September 2016

Sample Size Formulas for Estimating Intraclass Correlation Coefficients in Reliability Studies with Binary Outcomes

Mengxiao Xu
The University of Western Ontario

Supervisor
Dr. Guangyong Zou
The University of Western Ontario

Graduate Program in Epidemiology and Biostatistics

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

ABSTRACT

Introduction: Measurement errors can seriously affect quality of clinical practice and medical research. It is therefore important to assess such errors by conducting studies to estimate a coefficients reliability and assessing its precision. The intraclass correlation coefficient (ICC), defined on a model that an observation is a sum of information and random error, has been widely used to quantify reliability for continuous measurements. Sample formulas have been derived for explicitly incorporation of a prespecified probability of achieving the prespecified precision, i.e., the width or lower limit of a confidence interval for ICC. Although the concept of ICC is applicable to binary outcomes, existed sample size formulas for this case can only provide about 50% assurance probability to achieve the desired precision.

Methods: A common correlation model was adopted to characterize binary data arising from reliability studies. A large sample variance estimator for ICC was derived, which was then used to obtain an asymmetric confidence interval for ICC by the modified Wald method. Two sample size formulas were derived, one for achieving a prespecified confidence interval width and the other for requiring a prespecified lower confidence limit, both with given assurance probabilities. The accuracy of the formulas was evaluated using numerical studies. The utility of the formulas was assessed using example studies.

Results: Closed-form formulas were obtained. Numerical study results demonstrated that these formulas are fairly accurate in a wide range of scenarios. The examples showed that the formulas are simple to use in design reliability studies with binary outcomes.

Discussion: The formulas should be useful in the planning stage of a reliability study with binary outcomes in which the investigator wishes to obtain an estimate of ICC with prespecified precision in terms of width or lower limit of a confidence interval. It is no longer justified to conduct reliability studies on the basis of sub-

optimal formulas that provide only 50% assurance probability.

KEYWORDS: Agreement; Common Correlation Model; Confidence Intervals; Interrater; Reproducibility.

ACKNOWLEDGMENTS

I would like to express my gratitude to all the people who assisted and supported my study in University of Western Ontario. I would like to especially thank my supervisor, Dr. Guangyong Zou, for his generous and consistent guidance through out the two years to complete this the- sis. I am greatly impressed by his devotion and knowledgeability which become my strongest source of motivation to keep learning even in the future. His super- vision and financial support made this thesis possible.

I would also thank the faculty and students of the Department of Epidemiology and Biostatistics who helped me survive the abroad studies. I will always remember their encouragement and friendship, and this precious experiences in my life.

Lastly, I would give special thanks to my parents and my partner. The belief from family members accompanies me to face all the difficulties fearlessly.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

INTRODUCTION

The act of measurement is an essential part of any scientific inquiry. In contrast to many natural science disciplines, research in the medical, epidemiological and health sciences often relies on measurements obtained through subjective judgement. The need for reliable and valid measures in these situations has been clearly demonstrated by Marshall *et al.* (2000) who reported that compared to randomized trials of Schizophrenia using published measuring scales, those studies which used unpublished measuring scales were 30 to 40% more likely to report significant treatment results. Section 2.2.2 of the International Conference on Harmonization (ICH, 1998) E9 has emphasized the importance of using reliable and valid measures in clinical trials.

It is well known that assessment of reliability is the first necessary step. This is because before one can assess whether an instrument is measuring what is intended to be measured (i.e. valid), one must first gather evidence that the scale is measuring in a reproducible fashion.

Despite a large literature on statistical methods for reliability as reviewed by Shoukri (2010) and Shoukri and Donner (2009), there is a paucity of feasible formulas for calculating sample size for reliability studies with binary measurements. This may have contributed to the situation that rarely can one identify reliability studies with pre-specified sample sizes in the literature.

The objective of this thesis is to fill in this gap by proposing sample size formulas for reliability studies with binary measurements aimed at quantification of reliability.

In this introductory chapter we begin with the concept of reliability, followed by a section presenting a brief summary of the relationship between validity and reliability. In Section 1.3, we define reliability coefficient as intraclass correlation coefficient (ICC) for continuous measurements, followed by Section 1.4 where we introduce ICC as an reliability index for binary measurements. Section 1.5 discusses a scope of the thesis explaining why we have decided to focus on sample size estimation for reliability studies with binary measurements. Section 1.6 lays out the organization of the thesis.

