1925, p.76-210). In what follows, I will use ICC and reliability coefficient interchangeably.

Unlike the estimate introduced in the next chapter, the coefficient ρ is necessarily positive with values range from 0 to 1. In reliability context ICC will be equal to one if and only if $\sigma_e^2 = 0$, i.e., if and only if there is no measurement error. High ρ would suggest high reliability and low ρ may be due to either large measurement errors or small variability between subjects. In practise, guidelines given by Landis and Koch (1977) are usually adopted for interpretation, specifically, slight (0.00 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80) and almost perfect (0.81 to 1.00).

1.2 Statistical inference for a single intraclass correlation coefficient

Although statistical inference in general consists of significance testing and estimation, I will take the view expressed by McGraw and Wong (1996) that the tests of hypothesis $\rho = 0$ are not particularly informative, and thus focus on confidence interval estimation in this thesis.

In the context of reliability studies, the number of repeated measurements k made

on each subject i or class size is usually fixed. Here the repeated measurement refers to two or more measurements on the same subject taken in identical conditions. As a consequence of fixed class size, point estimators of different approaches do not differ materially. There are at least three different approaches to estimate ρ, namely, pairwise estimation, maximum likelihood estimation (MLE) and ANOVA estimation. Specifically, the MLE is the same as pairwise estimator which with reasonably large number of subjects is virtually indistinguishable from that of a ANOVA estimator (Donner and Koval, 1980). For this reason I will not further discuss point estimation of ρ in this thesis but refer to Donner (1986) for a review.

There are at least two approaches to confidence interval estimation for ρ. The most well known approach is to apply Fisher's Z-transformation, proposed because the fact that the sampling distribution for $\hat{\rho}$ is skewed, especially when ρ is far away from 0 (Fisher, 1925, p.214-223). Under this method a confidence interval for ρ is obtained by back transforming the confidence interval for Z. The other approach, commonly known as the 'exact method' is based on the F distribution (Donner, 1986).

A common feature of these confidence limits is that they are asymmetric to the point estimator, reflecting the shape of the underlying sampling distribution. In this thesis I will exploit the asymmetry feature of these limits and take the approach discussed by Zou and Donner (2008) to develop a method for constructing confidence intervals for a difference between two correlated intraclass correlation coefficients.

In reliability studies when the performances of two instruments need to be com-

pared, a common design is to take replicated measurements using both instruments from a single sample of subjects. The resulting reliability coefficients are inherently considered to be correlated. Therefore, the method by Zou and Donner (2008) is extended to incorporate the dependency between two estimated ICCs.

1.3 Motivating examples

Before getting in to the details of the statistical procedure, I now present two examples that have motivated this project.

1.3.1 Example-1

The first example is based on the data from a study conducted by Turner *et al.* (1986) and further analyzed by Dunn (Dunn, 1989, ch-5). These data were derived from measurements taken from 50 patients using two devices; an automated (PIX) and hand held (PLAN) device. It was noted without performing a hypothesis test $H_0 : \rho_1 = \rho_2$, that the automated device is more reliable than the hand-held device where ρ_1 and ρ_2 are ICCs for automated device and hand-held device respectively. Although one may apply a hypothesis testing using methods such as Donner and Zou (2002) or Alsawalmeh and Feldt (1994) to compare two correlated ICCs, a confidence interval approach may be more informative.

1.3.2 Example-2

Quantitative ultrasound scanners are used in the diagnosis of osteoporosis and several devices are already in use. Giraudeau *et al.* (2003) reported a study in which a comparison of two intraclass correlation coefficients was needed to study the reproducibility of a newly developed ultrasound bone matrix densitometer, the BEAM scanner, in comparison with a currently available device, the UBIS 3000 scanner. In their study 5 repeated measurements of the right heel of 34 subjects were measured using both devices in three different sessions. Giraudeau *et al.* (2003) suggested that a comparison of two intraclass correlation coefficients would be of interest. Gomez *et al.* (2002) provided more details of the study.

As the sample of 34 subjects was used to draw measurements by both devices, the two ICC estimates are correlated. Again, one may conduct a hypothesis test for the equality of two correlated ICCs, $H_0 : \rho_1 = \rho_2$. But reporting the confidence interval for the difference between $\rho_1 - \rho_2$ would be more meaningful.

1.4 The objective of the thesis

From the previous sections, it is clear that researchers often face the problem of comparing two correlated ICCs. Specifically, comparing two correlated ICCs when the same sample of subjects are chosen to draw measurements by the two devices in question. Although procedures for the hypothesis testing of the equality of two correlated ICCs have been developed (Donner and Zou, 2002; Alsawalmeh and Feldt,

1994) little work has been done in developing confidence intervals for a difference between two correlated ICCs.

The objective of this thesis is to develop a method for constructing a confidence interval for a difference between two correlated ICCs taking in to consideration the asymmetric property of ICCs. Specifically, the variance estimates needed for the lower and upper limits of a difference of two ICCs are recovered from those of separate ICCs. This is an extension of the study by Zou (2007) on the constructing confidence intervals to compare correlations and the study by Zou and Donner (2008) on the general approach of constructing confidence limits about effect measures. The advantage of this method is that it reflects the underlying sampling distribution. The central idea is to recover the variance estimates from readily available confidence limits for single ICCs. I thus refer to the method as MOVER which stands for method of variance estimates recovery. As the method will be developed based on large sample theory, I will use Monte Carlo simulations to assess the performances. Specified criteria for the evaluation include overall coverage, tail errors and width of the confidence interval.

1.5 Organization of the thesis

This thesis is comprised of six chapters. Chapter 2 presents the literature review and Chapter 3 describes the development of the confidence interval procedure for a difference between two correlated ICCs. Evaluation criteria and the simulation results are presented in Chapter 4. In Chapter 5, I illustrate the methodology by working

out the motivating examples. A discussion which includes final conclusions and few ideas for future research are presented in Chapter 6.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Intraclass correlation coefficient (ICC) as a measure of reliability index has evolved over several decades. Although it was initially applied in social, educational and psychological disciplines, it has now been used in a wide range of areas including biomedical research (Bartko, 1966; Shrout and Fleiss, 1979; Donner, 1986; Shoukri *et al.*, 2004). Thus, an extensive amount of literature on issues regarding this index has been published and in this review, I will only provide a summary on statistical inference for ICC, largely in the context of medical research.

In section 2.2, I present the role of reliability in medical research and in section 2.3 present the historical view of the intraclass correlation coefficient. In section 2.4, interval estimation of ICC is presented and section 2.5 is allocated for describing the sample size estimation of ICC.

2.2 *Role of reliability in medical research*

In medical research, most of the outcomes are measured with random errors. An outcome measure may be a clinical characteristic based on subjective or objective assessment. It may be a quantitative measure or a qualitative or categorical measure used for patient diagnosis or may be employed to assess the eligibility for entry for a clinical trial. For example, assessment of radiographs or biopsy readings or scores on questionnaires collecting information. The essential requirement for all outcome measures is that they should be reproducible or reliable.

Unreliable outcome measures can lead to a wrong patient diagnosis. In clinical trials when unreliable measurements are used as an assessment criteria for eligibility or exclusion, subjects can be assigned into the wrong arms of the trial. Furthermore, unreliable measurements decrease the reliability and as a result the increase in sample size is required to maintain the desired level of power. Many authors (e.g., Shrout and Fleiss 1979, Lachin 2004) have shown that there is a direct relation between the reliability of measurement and the sample size required to provide a desired level of power to detect a given effect size. As such, for any experiment assessment of reliability is extremely important and many texts on the design and analysis of experiments provide descriptions of assessment of reliability (Haggard, 1958; Dunn, 1989).

Satisfactory reliability is a fundamental requirement in instrument development. During this process the reliability of a new instrument is compared with the reliability

of the instrument already in use (Giraudeau *et al.*, 2005). This will enable the investigator to assess the performance of the new instrument. In some other situations reliability is used to evaluate the laboratory assays (He *et al.*, 2006; Pellis *et al.*, 2003).

2.3 Some historical aspects of intraclass correlation coefficient

The concept of intraclass correlation coefficient has evolved over a century. Although it was not explicitly stated in a formula, ICC was first used in 1896 to estimate the correlation among siblings (Pearson, 1896). In that, ICC was estimated by calculating Pearson product moment correlation over all possible pairs of observation within a class and later this estimator of ICC was termed pairwise estimator of ICC (Donner and Koval, 1980). However, the most known method of estimating ICC was introduced by Fisher in 1925. Estimated ICC under Fisher's method is known as the analysis of variance (ANOVA) estimator of ICC as it is estimated based on the variance components of a random effects model (Fisher, 1925, p.188). By introducing the ANOVA estimator of ICC, Fisher broadened the scope of application of the ICC, which led to its applications in reliability theory, cluster randomization trials and in sensitivity analysis.

The use of ICC in reliability studies was summarized by Haggard as early as in the 1950s (Haggard, 1958, ch.6). Bartko (1966) had further illustrated that estimator of ICC obtained from a one-way random effects ANOVA is indeed equal to the correlation coefficient between any two measurements on the same subject.

An extensive amount of literature on inference procedures of ICC were developed during the last three decades. Shrout and Fleiss (1979) described the ICC in context of reliability studies and discussed six forms of ICC which can be used to assess rater reliability. Although it was not in reliability context, McGraw and Wong (1996) expanded the discussion of Shrout and Fleiss (1979) by including another two forms of ICC. Others have also shown that ICC can be used as an inter-rater and intra-rater reliability coefficient (Eliasziw *et al.*, 1994; Rousson *et al.*, 2002; Kottner and Dassen, 2007) and quantifying the test-retest reliability or reproducibility studies (Nickerson, 1997; Giraudeau *et al.*, 2003; Schuck, 2004).

2.4 Interval estimation of intraclass correlation coefficient

There are several approaches for constructing confidence intervals for ρ. Methods based on Fisher's Z-transformation and F-distribution are the most common approaches. Constructing confidence intervals using the large sample variance of ρ are also familiar to researchers.

Recognizing the skewness of the sampling distribution of ρ, Fisher (1925, ch.7) introduced the variance stabilizing transformation known as Fisher's Z-transformation. Fisher's Z-transformation is used to transform the already skewed distribution of $\hat{\rho}$ to an approximately normal distribution. The confidence intervals for ρ is then obtained by back transforming the confidence intervals for Z. However, according to Konishi (1985), Fisher's Z-transformation does not simultaneously normalize the sampling

distribution and stabilize variance when the number of observations on each subject exceeds two. As a result, he introduced a modified version of Fisher's transformation which may produce better confidence intervals when the number of classes exceeds two.

The second method of constructing confidence intervals for ρ is based on the F-distribution (Haggard, 1958, p.23). This method of constructing confidence intervals for ρ was later introduced by Searle (1971) and it was sometimes termed as Searle's exact method. Searle's method of constructing confidence intervals to analysis of variance estimator of ICC was derived for use with normally distributed data from a balanced design. In this method the confidence intervals for ρ are constructed based on the assumption that the variance ratio statistic is distributed as multiple of the central F-distribution.

A third method of constructing confidence intervals for ρ uses the large sample approximation formula to estimate the standard error of $\hat{\rho}$ and confidence intervals for ρ is obtained by applying a simple asymptotic approach, i.e., point estimate plus/minus normal quantile multiplied by the estimate of standard error. Formulae for large sample approximation of the standard error of analysis of variance estimator of $\hat{\rho}$ were derived by several authors. Assuming the sample size is sufficiently large, Fisher derived the standard error of ICC (Fisher, 1925, p.187). But this formula has a limitation that it can not be applied in the neighborhood of $+1$ and $-1/(k-1)$ even when the sample size is large. As Fisher described this is not an accurate formula

to use in testing significance and thus for constructing confidence intervals. Smith (1956) and Swiger *et al.* (1964), both derived formulae for large sample approximate standard error of ICC. Although these two formulae are more suitable for variable class sizes it can also be readily used for fixed class sizes.

By Monte Carlo simulation, Donner and Wells (1986) compared the performance of assigning confidence intervals to the ICC for normally distributed data. From this study they found that the method based on Smith's approximation to the standard error of ICC provides consistently good coverage for all values of ρ. They also commented that Searle's exact method and the method based on large sample formula derived by Swiger *et al.* (1964), provide excellent coverage (coverage probability $\geq .95$) for $\rho \leq 0.3$ and adequate coverage (coverage probability $\geq .925$) for $\rho \leq 0.7$. They found that the method based on Fisher's transformation yields coverage probabilities in excess of 98% and is thus conservative. A similar study was performed by Ukoumunne (2002) in the context of cluster randomization trials and it demonstrated similar results as Donner and Wells (1986).

2.5 Sample size estimation for intraclass correlation coefficient

Sample size estimation is crucial in planning a reliability study. It decides both the number of subjects to be included (n) and the number of replicated measures to be performed on each subject (k). Increasing the number of subjects over the required minimal amount will increase the cost of research unnecessarily and expose

the subjects to unwanted burden. In some situations the number of replicates have to be limited because of the practical or logical constraints. Shoukri *et al.* (2004) cited several examples drawn from biomedical research with limited replicated observations.

The determination of sample size of a reliability study is based on acquiring a specified level of precision for the estimated ρ. Therefore it can be based on the statistical test comparing an estimated reliability to a theoretical value, i.e., based on testing the null hypothesis $H_0 : \rho = \rho_0$ against $H_1 : \rho > \rho_0$ where ρ_0 is a specific value of ρ or it can be based on the width of the confidence interval of ρ or else it can be based on the number of subjects required to minimize the variance of the estimated ρ.

Donner and Eliasziw (1987) investigated values of n and k required to test $H_0 : \rho = \rho_0$ against $H_1 : \rho > \rho_0$ and provided exact power contours to give guidance for planning of a reliability study where the value of ρ_0 depends on a choice of a minimum value of ρ that the investigators consider applicable. The value of ρ_0 would typically not be 0 as zero reliability is of no interest. Walter *et al.* (1998) extended the study by Donner and Eliasziw (1987) by developing an approximation that allows the calculation of required sample size for the number of subjects n, when the number of replicates k is fixed. The formula derived by Walter *et al.* (1998) permits the investigator to explore the design options for parameter values.

Giraudeau and Mary (2001) and Bonett (2002) developed formulae for calculating the approximate number of subjects required to obtain an exact confidence interval of

desired width. They pointed out that the approach based on hypothesis testing may not be appropriate while planning a reliability study since such an approach requires to specify both the values of ρ_0 and ρ.

Shoukri *et al.* (2003) developed a method to determine the number of replicates k per subject needed to minimize the variance of the estimated ρ in which they assumed that the total number of observations nk is fixed. In the same study Shoukri *et al.* (2003) incorporated the cost constraints into sample size determination and addressed the issues of obtaining combinations of (n, k) under these constraints.

Sample size requirements in designing a study to compare two or more reliability coefficients were addressed by Donner (1998). Such problems may arise as discussed in section 1.3 of Chapter 1 when comparing the reliability of two instruments. The underlying assumption for the sample size formulae presented in Donner (1998) is that the estimated reliability coefficients are statistically independent. For studies where the same subjects are used for the comparison this assumption is no longer valid as the two reliability coefficients are then related. Hypothesis tests for comparing dependent ICCs in the case of continuous outcome variable have been considered by several authors (Alsawalmeh and Feldt, 1994; Donner *et al.*, 1984; Donner and Zou, 2002). Therefore an approach of sample size estimation may be obtained by directly applying the methods described by these authors to compare two dependent ICCs.

Chapter 3

DEVELOPMENT OF CONFIDENCE INTERVAL PROCEDURE FOR A DIFFERENCE BETWEEN TWO CORRELATED INTRACLASS CORRELATION COEFFICIENTS

3.1 Introduction

Comparison of two correlated intraclass correlation coefficients (ICC) is a common statistical problem frequently encountered in instrument reliability studies. For example the performance of a new device is compared with the performance of an old device with the intention of replacing the old device. In such situations the dichotomous answer 'reject' or 'not reject' based on the p value obtained for testing the null hypothesis, $H_0 : \rho_1 = \rho_2$ can only provide limited information. In contrast, a confidence interval for the difference between $\rho_1 - \rho_2$ produces a range of values that are considered to be plausible for the parameter of interest $\rho_1 - \rho_2$. Thus, it shows both the magnitude and direction of the difference.

In this chapter, I present the procedures for setting approximate confidence intervals for a difference between two ICCs. Section 3.2 describes the general principles in confidence interval construction. Section 3.3 presents a review of confidence interval procedures for a single ICC as well as point estimates. In section 3.4, confidence interval construction for a difference between two ICCs, both simple asymptotic method (SA) and the new procedure that uses the idea of recovering variances from single ICCs will be discussed. In what follows, I refer to this new approach as MOVER, reflecting method of variance estimates recovery.

3.2 Definition of a confidence interval

Let θ denote an unknown population parameter and l be a $100(1-\alpha/2)\%$ lower limit for it. By definition l is a random variable, under repeated sampling, will fall below θ for $100(1-\alpha/2)\%$ of the time. In other words, the lower limit, l, satisfies

$$Pr(\theta \geq l) = 1 - \alpha/2.$$

Similarly, the $100(1-\alpha/2)\%$ upper limit, u, is a random variable that will fall above θ for $100(1-\alpha/2)\%$ of the time, i.e.,

$$Pr(\theta \leq u) = 1 - \alpha/2.$$

In combination we have

$$Pr(l \leq \theta \leq u) = 1 - \alpha,$$

i.e., the interval (l, u) constitutes a $100(1-\alpha)\%$ two sided confidence interval for θ. A $100(1-\alpha)\%$ two sided confidence interval for θ has a coverage probability of $(1-\alpha)$ which implies that a random interval constructed in this way will contain θ's true value $100(1-\alpha)\%$ of the time. A two sided $100(1-\alpha)\%$ confidence interval constructed as such has equal tails, i.e., the total coverage error α is divided up evenly between the lower and upper ends of the interval. This is to say that a confidence interval may be used to exclude extreme small and large values as the parameter of interest, in light of the data.

3.3 Inference procedure for a single intraclass correlation coefficient

I first review estimating ICC in the one-way random effects model. Then I consider four procedures for obtaining confidence intervals for a single ICC. These four procedures are, simple asymptotic method, confidence interval obtained using Fisher's Z-transformation (Fisher, 1925), a modified form of Fisher's Z-transformation (Konishi, 1985) and F-distribution based confidence interval (Searle, 1971; Haggard, 1958, p.23).

3.3.1 Point estimator

Suppose the subjects are chosen at random from large population. The measurements taken by an instrument on these subjects can be arranged as in Table 3.1.

Table 3.1: Layout of data in the one-way random effects model

Subjects	Measurement replicates					
	1	2	...	j	...	k
1	Y_{11}	Y_{12}	...	Y_{1j}	...	Y_{1k}
2	Y_{21}	Y_{22}	...	Y_{2j}	...	Y_{2k}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	Y_{i1}	Y_{i2}	...	Y_{ij}	...	Y_{ik}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Y_{n1}	Y_{n2}	...	Y_{nj}	...	Y_{nk}

An entry in Table 3.1, Y_{ij}, represents the jth observation taken from the ith randomly selected subject where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$. Assuming that the random row variable in Table 3.1, subjects represents the only source of variance, a one-way random effects model may be written as

$$Y_{ij} = \mu + a_i + e_{ij}, \tag{3.1}$$

where μ is the grand mean of all the observations in the population, the subject effect $\{a_i\}$ are normally distributed with mean 0 and variance σ_a^2, the residual effects $\{e_{ij}\}$ are normally distributed with mean 0 and variance σ_e^2, and the $\{a_i\}, \{e_{ij}\}$ are

completely independent. The corresponding analysis of variance (ANOVA) table is given in Table 3.2.

Table 3.2: Analysis of variance for the one-way random effects model

Source of variation	Degree of freedom(df)	Sum of squares	Mean square	Expected mean square
Among groups	$n-1$	SSA=$k\sum(\bar{Y}_{i.}-\bar{Y}_{..})^2$	MSA=$\frac{\text{SSA}}{n-1}$	$\sigma_e^2+k\sigma_a^2$
Within groups	$N-n$	SSE=$\sum\sum(Y_{ij}-\bar{Y}_{i.})^2$	MSE=$\frac{\text{SSE}}{N-n}$	σ_e^2
Total	$N-1$	SST=$\sum\sum(Y_{ij}-\bar{Y}_{..})^2$		

In accordance with assumptions made above for the model (3.1), variance of Y_{ij} is then given by $\sigma_Y^2=\sigma_a^2+\sigma_e^2$, and the intraclass correlation coefficient ρ is defined as $\rho=\sigma_a^2/\sigma_Y^2$. Since MSE and (MSA − MSE)/k, are unbiased estimators of σ_e^2 and σ_a^2 respectively, the ANOVA estimator of ρ can be written as

$$\hat{\rho}=r=\frac{\text{MSA}-\text{MSE}}{\text{MSA}+(k-1)\text{MSE}}, \tag{3.2}$$

where MSA and MSE are mean sum of squares between subjects and mean sum of squares within subjects (Donner and Koval, 1980).

Although the ANOVA approach is the most commonly used method to estimate ICC, other methods may also be used. Such methods include pairwise estimator, obtained by computing the Pearson product-moment correlation over all possible

pairs of observations within a class. Assuming a multivariate normal assumption, one can also apply maximum likelihood estimation (Donner and Koval, 1980).

In the context of reliability studies in which class sizes are constant, all these estimators are equivalent (Donner and Koval, 1980). Therefore I do not distinguish them in this thesis.

3.3.2 Confidence limits using the simple asymptotic method

An intuitive confidence interval constructed using simple asymptotic (SA) method makes use of the fact that when the sample size n, gets larger the distribution of $\hat{\rho}$ becomes more and more normal, i.e.,

$$\hat{\rho} \sim N(\rho, \hat{\sigma}^2) \tag{3.3}$$

where $\hat{\sigma}^2$ is the variance estimator of $\hat{\rho}$, may be estimated using the large sample variance formulae for $\hat{\rho}$ in Donner (1986) given by

$$\widehat{\text{var}}(\widehat{\rho}) = \frac{2(nk-1)(1-\widehat{\rho})^2[1+(k-1)\widehat{\rho}]^2}{k^2(k-1)n(n-1)} \tag{3.4}$$

where n, k, and N are number of subjects, number of repeated observation on each subject and total number of observations respectively.

From the statement (3.3), we can write

$$\frac{\hat{\rho}-\rho}{\sqrt{\hat{\sigma}^2}} \sim N(0,1). \tag{3.5}$$

If z_α indicates the $100.\alpha$th percentile point of a standard normal distribution, then from (3.5)

$$Pr\left\{-z_{\alpha/2} \leq \frac{\hat{\rho}-\rho}{\sqrt{\hat{\sigma}^2}} \leq z_{\alpha/2}\right\} = 1-\alpha.$$

Or, this can be written as

$$Pr\left\{\rho \in \left[\hat{\rho} - z_{\alpha/2}\sqrt{\hat{\sigma}^2}, \hat{\rho} + z_{\alpha/2}\sqrt{\hat{\sigma}^2}\right]\right\} = 1-\alpha.$$

Therefore a $100(1-\alpha)\%$ confidence interval for ρ is given by

$$(l,u) = \left(\hat{\rho} - z_{\alpha/2}\sqrt{\hat{\sigma}^2}, \hat{\rho} + z_{\alpha/2}\sqrt{\hat{\sigma}^2}\right). \tag{3.6}$$

3.3.3 Confidence limits based on Fisher's Z-transformation

Fisher (1925, ch.7), showed that

$$Z_F = \frac{1}{2}\ln\left\{\frac{1+(k-1)\hat{\rho}}{1-\hat{\rho}}\right\}$$

has an approximate normal distribution with mean

$$Z_F(\rho) = \frac{1}{2}\ln\left\{\frac{1+(k-1)\rho}{1-\rho}\right\}$$

and variance

$$V = \frac{k}{2(k-1)(N-2)}.$$

Then, a $(1-\alpha)100\%$ confidence interval for $Z_F(\rho)$ is given by

$$(Z_l, Z_u) = \left\{Z_F - z_{\alpha/2}\sqrt{V}, Z_F + z_{\alpha/2}\sqrt{V}\right\}$$

where $z_{\alpha/2}$ is the upper quantile of the standard normal distribution.

A $(1-\alpha)100\%$ confidence interval for ρ obtained by back transforming the above interval is given by

$$\left\{\frac{e^{2Z_l}-1}{e^{2Z_l}+(k-1)}, \frac{e^{2Z_u}-1}{e^{2Z_u}+(k-1)}\right\}. \tag{3.7}$$

An alternative to the classical Fisher's Z-transformation described above is to apply the inverse tanh (hyperbolic tangent) transformation to sample estimation and then apply the delta method to derive the variance of Z (Casella and Berger, 2002, p240-245). This approach sometimes has been referred to as the Z-transformation (Lachin, 2004; Rosner and Willett, 1988).

According to the delta method, when $\hat{\theta}$ has the mean and variance as θ and $\text{var}(\hat{\theta})$, the mean and variance of $g(\hat{\theta})$ are approximated by $g(\theta)$ and $[g'(\theta)]^2\text{var}(\hat{\theta})$,

where $g'(\theta)$ is the derivative of the function g evaluated at $\hat{\theta} = \theta$. I will use the large sample variance estimate of ICC derived by Smith (1956) and apply the delta method to obtain the variance of $Z(\rho)$ where $Z(\rho) = \frac{1}{2} log \frac{1+\rho}{1-\rho}$. The confidence interval for ρ based on the Smith's formula performed well for all values of the parameters (Donner and Wells, 1986).

Applying the delta method to $Z(\rho)$

$$\begin{aligned} \text{var}(Z(\hat{\rho})) &= [Z(\rho)'(\rho)]^2 \, \text{var}(\hat{\rho}) \\ &= \left[\frac{1}{(1-\rho)(1+\rho)}\right]^2 \text{var}(\hat{\rho}) \end{aligned} \tag{3.8}$$

Confidence interval for $Z(\rho)$ is then given by

$$(Z_l, Z_u) = \left\{ Z(\hat{\rho}) - z_{\alpha/2}\sqrt{\widehat{\text{var}}(Z(\hat{\rho}))}, Z(\hat{\rho}) + z_{\alpha/2}\sqrt{\widehat{\text{var}}(Z(\hat{\rho}))} \right\}$$

By back transforming the above CI, a $100(1-\alpha)\%$ CI for ρ can be obtained.

3.3.4 Confidence limits using a modified Z-transformation

A modified form of Fisher's Z-transformation was introduced by Konishi (1985) and Konishi and Gupta (1987) in which they showed that for $k > 2$, a more effective transformation is given by

$$Z_m = \sqrt{\frac{k-1}{2k}} \ln \left\{ \frac{1+(k-1)\hat{\rho}}{1-\hat{\rho}} \right\}$$

which is asymptotically normally distributed as

$$N\left[Z(\rho)_m + \frac{7-5k}{N\sqrt{18k(k-1)}}, \frac{1}{N}\right],$$

where the second term in the mean of Z_m is a bias correction factor.

Lower and upper bounds of $(1-\alpha)100\%$ confidence interval for $Z(\rho)_m$ are given by

$$\begin{aligned} Z_{m,l} &= Z_m - z_{\alpha/2}\sqrt{V_m} - \frac{7-5k}{N\sqrt{18k(k-1)}} \\ Z_{m,u} &= Z_m + z_{\alpha/2}\sqrt{V_m} - \frac{7-5k}{N\sqrt{18k(k-1)}}, \end{aligned}$$

where $V_m = 1/N$. Back transforming the above interval yields a $(1-\alpha)100\%$ confidence interval for ρ using the Konishi modified method given by

$$\left\{ \frac{e^{\left(Z_{m,l}\sqrt{\frac{2k}{k-1}}\right)} - 1}{e^{\left(Z_{m,l}\sqrt{\frac{2k}{k-1}}\right)} + (k-1)}, \frac{e^{\left(Z_{m,u}\sqrt{\frac{2k}{k-1}}\right)} - 1}{e^{\left(Z_{m,u}\sqrt{\frac{2k}{k-1}}\right)} + (k-1)} \right\}. \tag{3.9}$$

3.3.5 Confidence limits based on F-distribution

Let $F = \text{MSA/MSE}$ be the variance ratio statistic. From Table 3.2, equating observed and expected mean squares we can obtain

$$\begin{aligned} \text{MSA} &= \sigma_e^2 + k\sigma_a^2 \\ \text{MSE} &= \sigma_e^2 \end{aligned}$$

and it can be shown that

$$\frac{\text{MSA}/(k\sigma_a^2+\sigma_e^2)}{\text{MSE}/\sigma_e}$$

is distributed as $F_{n-1,n(k-1)}$ (Searle, 1971). Let F_L and F_U be the lower and upper limits which enclose $(1-\alpha)$ of the $F_{n-1,n(k-1)}$ distribution. Thus,

$$\begin{aligned} Pr\left(F_{(\alpha/2,n-1,nk-n)} \leq \frac{\text{MSA}}{\text{MSE}}\frac{\sigma_e^2}{k\sigma_a^2+\sigma_e^2} \leq F_{(1-\alpha/2,n-1,nk-n)}\right) &= 1-\alpha \\ \Rightarrow Pr\left(F_L \leq F\frac{\sigma_e^2}{k\sigma_a^2+\sigma_e^2} \leq F_U\right) &= 1-\alpha. \end{aligned}$$

Rearranging the above equation

$$Pr\left(\frac{F/F_U-1}{k+F/F_U-1} \leq \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2} \leq \frac{F/F_L-1}{k+F/F_L-1}\right) = 1-\alpha$$

Therefore a $(1-\alpha)100\%$ confidence interval for ρ is given by

$$\left\{\frac{F/F_U-1}{k+F/F_U-1}, \frac{F/F_L-1}{k+F/F_L-1}\right\}. \tag{3.10}$$

3.4 Confidence interval procedure for a difference between two intra-class correlation coefficients

3.4.1 Notation

Let

$$\mathbf{Y_i} = (Y_{1,i,1}, Y_{1,i,2}, \ldots, Y_{1,i,k_1}, Y_{2,i,k_1+1}, Y_{2,i,k_1+2}, \ldots, Y_{2,i,k_1+k_2})$$

where $i = 1, \ldots, n$, the p-vector of measures associated to subject i with $Y_{1,i,1}, \ldots, Y_{1,i,k_1}$ being the k_1 vector of measures realised with device 1 and $Y_{2,i,k_1+1}, \ldots, Y_{2,i,k_1+k_2}$ being the k_2 vector of measures realised with device 2. The total number of measures associated to each of the n subjects, p is fixed and equals to $p = k_1 + k_2$.

We assume that the following model holds:

$$\mathbf{Y_i} \sim N(\mu, \mathbf{\Sigma}), \tag{3.11}$$

where $\mu^T = (\mu_1 \mathbf{1}_{k_1}^T, \mu_2 \mathbf{1}_{k_2}^T)$ and

$$\mathbf{\Sigma} = \begin{pmatrix} \{(1-\rho_1)\mathbf{I}_{k_1} + \rho_1 \mathbf{J}_{k_1}\}\sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \mathbf{J}_{k_1 \times k_2} \\ \rho_{12}\sigma_1\sigma_2 \mathbf{J}_{k_1 \times k_2} & \{(1-\rho_2)\mathbf{I}_{k_2} + \rho_2 \mathbf{J}_{k_2}\}\sigma_2^2 \end{pmatrix}$$

Here $\mathbf{1}_p$ is a column vector with all the p elements equal to 1, $\mathbf{I}_p$ is a $p \times p$ identity matrix and $\mathbf{J}_p$ and $\mathbf{J}_{p \times q}$ are $p \times p$ and $p \times q$ matrices with all the elements equal to 1. This model assumes that the k_1 observations taken by the first device have common mean μ_1, common variance σ_1^2 and common intraclass correlation ρ_1, whereas the k_2 observations taken by the second device have common mean μ_2, common variance σ_2^2 and common intraclass correlation ρ_2. It is also assumed that the interclass correlation coefficient ρ_{12} between any pair of observations $Y_{1,i,j} (j = 1, 2, \ldots, k_1)$ and $Y_{2,i,k_1+j'} (j' = 1, 2, \ldots, k_2)$ be constant across all subjects in the population. For Σ to be positive definitive we must have $\rho_{12} < \rho_1 \rho_2$.

Let ρ_1 and ρ_2 are intraclass correlation coefficients for device 1 and device 2 respectively and estimate them as described in section 3.3.1. Therefore

$$\hat{\rho}_l = r_l = \frac{\text{MSA}_l - \text{MSE}_l}{\text{MSA}_l + (k_l - 1)\text{MSE}_l}, \qquad l = 1, 2.$$

3.4.2 Confidence interval for a difference between two intraclass correlation coefficients using simple asymptotic method

Let $\hat{\rho}_1$, $\hat{\rho}_2$ be the sample estimates of ρ_1, ρ_2 obtained from the measurements from device 1 and device 2 and let $\widehat{\text{var}}(\hat{\rho}_1)$, $\widehat{\text{var}}(\hat{\rho}_2)$ be the variance estimates of $\hat{\rho}_1$, $\hat{\rho}_2$ respectively. By the central limit theorem and Slutsky's theorem (Casella and Berger, 2002, p239), the $100(1-\alpha)\%$ confidence interval for a difference between two ICCs, $\rho_1 - \rho_2$ is given by

$$\left\{(\hat{\rho}_1 - \hat{\rho}_2) - z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\rho}_1 - \hat{\rho}_2)}, (\hat{\rho}_1 - \hat{\rho}_2) + z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\rho}_1 - \hat{\rho}_2)}\right\}, \tag{3.12}$$

where $\widehat{\text{var}}(\hat{\rho}_1 - \hat{\rho}_2)$ is the sample estimate of $\text{var}(\hat{\rho}_1 - \hat{\rho}_2)$.

Assuming $\hat{\rho}_1$ and $\hat{\rho}_2$ are independent, the variance estimate of $(\hat{\rho}_1 - \hat{\rho}_2)$ can be written as

$$\widehat{\text{var}}(\hat{\rho}_1 - \hat{\rho}_2) = \widehat{\text{var}}(\hat{\rho}_1) + \widehat{\text{var}}(\hat{\rho}_2).$$

Substituting $\widehat{\text{var}}(\hat{\rho}_1 - \hat{\rho}_2)$ in (3.12) will result a $100(1-\alpha)\%$ confidence interval for a

difference between two independent ICCs as

$$\left\{(\hat{\rho}_1 - \hat{\rho}_2) \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\rho}_1) + \widehat{\text{var}}(\hat{\rho}_2)}\right\}. \tag{3.13}$$

Assuming $\hat{\rho}_1$ and $\hat{\rho}_2$ are correlated

$$\widehat{\text{var}}(\hat{\rho}_1 - \hat{\rho}_2) = \widehat{\text{var}}(\hat{\rho}_1) + \widehat{\text{var}}(\hat{\rho}_2) - 2\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2).$$

Substituting $\widehat{\text{var}}(\hat{\rho}_1 - \hat{\rho}_2)$ in (3.12), a $100(1-\alpha)\%$ confidence interval for a difference between 2 correlated ICCs is given as

$$\left\{(\hat{\rho}_1 - \hat{\rho}_2) \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\rho}_1) + \widehat{\text{var}}(\hat{\rho}_2) - 2\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2)}\right\}. \tag{3.14}$$

In the above confidence interval $\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2)$ may be estimated from the formula derived by (Elston, 1975, p.136) as

$$\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2) = \frac{2\hat{\rho}_{12}^2}{n}(1 - \hat{\rho}_1)(1 - \hat{\rho}_2). \tag{3.15}$$

The term $\hat{\rho}_{12}$ is the estimated interclass correlation between any two pairs of observation of device 1 and device 2 which may be obtained using the formula given in

Rosner (1982),

$$\hat{\rho}_{12} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_1}\sum_{m=k_1+1}^{k_1+k_2}(y_{ij}-\bar{\bar{y}}_p)(y_{im}-\bar{\bar{y}}_c)}{\left[\sum_{i=1}^{n}k_1\sum_{j=1}^{k_1}(y_{ij}-\bar{\bar{y}}_p)^2\right]\left[\sum_{i=1}^{n}k_2\sum_{j=k_1+1}^{k_1+k_2}(y_{ij}-\bar{\bar{y}}_c)^2\right]}, \tag{3.16}$$

where

$$\bar{\bar{y}}_p = \frac{\sum_{i=1}^{n}\bar{y}_{ip}}{n}, \qquad \bar{\bar{y}}_c = \frac{\sum_{j=k_1+1}^{k_1+k_2}\bar{y}_{ic}}{n}$$

$$\bar{y}_{ip} = \frac{\sum_{j=1}^{k_1}y_{ij}}{k_1}, \qquad \bar{y}_{ic} = \frac{\sum_{j=k_1+1}^{k_1+k_2}y_{ij}}{k_2}.$$

3.4.3 Confidence intervals for a difference between two intraclass correlation coefficients using method of variance recovery

Confidence intervals constructed using simple asymptotic (SA) method are accurate only when the sample size is large or when the sampling distribution is close to normal. It is a well known that the distribution of $\hat{\rho}$ is highly skewed. The poor performance of the simple asymptotic method is that it does not adjust for skewness of the sampling distribution of $\hat{\rho}$ and as a result, the simple asymptotic method does not reflect the asymmetry of the underlying sampling distribution. The simple asymptotic method forces the confidence intervals to be symmetrical over the parameter of interest, i.e., $\hat{\rho}-l = u-\hat{\rho}$ where l and u are the $100(1-\alpha)\%$ lower and upper limit of ρ respectively. Therefore variance estimates for $\hat{\rho}$ at $\rho = l$ and at $\rho = u$ are equal and hence the variance estimate is independent of $\hat{\rho}$.

In the case of intraclass correlation coefficient it is known that the variance of $\hat{\rho}$

is a function of itself, i.e., $\text{var}(\hat{\rho})$ depends on ρ (Fisher, 1925, p.187). As a result, symmetrical confidence intervals for ρ which assume a variance estimate independent of $\hat{\rho}$ do not adjust for the asymmetry of $\hat{\rho}$. Confidence intervals for ρ accounting for the skewness in the underlying distribution of $\hat{\rho}$ do so, as shown below, through different variance estimates at the lower (l) and upper (u) limits of $\hat{\rho}$ (Zou, 2007). This is the same logic of the score-type confidence intervals which are known to perform well in practice (Bartlett, 1953).

Suppose a $100(1-\alpha)\%$ confidence interval for ρ is given by (l, u) where

$$l = \hat{\rho} - z_{\alpha/2}\sqrt{\text{var}(\hat{\rho})}, \qquad u = \hat{\rho} + z_{\alpha/2}\sqrt{\text{var}(\hat{\rho})},$$

Without assuming l and u are symmetrical about $\hat{\rho}$, we can recover variance estimates for $\hat{\rho}$

at $\rho = l$ as

$$\widehat{\text{var}}_l(\hat{\rho}) = \frac{(\hat{\rho} - l)^2}{z^2_{\alpha/2}}, \tag{3.17}$$

and at $\rho = u$ as

$$\widehat{\text{var}}_u(\hat{\rho}) = \frac{(u - \hat{\rho})^2}{z^2_{\alpha/2}}. \tag{3.18}$$

Therefore in this thesis I use different variance estimates for lower and upper limits of confidence interval (CI) to develop a new method of constructing CI for a difference between two ICCs. In doing so I adjust the required CI for the skewness

of the underlying sampling distribution. Zou and Donner (2008) have extended the idea to a wide range of applications.

Let L and U be the lower and upper limits of $100(1-\alpha)\%$ confidence interval for a difference between two ICCs. And also let (l_1, u_1) and (l_2, u_2) be separate $100(1-\alpha)\%$ confidence intervals which contain the plausible values of ρ_1 and ρ_2 respectively. Again in the spirit of score-type confidence intervals, one can estimate the variance of $(\hat{\rho}_1 - \hat{\rho}_2)$ in the neighborhood of L when obtaining L and estimate variance in the neighborhood of U when obtaining U.