1.1 Reliability

Any measurement inherently consists of random and systematic errors. The concept of reliability is a fundamental way to reflect the amount of error. Reliability concerns the extent to which an instrument measures in a reproducible fashion the same individuals on different occasions, or by different observers, or by similar tests. For measures with concrete meaningful units, e.g., a bathroom scale, an indication of measurement error of $\pm 1kg$ would be sufficient for one to conclude that measurements obtained using these scales would be reliable for assessing weight gain of adults, but unreliable for assessing growth of an infant. However, a subjective scale with ± 2 units alone provides no information on whether it can be used to distinguish individuals unless we have some idea about the likely range of scores. To overcome this difficulty, reliability is usually defined as a ratio of the variability between individuals to the total variability in the scores. By so defined, it reflects the extent to which a measurement instrument can differentiate among individuals. From this definition, it is also clear that reliability of an instrument depends

not only on its characteristics, but also on the underlying context. In other words, there is no such a thing as "the reliability of the scale", but rather "the reliability of the scale with this population". This implies that a measurement scale may need to be assessed for reliability in a research study even if it has been evaluated in another population, unless it can be assured that the two populations are similar.

1.2 Validity

The validity of a measurement scale refers to the relationship of the measured score to its purported underlying attribute. Operationally, validity can be defined as the proportion of the observed variance that reflects variance in the construct the measure was intended to measure (Carey and Gottesman, 1978).

Specifically, validity may be defined as decomposing the variability of observations as:

$$\sigma^2_{\text{obs}} = \sigma^2_{\text{construct}} + \sigma^2_{\text{systematic}} + \sigma^2_{\text{random}}.$$

and thus:

$$\text{validity} = \frac{\sigma^2_{\text{construct}}}{\sigma^2_{\text{obs}}}$$

in comparison to

$$\text{reliability} = \frac{\sigma^2_{\text{construct}} + \sigma^2_{\text{systematic}}}{\sigma^2_{\text{obs}}}.$$

Thus, it is clear that

$$\text{reliability} \geqslant \text{validity}.$$

Thus, reliability places an upper limit on validity. Note that the validity of a scale also depends on the population of interest and the specific context.

1.3 Intraclass Correlation Coefficient as Reliability Index

Calculation of reliability is possible only if repeated observations on each of N subjects are derived by the same observer over a period of time in a way that ensures blindness (intra-observer reliability), or randomly selected from a pool of observers who independently observe the subject at one point in time (inter-observer reliability). An additional situation is that observations are made by randomly selected observers from a pool of observers, each of which observes the subject at one of several randomly selected time points over a span of time in which the characteristic of interest of the subjects is unlikely to change. This will result in test-retest reliability.

The underlying model for observations corresponding to the above three types of reliability can be written as

$$Y_{ij} = \mu + s_i + e_{ij},$$

for $i = 1, 2, \ldots, N$, denoting subjects, and $j = 1, 2, \ldots, n$, denoting repeated observations. The usual assumptions for estimation purposes are that s_i are independently and identically distributed (iid) with mean 0 and variance σ_s^2, and e_{ij} are iid having mean 0 and variance σ_e^2. The sets $\{s_i\}$ and $\{e_{ij}\}$ are also assumed to be mutually independent. This model is commonly referred to as one-way random effects model (Bartko, 1966).

In a typical inter-observer reliability study in which the n separate measurements correspond to the values recorded by each of the observers, the above model represents the "no observer effect" situation. Despite its simplicity, this model is of considerable interest when the focus is directed at the reliability of the measurement process itself, where reliability is defined as

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$$

and commonly referred to as the (within subject) intraclass correlation coefficient (ICC) since Fisher (1925). In the present context, it indicates that $\rho \times 100\%$ of variance in the scores results from "true" variance among subjects. Thus, the higher the value of ρ, the easier it is to distinguish or tell subjects apart based on the measurements. As pointed out by Bartko (1966), ρ can also be interpreted as a correlation between any two observations within a subject.

Statistical methods for the ICC in one-way random effects model have been reviewed by Donner (1986), with an emphasis on procedures for point and confidence interval estimation, as well as hypothesis testing for nonzero values of ICC. Due to the severe left skewness of the sampling distribution, inference for the ICC is usually conducted on a transformed scale first suggested by Fisher (1925), and thus commonly known as Fisher's Z-transformation, given by

$$Z = \frac{1}{2}\ln\{[1 + (n-1)\rho]/[1-\rho]\}.$$

Using the ICC as reliability index for more complex designs is discussed by Shrout and Fleiss (1979) and McGraw and Wong (1996). Sample size requirements for reliability studies with continuous measurements have been given by Zou (2012).

1.4 ICC for Binary Measurement

Assessment of reliability for binary measurement has historically begun with consideration of the inter-observer agreement, due largely to the chance corrected agreement index proposed by Cohen (1960), which is given by

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e}$$

where π_0 is the observed probability of agreement and π_e is a hypothetical expected probability of agreement by chance. Strictly speaking, κ is an index for reliability, not agreement (Kraemer *et al.*, 2002; de Vet *et al.*, 2006). These two terms

have sometimes been used interchangeably, primarily because of Cohen's κ for inter-rater agreement.

Fleiss (1971) generalized Cohen's kappa to the case where each of a sample of subjects is rated on a nominal scale by the same number of raters, but where the raters rating one subject are not necessarily the same as those rating another subject. Essentially, this corresponds to the assumption of "no observer effects", and "is of main interest where the main emphasis is directed at the reliability of the measurement process itself, rather than at potential differences among observers" (Landis and Koch, 1977). This situation will also be the focus of this thesis.

Despite the popularity since its inception, the kappa coefficient is defined without a population model until Landis and Koch (1977) who adopted a one-way random effects model for categorical data. In the light of this model, the intraclass correlation coefficient is virtually identical to Fleiss's kappa coefficient, with the difference arising from using $N-1$ instead of N in the calculation of mean square between subjects.

Let N be the total number of subjects, n be number of ratings each subject received. Under the one-way random effects model for binary measurement (=1 Yes; 0 No)

$$Y_{ij} = \mu + s_i + e_{ij},$$

where $i = 1, 2, \ldots, N$ denotes subjects and $j = 1, 2, \ldots, n$ denotes raters. Let Y_i denote the number of 1s subject i received. The mean square within subjects (MSW) is given by

$$\text{MSW} = \frac{1}{Nn(n-1)} \sum_{i=1}^{N} Y_i(n - Y_i)$$

and the mean square between subjects (MSB) is

$$\text{MSB} = \frac{1}{(N-1)n} \sum_{i=1}^{N} (Y_i - n\hat{p})^2$$

with $\hat{p}$ denoting overall proportion of 1s, i.e., $\hat{p} = \sum_{i=1}^{N} Y_i/(Nn)$ and the ICC

$$\begin{aligned}\hat{\rho} &= \frac{\text{MSB} - \text{MSW}}{\text{MSB} + (n-1)\text{MSW}} \\ &= 1 - \frac{\sum_{i=1}^{N} Y_i(n - Y_i)}{(N-1)n(n-1)\hat{p}(1-\hat{p})} \\ &\cong 1 - \frac{\sum_{i=1}^{N} Y_i(n - Y_i)}{Nn(n-1)\hat{p}(1-\hat{p})}.\end{aligned}$$

Further insight on the ICC for binary measurements has been provided by Kraemer (1979). For example, under the population model, it is possible to identify factors which influence its magnitude, which in turn suggest strategies to increase reliability of the measurement (Kraemer *et al.*, 2002).

A reliable inference procedure, specifically for constructing confidence interval estimation for ICC with binary data was not available until Zou and Donner (2004) derived the closed form variance estimator for $\hat{\rho}$. Simulation results by Zou and Donner (2004) suggest that a modified Wald-type confidence interval procedure performs well over a wide range of parameter combinations. The issue of sample size estimation for reliability studies with binary measurement has usually been discussed for the case of Cohen's κ, with the exception of Donner and Rotondi (2010) who proposed an iterative procedure based on the Goodness of fit (Donner and Eliasziw, 1992). In addition to inconvenience in computation, the resulting sample size can only assure to achieve the pre-specified precision with 50% probability.

1.5 Scope of the Thesis

There exists a large literature on statistical methods for reliability studies. Even the intraclass correlation coefficient has a variety of versions to quantify reliability in different situations (Shrout and Fleiss, 1979; McGraw and Wong, 1996).

The primary focus of this thesis is on reliability studies whose objective is to assess the reliability of the measurement process itself as discussed by Landis and Koch (1977) and Kraemer (1979).

The specific objective of this thesis is to derive and evaluate closed-form sample size formulas for planning reliability studies focusing on the estimation of ICC. The model that we rely on is the one-way model as discussed by Landis and Koch (1977). In contrast to approaches currently available in the literature (e.g. Donner and Rotondi 2010), we follow the approach by Zou (2012) who explicitly incorporated a pre-specified assurance probability of achieving a desired precision in estimating ICC. Two advantages of our approaches are as follows. First, calculating sample size on the basis of ICC estimation can directly focus on precision of the estimates, rather than the probability of observing values of ICC that are more extreme than the estimate when the true value of ICC is zero. Second, incorporating explicitly the assurance probability in the calculation can increase the chance of achieving the desired precision.

1.6 Organization of the Thesis

This thesis consists of six chapters. Chapter 2 provides a review for measures of reliability and the corresponding statistical methods when the measurements are binary, starting from Scott's (1955) π and Cohen's (1960) κ for cases of two raters to ICC for multiple raters (Altaye *et al.*, 2001). In Chapter 3, we first adopt the common correlation model for correlated binary data to derive a variance estimator for the resulting ICC estimator, which is followed by derivation of sample size formulas for estimating the ICC with precision and assurance probability. Since the results in Chapter 3 are asymptotic, we assess small sample properties in Chapter 4. Chapter 5 provides illustrative examples for application of the sample size formulas. We finish the thesis with Chapter 6 where we present some general con-

clusions and possible future research directions.