We can reasonably argue that among the plausible values provided by the two pairs of confidence limits, $l_1 - u_2$ is the value near L and $u_1 - l_2$ is the value near U. Therefore for the lower limit for $(\rho_1 - \rho_2)$, the variance of $(\hat{\rho}_1 - \hat{\rho}_2)$ is estimated under the condition $\rho_1 = l_1$ and $\rho_2 = u_2$. Similarly for the upper limit for $(\rho_1 - \rho_2)$, variance of $(\hat{\rho}_1 - \hat{\rho}_2)$ is estimated under the condition $\rho_1 = u_1$ and $\rho_2 = l_2$ (Zou and Donner, 2008; Zou, 2007).

By the central limit theorem and under the assumption that $\hat{\rho}_1$ and $\hat{\rho}_2$ are independent, we have a $100(1-\alpha)\%$ confidence interval for a difference between two ICCs, $\rho_1 - \rho_2$ as

$$\begin{aligned} L &= (\hat{\rho}_1 - \hat{\rho}_2) - z_{\alpha/2}\sqrt{\mathrm{var}_L(\hat{\rho}_1 - \hat{\rho}_2)} \\ &= (\hat{\rho}_1 - \hat{\rho}_2) - z_{\alpha/2}\sqrt{\mathrm{var}_{l_1}(\hat{\rho}_1) + \mathrm{var}_{u_2}(\hat{\rho}_2)} \\ &= (\hat{\rho}_1 - \hat{\rho}_2) - z_{\alpha/2}\sqrt{\frac{(\hat{\rho}_1 - l_1)^2}{z^2_{\alpha/2}} + \frac{(u_2 - \hat{\rho}_2)^2}{z^2_{\alpha/2}}} \end{aligned} \tag{3.19}$$

Similarly, estimated variance of $(\hat{\rho}_1 - \hat{\rho}_2)$ for the upper limit of $(\rho_1 - \rho_2)$ uses the variance estimates when $\rho_1 = u_1$ and $\rho_2 = l_2$.

$$\begin{aligned} \text{var}_U(\hat{\rho}_1 - \hat{\rho}_2) &= \text{var}_{u_1}(\hat{\rho}_1) + \text{var}_{l_2}(\hat{\rho}_2), \\ \widehat{\text{var}}_{u_1}(\hat{\rho}_1) &= \frac{(u_1 - \hat{\rho}_1)^2}{z^2_{\alpha/2}}, \\ \widehat{\text{var}}_{l_2}(\hat{\rho}_2) &= \frac{(\hat{\rho}_2 - l_2)^2}{z^2_{\alpha/2}}. \end{aligned}$$

Therefore a $100(1-\alpha)\%$ upper confidence interval for a difference between two ICCs, $\rho_1 - \rho_2$ can be written as

$$U = (\hat{\rho}_1 - \hat{\rho}_2) + z_{\alpha/2}\sqrt{\frac{(u_1 - \hat{\rho}_1)^2}{z^2_{\alpha/2}} + \frac{(\hat{\rho}_2 - l_2)^2}{z^2_{\alpha/2}}}. \tag{3.20}$$

The above argument may be extended to the case where $\hat{\rho}_1$ and $\hat{\rho}_2$ are dependent. Specifically,

$$\begin{aligned} \text{var}_L(\hat{\rho}_1 - \hat{\rho}_2) &= \text{var}_{l_1}(\hat{\rho}_1) + \text{var}_{u_2}(\hat{\rho}_2) - 2\text{cov}(\hat{\rho}_1, \hat{\rho}_2) \\ \widehat{\text{var}}_L(\hat{\rho}_1 - \hat{\rho}_2) &= \frac{(\hat{\rho}_1 - l_1)^2}{z^2_{\alpha/2}} + \frac{(u_2 - \hat{\rho}_2)^2}{z^2_{\alpha/2}} - 2\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2) \end{aligned}$$

and

$$\begin{aligned}
\text{var}_U(\hat{\rho}_1 - \hat{\rho}_2) &= \text{var}_{u_1}(\hat{\rho}_1) + \text{var}_{l_2}(\hat{\rho}_2) - 2\text{cov}(\hat{\rho}_1, \hat{\rho}_2) \\
\widehat{\text{var}}_U(\hat{\rho}_1 - \hat{\rho}_2) &= \frac{(u_1 - \hat{\rho}_1)^2}{z^2_{\alpha/2}} + \frac{(\hat{\rho}_2 - l_2)^2}{z^2_{\alpha/2}} - 2\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2).
\end{aligned}$$

Therefore a $100(1-\alpha)\%$ confidence interval for a difference between two dependent ICCs, $\rho_1 - \rho_2$ can be written as

$$L = (\hat{\rho}_1 - \hat{\rho}_2) - z_{\alpha/2}\sqrt{\frac{(\hat{\rho}_1 - l_1)^2}{z^2_{\alpha/2}} + \frac{(u_2 - \hat{\rho}_2)^2}{z^2_{\alpha/2}} - 2\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2)}, \tag{3.21}$$

$$U = (\hat{\rho}_1 - \hat{\rho}_2) + z_{\alpha/2}\sqrt{\frac{(u_1 - \hat{\rho}_1)^2}{z^2_{\alpha/2}} + \frac{(\hat{\rho}_2 - l_2)^2}{z^2_{\alpha/2}} - 2\widehat{\text{cov}}(\hat{\rho}_1, \hat{\rho}_2)}. \tag{3.22}$$

The $\text{cov}(\hat{\rho}_1, \hat{\rho}_2)$ can be estimated from the well known formula

$$\text{corr}(\hat{\rho}_1, \hat{\rho}_2) = \frac{\text{cov}(\hat{\rho}_1, \hat{\rho}_2)}{\sqrt{\text{var}(\hat{\rho}_1) \times \text{var}(\hat{\rho}_2)}} \tag{3.23}$$

Then equation (3.21) gives

$$\begin{aligned}
L &= (\hat{\rho}_1 - \hat{\rho}_2) - z_{\alpha/2}\sqrt{\frac{(\hat{\rho}_1 - l_1)^2}{z^2_{\alpha/2}} + \frac{(u_2 - \hat{\rho}_2)^2}{z^2_{\alpha/2}} - 2\widehat{\text{corr}}(\hat{\rho}_1, \hat{\rho}_2)\sqrt{\frac{(\hat{\rho}_1 - l_1)^2}{z^2_{\alpha/2}}}\sqrt{\frac{(u_2 - \hat{\rho}_2)^2}{z^2_{\alpha/2}}}} \\
&= (\hat{\rho}_1 - \hat{\rho}_2) - \sqrt{(\hat{\rho}_1 - l_1)^2 + (u_2 - \hat{\rho}_2)^2 - 2\widehat{\text{corr}}(\hat{\rho}_1, \hat{\rho}_2)(\hat{\rho}_1 - l_1)(u_2 - \hat{\rho}_2)}.
\end{aligned}$$

From equation (3.22) we can write

$$\begin{aligned} U &= (\hat{\rho}_1 - \hat{\rho}_2) + z_{\alpha/2}\sqrt{\frac{(u_1-\hat{\rho}_1)^2}{z^2_{\alpha/2}} + \frac{(\hat{\rho}_2 - l_2)^2}{z^2_{\alpha/2}} - 2\widehat{\text{corr}}(\hat{\rho}_1,\hat{\rho}_2)\sqrt{\frac{(u_1-\hat{\rho}_1)^2}{z^2_{\alpha/2}}}\sqrt{\frac{(\hat{\rho}_2-l_2)^2}{z^2_{\alpha/2}}}} \\ &= (\hat{\rho}_1 - \hat{\rho}_2) + \sqrt{(u_1-\hat{\rho}_1)^2 + (\hat{\rho}_2-l_2)^2 - 2\widehat{\text{corr}}(\hat{\rho}_1,\hat{\rho}_2)(u_1-\hat{\rho}_1)(\hat{\rho}_2-l_2)}. \end{aligned}$$

Therefore a $100(1-\alpha)\%$ confidence interval for a difference between two correlated ICCs, $\rho_1 - \rho_2$ can be written as

$$L = (\hat{\rho}_1 - \hat{\rho}_2) - \sqrt{(\hat{\rho}_1 - l_1)^2 + (u_2-\hat{\rho}_2)^2 - 2\widehat{\text{corr}}(\hat{\rho}_1,\hat{\rho}_2)(\hat{\rho}_1-l_1)(u_2-\hat{\rho}_2)} \quad (3.24)$$

$$U = (\hat{\rho}_1 - \hat{\rho}_2) + \sqrt{(u_1 - \hat{\rho}_1)^2 + (\hat{\rho}_2-l_2)^2 - 2\widehat{\text{corr}}(\hat{\rho}_1,\hat{\rho}_2)(u_1-\hat{\rho}_1)(\hat{\rho}_2-l_2)} \quad (3.25)$$

In equation (3.23) and (3.24), $\text{corr}(\hat{\rho}_1,\hat{\rho}_2)$ can be obtain using the following asymptotic formulas derived in Elston (1975).

$$\begin{aligned} \text{var}(\hat{\rho}_l) &= \frac{2(k_l-1)}{nk_l}(1-\rho_l)^2\left(\frac{1}{k_l-1}+\rho_l\right)^2; l = 1,2, \\ \text{cov}(\hat{\rho}_1,\hat{\rho}_2) &= \frac{2\rho_{12}^2}{n}(1-\rho_1)(1-\rho_2) \end{aligned}$$

Substituting $\text{var}(\hat{\rho}_l)$ and $\text{cov}(\hat{\rho}_1,\hat{\rho}_2)$ in equation (3.23)

$$\begin{aligned} \text{corr}(\hat{\rho}_1,\hat{\rho}_2) &= \frac{\frac{2\rho_{12}^2}{n}(1-\rho_1)(1-\rho_2)}{\left[\frac{2(k_1-1)}{nk_1}(1-\rho_1)^2\left(\frac{1}{k_1-1}+\rho_1\right)^2\right]\left[\frac{2(k_2-1)}{nk_2}(1-\rho_2)^2\left(\frac{1}{k_2-1}+\rho_2\right)^2\right]} \\ &= \frac{\hat{\rho}_{12}^2\left[k_1k_2(k_1-1)(k_2-1)\right]^{1/2}}{\left[1+(k_1-1)\rho_1\right]\left[1+(k_2-1)\rho_2\right]}. \end{aligned} \quad (3.26)$$

By substituting $\rho_1 = l_1$ and $\rho_2 = u_2$ in the above equation, $\widehat{\text{corr}}(\hat{\rho}_1, \hat{\rho}_2)$ for the $100(1-\alpha)\%$ lower limit for $(\rho_1 - \rho_2)$ is obtained. Similarly, $\widehat{\text{corr}}(\hat{\rho}_1, \hat{\rho}_2)$ for the $100(1-\alpha)\%$ upper limit for $(\rho_1 - \rho_2)$ is obtained by substituting $\rho_1 = u_1$ and $\rho_2 = l_2$.

Confidence intervals given by (3.23) and (3.24) reduce to (3.13) when SA method has been used to obtain the CI for a single ICC. This can be shown as below.

From equation (3.6)

$$\begin{aligned} l_1 &= \hat{\rho}_1 - z_{\alpha/2}\sqrt{\text{var}(\hat{\rho}_1)} \\ (\hat{\rho}_1 - l_1)^2 &= z^2_{\alpha/2}\text{var}(\hat{\rho}_1) \end{aligned}$$

and

$$\begin{aligned} u_1 &= \hat{\rho}_1 + z_{\alpha/2}\sqrt{\text{var}(\hat{\rho}_1)} \\ (u_1 - \hat{\rho}_1)^2 &= z^2_{\alpha/2}\text{var}(\hat{\rho}_1) \end{aligned}$$

According to the SA method confidence limits are symmetrical around the parameter of estimate and hence

$$(\hat{\rho}_1 - l_1) = (u_1 - \hat{\rho}_1)$$

and therefore

$$(\hat{\rho}_1 - l_1)^2 = (u_1 - \hat{\rho}_1)^2 = z^2_{\alpha/2}\text{var}(\hat{\rho}_1).$$

Similarly

$$(\hat{\rho}_2 - l_2)^2 = (u_2 - \hat{\rho}_2)^2 = z^2_{\alpha/2}\text{var}(\hat{\rho}_2).$$

This again shows the key feature of the MOVER method which recognizes the asymmetric nature of the sampling distribution of the single ICC, in contrast to the SA method which ignores this fact.

3.4.4 Summary

The simple asymptotic method of constructing confidence intervals may perform poorly when the underlying sampling distribution is skewed. Since the sampling distribution of $\hat{\rho}$ is skewed, I applied the idea of recovering variance estimates from single ICCs to construct confidence intervals for a difference between two ICCs. The resulting confidence intervals in general are not symmetrical such that it reflects the sampling distribution for single ICCs which are well known to be left skewed (Fisher, 1925, p.180).

Chapter 4

SIMULATION

4.1 Introduction

The theoretical properties of the proposed confidence interval estimation procedures discussed in the last chapter are asymptotic and intractable in finite samples. I therefore used simulation studies to evaluate the performance of the proposed method. Simulation studies provide empirical estimation of the sampling distribution of parameters of interest. Therefore this technique is employed to evaluate statistical methods which can not be achieved with studies of real data alone.

In section 4.2 of this chapter, I describe the parameter selection, data generation, statistical methods to be evaluated, number of simulations needed to be performed and the criteria for evaluation of the simulation procedure. Section 4.3 presents the simulation results and section 4.4 presents discussion and conclusion of simulation results. The summary of chapter 4 is given in section 4.5.

4.2 Study design

4.2.1 Parameter selection and data generation

The parameters for the simulation of this thesis include the number of subjects (n), number of measurements from each device (k_1, k_2), values for ρ_1, ρ_2 and ρ_{12}. Values for ρ_1 and ρ_2 were selected based on the arbitrary benchmark values of reliability coefficient suggested by Landis and Koch (1977) as slight (0.00 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80) and almost perfect (0.81 to 1.00). Therefore the parameter values of ρ_1 range from 0.1 to 0.9 with increments of 0.2, i.e., the values are 0.1, 0.3, 0.5, 0.7, 0.9 and ρ_2 was selected such that $\rho_2 = \rho_1 + \delta$ where δ range from 0 to 0.06 with increment of 0.02. Since the primary objective of this thesis is to propose a method of constructing confidence interval for $\rho_1 - \rho_2$ taking into account the dependence induced by the positive values of ρ_{12}, it was chosen such that $\rho_{12} = \sqrt{\rho_1 \rho_2} - 0.05$ (range from 0.05 to 0.87) so that it would satisfy the condition that Σ is a positive definitive matrix. For the sample size n, considered $n = 15, 50, 100$ to represent small, medium and large sample sizes. As for k_1 and k_2, five parameter combinations such as $k_1 = k_2 = 2$; $k_1 = 2, k_2 = 4$; $k_1 = k_2 = 4$; $k_1 = 6, k_2 = 3$; $k_1 = k_2 = 6$ were chosen, so that the different methods can be evaluated with small number of observations as well as with large number of observations. Without loss of generality, $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$ were set where μ_1, μ_2 and σ_1^2, σ_2^2 are the common mean and common variance of the measurements taken from device 1 and

device 2 respectively.

For each of 300 parameter combinations, I generated 10000 observations from a multivariate normal distribution with correlation structure as defined in Chapter 3, expression (3.10). The minimum number of simulations required was calculated based on 0.4% margin of error and expected the empirical coverage vary between 94.6% to 95.4%. All simulations were performed using exact computations in SAS 9.1 Proc IML (SAS Institute, NC).

4.2.2 Confidence interval procedures compared

Simulation studies were used to study the performances of the MOVER method. Four different confidence interval estimation methods were used to obtain a CI for single ICC and then applied to the MOVER for a difference between ICCs. Confidence intervals for single ICCs were obtained using, 1) simple asymptotic method, 2) Fisher's Z-transformation (Fisher, 1925, p.185) with variance derived using Delta method, 3) modified Z-transformation by Konishi (Konishi, 1985) and 4) based on the F distribution (Donner, 1986). In this thesis I named these four methods as 'SA', 'Fisher', 'Konishi' and 'Exact' respectively. For all 300 parameter combinations, 95% confidence intervals were constructed using all four methods.

4.2.3 Evaluation criteria

In simulation studies, comparison of simulated results with the true values used to simulate the data, provides a measure of performance. When describing performance of methods of constructing confidence intervals, I focus on the following three criteria.

Coverage

The coverage of a confidence interval is the proportion of times, in repeated sampling, that the obtained confidence interval contains the true parameter value. If a procedure is working well, we expect the empirical coverage of the confidence interval constructed to be approximately equal to nominal coverage of 95%. To evaluate the extent to which the empirical coverage of the confidence interval constructed matched with the nominal coverage of 95%, I use three criteria adapted by many authors as: strict criterion (94.5% to 95.5%); moderate criterion (93.75% to 96.25%); liberal criterion (92.5% to 97.5%) (Zou, 2007; Robey and Barcikowski, 1992). Over coverage is an indication that the results are too conservative which leads to a loss of statistical power with too many type II errors. Similarly under coverage is an indication that the results are too liberal which leads to too many type I errors. The empirical coverage percentage was estimated by the relative frequency out of 10000 intervals that contains the true parameter.

Tail errors

If the confidence intervals constructed are appropriate, then the two sided overall error rate shall approximately be equal to α, which in our case is 0.05. As Efron and

Tibshirani (1993, p.156) pointed out, for a two tailed confidence interval it requires that one sided miscoverage of the interval be $\alpha/2$ on each side, left and right, rather than just an overall error of α. According to this point of view, if the entire error probability was contained in one tail, reporting overall error will mask the true picture. Therefore when constructing confidence intervals using different methods I recorded the number of intervals that 'miss left' and 'miss right' where 'miss left' occurs when the interval is completely to the left of the parameter and 'miss right' occurs when the interval is completely to the right of the parameter. Therefore in this thesis I consider confidence interval procedures which produce confidence intervals with coverage probability approximately close to the nominal coverage of 95% with left miss and right miss approximately close to 2.5% as procedures with better performances. And also when two CI procedures satisfy the same coverage criterion I prefer the procedure that yields the least difference between miss left and miss right.

Confidence interval width

A narrower confidence interval with satisfactory coverage provides greater accuracy. Therefore in a given simulation study, when two procedures both have good coverage, I prefer the procedure which yields the substantially shorter confidence interval. The length of a confidence interval was calculated by subtracting the lower limit from the upper limit.

4.3 Simulation results

Empirical coverage percentage based on 10000 runs for 95% nominal confidence intervals for a difference between two dependent ICCs using MOVER method are graphically presented in Fig 4.1 to Fig 4.5. Each boxplot was drawn from coverage percentages of 20 parameter combinations and presented based on the sample size. A reference line has been drawn at 95% of the empirical coverage percentage axis (y-axis) to aid in the interpretations. These figures illustrate four summaries; location, spread, skewness and longtailedness of the results obtained. Figure 4.6 presents the imbalance of tail errors, quantify by relative bias percentage obtained as

$$\frac{|MR - ML|}{MR + ML} \times 100$$

and Fig 4.7 presents the CI width for 95% nominal confidence intervals for a difference between two dependent ICCs using MOVER method. Each boxplot was drawn from coverage percentages of 100 parameter combinations. In reliability context, higher reliability coefficients are of great importance. Therefore I present the simulation results obtained for high reliability coefficients, i.e., ρ_1 and ρ_2 are greater than or equal to 0.5 and are presented in Table 4.1, 4.2 and 4.3. A summary of the performances of SA, Fisher, Konishi and Exact methods to construct CI for single ρ in terms of coverage, tail errors and width which can be used as an aid to explain the good performance of MOVER method, are given in Table 4.4.