Chapter 2

LITERATURE REVIEW

We have alluded to in the last chapter that reliability and agreement are related but distinct concepts. Agreement indicates how close the repeated measurements are, while reliability suggests whether subjects can be distinguished on the basis of the measurement. Methods for reliability of binary measurements have a long history of using the term "agreement". In this chapter, we review this literature, beginning with precursors of kappas in Section 2.1, followed by Scott's π and Cohen's kappa. In Section 2.3, we review the literature on intraclass correlation coefficient computed from binary measurements. The relationship between kappa coefficient and intraclass correlation coefficient is reviewed in Section 2.4. Relating kappa to ICC is important for understanding the relevance of assessing reliability in research. We finish this chapter with a review of methods for sample size estimation, which provides a justification for Chapter 3.

2.1 Precursors of kappa

Methods for reliability of binary measurements have historically originated from agreement of two raters. Consider a 2×2 frequency table in Table 2.1 giving a binary score to N subjects. Each subject is classified as either positive, denoted as x=1, or negative, denoted as x=0, by two raters. In the table, $n_{ij}(i = 1, 2, j = 1, 2)$ represents the number of subjects rated in the ith row by rater 1 and jth column

Table 2.1: Frequency Table for Two Raters Rating N subjects

		Rater 2		
		x=1	x=0	Total
Rater 1	x=1	$n_{11}(p_{11})$	$n_{12}(p_{12})$	$n_{1.}(p_{1.})$
	x=0	$n_{21}(p_{21})$	$n_{22}(p_{22})$	$n_{2.}(p_{2.})$
	Total	$n_{.1}(p_{.1})$	$n_{.2}(p_{.2})$	N

by rater 2 and $p_{ij}(i = 1, 2, j = 1, 2)$ represents the corresponding probability. Both Fleiss (1975) and Landis and Koch (1975) provide summaries of early agreement indices arised under this setting.

At first glance of the data collected this way, agreement could be obtained namely as $(n_{11} + n_{22})/N$, representing the "index of crude agreement" (Rogot and Goldberg, 1966).

Armitage *et al.* (1966) also proposed two measurements identical to the index of crude agreement: mean majority agreement index and mean pair agreement index.

While Goodman and Kruskal (1979, p.758) claimed that agreement should be measured as a function of $p_{11} + p_{22}$, there are other indices only incorporating p_{11} or p_{22} so as to treat positive ratings and negative ratings unequally.

Dice (1945) proposed

$$S_D = p_{11}/\bar{p}$$

where $\bar{p} = (p_{1.} + p_{.1})/2$ measures the probability of positive ratings conditional on at least one of the raters rating positively as well. If one needs to measure the probability of negative ratings conditional on all the negative ratings, the corre-

sponding index is

$$S'_D = p_{22}/\bar{q}$$

where $\bar{q} = (p_{2.} + p_{.2})/2$.

Rogot and Goldberg (1966) simply took the average of S_D and S'_D and proposed the index

$$A_2 = \frac{p_{11}}{p_{1.} + p_{.1}} + \frac{p_{22}}{p_{2.} + p_{.2}}.$$

A_2 ranges from 0, complete disagreement to 1, complete agreement. They also proposed another index as a function of four conditional probabilities, that is

$$A_1 = \frac{1}{4}\left(\frac{p_{11}}{p_{1.}} + \frac{p_{11}}{p_{.1}} + \frac{p_{22}}{p_{2.}} + \frac{p_{22}}{p_{.2}}\right).$$

The first term $\frac{p_{11}}{p_{1.}}$ can be interpreted as the probability of rater 2 rating positive conditional on rater 1 rating positive. A_1 has a minimum value of 0 when there is complete disagreement and a maximum value of 1 when there is complete agreement.

Armitage *et al.* (1966) also suggested an agreement index in the form of a standardized deviation of subjects' total ratings scores. The index, known as "standard deviation agreement index (SDAI)", is given as the square root of

$$SDAI^2 = \frac{N}{N-1}[p_{11} + p_{22} - (p_{11} - p_{22})^2].$$

However, since the maximum value of $SDAI$ equals $\sqrt{N/(N-1)}$ only under the condition that $p_{1.} = p_{.1} = 1/2$, $SDAI$ was rescaled to

$$RSD^2 = \frac{p_{11} + p_{22} - (p_{11} - p_{22})^2}{1 - (\bar{p} - \bar{q})^2} \tag{2.1}$$

which ranges from 0 to unity.

Goodman and Kruskal (1954) proposed an index noted as λ_r, for the case when there are more negative ratings than positive ratings, as

$$\lambda_r = \frac{(p_{11} + p_{22}) - \bar{q}}{\bar{p}} = \frac{2p_{11} - (p_{12} + p_{21})}{2p_{11} + (p_{12} + p_{21})}.$$

It also has a maximum value of 1 with complete agreement, but can reach the minimum value of -1 whenever $a = 0$, irrespective of p_{22}. Besides, it has been noted that $\lambda_r = 2S_D - 1$. Thus λ_r is also feasible for conditional probabilities.