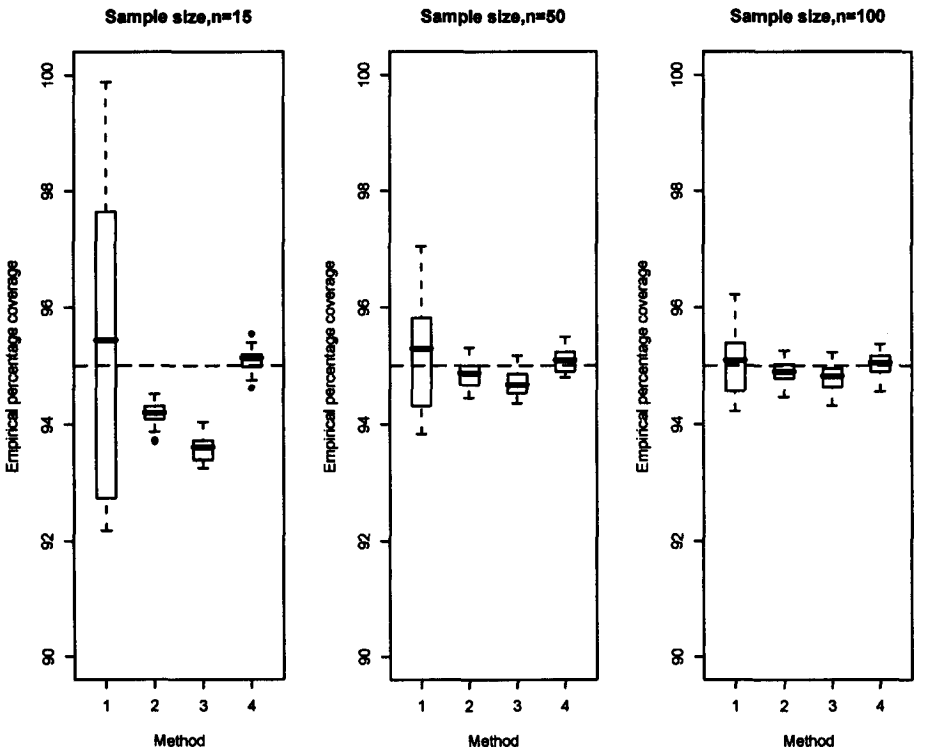

Figure 4.1: Mean coverage percentage based on 10,000 runs for nominal 95% confidence intervals for a difference between two correlated ICCs using 4 confidence interval methods for single ICCs when $k_1 = k_2 = 2$. Methods 1, 2, 3, and 4 represents SA, Fisher, Konishi and Exact procedures respectively. Each boxplot was drawn from coverage percentage of 20 parameter combinations.

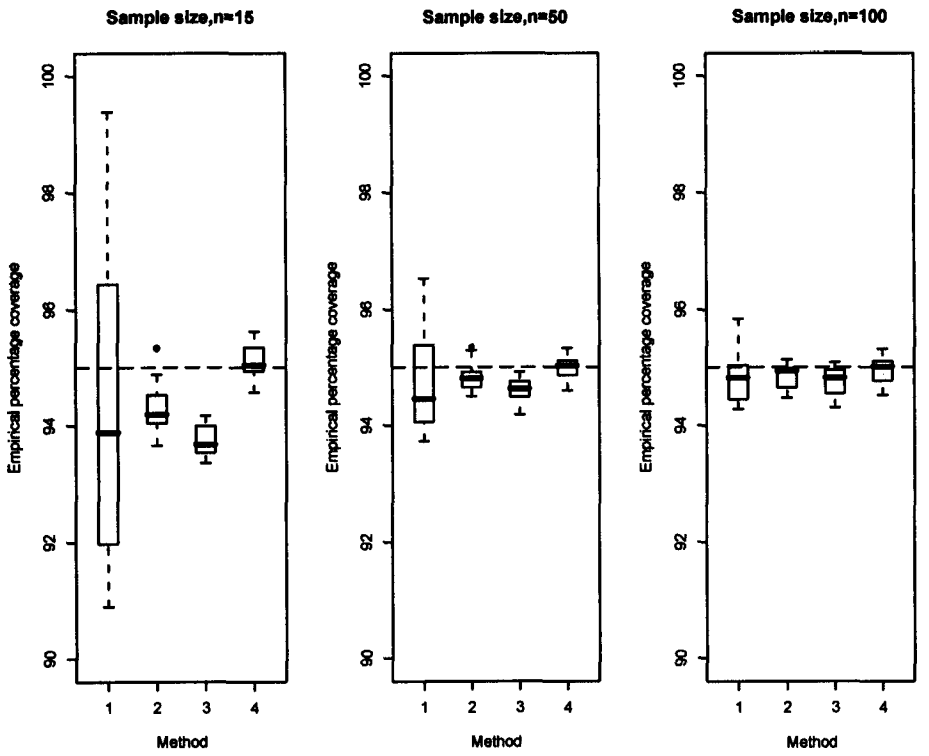

Figure 4.2: Mean coverage percentage based on 10,000 runs for nominal 95% confidence intervals for a difference between two correlated ICCs using 4 confidence interval methods for single ICCs when $k_1 = 4, k_2 = 2$. Methods 1, 2, 3, and 4 represents SA, Fisher, Konishi and Exact procedures respectively. Each boxplot was drawn from coverage percentage of 20 parameter combinations.

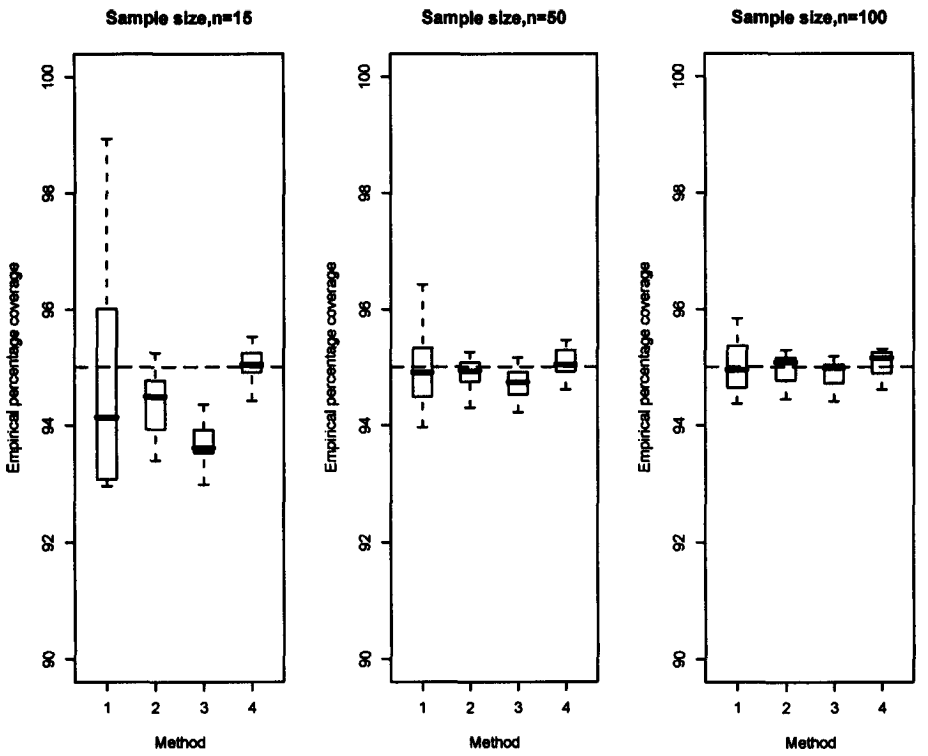

Figure 4.3: Mean coverage percentage based on 10,000 runs for nominal 95% confidence intervals for a difference between two correlated ICCs using 4 confidence interval methods for single ICCs when $k_1 = k_2 = 4$. Methods 1, 2, 3, and 4 represents SA, Fisher, Konishi and Exact procedures respectively. Each boxplot was drawn from coverage percentage of 20 parameter combinations.

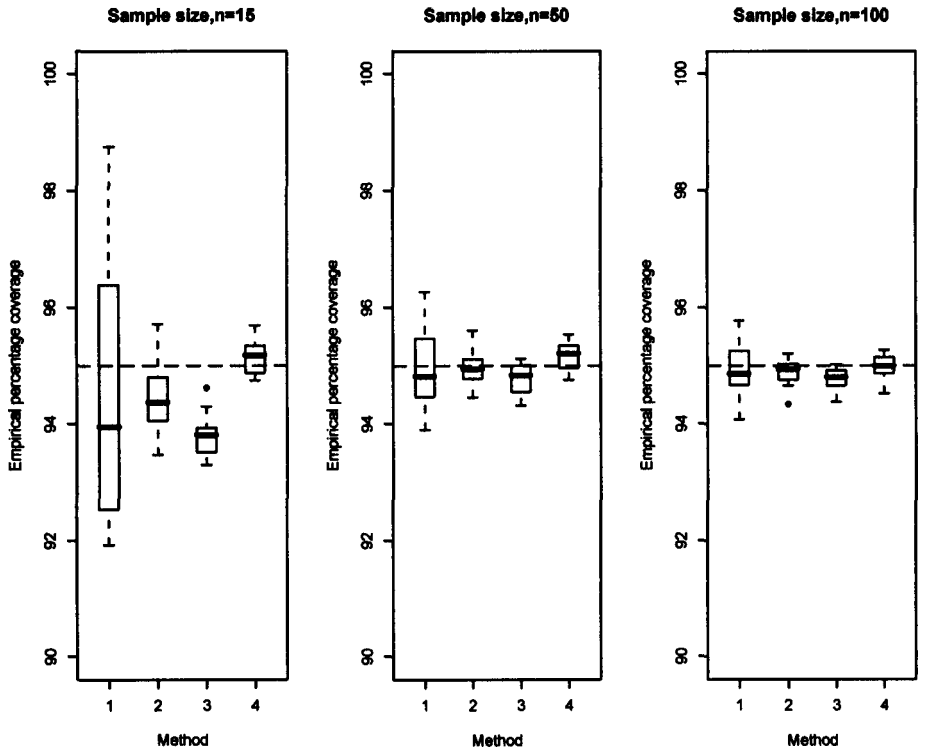

Figure 4.4: Mean coverage percentage based on 10,000 runs for nominal 95% confidence intervals for a difference between two correlated ICCs using 4 confidence interval methods for single ICCs when $k_1 = 6, k_2 = 3$. Methods 1, 2, 3, and 4 represents SA, Fisher, Konishi and Exact procedures respectively. Each boxplot was drawn from coverage percentage of 20 parameter combinations.

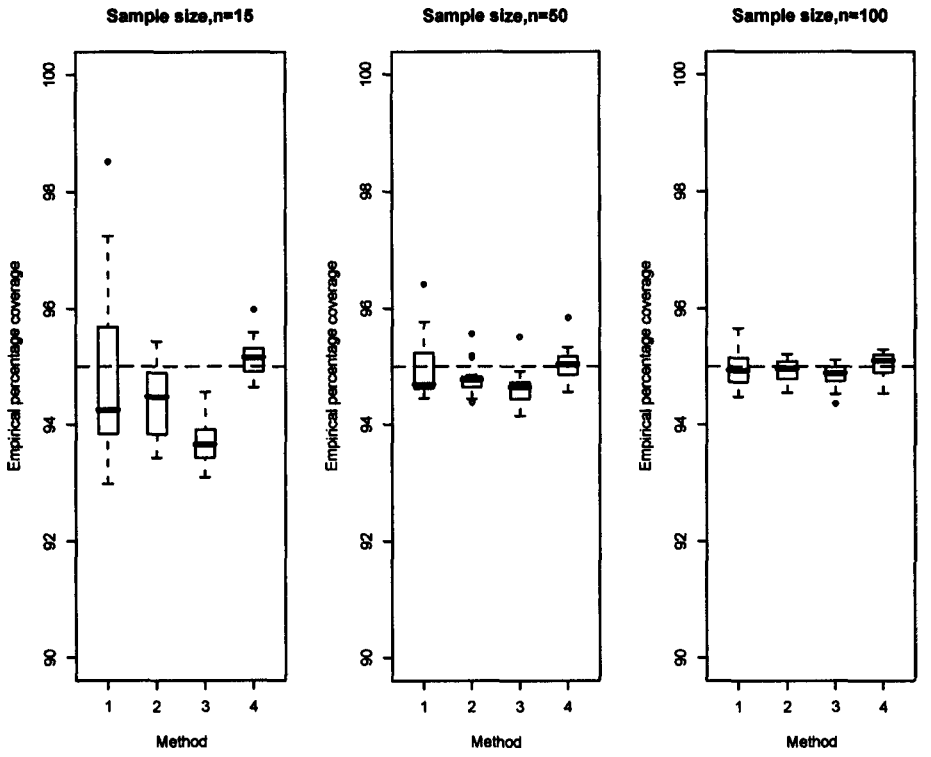

Figure 4.5: Mean coverage percentage based on 10,000 runs for nominal 95% confidence intervals for a difference between two correlated ICCs using 4 confidence interval methods for single ICCs when $k_1 =, k_2 = 6$. Methods 1, 2, 3, and 4 represents SA, Fisher, Konishi and Exact procedures respectively. Each boxplot was drawn from coverage percentage of 20 parameter combinations.

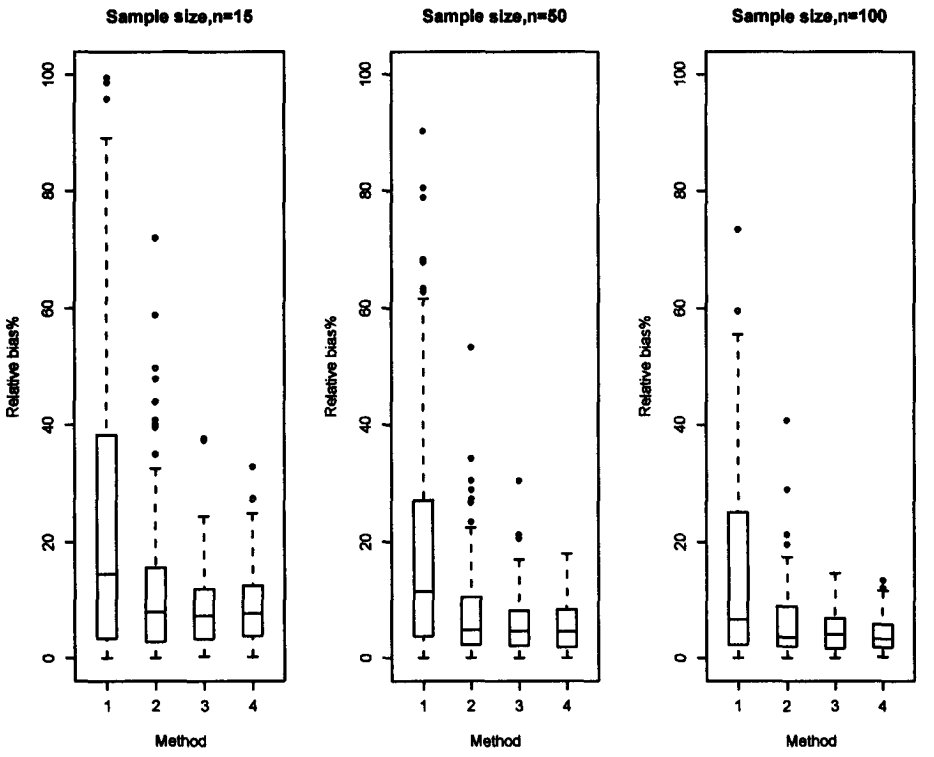

Figure 4.6: Imbalance of tail errors, quantified by the relative bias % $[100|\mathrm{MR}-\mathrm{ML}|/(\mathrm{MR}+\mathrm{ML})]$, of 95% nominal confidence intervals for a difference between two correlated ICCs using 4 confidence interval methods for single ICCs. Methods 1, 2, 3, and 4 represents SA, Fisher, Konishi and Exact procedures respectively. Each boxplot was drawn from coverage percentage of 100 parameter combinations.

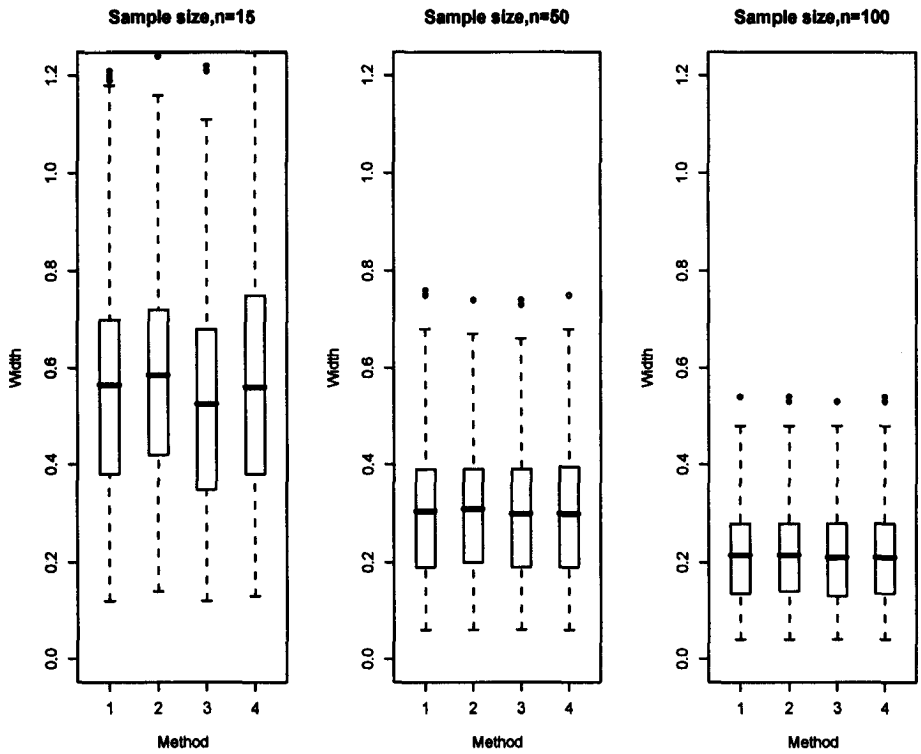

Figure 4.7: Confidence interval width of 95% nominal confidence intervals for a difference between two correlated ICCs using 4 confidence interval methods for single ICCs. Methods 1, 2, 3, and 4 represents SA, Fisher, Konishi and Exact procedures respectively. Each boxplot was drawn from coverage percentage of 100 parameter combinations.

Table 4.1: The performance of the new approach for constructing two-sided 95% confidence intervals (CI) for a difference between two correlated intraclass correlation coefficients based on 10000 runs when sample size $n = 15$. The lower and upper bound of single ICCs were calculated using SA, Fisher, Konishi and Exact method. Ideally missing left (ML) and missing right (MR) should be 2.50%.

$n = 15$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width
$k_1 = 2$	$k_2 = 2$								
0.50	0.50	95.60 (2.13, 2.27)	0.99	94.50 (2.64, 2.86)	0.99	94.04 (2.87, 3.09)	0.94	95.54 (2.12, 2.34)	1.02
	0.52	95.43 (2.30, 2.27)	0.98	94.21 (2.91, 2.88)	0.98	93.61 (3.27, 3.12)	0.93	95.13 (2.51, 2.36)	1.01
	0.54	95.44 (2.53, 2.03)	0.96	94.14 (3.21, 2.65)	0.96	93.60 (3.56, 2.84)	0.91	95.13 (2.71, 2.16)	1.00
	0.56	95.59 (2.20, 2.21)	0.95	94.25 (3.03, 2.72)	0.95	93.67 (3.45, 2.88)	0.90	95.18 (2.55, 2.27)	0.98
0.70	0.70	97.98 (0.93, 1.09)	0.67	94.08 (2.83, 3.09)	0.73	93.37 (3.21, 3.42)	0.67	94.97 (2.42, 2.61)	0.76

Continued on next page

Table 4.1 – continued from previous page

$n = 15$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	97.75 (1.27, 0.98)	0.66	94.09 (3.03, 2.88)	0.71	93.39 (3.46, 3.15)	0.66	94.99 (2.58, 2.43)	0.74
	0.74	97.85 (1.33, 0.82)	0.64	94.30 (3.20, 2.50)	0.70	93.71 (3.66, 2.63)	0.64	95.16 (2.72, 2.12)	0.72
	0.76	97.53 (1.65, 0.82)	0.62	94.18 (3.07, 2.75)	0.68	93.48 (3.73, 2.79)	0.63	95.10 (2.67, 2.23)	0.70
0.90	0.90	99.88 (0.07, 0.05)	0.26	94.19 (2.99, 2.82)	0.32	93.31 (3.38, 3.31)	0.29	95.12 (2.47, 2.41)	0.34
	0.92	99.25 (0.70, 0.05)	0.23	94.36 (3.01, 2.63)	0.29	93.42 (3.83, 2.75)	0.26	95.17 (2.60, 2.23)	0.30
	0.94	95.59 (4.38, 0.03)	0.22	94.52 (3.28, 2.20)	0.27	93.62 (4.38, 2.00)	0.24	95.13 (3.10, 1.77)	0.28
	0.96	92.18 (7.80, 0.02)	0.20	94.49 (3.14, 2.37)	0.25	93.68 (4.35, 1.97)	0.22	95.36 (2.75, 1.89)	0.26
$k_1 = 4$	$k_2 = 2$								
0.50	0.50	93.26 (1.64, 5.10)	0.81	94.09 (2.50, 3.41)	0.81	93.70 (2.65, 3.65)	0.75	94.98 (2.55, 2.47)	0.81
	0.52	94.09 (1.50, 4.41)	0.79	94.51 (2.51, 2.98)	0.80	94.16 (2.68, 3.16)	0.74	95.37 (2.53, 2.10)	0.80
	0.54	93.68 (1.51, 4.81)	0.78	94.10 (2.51, 3.39)	0.78	93.76 (2.70, 3.54)	0.72	95.05 (2.61, 2.34)	0.78
	0.56	94.26 (1.27, 4.47)	0.76	94.43 (2.25, 3.32)	0.77	93.98 (2.58, 3.44)	0.71	95.37 (2.46, 2.17)	0.77
0.70	0.70	95.41 (0.67, 3.92)	0.55	94.34 (2.38, 3.28)	0.60	93.62 (2.56, 3.82)	0.54	95.31 (2.43, 2.26)	0.59

Continued on next page

Table 4.1 – continued from previous page

$n = 15$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	96.19 (0.44, 3.37)	0.53	94.62 (2.10, 3.28)	0.58	94.13 (2.33, 3.54)	0.52	95.32 (2.28, 2.40)	0.57
	0.74	96.70 (0.42, 2.88)	0.51	94.91 (2.17, 2.92)	0.56	94.17 (2.72, 3.11)	0.50	95.56 (2.43, 2.01)	0.55
	0.76	96.80 (0.58, 2.62)	0.49	94.21 (2.61, 3.18)	0.55	93.50 (3.34, 3.16)	0.48	94.77 (3.03, 2.20)	0.53
0.90	0.90	98.09 (0.04, 1.87)	0.21	94.38 (2.27, 3.35)	0.26	93.40 (2.50, 4.10)	0.23	94.97 (2.34, 2.69)	0.26
	0.92	99.39 (0.06, 0.55)	0.19	94.69 (2.32, 2.99)	0.24	93.54 (3.32, 3.14)	0.20	95.03 (2.91, 2.06)	0.23
	0.94	98.14 (1.59, 0.27)	0.17	95.30 (2.05, 2.65)	0.21	94.02 (3.57, 2.41)	0.18	95.48 (3.00, 1.52)	0.20
	0.96	95.54 (4.14, 0.32)	0.16	95.33 (1.40, 3.27)	0.19	94.04 (3.40, 2.56)	0.16	95.62 (2.51, 1.87)	0.18
$k_1 = 4$	$k_2 = 4$								
0.50	0.50	94.08 (2.98, 2.94)	0.58	94.42 (2.80, 2.78)	0.59	93.93 (3.06, 3.01)	0.54	95.10 (2.42, 2.48)	0.56
	0.52	94.53 (2.74, 2.73)	0.57	94.89 (2.61, 2.50)	0.59	94.36 (2.93, 2.71)	0.53	95.52 (2.40, 2.08)	0.56
	0.54	94.30 (2.95, 2.75)	0.57	94.83 (2.63, 2.54)	0.58	94.02 (3.24, 2.74)	0.52	95.35 (2.57, 2.08)	0.55
	0.56	94.18 (3.15, 2.67)	0.56	94.63 (2.95, 2.42)	0.57	93.91 (3.60, 2.49)	0.51	95.27 (2.92, 1.81)	0.54
0.70	0.70	96.07 (1.92, 2.01)	0.41	94.90 (2.42, 2.68)	0.45	93.91 (2.94, 3.15)	0.38	95.20 (2.29, 2.51)	0.41

Continued on next page

Table 4.1 – continued from previous page

$n = 15$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.95 (2.26, 1.79)	0.40	94.81 (2.70, 2.49)	0.44	93.60 (3.54, 2.86)	0.37	95.07 (2.85, 2.08)	0.40
	0.74	96.12 (2.23, 1.65)	0.39	94.90 (2.47, 2.63)	0.43	93.87 (3.42, 2.71)	0.36	95.06 (2.78, 2.16)	0.39
	0.76	96.27 (2.43, 1.30)	0.38	95.38 (2.28, 2.34)	0.42	94.18 (3.62, 2.20)	0.35	95.34 (2.91, 1.75)	0.38
0.90	0.90	98.93 (0.51, 0.56)	0.16	95.09 (2.26, 2.65)	0.20	93.33 (3.12, 3.55)	0.16	94.72 (2.46, 2.82)	0.17
	0.92	97.80 (1.99, 0.21)	0.15	95.34 (2.05, 2.61)	0.19	93.52 (3.77, 2.71)	0.15	95.02 (2.91, 2.07)	0.16
	0.94	94.84 (4.87, 0.29)	0.14	95.00 (1.69, 3.31)	0.17	93.70 (3.77, 2.53)	0.14	95.18 (2.84, 1.98)	0.15
	0.96	93.03 (6.59, 0.38)	0.14	94.35 (1.42, 4.23)	0.16	94.18 (3.31, 2.51)	0.14	95.38 (2.58, 2.04)	0.15
$k_1 = 6$	$k_2 = 3$								
0.50	0.50	93.58 (2.47, 3.95)	0.59	94.24 (3.01, 2.75)	0.60	93.74 (2.78, 3.48)	0.55	95.04 (2.28, 2.68)	0.58
	0.52	93.82 (2.36, 3.82)	0.59	94.56 (2.93, 2.51)	0.60	93.86 (2.80, 3.34)	0.54	95.33 (2.29, 2.38)	0.57
	0.54	94.57 (2.26, 3.17)	0.58	95.02 (2.79, 2.19)	0.59	94.62 (2.74, 2.64)	0.53	95.69 (2.31, 2.00)	0.56
	0.56	94.20 (2.29, 3.51)	0.56	94.52 (3.07, 2.41)	0.58	93.90 (3.16, 2.94)	0.52	95.45 (2.56, 1.99)	0.55
0.70	0.70	95.55 (1.36, 3.09)	0.42	94.47 (2.78, 2.75)	0.45	93.64 (2.83, 3.53)	0.39	94.87 (2.39, 2.74)	0.41

Continued on next page

Table 4.1 – continued from previous page

$n = 15$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	96.50 (1.24, 2.26)	0.40	94.95 (2.81, 2.24)	0.44	94.03 (3.10, 2.87)	0.38	95.35 (2.63, 2.02)	0.40
	0.74	96.29 (1.24, 2.47)	0.39	94.94 (2.37, 2.69)	0.43	93.99 (2.87, 3.14)	0.37	95.20 (2.43, 2.37)	0.39
	0.76	96.64 (1.30, 2.06)	0.38	94.95 (2.47, 2.58)	0.42	93.83 (3.40, 2.77)	0.35	95.21 (2.87, 1.92)	0.38
0.90	0.90	98.36 (0.20, 1.44)	0.16	95.13 (2.37, 2.50)	0.20	93.40 (2.93, 3.67)	0.17	94.78 (2.51, 2.71)	0.18
	0.92	98.75 (0.61, 0.64)	0.14	95.31 (1.96, 2.73)	0.18	93.29 (3.67, 3.04)	0.15	94.75 (2.96, 2.29)	0.16
	0.94	96.59 (2.94, 0.47)	0.13	95.81 (1.24, 2.95)	0.17	94.30 (3.26, 2.44)	0.14	95.61 (2.52, 1.87)	0.15
	0.96	94.08 (5.41, 0.51)	0.13	94.66 (1.10, 4.24)	0.16	93.82 (3.43, 2.75)	0.13	95.27 (2.68, 2.05)	0.14
$k_1 = 6$	$k_2 = 6$								
0.50	0.50	93.82 (3.19, 2.99)	0.47	94.67 (2.77, 2.56)	0.48	93.87 (3.18, 2.95)	0.44	95.35 (2.40, 2.25)	0.46
	0.52	94.25 (3.12, 2.63)	0.47	94.95 (2.70, 2.35)	0.48	94.08 (3.29, 2.63)	0.43	95.56 (2.45, 1.99)	0.46
	0.54	94.72 (2.56, 2.72)	0.46	95.41 (2.27, 2.32)	0.47	94.56 (2.90, 2.54)	0.43	95.98 (2.17, 1.85)	0.45
	0.56	94.25 (2.90, 2.85)	0.45	94.99 (2.47, 2.54)	0.47	93.93 (3.38, 2.69)	0.42	95.53 (2.45, 2.02)	0.44
0.70	0.70	95.68 (2.02, 2.30)	0.34	95.14 (2.39, 2.47)	0.37	93.71 (3.12, 3.17)	0.31	95.16 (2.39, 2.45)	0.33

Continued on next page

Table 4.1 – continued from previous page

$n = 15$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.89 (2.29, 1.82)	0.33	95.11 (2.44, 2.45)	0.36	93.55 (3.48, 2.97)	0.30	95.10 (2.76, 2.14)	0.32
	0.74	95.84 (2.50, 1.66)	0.32	95.36 (2.28, 2.36)	0.36	93.88 (3.62, 2.50)	0.30	95.20 (2.93, 1.87)	0.31
	0.76	95.68 (2.63, 1.69)	0.31	95.42 (1.88, 2.70)	0.35	93.70 (3.82, 2.48)	0.29	95.27 (2.88, 1.85)	0.31
0.90	0.90	98.51 (0.72, 0.77)	0.13	95.68 (2.05, 2.27)	0.17	93.08 (3.32, 3.60)	0.13	94.64 (2.56, 2.80)	0.14
	0.92	97.24 (2.39, 0.37)	0.12	95.62 (1.63, 2.75)	0.16	93.41 (3.75, 2.84)	0.12	94.72 (3.06, 2.22)	0.13
	0.94	94.20 (5.39, 0.41)	0.12	94.78 (1.36, 3.86)	0.15	93.44 (3.73, 2.83)	0.12	94.81 (3.06, 2.13)	0.13
	0.96	93.54 (5.98, 0.48)	0.13	94.16 (0.82, 5.02)	0.14	94.27 (2.95, 2.78)	0.12	95.58 (2.35, 2.07)	0.13

Table 4.2: The performance of the new approach for constructing two-sided 95% confidence intervals (CI) for a difference between two correlated intraclass correlation coefficients based on 10000 runs when sample size $n = 50$. The lower and upper bound of single ICCs were calculated using SA, Fisher, Konishi and Exact method. Ideally missing left (ML) and missing right (MR) should be 2.50%.

$n = 50$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width
$k_1 = 2$	$k_2 = 2$								
0.50	0.50	95.45 (2.38, 2.17)	0.54	95.06 (2.60, 2.34)	0.54	94.93 (2.66, 2.41)	0.53	95.38 (2.44, 2.18)	0.55
	0.52	95.40 (2.35, 2.25)	0.53	95.08 (2.46, 2.46)	0.53	94.98 (2.50, 2.52)	0.52	95.26 (2.37, 2.37)	0.54
	0.54	95.41 (2.51, 2.08)	0.52	95.11 (2.63, 2.26)	0.52	94.91 (2.76, 2.33)	0.51	95.33 (2.53, 2.14)	0.53
	0.56	95.30 (2.54, 2.16)	0.51	94.94 (2.69, 2.37)	0.52	94.81 (2.81, 2.38)	0.50	95.20 (2.57, 2.23)	0.52
0.70	0.70	95.92 (2.04, 2.04)	0.35	94.70 (2.73, 2.57)	0.36	94.52 (2.83, 2.65)	0.35	94.90 (2.62, 2.48)	0.36

Continued on next page

Table 4.2 – continued from previous page

$n = 50$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	96.25 (1.98, 1.77)	0.34	94.89 (2.52, 2.59)	0.35	94.74 (2.59, 2.67)	0.34	95.09 (2.46, 2.45)	0.35
	0.74	95.93 (2.38, 1.69)	0.33	94.92 (2.74, 2.34)	0.34	94.70 (2.95, 2.35)	0.33	95.15 (2.63, 2.22)	0.34
	0.76	95.71 (2.77, 1.52)	0.32	94.67 (2.91, 2.42)	0.33	94.48 (3.12, 2.40)	0.32	94.86 (2.82, 2.32)	0.33
0.90	0.90	97.05 (1.55, 1.40)	0.12	94.55 (2.95, 2.50)	0.13	94.36 (3.02, 2.62)	0.13	94.80 (2.80, 2.40)	0.14
	0.92	96.60 (2.68, 0.72)	0.11	94.91 (2.85, 2.24)	0.12	94.70 (3.10, 2.20)	0.12	95.09 (2.74, 2.17)	0.12
	0.94	95.28 (4.26, 0.46)	0.10	94.88 (2.83, 2.29)	0.11	94.64 (3.25, 2.11)	0.11	95.19 (2.67, 2.14)	0.11
	0.96	93.83 (5.87, 0.30)	0.10	95.24 (2.78, 1.98)	0.11	95.06 (3.22, 1.72)	0.10	95.49 (2.66, 1.85)	0.11
$k_1 = 4$	$k_2 = 2$								
0.50	0.50	94.47 (1.69, 3.84)	0.44	94.87 (2.29, 2.84)	0.44	94.73 (2.35, 2.92)	0.43	95.21 (2.44, 2.35)	0.44
	0.52	94.44 (1.55, 4.01)	0.43	94.68 (2.38, 2.94)	0.43	94.53 (2.44, 3.03)	0.42	94.95 (2.56, 2.49)	0.43
	0.54	94.40 (1.67, 3.93)	0.42	94.72 (2.46, 2.82)	0.42	94.61 (2.57, 2.82)	0.41	94.88 (2.70, 2.42)	0.42
	0.56	94.59 (1.65, 3.76)	0.41	94.65 (2.47, 2.88)	0.41	94.45 (2.70, 2.85)	0.40	94.66 (2.84, 2.50)	0.41
0.70	0.70	95.16 (1.04, 3.80)	0.29	94.81 (2.31, 2.88)	0.30	94.64 (2.33, 3.03)	0.29	95.08 (2.45, 2.47)	0.29

Continued on next page

Table 4.2 – continued from previous page

$n = 50$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.24 (1.10, 3.66)	0.27	94.68 (2.31, 3.01)	0.28	94.56 (2.39, 3.05)	0.27	94.96 (2.49, 2.55)	0.28
	0.74	95.50 (1.17, 3.33)	0.26	94.77 (2.34, 2.89)	0.27	94.60 (2.48, 2.92)	0.26	94.96 (2.49, 2.55)	0.27
	0.76	95.74 (1.19, 3.07)	0.25	94.99 (2.16, 2.85)	0.26	94.82 (2.39, 2.79)	0.25	95.16 (2.41, 2.43)	0.26
0.90	0.90	96.12 (0.71, 3.17)	0.10	94.63 (2.57, 2.80)	0.11	94.36 (2.61, 3.03)	0.11	94.76 (2.70, 2.54)	0.11
	0.92	96.54 (1.34, 2.12)	0.09	94.60 (2.61, 2.79)	0.10	94.18 (3.03, 2.79)	0.09	94.60 (3.01, 2.39)	0.10
	0.94	96.21 (2.67, 1.12)	0.08	95.28 (2.25, 2.47)	0.09	94.70 (3.07, 2.23)	0.08	95.13 (2.84, 2.03)	0.09
	0.96	95.27 (3.85, 0.88)	0.08	95.28 (1.73, 2.99)	0.08	94.79 (2.82, 2.39)	0.08	95.26 (2.50, 2.24)	0.08
$k_1 = 4$	$k_2 = 4$								
0.50	0.50	94.98 (2.44, 2.58)	0.32	95.12 (2.38, 2.50)	0.32	94.99 (2.45, 2.56)	0.31	95.31 (2.30, 2.39)	0.31
	0.52	95.08 (2.47, 2.45)	0.31	95.16 (2.42, 2.42)	0.31	94.99 (2.57, 2.44)	0.30	95.37 (2.37, 2.26)	0.31
	0.54	94.82 (2.60, 2.58)	0.31	94.93 (2.53, 2.54)	0.31	94.77 (2.70, 2.53)	0.30	95.06 (2.53, 2.41)	0.30
	0.56	95.24 (2.46, 2.30)	0.30	95.31 (2.40, 2.29)	0.30	95.16 (2.57, 2.27)	0.29	95.46 (2.45, 2.09)	0.30
0.70	0.70	95.35 (2.24, 2.41)	0.21	94.98 (2.46, 2.56)	0.22	94.71 (2.62, 2.67)	0.21	95.02 (2.44, 2.54)	0.21

Continued on next page

Table 4.2 – continued from previous page

$n = 50$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.59 (2.27, 2.14)	0.20	95.25 (2.36, 2.39)	0.21	94.85 (2.75, 2.40)	0.20	95.25 (2.46, 2.29)	0.20
	0.74	95.69 (2.39, 1.92)	0.20	95.31 (2.31, 2.38)	0.21	94.96 (2.72, 2.32)	0.20	95.35 (2.49, 2.16)	0.20
	0.76	95.30 (2.73, 1.97)	0.19	95.01 (2.41, 2.58)	0.20	94.65 (2.98, 2.37)	0.19	94.98 (2.75, 2.27)	0.19
0.90	0.90	96.43 (1.85, 1.72)	0.08	95.09 (2.56, 2.35)	0.08	94.48 (2.94, 2.58)	0.08	94.92 (2.65, 2.43)	0.08
	0.92	95.99 (2.83, 1.18)	0.07	94.93 (2.33, 2.74)	0.08	94.53 (2.88, 2.59)	0.07	94.93 (2.68, 2.39)	0.07
	0.94	95.08 (4.40, 0.52)	0.07	94.98 (2.13, 2.89)	0.07	94.76 (3.16, 2.08)	0.07	95.26 (2.77, 1.97)	0.07
	0.96	94.02 (5.03, 0.95)	0.07	94.63 (1.87, 3.50)	0.07	94.42 (2.94, 2.64)	0.07	94.75 (2.65, 2.60)	0.07
$k_1 = 6$	$k_2 = 3$								
0.50	0.50	94.75 (2.16, 3.09)	0.32	94.87 (2.64, 2.49)	0.32	94.84 (2.37, 2.79)	0.31	95.18 (2.34, 2.48)	0.32
	0.52	94.93 (2.12, 2.95)	0.32	95.01 (2.62, 2.37)	0.32	94.90 (2.48, 2.62)	0.31	95.24 (2.40, 2.36)	0.31
	0.54	94.88 (2.22, 2.90)	0.31	94.92 (2.73, 2.35)	0.31	94.83 (2.63, 2.54)	0.30	95.16 (2.56, 2.28)	0.31
	0.56	95.02 (2.10, 2.88)	0.30	95.22 (2.47, 2.31)	0.31	95.11 (2.42, 2.47)	0.30	95.39 (2.40, 2.21)	0.30
0.70	0.70	95.50 (1.72, 2.78)	0.22	95.23 (2.43, 2.34)	0.22	95.05 (2.37, 2.58)	0.21	95.32 (2.32, 2.36)	0.22