2.2 Chance-corrected Agreement Indices

Since the observed agreement is contributed partly by chance, it is reasonable to exclude agreement caused by chance in measuring the real agreement. When marginal probabilities are $p_{1.}$ and $p_{.1}$ for rater 1 and rater 2 respectively rating subjects as positive, there is a probability of $p_{1.} \times p_{.1}$ that is expected by chance alone. It is reasonable to exclude this portion to account for agreement.

There exists a natural means to correct for chance. Let I_o represent the observed agreement proportion and I_e represent the agreement proportion expected by chance alone. $I_o - I_e$ is then the excess agreement beyond chance and $1 - I_e$ is the maximum potential excess agreement beyond chance. The index is defined as a ratio of these two differences,

$$M(I) = \frac{I_o - I_e}{1 - I_e}.$$

This $M(I)$ is considered to be the standardized index corrected for chance as a measurement for agreement. It reaches the maximum value of 1 when there is perfect agreement, 0 when the observed agreement and the expected agreement are numerically same. The minimum value is equal to $-p_e/(1 - p_e)$, which becomes -1 only when $I_e = 1/2$. Otherwise its minimum value ranges from -1 to 0.

Following this form, there are many indices in this form as follows.

2.2.1 Scott's π

Scott (1955) is among the first to propose an index correcting for chance. The index was designed under the assumptions of independence and marginal homogeneity.

If x_{i1} represents the rating outcome from rater 1 for the ith subject, x_{i2} represents the rating outcome from rater 2 for the ith subject, and $\Pr(x_{i1} = 1) = p_1$, $\Pr(x_{i2} = 1) = p_2$, the assumption of marginal homogeneity implies $p_1 = p_2 = p$, meaning two raters have the same probabilities of measuring subjects into the same categories.

Scott (1955) assumed that in the probability sense, the raters' rating tendencies are identical and can be estimated as the average of two raters' marginal probabilities. Thus, the observed and expected agreement can be written as

$$I_o = \frac{n_{11} + n_{22}}{N}, \;\; I_e = p^2 + (1-p)^2$$

where $p = (2n_{11} + n_{12} + n_{21})/(2N)$, and Scott's π can be given by

$$\begin{aligned} \pi &= \frac{I_o - I_e}{1 - I_e} \\ &= \frac{I_o - [p^2 + (1-p)^2]}{1 - [p^2 + (1-p)^2]} \\ &= \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})} \end{aligned}$$

2.2.2 Cohen's Kappa

While Scott's π assumes marginal homogeneity, Cohen (1960) argued the raters' different proclivities of ratings should be considered as a source of disagreement and also corrected for accordingly.

In proposing the kappa statistic, there were no restrictions concerning the marginal distributions but only independence. The definition of observed agreement is still defined the same as previously and the expected agreement $I_e = \frac{n_{1\cdot}}{N}\frac{n_{\cdot 1}}{N} + \frac{n_{2\cdot}}{N}\frac{n_{\cdot 2}}{N}$.

Cohen's kappa, which subsequently became one of the mostly commonly applied statistics for agreement measurement, is given by

$$\hat{\kappa} = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{(n_{11} + n_{12})(n_{12} + n_{22}) + (n_{11} + n_{21})(n_{21} + n_{22})}$$

Fleiss *et al.* (1969) provided an approximate asymptotic variance formula for $\hat{\kappa}$, given by

$$\widehat{\text{var}(\kappa)} = \frac{1}{N(1 - I_e)^2} \times$$
$$\left[\sum_{i=1}^{2} \hat{p_{ii}}[1 - (\hat{p_{i.}} + \hat{p_{.i}})(1 - \hat{\kappa})]^2 + (1 - \hat{\kappa})^2 \sum_{i \neq j}^{2} \hat{p_{ii}}(\hat{p_{i.}} + \hat{p_{.j}})^2 - [\hat{\kappa} - I_e(1 - \hat{\kappa})]^2 \right].$$

Under marginal homogeneity, Cohen's kappa equals Scott's π.

There has been a heated debate on κ's limitations and disadvantages. Kraemer (1979) clarified that the prevalence of the outcome can alter the results of kappa. Shrout *et al.* (1987) considered this to be a desired property, but the dependence on the true prevalence of the characteristic of interest actually complicates the interpretation of the agreement index, for it is difficult to compare two kappa values when the prevalences differ between studies.

Another issue was discussed by Feinstein and Cicchetti (1990) and Cicchetti and Feinstein (1990). Investigators sometimes find a striking paradox that despite a high crude proportion of agreement I_o, the kappa value may be relatively low. They provided the explanation that if $n_{1.} = n_{2.}$ or $n_{.1} = n_{.2}$ is considered to be perfect balance, this phenomena occurs only when the marginal numbers in a 2 $\times$ 2 table are highly symmetrically unbalanced, e.g., $n_{1.}$ is very different from $n_{2.}$, or $n_{.1}$ is very different from $n_{.2}$. There is also another paradox that unbalanced marginal totals can produce higher values of kappa than balanced marginal totals. This situation occurs when $n_{1.}$ is much larger than $n_{2.}$ while $n_{.1}$ is much smaller than $n_{.2}$, or vice versa and is considered as asymmetrical unbalanced marginals.

In summary, low kappa values may occur despite relatively high I_o. Kappa sometimes increases only because of the departure from the symmetry in the marginal totals. Vach (2005) attributed those limitations to a consequence of the definition of kappa whose objective is to correct the crude agreement to the expected agreement by chance. He believed that it makes no sense to criticize kappa for exactly fulfilling this property and wouldn't regard the dependence on the marginal totals as a drawback.

2.2.3 Intraclass Kappa for Two Raters

Kraemer (1979) pointed out that it is difficult to adopt specific strategies to improve the measurement of agreement for Cohen's kappa without a clear population characteristic in model. Bloch and Kraemer (1989) proposed an alternative version of Cohen's kappa under the assumption that all the raters are characterized by the same marginal probabilities of rating the subjects into the same category. It is a simplified version of intraclass kappa for binary responses and multiple raters based on Kraemer (1979) and Mak (1988) under the assumption of interchangeability, that is, the distribution of ratings for each subject is invariant under the permutation of the raters. This model is known as the common correlation model (Donner *et al.*, 1981; Donner and Eliasziw, 1992).

When there are two raters only, let x_{ij} denote the measurement for the ith subject rated by the jth rater, as the name suggests, the correlation between any pair of ratings has the same value of κ_I, that is,

$$\kappa_I = \text{corr}(x_{i1}, x_{i2}) = \frac{\text{cov}(x_{i1}, x_{i2})}{\sqrt{\text{var}(x_{i1})\text{var}(x_{i1})}}.$$

Let the probability of the measurement being positive ($x_{ij} = 1$) be denoted by p_i for the ith subject and $p_i' = 1 - p_i$. For the population model, let $E(p_i) = p$ and $\text{var}(p_i) = \sigma_p^2$.

Table 2.2: The Theoretical Model for Two Raters ($p_i' = 1 - p_i$, $p' = 1 - p$)

		Rater 2		
		$x_{i2} = 1$	$x_{i2} = 0$	Total
Rater 1	$x_{i1} = 1$	$\mathrm{E}(p_i^2)$	$\mathrm{E}(p_i p_i')$	p
	$x_{i1} = 0$	$\mathrm{E}(p_i p_i')$	$\mathrm{E}(p_i'^2)$	p'
	Total	p	p'	1

The agreement between two raters for the ith subject is $p_i^2 + (1 - p_i)^2$ with an expectation of $E[p_i^2 + (1 - p_i)^2] = 2\sigma_p^2 + p^2 + p'^2$. The random agreement is $p^2 + p'^2$. Thus, according to the chance-corrected index definition, the intraclass kappa can be defined as

$$\kappa_I = \frac{I_o - I_e}{1 - I_e} = \frac{\sigma_p^2}{pp'}.$$

The theoretical model for the above case where there are only two raters and binary outcomes are shown in Table 2.2.

In the model for agreement, the intraclass kappa can be expressed as

$$\kappa_I = \frac{\mathrm{E}(p_i^2) - p^2}{pp'}.$$

Hence equivalently, the model as given by Table2.2 can be written in terms of probabilities of joint responses as

$$p_1(\kappa_I) = \Pr(x_{i1} = 1, x_{i2} = 1) = p^2 + p(1 - p)\kappa_I$$
$$p_2(\kappa_I) = \Pr(x_{i1} = 1, x_{i2} = 0 \text{ or } x_{i1} = 0, x_{i2} = 1) = 2p(1 - p)(1 - \kappa_I)$$
$$p_3(\kappa_I) = \Pr(x_{i1} = 0, x_{i2} = 0) = (1 - p)^2 + p(1 - p)\kappa_I.$$

Bloch and Kraemer (1989) also provided the maximum likelihood estimates of

κ_I and p. The log likelihood function can be written as

$$\begin{aligned} &ln[L(p,\kappa_I|n_{11},n_{12},n_{21},n_{22})] \\ &= n_{11}ln(p^2+\kappa_I pp') + (n_{12}+n_{21})ln[pp'(1-\kappa_I)] + n_{22}ln(p'^2+\kappa_I pp'). \end{aligned}$$

After taking partial derivatives with respect to κ_I and p, setting them to be 0 and solving the equations, the maximum likelihood estimates are given by

$$\begin{aligned} \hat{\kappa_I} &= \frac{4(n_{11}n_{22}-n_{12}n_{21})-(n_{12}-n_{21})^2}{(2n_{11}+n_{12}+n_{21})(2n_{22}+n_{12}+n_{21})}, \\ \hat{p} &= \frac{2n_{11}+n_{12}+n_{21}}{2N} \end{aligned}$$

and its asymptotic variance is given by

$$\text{var}(\hat{\kappa_I}) = \frac{1-\kappa_I}{\kappa_I}\left[(1-\kappa_I)(2-\kappa_I) + \frac{\kappa_I(2-\kappa_I)}{2p(1-p)}\right].$$

As pointed out by Bloch and Kraemer (1989), the MLE of κ_I applied to two raters is identical to Scott's π under the assumption of marginal homogeneity and independence, and is proved to be identical to the intraclass correlation coefficient estimator when applied to 0-1 data (Winer, 1971, pp. 294-296). Furthermore, $\hat{\kappa_I}$ and Cohen's kappa estimator are asymptotically equivalent (Blackman and Koval, 2000).