Continued on next page

Table 4.2 – continued from previous page

$n = 50$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.71 (1.61, 2.68)	0.21	95.28 (2.39, 2.33)	0.22	94.97 (2.43, 2.60)	0.20	95.35 (2.37, 2.28)	0.21
	0.74	95.44 (1.89, 2.67)	0.20	94.93 (2.53, 2.54)	0.21	94.64 (2.71, 2.65)	0.20	95.01 (2.61, 2.38)	0.20
	0.76	95.27 (1.97, 2.76)	0.19	94.73 (2.50, 2.77)	0.20	94.45 (2.78, 2.77)	0.19	94.91 (2.66, 2.43)	0.19
0.90	0.90	95.88 (1.31, 2.81)	0.08	94.84 (2.51, 2.65)	0.09	94.37 (2.61, 3.02)	0.08	94.77 (2.51, 2.72)	0.08
	0.92	96.28 (2.07, 1.65)	0.07	95.09 (2.39, 2.52)	0.08	94.51 (2.99, 2.50)	0.07	94.89 (2.86, 2.25)	0.07
	0.94	95.73 (3.24, 1.03)	0.06	95.62 (1.74, 2.64)	0.07	95.12 (2.76, 2.12)	0.07	95.54 (2.54, 1.92)	0.07
	0.96	94.68 (4.30, 1.02)	0.06	94.93 (1.67, 3.40)	0.07	94.83 (2.75, 2.42)	0.06	95.22 (2.55, 2.23)	0.07
$k_1 = 6$	$k_2 = 6$								
0.50	0.50	95.44 (2.14, 2.42)	0.25	95.70 (2.05, 2.25)	0.26	95.51 (2.11, 2.38)	0.25	95.84 (1.99, 2.17)	0.25
	0.52	94.65 (2.66, 2.69)	0.25	94.87 (2.54, 2.59)	0.25	94.68 (2.70, 2.62)	0.24	95.06 (2.50, 2.44)	0.25
	0.54	94.64 (2.77, 2.59)	0.25	94.91 (2.60, 2.49)	0.25	94.63 (2.87, 2.50)	0.24	95.03 (2.61, 2.36)	0.25
	0.56	94.70 (2.79, 2.51)	0.24	94.94 (2.66, 2.40)	0.25	94.71 (2.93, 2.36)	0.24	95.05 (2.77, 2.18)	0.24
0.70	0.70	95.49 (2.31, 2.20)	0.17	95.31 (2.40, 2.29)	0.18	94.92 (2.59, 2.49)	0.17	95.33 (2.38, 2.29)	0.17

Continued on next page

Table 4.2 – continued from previous page

$n = 50$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.29 (2.63, 2.08)	0.17	94.93 (2.60, 2.47)	0.18	94.51 (2.96, 2.53)	0.16	95.00 (2.77, 2.23)	0.17
	0.74	95.17 (2.75, 2.08)	0.16	95.00 (2.44, 2.56)	0.17	94.52 (3.00, 2.48)	0.16	94.96 (2.83, 2.21)	0.16
	0.76	94.75 (3.10, 2.15)	0.16	94.60 (2.73, 2.67)	0.17	94.15 (3.41, 2.44)	0.16	94.56 (3.18, 2.26)	0.16
0.90	0.90	96.42 (1.74, 1.84)	0.06	95.26 (2.40, 2.34)	0.07	94.46 (2.88, 2.66)	0.06	94.86 (2.62, 2.52)	0.07
	0.92	95.77 (2.94, 1.29)	0.06	94.85 (2.11, 3.04)	0.07	94.41 (2.81, 2.78)	0.06	94.76 (2.67, 2.57)	0.06
	0.94	94.67 (4.47, 0.86)	0.06	94.86 (1.87, 3.27)	0.06	94.45 (3.22, 2.33)	0.06	94.88 (3.01, 2.11)	0.06
	0.96	94.92 (4.06, 1.02)	0.06	94.48 (1.29, 4.23)	0.06	94.73 (2.42, 2.85)	0.06	95.20 (2.26, 2.54)	0.06

Table 4.3: The performance of the new approach for constructing two-sided 95% confidence intervals (CI) for a difference between two correlated intraclass correlation coefficients based on 10000 runs when sample size $n = 100$. The lower and upper bound of single ICCs were calculated using SA, Fisher, Konishi and Exact method. Ideally missing left (ML) and missing right (MR) should be 2.50%.

$n = 100$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width	Coverage (%) (ML, MR) (%)	CI Width
$k_1 = 2$	$k_2 = 2$								
0.50	0.50	95.12 (2.43, 2.45)	0.38	94.94 (2.50, 2.56)	0.38	94.85 (2.56, 2.59)	0.38	95.05 (2.44, 2.51)	0.38
	0.52	95.31 (2.31, 2.38)	0.37	95.02 (2.42, 2.56)	0.37	94.97 (2.45, 2.58)	0.37	95.21 (2.33, 2.46)	0.38
	0.54	95.31 (2.43, 2.26)	0.37	95.08 (2.49, 2.43)	0.37	94.99 (2.56, 2.45)	0.36	95.21 (2.42, 2.37)	0.37
	0.56	95.23 (2.38, 2.39)	0.36	95.06 (2.43, 2.51)	0.36	94.96 (2.53, 2.51)	0.36	95.18 (2.37, 2.45)	0.36
0.70	0.70	95.63 (2.15, 2.22)	0.24	95.00 (2.47, 2.53)	0.25	94.91 (2.55, 2.54)	0.24	95.15 (2.38, 2.47)	0.25

Continued on next page

Table 4.3 – continued from previous page

$n = 100$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.67 (2.31, 2.02)	0.23	95.03 (2.49, 2.48)	0.24	94.94 (2.56, 2.50)	0.24	95.13 (2.42, 2.45)	0.24
	0.74	95.46 (2.64, 1.90)	0.23	94.84 (2.72, 2.44)	0.23	94.68 (2.88, 2.44)	0.23	94.91 (2.70, 2.39)	0.23
	0.76	95.07 (3.06, 1.87)	0.22	94.47 (3.03, 2.50)	0.23	94.32 (3.21, 2.47)	0.22	94.56 (3.00, 2.44)	0.23
0.90	0.90	96.14 (1.93, 1.93)	0.08	94.85 (2.60, 2.55)	0.09	94.74 (2.61, 2.65)	0.09	95.05 (2.49, 2.46)	0.09
	0.92	96.22 (2.57, 1.21)	0.08	95.15 (2.48, 2.37)	0.08	95.09 (2.59, 2.32)	0.08	95.27 (2.44, 2.29)	0.08
	0.94	95.11 (3.90, 0.99)	0.07	94.80 (2.72, 2.48)	0.08	94.60 (3.06, 2.34)	0.07	94.89 (2.65, 2.46)	0.08
	0.96	94.52 (4.75, 0.73)	0.07	94.94 (2.53, 2.53)	0.07	94.93 (2.81, 2.26)	0.07	95.09 (2.46, 2.45)	0.07
$k_1 = 4$	$k_2 = 2$								
0.50	0.50	94.68 (1.84, 3.48)	0.31	94.89 (2.46, 2.65)	0.31	94.83 (2.48, 2.69)	0.31	95.07 (2.52, 2.41)	0.31
	0.52	94.81 (1.63, 3.56)	0.30	95.02 (2.15, 2.83)	0.30	94.97 (2.17, 2.86)	0.30	95.15 (2.29, 2.56)	0.30
	0.54	94.91 (1.83, 3.26)	0.29	94.94 (2.47, 2.59)	0.30	94.85 (2.56, 2.59)	0.29	95.01 (2.67, 2.32)	0.30
	0.56	94.83 (1.66, 3.51)	0.29	95.01 (2.19, 2.80)	0.29	94.95 (2.26, 2.79)	0.29	95.11 (2.37, 2.52)	0.29
0.70	0.70	94.84 (1.62, 3.54)	0.20	94.62 (2.58, 2.80)	0.20	94.55 (2.58, 2.87)	0.20	94.75 (2.68, 2.57)	0.20

Continued on next page

Table 4.3 – continued from previous page

$n = 100$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	94.87 (1.81, 3.32)	0.19	94.48 (2.77, 2.75)	0.20	94.40 (2.82, 2.78)	0.19	94.52 (2.90, 2.58)	0.19
	0.74	95.16 (1.66, 3.18)	0.18	94.55 (2.64, 2.81)	0.19	94.46 (2.77, 2.77)	0.18	94.67 (2.82, 2.51)	0.19
	0.76	95.40 (1.74, 2.86)	0.17	95.08 (2.40, 2.52)	0.18	94.98 (2.57, 2.45)	0.17	95.15 (2.62, 2.23)	0.18
0.90	0.90	95.75 (1.10, 3.15)	0.07	95.10 (2.28, 2.62)	0.07	94.96 (2.28, 2.76)	0.07	95.17 (2.35, 2.48)	0.07
	0.92	95.84 (1.73, 2.43)	0.06	94.95 (2.45, 2.60)	0.06	94.83 (2.61, 2.56)	0.06	94.96 (2.63, 2.41)	0.06
	0.94	95.46 (2.72, 1.82)	0.05	94.62 (2.37, 3.01)	0.06	94.31 (2.95, 2.74)	0.06	94.65 (2.83, 2.52)	0.06
	0.96	94.88 (3.72, 1.40)	0.05	94.95 (2.18, 2.87)	0.05	94.78 (2.86, 2.36)	0.05	94.96 (2.72, 2.32)	0.05
$k_1 = 4$	$k_2 = 4$								
0.50	0.50	95.18 (2.44, 2.38)	0.22	95.21 (2.44, 2.35)	0.22	95.18 (2.45, 2.37)	0.22	95.29 (2.38, 2.33)	0.22
	0.52	94.61 (2.92, 2.47)	0.22	94.73 (2.90, 2.37)	0.22	94.59 (3.01, 2.40)	0.22	94.84 (2.89, 2.27)	0.22
	0.54	95.01 (2.52, 2.47)	0.22	95.02 (2.52, 2.46)	0.22	94.99 (2.55, 2.46)	0.21	95.12 (2.52, 2.36)	0.21
	0.56	94.68 (2.72, 2.60)	0.21	94.68 (2.70, 2.62)	0.21	94.67 (2.80, 2.53)	0.21	94.86 (2.73, 2.41)	0.21
0.70	0.70	95.30 (2.45, 2.25)	0.15	95.18 (2.51, 2.31)	0.15	95.03 (2.57, 2.40)	0.15	95.20 (2.49, 2.31)	0.15

Continued on next page

Table 4.3 – continued from previous page

$n = 100$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.43 (2.35, 2.22)	0.14	95.16 (2.37, 2.47)	0.15	94.99 (2.55, 2.46)	0.14	95.19 (2.45, 2.36)	0.14
	0.74	95.52 (2.61, 1.87)	0.14	95.28 (2.47, 2.25)	0.14	95.05 (2.77, 2.18)	0.14	95.27 (2.65, 2.08)	0.14
	0.76	95.54 (2.51, 1.95)	0.14	95.27 (2.29, 2.44)	0.14	95.13 (2.61, 2.26)	0.13	95.31 (2.49, 2.20)	0.14
0.90	0.90	95.84 (2.07, 2.09)	0.05	95.17 (2.36, 2.47)	0.06	94.98 (2.50, 2.52)	0.05	95.12 (2.38, 2.50)	0.05
	0.92	95.78 (2.73, 1.49)	0.05	95.30 (2.03, 2.67)	0.05	95.05 (2.52, 2.43)	0.05	95.24 (2.41, 2.35)	0.05
	0.94	94.94 (3.80, 1.26)	0.05	95.10 (1.97, 2.93)	0.05	94.78 (2.79, 2.43)	0.05	95.05 (2.61, 2.34)	0.05
	0.96	94.61 (4.19, 1.20)	0.05	95.01 (2.06, 2.93)	0.05	94.92 (2.80, 2.28)	0.05	95.15 (2.58, 2.27)	0.05
$k_1 = 6$	$k_2 = 3$								
0.50	0.50	94.88 (2.16, 2.96)	0.23	94.89 (2.54, 2.57)	0.23	94.89 (2.34, 2.77)	0.22	95.09 (2.34, 2.57)	0.23
	0.52	94.85 (2.23, 2.92)	0.22	95.00 (2.56, 2.44)	0.22	94.80 (2.48, 2.72)	0.22	95.13 (2.46, 2.41)	0.22
	0.54	94.88 (2.61, 2.51)	0.22	94.89 (2.90, 2.21)	0.22	94.89 (2.85, 2.26)	0.22	94.99 (2.84, 2.17)	0.22
	0.56	94.72 (2.34, 2.94)	0.21	94.93 (2.57, 2.50)	0.21	94.80 (2.56, 2.64)	0.21	95.04 (2.56, 2.40)	0.21
0.70	0.70	95.34 (1.79, 2.87)	0.15	95.23 (2.38, 2.39)	0.15	95.02 (2.34, 2.64)	0.15	95.27 (2.33, 2.40)	0.15

Continued on next page

Table 4.3 – continued from previous page

$n = 100$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.24 (2.05, 2.71)	0.14	95.13 (2.50, 2.37)	0.15	95.00 (2.49, 2.51)	0.14	95.22 (2.48, 2.30)	0.14
	0.74	94.90 (2.15, 2.95)	0.14	94.76 (2.48, 2.76)	0.14	94.56 (2.64, 2.80)	0.14	94.73 (2.62, 2.65)	0.14
	0.76	95.38 (2.24, 2.38)	0.13	95.06 (2.58, 2.36)	0.14	94.87 (2.82, 2.31)	0.13	95.00 (2.79, 2.21)	0.13
0.90	0.90	95.43 (1.62, 2.95)	0.05	94.82 (2.47, 2.71)	0.06	94.65 (2.47, 2.88)	0.05	94.80 (2.47, 2.73)	0.05
	0.92	95.78 (2.21, 2.01)	0.05	95.02 (2.43, 2.55)	0.05	94.74 (2.81, 2.45)	0.05	94.90 (2.77, 2.33)	0.05
	0.94	95.28 (3.13, 1.59)	0.04	94.86 (2.27, 2.87)	0.05	94.67 (2.82, 2.51)	0.04	94.87 (2.73, 2.40)	0.05
	0.96	94.82 (3.65, 1.53)	0.05	95.05 (1.76, 3.19)	0.05	94.98 (2.54, 2.48)	0.04	95.13 (2.48, 2.39)	0.05
$k_1 = 6$	$k_2 = 6$								
0.50	0.50	95.04 (2.47, 2.49)	0.18	95.11 (2.45, 2.44)	0.18	95.04 (2.46, 2.50)	0.18	95.20 (2.39, 2.41)	0.18
	0.52	94.79 (2.58, 2.63)	0.18	94.89 (2.55, 2.56)	0.18	94.80 (2.60, 2.60)	0.17	95.00 (2.55, 2.45)	0.18
	0.54	94.97 (2.50, 2.53)	0.17	95.06 (2.44, 2.50)	0.18	94.96 (2.55, 2.49)	0.17	95.13 (2.46, 2.41)	0.17
	0.56	94.71 (2.71, 2.58)	0.17	94.80 (2.65, 2.55)	0.17	94.64 (2.85, 2.51)	0.17	94.83 (2.75, 2.42)	0.17
0.70	0.70	95.35 (2.39, 2.26)	0.12	95.28 (2.41, 2.31)	0.12	95.11 (2.46, 2.43)	0.12	95.28 (2.41, 2.31)	0.12

Continued on next page

Table 4.3 – continued from previous page

$n = 100$		SA		Fisher		Konishi		Exact	
ρ_1	ρ_2	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width	Coverage (%) (ML, MR) (%)	Width
	0.72	95.05 (2.62, 2.33)	0.12	94.98 (2.55, 2.47)	0.12	94.76 (2.78, 2.46)	0.12	94.94 (2.67, 2.39)	0.12
	0.74	94.65 (2.84, 2.51)	0.11	94.60 (2.63, 2.77)	0.12	94.36 (2.96, 2.68)	0.11	94.54 (2.88, 2.58)	0.11
	0.76	95.22 (2.70, 2.08)	0.11	95.21 (2.26, 2.53)	0.11	94.93 (2.82, 2.25)	0.11	95.17 (2.68, 2.15)	0.11
0.90	0.90	95.42 (2.37, 2.21)	0.04	94.98 (2.64, 2.38)	0.05	94.79 (2.74, 2.47)	0.04	94.85 (2.70, 2.45)	0.04
	0.92	95.66 (2.85, 1.49)	0.04	95.23 (2.04, 2.73)	0.04	94.98 (2.63, 2.39)	0.04	95.20 (2.54, 2.26)	0.04
	0.94	94.94 (3.78, 1.28)	0.04	95.12 (1.92, 2.96)	0.04	94.91 (2.79, 2.30)	0.04	95.19 (2.65, 2.16)	0.04
	0.96	94.79 (3.91, 1.30)	0.04	94.87 (1.52, 3.61)	0.04	94.91 (2.56, 2.53)	0.04	95.16 (2.43, 2.41)	0.04

Table 4.4: Comparative performance of the four procedures for constructing a 95% two-sided confidence interval for single ICC (summary of 75 parameter combinations with 10000 runs for each combination)

		n = 15			n = 50			n = 100		
Method		Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
SA	C	91.63	89.85	94.50	93.96	93.00	95.30	94.47	93.92	95.37
	ML	3.23	0.00	8.52	2.38	0.31	5.29	2.24	0.71	4.14
	MR	5.13	0.64	9.52	3.68	0.94	6.40	3.28	1.29	5.08
	W	0.50	0.07	0.96	0.27	0.04	0.54	0.19	0.02	0.39
F	C	93.39	90.90	94.59	94.53	93.52	95.14	94.75	94.11	95.34
	ML	4.73	2.73	8.68	3.44	2.38	5.37	3.09	2.19	4.40
	MR	1.87	0.38	3.11	2.02	0.77	3.16	2.15	1.20	2.96
	W	0.49	0.08	0.89	0.27	0.04	0.53	0.19	0.02	0.38
K	C	93.56	93.03	94.02	94.60	93.97	95.07	94.77	94.08	95.30
	ML	2.88	2.13	3.68	2.49	1.99	2.99	3.09	2.19	4.40
	MR	3.56	2.77	4.29	2.90	2.44	3.70	2.77	2.35	3.29
	W	0.46	0.07	0.87	0.27	0.03	0.53	0.19	0.02	0.38
E	C	95.02	94.56	95.41	95.01	94.59	94.49	94.48	94.37	95.47
	ML	2.48	2.09	2.89	2.50	2.18	2.82	2.49	2.13	2.96
	MR	2.49	2.17	2.92	2.49	2.15	3.00	2.52	2.23	2.95
	W	0.50	0.07	0.92	0.27	0.04	0.53	0.19	0.02	0.38

SA: simple asymptotic method; F: Fisher's method; K: Konishi method; E: Exact method C: coverage; ML: the confidence interval lies completely below the parameter; MR: the confidence interval lies completely above the parameter; W: average confidence interval width; Min: minimum; Max: Maximum

4.4 Discussion of simulation results

4.4.1 Coverage

Sample size of n = 15

Simulation results in Table 4.1 indicate that empirical coverage provided by simple

asymptotic method was somewhat erratic. For example, when $\rho_1 = 0.7$ to 0.9 and $\rho_2 = 0.7$ to 0.94, the CIs provided by simple asymptotic method are too conservative and for $\rho_1 = 0.9$ and $\rho_2 = 0.96$, it is too liberal. But for $\rho_1, \rho_2 \leq 0.7$, simple asymptotic method provided adequate coverage close to the nominal level of 95%. It can also be seen that Fisher's method provided an empirical coverage within the desired interval of 94.6% to 95.4% for all parameter values but the coverage provided by this method were always less than the coverage provided by Exact method. Empirical coverage percentages provided by Konishi method were outside the desired interval of 94.6% to 95.4% and for some parameter values it even falls short of the moderate criterion of 93.75%. In contrast to the coverage percentages provided by these three methods, Exact method consistently provided excellent coverage percentages within the desired interval for all values of ρ_1 and ρ_2. Empirical coverage percentages provided by Exact method are much closer to the nominal level of 95%.

Sample size of $n = 50$

Results in Table 4.2 shows, when $k_1, k_2 = 2$, SA method provided empirical coverage close to the nominal coverage of 95% for ρ_1, $\rho_2 < 0.7$, but for $\rho_2 \geq 0.7$, the simple asymptotic method provided over coverage. The simulation results in the Table 4.2 shows that both Fisher and Konishi methods provided coverage within the desired interval of 94.6% to 95.4% but the coverage percentages provided by Konishi were always less than that provided by Fisher's method. Again Exact method provided an excellent empirical coverage percentages very close to the nominal coverage of 95%.

Sample size of n=100

The simulation results in the Table 4.3 shows that Fisher, Konish and Exact methods provided excellent empirical coverage within the desired interval which were very close to the nominal coverage of 95%. Simple asymptotic method provided empirical coverage within the desired interval except for some parameter values ($\rho_1, \rho_2 \geq 0.94$ for $k_1, k_2 = 2$) where the empirical coverage is outside the desired interval.

4.4.2 Tail errors

Sample size of n = 15

The simulation results in Table 4.1 show that the tail errors provided by the simple asymptotic method has a better balance only when $k_1 = k_2$, otherwise the tail errors were severely unbalanced, i.e., the tails errors were concentrated in one tail either left or right. Also the tail errors (ML and MR) were not close to the nominal level of 2.5%. The tail errors provided by the Fisher and Konishi methods also exceeded the desired level of 2.5% and the errors were not equally divided between the two tails. The Exact method provided tail errors close to nominal level of 2.5% for most parameter combinations.

Sample size of n = 50

The simulation results in Table 4.2 indicate that, the simple asymptotic method provided tail errors with an erratic pattern. The errors were not equally divided between two tails and ML and MR were not close to the nominal level. From this

table it can be seen that although Fisher method provided tail errors close to nominal level of 2.5% for $k_1 = k_2 = 2$, the errors were concentrated in one tail when k_1, $k_2 > 2$ and ρ_1, $\rho_2 > 0.9$. Konishi method always provided tail errors in excess of 2.5%. In contrast to the simple asymptotic, Fisher and Konishi methods, Exact method consistently provided tail errors close to nominal level of 2.5% except for $\rho_1, \rho_2 > 0.9$.

Sample size of $n = 100$

Results in Table 4.3 shows that for most parameter combinations, tail errors provided by the simple asymptotic method were not close to the nominal level of 2.5%. For higher values of ρ_1, ρ_2 errors were concentrated in one tail. Fisher and Exact methods provided tail errors very close to the nominal level while Konishi method exceeded the nominal level.

4.4.3 Confidence interval width

Sample size of $n = 15$

Table 4.1 shows that for $\rho_1, \rho_2 > 0.9$, CI width provided by the simple asymptotic method were always narrower than that provided by other three methods. Confidence interval widths provided by the Fisher method were narrower than that provided by the Exact method for k_1, $k_2 = 2$. However, for k_1, $k_2 > 2$, the CI width provided by the Fisher method were equal to or wider than that provided by the Exact method. Konishi method always provided a narrower CI than that provided by the other three methods except for $\rho_1, \rho_2 > 0.9$.

Sample size of $n = 50$

It can be observed that confidence interval width provided by all four methods were having near similar values once the sample size was increased to 50 except that the Fisher method provided narrower CI than the Exact method for k_1, $k_2 = 2$. Also the simple asymptotic method provides a wider CI for ρ_1, $\rho_2 > 0.9$.

Sample size of $n = 100$

From Table 4.3, it can be seen that the CI width provided by all four methods were almost similar. As noted above the Fisher method provided narrow confidence intervals than the Exact method for k_1, $k_2 = 2$. As for the simple asymptotic method, it provided narrower CIs for ρ_1, $\rho_2 > 0.9$ when compared with that provided by the other three methods.

4.5 Summary

Simulation results indicate that the simple asymptotic method provided coverage outside the desired interval of 94.6% to 95.4% for sample size $n = 15$ and the resulting CIs were wide. As the sample size increases, the coverage percentages tend to reach the nominal level of 95%. The box plots given in Figure 4.1 to 4.5, clearly show this behavior. These box plots also show that coverage percentage provided by the simple asymptotic method for small sample size, spread over a wide range and became narrow when the sample size increased. Although the simple asymptotic method provided coverage percentage close to nominal level when the sample size increased, the tail

errors provided by this method were not close to the nominal level and the tail errors were severely unbalanced. The problem with the simple asymptotic method is that for most parameter values the errors were concentrated in one tail (Fig 4.6). The poor performance of the simple asymptotic method is due to that it does not adjusted for the skewness of the underlying sampling distribution when calculating the CIs for single ICCs. The summary statistic given in Table 4.4 for a single ICC shows the poor performance of the simple asymptotic method.

The coverage percentages provided by Konishi method were always outside of the desired interval and CIs were overly narrow. The box plots shows that the empirical coverage percentages provided by this method were not as wide as in the simple asymptotic method. The tail errors provided by Konishi method always exceeded the nominal level of 2.5%.

Both the Fisher and Exact methods perform well except that the coverage provided by the Fisher method is always less than that of the Exact method. The Exact method provided empirical coverage closer to the nominal level of 95%. The box plots show that the empirical coverage percentages provided by the Exact method had a smaller range than other three methods. Considering the tail errors the Exact method achieved a better balance in the left and right tails (Fig 4.6) and it consistently provided tail errors close to the nominal level of 2.5% except for parameter values ρ_1, $\rho_2 > 0.9$. The Fisher method did not provide balanced tail errors consistently but for some parameter values (k_1, $k_2 = 2$) the Fisher method provide narrower CIs than the

Exact method. Table 4.4 shows the excellent performance of the Exact method for constructing CIs for a single ICC, this excellent performances for a single ICC explains the simulation results for the difference between two ICCs. But the performance of the Exact method for parameter values ρ_1, $\rho_2 > 0.9$ are consistent with the results obtained by Donner and Wells (1986) that performance of the Searle's exact method decreases for $\rho > 0.7$.

Chapter 5

WORKED EXAMPLES

In this chapter, I describe the use of the MOVER method as applied to the motivating examples discussed in Chapter 1. In the two examples discussed below, the data from studies reported by Turner *et al.* (1986) and Gomez *et al.* (2002) are used to calculate the confidence intervals for a difference between two correlated ICCs using the MOVER method. The common features of both these studies are that both were carried out to compare the performances of two devices and they used the same sample of subjects to draw measurements from the two devices. Hence each of these two studies leads to a comparison of two correlated reliability coefficients.

5.1 Example 1

As the first example, I consider the data derived from computer-aided tomographic scans (CAT scans) of the heads of 50 psychiatric patients (Turner *et al.*, 1986). The purpose of this study was to determine the size of the brain ventricle relative to that of the patient's skull given by,

Ventricle brain ratio or VBR= (ventricle size/brain size)x100.

For a given scan VBR was determined from measurements of the perimeter of the patient's ventricle together with the perimeter of the inner surface of the skull. These measurements were made using a hand held planimeter (PLANN) on a projection of the X-ray image or using an automated pixel count (PIX) based on the image displayed on a television screen. As described in Chapter 1, section 1.3.2 of this thesis, Dunn (Dunn, 1989, ch.5) used these data to compare the performance of the two devices PIX and PLANN by examining the raw data visually. Subsequent to this, these data were analyzed by Donner and Zou (2002) using a proper hypothesis test for testing the equality of two dependent ICCs. The logged VBRs for single scans from 50 patients are given in Table 5.1. The first two columns corresponds to repeated determinations based on pixel counts and the second two columns corresponds to repeated determinations based on the use of a planimeter (Dunn, 1989, ch.5).

Table 5.1: CAT scan data; log(VBR) on 50 patients.

Subject	PIX1	PIX3	PLAN1	PLAN3	Subject	PIX	PINX3	PLAN1	PLAN3
1	1.79	1.77	2.05	2.13	26	2.33	2.37	2.24	2.03
2	0.00	0.00	1.72	1.28	27	1.22	1.19	1.63	1.76
3	1.53	1.55	1.93	1.79	28	1.63	1.39	1.55	1.53
4	1.57	1.57	2.16	1.96	29	1.87	1.84	2.12	2.30
5	1.65	1.70	2.27	1.95	30	1.19	1.10	1.63	1.34
6	2.05	2.12	2.53	2.17	31	0.34	0.34	1.46	0.96

Continued on next page

Table 5.1 – continued from previous page

Subject	PIX1	PIX3	PLAN1	PLAN3	Subject	PIX	PINX3	PLAN1	PLAN3
7	1.59	1.65	1.79	1.67	32	1.19	1.25	1.87	1.41
8	1.03	1.03	1.87	1.48	33	1.53	1.53	1.79	1.84
9	0.69	0.74	1.57	1.57	34	1.63	1.65	2.33	1.84
10	1.69	1.79	1.39	1.39	35	0.83	0.88	1.39	1.16
11	1.50	1.55	1.89	1.84	36	1.10	1.10	1.96	1.53
12	1.74	1.72	2.39	2.26	37	0.76	1.76	2.40	2.30
13	1.50	1.63	1.67	1.72	38	1.41	1.44	2.09	1.89
14	0.74	0.74	1.57	1.39	39	0.92	0.96	1.39	1.41
15	1.67	1.69	2.30	2.25	40	1.63	1.65	2.22	1.89
16	1.61	1.59	2.03	1.93	41	0.74	0.79	1.67	1.34
17	1.03	0.99	1.19	1.70	42	0.74	0.79	2.03	1.46
18	0.88	0.96	1.13	0.41	43	1.36	1.36	2.26	2.12
19	1.25	1.28	1.63	1.22	44	1.28	1.31	1.69	1.63
20	1.79	1.77	1.93	2.03	45	2.30	2.29	2.30	2.50
21	1.84	1.89	1.89	1.50	46	1.39	1.34	2.01	1.50
22	1.22	1.22	1.63	2.03	47	1.16	1.16	1.57	1.59
23	1.90	1.99	1.70	1.96	48	0.69	0.69	2.08	1.55
24	2.91	2.93	2.82	2.84	49	1.95	1.95	2.13	2.09
25	1.19	1.10	0.53	0.99	50	1.57	1.55	1.69	1.13

The intraclass correlation coefficient for the measurements obtained from the two devices were estimated using equation (3.2) in Chapter 3, section 3.3.1. For PIX, $\text{MSA}_1 = 0.546$, $\text{MSE}_1 = 0.001$ and $k = 2$. Therefore from equation (3.2)

$$\begin{aligned}\hat{\rho}_1 &= \frac{\text{MSA} - \text{MSE}}{\text{MSA} + (k-1)\text{MSE}} \\ &= \frac{0.546 - 0.001}{0.546 + (2-1)0.001} \\ &= 0.994.\end{aligned}$$

Similarly for PLANN, $\text{MSA}_2 = 0.315$, $\text{MSE}_2 = 0.049$ and $k = 2$.

$$\begin{aligned}\hat{\rho}_2 &= \frac{\text{MSA} - \text{MSE}}{\text{MSA} + (k-1)\text{MSE}} \\ &= \frac{0.315 - 0.049}{0.315 + (2-1)0.049} \\ &= 0.730.\end{aligned}$$

5.1.1 Estimating confidence limits for single ICC

Here I used the method described in Chapter 3, to obtain the required confidence limits for ρ_1 and ρ_2.

A 95% confidence limits for ρ_1 and ρ_2 obtained using simple asymptotic method described in section 3.3.2 of Chapter 3.

For PIX,

$$l_1 = 0.991$$

$$u_1 = 0.997$$

and for PLANN,

$$l_2 = 0.599$$

$$u_2 = 0.860$$

A 95% confidence limits for ρ_1 and ρ_2 obtained using Fisher's Z-transformation and variance obtained using delta method as described in section 3.3.3 of Chapter 3.

For PIX,

$$\begin{aligned} Z_1 &= \frac{1}{2}\ln\left\{\frac{1+\hat{\rho}_1}{1-\hat{\rho}_1}\right\} \\ &= \frac{1}{2}\ln\left\{\frac{1+0.994}{1-0.994}\right\} \\ &= 2.903 \\ \text{var}(Z(\hat{\rho}_1)) &= \frac{2(nk-1)(1-\widehat{\rho})^2[1+(k-1)\widehat{\rho}]^2}{k^2(k-1)n(n-1)} \\ \text{var}(Z_1(\hat{\rho}_1)) &= \left[\frac{1}{2(1-\rho)[1+(k-1)\rho]}\right]^2 \text{var}(\hat{\rho}_1) \\ \widehat{\text{var}}(Z_1(\hat{\rho}_1)) &= 0.020 \end{aligned}$$

Using,

$$\begin{aligned}(Z_l, Z_u) &= \left\{Z_{\rho_1} - z_{\alpha/2}\sqrt{V}, Z_{\rho_1} + z_{\alpha/2}\sqrt{V}\right\} \\ (Z_l, Z_u) &= (2.624, 3.181)\end{aligned}$$

Therefore using

$$\begin{aligned}(l, u) &= \left\{\frac{e^{2Z_l} - 1}{e^{2Z_l} + 1}, \frac{e^{2Z_u} - 1}{e^{2Z_u} + 1}\right\} \\ l_1 &= \frac{e^{2Z_l} - 1}{e^{2Z_l} + 1} \\ &= \frac{e^{2\times 2.624} - 1}{e^{2\times 2.624} + 1} \\ &= 0.989 \\ u_1 &= \frac{e^{2Z_u} - 1}{e^{2Z_u} + 1} \\ &= \frac{e^{2\times 3.181} - 1}{e^{2\times 3.181} + 1} \\ &= 0.997.\end{aligned}$$

Similarly,

$$\begin{aligned}Z_2 &= 0.929 \\ \widehat{\text{var}}(Z_2(\hat{\rho}_2)) &= 0.020 \\ (Z_l, Z_u) &= (0.650, 1.207) \\ (l_2, u_2) &= (0.572, 0.836)\end{aligned}$$

A 95% confidence limits for ρ_1 and ρ_2 obtained using Konishi modified Z-transformation described in section 3.3.4 of Chapter 3.

For PIX,

$$\begin{aligned} Z_m &= \sqrt{\frac{k-1}{2k}} \ln\left\{\frac{1+(k-1)\hat{\rho}}{1-\hat{\rho}}\right\} \\ Z_{1m} &= \sqrt{\frac{2-1}{4}} \ln\left\{\frac{1+(2-1)0.994}{1-0.994}\right\} \\ &= 2.903 \\ \mathrm{V}(Z_{1m}) &= \frac{1}{N} \\ &= 0.02 \end{aligned}$$

Therefore

$$\begin{aligned} Z_{m,l} &= Z_m - z_{\alpha/2}\sqrt{V_m} - \frac{7-5k}{N\sqrt{18k(k-1)}} \\ Z_{1,m,l} &= 2.903 - z_{.05/2}\sqrt{0.02} - \frac{7-10}{50\sqrt{36(2-1)}} \\ Z_{1,m,l} &= 2.635 \\ Z_{m,u} &= Z_m + z_{\alpha/2}\sqrt{V_m} - \frac{7-5k}{N\sqrt{18k(k-1)}} \\ Z_{1,m,u} &= 2.903 + z_{.05/2}\sqrt{0.02} - \frac{7-10}{50\sqrt{36(2-1)}} \\ Z_{1,m,u} &= 3.190 \end{aligned}$$

Therefore using (3.8) in Chapter 3 confidence limits for PIX is obtained as follows.

$$\begin{aligned} l_1 &= \frac{e^{\left(Z_{1,m,l}\sqrt{\frac{2k}{k-1}}\right)} - 1}{e^{\left(Z_{1,m,l}\sqrt{\frac{2k}{k-1}}\right)} + (k-1)} \\ &= \frac{e^{\left(2.635\sqrt{\frac{4}{2-1}}\right)} - 1}{e^{\left(2.635\sqrt{\frac{4}{2-1}}\right)} + (2-1)} \\ &= 0.989 \end{aligned}$$

and

$$\begin{aligned} u_1 &= \frac{e^{\left(Z_{m,u}\sqrt{\frac{2k}{k-1}}\right)} - 1}{e^{\left(Z_{m,u}\sqrt{\frac{2k}{k-1}}\right)} + (k-1)} \\ &= \frac{e^{\left(3.190\sqrt{\frac{4}{2-1}}\right)} - 1}{e^{\left(3.190\sqrt{\frac{4}{4-1}}\right)} + (4-1)} \\ &= 0.997. \end{aligned}$$

Similarly, for PLANN,

$$\begin{aligned} Z_{2m} &= 0.929 \\ Z_{2,m,l} &= 0.662 \\ Z_{2,m,u} &= 1.216 \\ l_2 &= 0.579 \\ u_2 &= 0.838 \end{aligned}$$

A 95% confidence limits for ρ_1 and ρ_2 obtained using Exact method as described in section 3.3.5 of Chapter 3, variance ratio statistic F obtained for PIX and PLANN are given below.

For PIX,

$$\begin{aligned} \text{MSA}_1 &= 0.546 \\ \text{MSE}_1 &= 0.001 \\ F_1 &= 339.447 \end{aligned}$$

For PLANN,

$$\begin{aligned} \text{MSA}_2 &= 0.315 \\ \text{MSE}_2 &= 0.049 \\ F_2 &= 6.042 \end{aligned}$$

Confidence limits for ρ_1 and ρ_2 were obtained using the formula (3.9) given in Chapter 3

$$\left\{ \frac{F/F_U - 1}{k + F/F_U - 1}, \frac{F/F_L - 1}{k + F/F_U - 1} \right\}$$

where

$$F_L = F_{(\alpha/2, n-1, nk-n)} = F_{(.025,49,50)} = 0.568$$

and

$$F_U = F_{(1-\alpha/2,n-1,nk-n)} = F_{(0.975,49,50)} = 1.755$$

By substituting the values for F, F_U and F_L, the exact 95% confidence limits for ρ_1 and ρ_2 obtained are as follows.

For PIX,

$$l_1 = 0.989$$

$$u_1 = 0.997$$

and for PLANN,

$$l_2 = 0.570$$

$$u_2 = 0.837$$

5.1.2 Estimating 95% confidence limits for a difference between two ICCs using MOVER method.

Interclass correlation coefficient $\hat{\rho}_{12}$ of PIX and PLANN was estimated using (3.15) in section 3.4.2 of Chapter 3. The value of $\hat{\rho}_{12}$ is substitute in (3.25) to obtain the

$\text{corr}(\hat{\rho}_1, \hat{\rho}_2)$ as follows.

$$\begin{aligned}
\text{corr}(\hat{\rho}_1, \hat{\rho}_2) &= \frac{\hat{\rho}_{12}^2 \left[k_1 k_2 (k_1 - 1)(k_2 - 1)\right]^{1/2}}{\left[1 + (k_1 - 1)\rho_1\right]\left[1 + (k_2 - 1)\rho_2\right]} \\
&= \frac{0.647^2 \left[2 \times 2(2 - 1)(2 - 1)\right]^{1/2}}{\left[1 + (2 - 1)0.994\right]\left[1 + (2 - 1)0.730\right]} \\
&= 0.246
\end{aligned}$$

A 95% confidence limits for $(\rho_1 - \rho_2)$ was obtained using (3.23) and (3.24) of Chapter 3.

$$L = (\hat{\rho}_1 - \hat{\rho}_2) - \sqrt{(\hat{\rho}_1 - l_1)^2 + (u_2 - \hat{\rho}_2)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(\hat{\rho}_1 - l_1)(u_2 - \hat{\rho}_2)} \quad (5.1)$$

$$U = (\hat{\rho}_1 - \hat{\rho}_2) + \sqrt{(u_1 - \hat{\rho}_1)^2 + (\hat{\rho}_2 - l_2)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(u_1 - \hat{\rho}_1)(\hat{\rho}_2 - l_2)} \quad (5.2)$$

A 95% confidence limits for $(\rho_1 - \rho_2)$ using MOVER method when 95% confidence limits for (ρ_1 and ρ_2) are obtained by simple asymptotic method.

$$\begin{aligned}
(l_1, u_1) &= (0.991, 0.997) \\
(l_2, u_2) &= (0.599, 0.860)
\end{aligned}$$

Substituting these values in equation (5.1) and (5.2) will result a 95% confidence limits for $(\rho_1 - \rho_2)$ as given below.

$$\begin{aligned} L &= (\hat{\rho}_1 - \hat{\rho}_2) - \sqrt{(\hat{\rho}_1 - l_1)^2 + (u_2 - \hat{\rho}_2)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(\hat{\rho}_1 - l_1)(u_2 - \hat{\rho}_2)} \\ &= 0.034 - \sqrt{(0.994 - 0.991)^2 + (0.860 - 0.730)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(0.003)(0.130)} \\ &= 0.135 \\ U &= (\hat{\rho}_1 - \hat{\rho}_2) + \sqrt{(u_1 - \hat{\rho}_1)^2 + (\hat{\rho}_2 - l_2)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(u_1 - \hat{\rho}_1)(\hat{\rho}_2 - l_2)} \\ &= 0.034 + \sqrt{(0.997 - 0.994)^2 + (0.730 - 0.599)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(0.003)(0.131)} \\ &= 0.393 \end{aligned}$$

A 95% confidence limits for $(\rho_1 - \rho_2)$ using MOVER method when 95% when confidence limits for ρ_1 and ρ_2 are obtained by Fisher's Z-transformation, variance obtained using Delta method.

$$\begin{aligned} (l_1, u_1) &= (0.989, 0.997) \\ (l_2, u_2) &= (0.572, 0.836) \end{aligned}$$

Substituting these values in equation (5.1) and (5.2) will result a 95% confidence limits for $(\rho_1 - \rho_2)$ as

$$(L, U) = (0.159, 0.421)$$

A 95% confidence limits for $(\rho_1 - \rho_2)$ using MOVER method when 95% when

confidence limits for ρ_1 and ρ_2 are obtained by Konishi modified Z-transformation.

$$\begin{aligned}(l_1, u_1) &= (0.989, 0.997)\\ (l_2, u_2) &= (0.579, 0.838)\end{aligned}$$

Substituting these values in equation (5.1) and (5.2) will result a 95% confidence limits for $(\rho_1 - \rho_2)$ as

$$(L, U) = (0.157, 0.414)$$

A 95% confidence limits for $(\rho_1 - \rho_2)$ using MOVER method when 95% when confidence limits for ρ_1 and ρ_2 are obtained by Exact method.

$$\begin{aligned}(l_1, u_1) &= (0.989, 0.997)\\ (l_2, u_2) &= (0.570, 0.837)\end{aligned}$$

Substituting these values in equation (5.1) and (5.2) will result a 95% confidence limits for $(\rho_1 - \rho_2)$ as given below.

$$(L, U) = (0.158, 0.423)$$

Confidence intervals for $(\rho_1 - \rho_2)$ obtained using MOVER method are summarized in Table 5.2 below.