If it is assumed that $\hat{\kappa_I}$ is asymptotically normally distributed with mean κ_I and standard error $SE(\hat{\kappa_I}) = \sqrt{\text{var}(\hat{\kappa_I})}$. The 100(1-$\alpha$)% confidence interval can be obtained by the Wald method, $\hat{\kappa_I} \pm Z_{1-\alpha/2}SE(\hat{\kappa_I})$, where $Z_{1-\alpha/2}$ is the 100(1-$\alpha/2$) percentile point of the standard normal distribution. This method has poor performance even with sample size as large as 100 (Blackman and Koval, 2000; Donner and Eliasziw, 1992).

Donner and Eliasziw (1992) proposed a confidence interval approach based on a goodness-of-fit statistic which was shown to perform well in small samples. Essentially, the approach was to equate the computed chi-square statistic with one

degree of freedom to the targeted critical value and solve for the two roots as the upper and lower confidence interval limits. Under the null hypothesis of $H_0 : \kappa_I = \kappa_0$, the equation was given by

$$\chi_1^2 = \frac{[n_{22} - N\hat{p}_1(\kappa_0)]^2}{N\hat{p}_1(\kappa_0)} + \frac{[n_{12} + n_{21} - 2N\hat{p}_2(\kappa_0)]^2}{2N\hat{p}_2(\kappa_0)} + \frac{[n_{11} - N\hat{p}_3(\kappa_0)]^2}{N\hat{p}_3(\kappa_0)},$$

equal to the targeted critical value with one degree of freedom at chosen significance level α, where $\hat{p}_i(\kappa_0), i = 1, 2, 3$ are obtained by replacing p with $\hat{p}$. The expressions for the upper and lower 95% confidence interval limits are given by

$$\kappa_U = \left(\frac{1}{9}y_3^2 - \frac{1}{3}y_2\right)^2 \left(\cos\frac{\theta + 5\pi}{3} + \sqrt{3}\sin\frac{\theta + 2\pi}{3}\right) - \frac{1}{3}y_3,$$

$$\kappa_L = 2\left(\frac{1}{9}y_3^2 - \frac{1}{3}y_2\right)^{\frac{1}{2}} \cos\frac{\theta + 5\pi}{3} - \frac{1}{3}y_3,$$

where

$$\theta = \cos_{-1}\frac{V}{W},\ V = \frac{1}{27}y_3^2 - \frac{1}{6}(y_2y_3 - 3y_1),\ W = \left(\frac{1}{9}y_3^2 - \frac{1}{3}y_2\right)^{\frac{3}{2}},$$

and

$$y_1 = \frac{[n_{12} + n_{21} - 2N\hat{p}(1-\hat{p})]^2 + 4N^2\hat{p}^2(1-\hat{p})^2}{4N\hat{p}^2(1-\hat{p})^2(N + 3.84)} - 1,$$

$$y_2 = \frac{(n_{12} + n_{21})^2 - 4(3.84)N\hat{p}(1-\hat{p})[1 - 4\hat{p}(1-\hat{p})]}{2N\hat{p}^2(1-\hat{p})^2(N + 3.84)} - 1,$$

$$y_3 = \frac{n_{12} + n_{21} + 3.84[1 - 2\hat{p}(1-\hat{p})]}{\hat{p}(1-\hat{p})(N + 3.84)} - 1.$$

An alternative procedure that recognizes the property that the variance estimator for $\hat{\kappa_I}$ is a function of κ_I itself was proposed by Donner and Zou (2002). Simulation results suggest that this procedure performed equally as well as the goodness-of-fit method by Donner and Eliasziw (1992).

2.2.4 Connection with Early Agreement Indices

The application of the chance-corrected agreement index form $M(I)$ succeeds in unifying most of indices introduced in Section 3.1 (Fleiss, 1975).

The chance-expected value of index S_D is estimated as $E(S_D) = p_{1.}p_{.1}/\bar{p}$ and $M(S_D)$ is then

$$M(S_D) = \frac{S_D - E(S_D)}{1 - E(S_D)} = \frac{2(p_{11}p_{22} - p_{12}p_{21})}{p_{1.}p_{.2} + p_{.1}p_{2.}},$$

identical to Cohen's κ. In addition, since $\lambda_r = 2S_D - 1$, $M(\lambda_r) = \kappa$.