Table 5.2: A 95% two sided confidence interval for a difference between two ICCs, confidence intervals for single ICCs were obtained using four different methods.

Method for single ICC	95% CI for ρ_1 (l_1, u_1)	95% CI for ρ_2 (l_2, u_2)	95% CI for $(\rho_1 - \rho_2)$ (L, U)
Simple asymptotic	(0.991,0.997)	(0.599,0.860)	(0.135,0.393)
Fisher's Z transformation	(0.989,0.997)	(0.572,0.836)	(0.159,0.421)
Fisher's modified Z transformation	(0.989,0.997)	(0.579,0.838)	(0.157,0.414)
Exact	(0.989,0.997)	(0.570,0.837)	(0.158,0.423)

CI: confidence interval.

5.1.3 Summary

The difference between the two reliability coefficients of the two instruments PIX and PLANN could be as high as 0.422 based on the MOVER as applied to exact limits for single ICCs. The obtained results suggests that there is statistically significant difference exists between the two instruments at 5% level of significance. These results are in consistent with the results obtained by Donner and Zou (2002). The confidence intervals and widths obtained in this example can be explained based on the simulation results as follows. When $\rho \geq 0.9$, SA method provide CI even narrower than that provided by the Konishi method and when $k_1 = k_2 = 2$, Fisher method provide CI narrower than that provided by the Exact method. In light of simulation results, the interval based on Exact method for single ICC should be the standard.

5.2 Example 2

As the second example I consider the study reported by Gomez *et al.* (2002) on comparing performances of two qualitative ultrasound scanners. Qualitative ultrasound devices are mainly used in osteoporosis diagnosis and several are already available. Gomez *et al.* (2002) compared the performance of a newly developed scanner (BEAM scanner) with the performance of a scanner (UBIS 3000) already in use with regard to broadband ultrasound attenuation (BUA) and speed of sound (SOS). Broadband ultrasound attenuation and speed of sound are the two parameters used in clinical bone investigations. The study included 34 healthy volunteers as subjects and five repeated measurements of the right heel with interim repositioning were performed using the two devices.

In this thesis I used the intraclass correlation coefficients given in Giraudeau *et al.* (2005) which are estimated for the BUA data set of the study by Gomez *et al.* (2002). The estimated values are $\hat{\rho}_1 = 0.982$, $\hat{\rho}_2 = 0.948$ and $\hat{\rho_{12}} = 0.915$. When number of repeated measurements k and the value of the estimated ICC is available, we can easily calculate the variance ratio statistic F as below.

$$\begin{aligned} \hat{\rho} &= \frac{MSA - MSE}{MSA + (k-1)MSE} \\ &= \frac{F-1}{F+(k-1)} \\ F &= \frac{(k-1)\hat{\rho}+1}{1-\hat{\rho}} \end{aligned}$$

5.2.1 *Estimating 95% confidence limits for a difference between two ICCs using MOVER method.*

As for Example 1, described earlier in this chapter the required confidence limits for single ICCs were obtained.

A 95% confidence limits for $(\rho_1 - \rho_{2)}$ using MOVER method when confidence limits for single ICCs obtained using simple asymptotic method are given below.

$$\begin{aligned}(l_1, u_1) &= (0.972, 0.992)\\ (l_2, u_2) &= (0.921, 0.975)\end{aligned}$$

Substituting these values in equation (5.1) and (5.2) will result a 95% confidence limits for $(\rho_1 - \rho_2)$ as

$$\begin{aligned}L &= (\hat{\rho}_1 - \hat{\rho}_2) - \sqrt{(\hat{\rho}_1 - l_1)^2 + (u_2 - \hat{\rho}_2)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(\hat{\rho}_1 - l_1)(u_2 - \hat{\rho}_2)}\\ &= 0.01 - \sqrt{(0.982 - 0.972)^2 + (0.975 - 0.948)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(0.010)(0.027)}\\ &= 0.013\\ U &= (\hat{\rho}_1 - \hat{\rho}_2) + \sqrt{(u_1 - \hat{\rho}_1)^2 + (\hat{\rho}_2 - l_2)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(u_1 - \hat{\rho}_1)(\hat{\rho}_2 - l_2)}\\ U &= 0.01 + \sqrt{(0.992 - 0.982)^2 + (0.948 - 0.921)^2 - 2\text{corr}(\hat{\rho}_1, \hat{\rho}_2)(0.010)(0.027)}\\ &= 0.055\end{aligned}$$

A 95% confidence limits for $(\rho_1 - \rho_{2)}$ using MOVER method when confidence limits

for single ICCs obtained using Fisher's Z-transformation are given below. For BEAM scanner,

$$\begin{aligned} Z_1 &= 2.350 \\ \widehat{\text{var}}(Z_1(\hat{\rho}_1)) &= 0.018 \\ (Z_l, Z_u) &= (2.083, 2.618) \\ (l_1, u_1) &= (0.969, 0.989) \end{aligned}$$

For UBIS 3000 scanner,

$$\begin{aligned} Z_2 &= 1.811 \\ \widehat{\text{var}}(Z_2(\hat{\rho}_2)) &= 0.018 \\ (Z_l, Z_u) &= (1.547, 2.076) \\ (l_2, u_2) &= (0.913, 0.969) \end{aligned}$$

Therefore a 95% confidence limits for $(\rho_1 - \rho_{2)}$ obtained following the similar steps as described in Example 1 are given below.

$$(L, U) = (0.019, 0.064)$$

A 95% confidence limits for $(\rho_1 - \rho_{2)}$ using MOVER method when confidence limits

for single ICCs obtained using Konishi modified Z- transformation are given below.

For BEAM scanner,

$$\begin{aligned} Z_{1,m} &= 3.549 \\ (Z_{m,l}, Z_{m,u}) &= (3.241, 3.913) \\ (l_1, u_1) &= (0.971, 0.989) \end{aligned}$$

For UBIS 3000 scanner,

$$\begin{aligned} Z_{2,m} &= 2.860 \\ (Z_{m,l}, Z_{m,u}) &= (2.552, 3.224) \\ (l_2, u_2) &= (0.917, 0.970) \end{aligned}$$

Therefore a 95% confidence limits for $(\rho_1 - \rho_{2)}$ obtained following the similar steps as described in Example 1 are given below.

$$(L, U) = (0.018, 0.060)$$

A 95% confidence limits for $(\rho_1 - \rho_{2)}$ using MOVER method when confidence limits for single ICCs obtained using exact method are given below.

For BEAM scanner,

$$\begin{aligned}
\hat{\rho} &= 0.982 \\
k_1 &= 5 \\
F_1 &= \frac{(k-1)\hat{\rho}+1}{1-\hat{\rho}} \\
&= \frac{(5-1)0.982+1}{1-0.982} \\
&= 273.777
\end{aligned}$$

For UBIS 3000 scanner,

$$\begin{aligned}
\hat{\rho} &= 0.948 \\
k_2 &= 5 \\
F_2 &= \frac{(k-1)\hat{\rho}+1}{1-\hat{\rho}} \\
&= \frac{(5-1)0.948+1}{1-0.948} \\
&= 92.1538 \\
(l_1, u_1) &= (0.971, 0.989) \\
(l_2, u_2) &= (0.917, 0.971)
\end{aligned}$$

Therefore a 95% confidence limits for $(\rho_1 - \rho_{2)}$ obtained following the similar steps as

described in Example 1 are given below.

$$\begin{aligned} \hat{\rho}_{12} &= 0.915 \\ \text{corr}(\hat{\rho}_1, \hat{\rho}_2) &= 0.537 \\ (L, U\ ) &= (0.017, 0.060) \end{aligned}$$

Confidence intervals for $(\rho_1 - \rho_2)$ obtained using MOVER method are summarized in Table 5.3 below.

Table 5.3: A 95% two sided confidence interval for a difference between two ICCs, confidence intervals for single ICCs were obtained using four different methods.

Method for single ICC	95% CI for ρ_1 (l_1, u_1)	95% CI for ρ_2 (l_2, u_2)	95% CI for $(\rho_1 - \rho_2)$ (L, U)
Simple asymptotic	(0.972,0.992)	(0.921,0.975)	(0.013,0.055)
Fisher's Z transformation	(0.969,0.989)	(0.913,0.969)	(0.019,0.064)
Fisher's modified Z transformation	(0.971,0.989)	(0.917,0.970)	(0.018,0.060)
Exact	(0.971,0.989)	(0.917,0.971)	(0.017,0.060)

CI: confidence interval.

5.2.2 Summary

The difference between the two reliability coefficients of the two scanners BEAM and UBIS 3000 could be as high as 0.060. The obtained results suggests that there is statistically significant difference exists between the two instruments at 5% level of significance. These results are in consistent with the results obtained by Giraudeau

et al. (2005). In this example too, the confidence intervals and CI widths obtained can be explained based on the simulation results as well. When $\rho \geq 0.9$, SA method provide CI even narrower than that provided by the Konishi method and when $k_1, k_2 > 2$, CI provided by the Exact method are narrower than that provided by the Fisher method.

Chapter 6

DISCUSSION

In this thesis, I have focused on setting approximate CI for a difference between two ICCs. As noted by several authors (Giraudeau *et al.*, 2005; Donner and Zou, 2002; Alsawalmeh and Feldt, 1994; Feldt, 1980) there are many situations in which a comparison of two ICCs is required. Although one can perform a statistical hypothesis testing for this purpose, reporting confidence intervals may be more informative.

Reporting CI as a supplement of the p value or instead of only the p value has been advocated (Rothman, 1986; Wilkinson, 1999; Altman and Garner, 1992; Cummings and Rivara, 2003; Altman, 2005). Confidence intervals are regarded as more informative than hypothesis testing because they provide a range of values that are considered plausible for the parameter of the population. Confidence interval encompasses the hypothesis testing by capturing the value reflecting 'no difference' and hence it can always answer the questions that the p value answers.

The confidence interval constructing procedure discussed in this thesis takes the skewness of the sampling distribution of $\hat{\rho}$ into account, hence this procedure derives its validity from the validity of the confidence limits for a single ICC. Simulation results have shown that, even for small sample size ($N = 15$), the MOVER method

performs excellent in terms of coverage, tail errors and CI width when the Exact method is used to obtain the CIs for single ICCs. For large sample sizes, all four methods perform well in terms of coverage. However, even for large sample sizes, the SA method provides severely unbalanced tail errors. This is because the SA method ignores the skewness of the sampling distribution. Therefore based on the simulation results, the Exact method is recommended for obtaining the required CI for single ICC. One should note that, if number of observations on each subject is two, Fisher's Z transformation provides narrower CIs than that provided by the Exact method. However, the tail errors provided by the Exact method has better balance than provided by the Fisher method. Looking at coverage rates only would lead a one to say that there is no difference among the four methods when the sample size are large, but it should be noted that the Exact method performs much better in terms of tail errors.

The procedure discussed in this thesis on setting an approximate CI for a difference between two ICCs, used the ICCs based on the one-way random effects ANOVA. But, the procedure can be readily used for ICCs obtained based on two-way ANOVA models (two-way random effects or two way mixed effects model).

The procedure developed in this thesis for a difference between two ICCs may also be extended to construct confidence intervals for a difference between two Cronbach's alpha. Cronbach's alpha is another measure of reliability and is a one to one function

of ICC. Estimated Cronbach's alpha ($\hat{\rho}_\alpha$) can be given as

$$\hat{\rho}_\alpha = \frac{k\hat{\rho}_I}{1 + (k-1)\hat{\rho}_I},$$

where $\hat{\rho}_I$ is the intraclass correlation coefficient (Kistner and Muller, 2004). Therefore application of the Delta method can provide the estimated variances required for the construction of the lower and upper CI for a difference between two Cronbach's alpha.

An assumption underlying the procedure discussed in this thesis is that the number of observations k_l is constant across all the subjects for $l = 1, 2$. Although in reliability context, variable number of observations for each subject is uncommon it could happen due to reasons such as investigator fatigue. In this case, k_l may be replaced by the harmonic mean of class sizes; Thomas and Hultquist (1978) have shown that this substitution in the F-distribution based formula for single ICCs works well, provided $\rho > 0.3$, which is reasonable in reliability studies. One could then apply MOVER to this situation. Evaluation in this case is needed and left out for future research.

BIBLIOGRAPHY

Alsawalmeh, Y. M. and Feldt, L. S. (1994). Testing the equality of two related intraclass reliability coefficients. *Applied Psychological Measurement* **18**, 183–190.

Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery* **29**, 554–556.

Altman, D. G. and Garner, M. J. (1992). Confidence intervals for research findings. *British Journal of Obstetrics and Gynaecology* **99**, 90–91.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **19**, 3–11.

Bartlett, M. S. (1953). Approximate confidence intervals. 2. more than one unknown parameter. *Biometrika* **40**, 306–317.

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlation with desired precision. *Statistics in Medicine* **21**, 1331–1335.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Duxbury, Thomson Learning.

Cummings, P. and Rivara, F. P. (2003). Reporting statistical information in medical articles. *Archives of Pediatrics and Adolescent Medicine* **157**, 321–324.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* **54**, 67–82.

Donner, A. (1998). Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine* **17**, 1157–1168.

Donner, A. and Eliasziw, M. (1987). Sample size requirements for reliability studies. *Statistics in Medicine* **6**, 441–448.

Donner, A. and Koval, J. J. (1980). The estimation of intraclass correlation in the analysis of family data. *Biometrics* **36**, 19–25.

Donner, A., Koval, J. J. and Bull, S. (1984). Testing the effect of sex differences on sib-sib correlations. *Biometrics* **40**, 349–356.

Donner, A. and Wells, G. (1986). A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics* **42**, 401–412.

Donner, A. and Zou, G. (2002). Testing the equality of dependent intraclass correlation coefficients. *The Statistician* **51**, 367–379.

Dunn, G. A. (1989). *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors.* Edward Arnold.

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

Eliasziw, M., Young, S. L., Woodbury, M. G. and Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using Goniometric measurements as an example. *Physical Therapy* **74**, 777–788.

Elston, R. C. (1975). On the correlation between correlations. *Biometrika* **62**, 133–140.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika* **45**, 99–105.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Giraudeau, B., Gomez, M. A. and Defontaine, M. (2003). Assessing the reproducibility of quantitative ultrasound parameters with standardized coefficient of variation or intraclass correlation coefficient: a unique approach. *Osteoporosis International* **14**, 614–615.

Giraudeau, B. and Mary, J. Y. (2001). Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Statistics in Medicine* **20**, 3205–3214.

Giraudeau, B., Porcher, R. and Mary, J. Y. (2005). Power calculation for the likelihood ratio-test when comparing two dependent intraclass correlation coefficients. *Computer Methods and Programs in Biomedicine* **77**, 165–173.

Gomez, M. A., Defontaine, M., Giraudeau, B., Camus, E., Colin, L., Laugier, P. and Patat, F. (2002). In vivo performance of a matrix-based quantitative ultrasound imagine device dedicated to calcaneus investigation. *Ultrasound in Medicine and Biology* **28**, 1285–1293.

Haggard, E. A. (1958). *Intraclass Correlation and the Analysis of Variance.* The Dryden Press, Inc, New York.

He, W., Bull, S. B., Gokgoz, N., Andrulis, I. and Wunder, J. (2006). Application of reliability coefficients in cDNA microarray data analysis. *Statistics in Medicine* **25**, 1051–1066.

Kistner, E. O. and Muller, K. E. (2004). Exact distribution of intraclass correlation and cronbach's alpha with gaussian and general covariance. *Psychometrica* **69**, 459–474.

Konishi, S. (1985). Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics* **37**, 87–94.

Kottner, J. and Dassen, T. Interpreting interrater reliability coefficients of the Braden scale: A discussion paper. Article in press.

Lachin, J. M. (2004). The role of measurement reliability in clinical trials. *Clinical Trials* **1**, 553–566.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

McGraw, K. O. and Wong, S. (1996). Forming inference about some intraclass correlation coefficients. *Psychological Methods* **1**, 30–46.

Nickerson, C. A. E. (1997). A note on 'A concordance correlation coefficient to evaluate reproducibility'. *Biometrics* **53**, 1503–1507.

Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution - III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society, Series A* **187**, 253–318.

Pellis, L., van Hal, N. L. W. F., Burema, J. and Keijer, J. (2003). The intraclass correlation coefficient applied for evaluation of data correction, labeling methods and rectal biopsy sampling in DNA microarray experiments. *Physiological Genomics* **16**, 99–106.

Robey, R. R. and Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology* **45**, 283–288.

Rosner, B. (1982). On the estimation and testing of intraclass correlations: The general case of multiple replicates for each variable. *American Journal of Epidemiology* **116**, 722–730.

Rosner, B. and Willett, W. C. (1988). Interval estimates for correlation coefficients correacted for within-person varaition : implications for study design and hypothesis testing. *American Journal of Epidemiology* **127**, 377–386.

Rothman, K. J. (1986). Significance questing (editorial). *Annals of Internal Medicine* **105**, 445–447.

Rousson, V., Gasser, T. and Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in Medicine* **21**, 3431–3446.

Schuck, P. (2004). Assessing reproducibility for interval data in health related quality of the life questionnaires: Which coefficient should be used? *Quality of Life Research* **13**, 571–586.

Searle, S. R. (1971). Topics in variance component estimation. *Biometrics* **27**, 1–76.

Shoukri, M. M., Asyali, M. H. and Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research* **13**, 251–271.

Shoukri, M. M., Asyali, M. H. and Walter, S. D. (2003). Issues of cost and efficiency in the design of reliability studies. *Biometrics* **59**, 1107–1112.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research* **7**, 301–317.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86**, 420–428.

Smith, C. A. B. (1956). On the estimation of intraclass correlation. *Annals of Human Genetics* **105**, 445–447.

Swiger, L. A., Harvey, W. R., Everson, D. O. and Gregory, K. E. (1964). The variance of intraclass correlation involving groups with one observation. *Biometrics* **20**, 818–826.

Thomas, J. D. and Hultquist, R. A. (1978). Interval estimation for the unbalanced case of the one-way random effects model. *The Annals of Statistics* **6**, 582–587.

Turner, S. W., Toone, B. K. and Brett-Jones, J. R. (1986). Computerized tomographic scan changes in early schizophrenia-preliminary finding. *Psychological Methods* **16**, 219–225.

Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine* **21**, 3757–3774.

Walter, S. D., Eliasziw, M. and Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine* **17**, 101–110.

Wilkinson, L. (1999). Statistical methods in psychological journals guidelines and explanations. *American Psychologist* **54**, 594–604.

Zou, G. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods* **12**, 399–413.

Zou, G. and Donner, A. (2008). Construction of confidence limits about effect measures: A general approach. *Statistics in Medicine* **27**, 1693–1702.