For Rogot and Goldberg's (1966) A_2, its chance-expected value can be estimated as

$$E(A_2) = \frac{p_{1.}p_{.1}}{p_{1.} + p_{.1}} + \frac{p_{2.}p_{.2}}{p_{2.} + p_{.2}},$$

and $M(A_2)$ is then

$$M(A_2) = \frac{A_2 - E(A_2)}{1 - E(A_2)} = \frac{2(p_{11}p_{22} - p_{12}p_{21})}{p_{1.}p_{.2} + p_{.1}p_{2.}},$$

identical to Cohen's κ again.

$SDAI$'s maximum value doesn't necessarily equal to 1 with complete agreement, but RSD^2 does. If N is large enough to ignore the term $1/N$, the chance-expected value of RSD^2 is estimated as

$$E(ESD^2) = \frac{p_{1.}p_{2.} + p_{.1}p_{.2}}{1 - (\bar{p} - \bar{q})^2}.$$

and it is easily checked that $M(RSD^2)$ also equals to Cohen's κ.

2.3 Intraclass Correlation Coefficient

2.3.1 ANOVA Estimator

When subjects are rated by multiple raters, agreement is usually measured by intraclass correlation coefficient. Traditionally, analysis of variance (ANOVA) estimators, which were originally proposed for continuous measurements, can also be applied to binary and even unbalanced data (Ridout *et al.*, 1999; Landis and Koch, 1977).

Consider the data collected from a random sample of N subjects rated by varying sets of raters. Let $x_{ij} = 1$ for a positive rating outcome and $x_{ij} = 0$ for negative rating outcome. The one-way random effects model is given by

$$x_{ij} = \mu_i + s_i + e_{ij}$$

where $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, n_i$ and $\{s_i\}$ are iid with mean 0 and variance σ_s^2, $\{e_i\}$ are iid with mean 0 and variance σ_e^2, $\{s_i\}$ and $\{e_i\}$ are independent. It can be written that $E(x_{ij}) = p = \Pr(x_{ij} = 1)$ and since x_{ij} are binary data, $\sigma^2 = \text{var}(x_{ij}) = p(1-p)$.

Let $\delta = \Pr(x_{ij} = 1, x_{il} = 1) = \mathrm{E}(x_{ij}x_{il})$, for $j \neq l$, it then shows

$$\delta = \text{cov}(x_{ij}, x_{il}) + \mathrm{E}(x_{ij})\mathrm{E}(x_{il}) = \rho p(1-p) + p^2,$$

where $\rho = (\delta - p^2)/[p(1-p)]$. Then the probability of subjects receiving same measurements is $p_o = p^2 + (1-p)^2 + 2\rho p(1-p)$. The chance-expected agreement can be obtained by substituting $\rho = 0$ as $p_e = p^2 + (1-p)^2$. The chance-corrected agreement index can be given as

$$\rho = \frac{p_o - p_e}{1 - p_e}.$$

Thus it is clear that ρ can also be interpretable as a chance corrected index.

For the ANOVA estimator, let

$$\sigma_s^2 = \rho p(1-p)\sigma_e^2 = (1-\rho)p(1-p),$$

then $\sigma^2 = \sigma_s^2 + \sigma_e^2$. The corresponding estimator is given by

$$\hat{\rho} = \frac{MSB - MSW}{MSB + (\bar{n} - 1)MSW}$$

where

$$MSB = \frac{1}{N-1}\left[\sum_i^N \frac{x_{i.}^2}{n_i} - \frac{\left(\sum_i^N x_{i.}\right)^2}{\sum_i^N n_i}\right],$$

$$MSW = \frac{1}{\sum_i^N n_i - N}\left[\sum_i^N x_{i.} - \sum_i^N \frac{x_{i.}^2}{n_i}\right],$$

$$\bar{n} = \frac{1}{N-1}\left[\sum_i^N n_i - \sum_i^N \frac{n_i^2}{\sum_i^N n_i}\right].$$

These expressions can be simplified in the context of reliability studies where usually $n_i = n$ for all i.

2.3.2 *Fleiss-Cuzick Estimator and Mak's ρ*

Ridout *et al.* (1999) summarized that Fleiss-Cuzick (1979) estimator and Mak's (1988) ρ both have a direct probabilistic interpretation.

Let α represent the probability that two measurements are equal when they come from the same subject, β represent the the probability of two identical measurements when they come from different subjects. It is shown that $\alpha = 1 - 2p(1-p)(1-\rho)$ and $\beta = 1 - 2p(1-p)$, and hence

$$\rho = \frac{\alpha - \beta}{1 - \beta}.$$

An unbiased estimator of α for the ith subject is

$$1 - \frac{2x_{i.}(n_i - x_{1.})}{n_i(n_i - 1)}.$$

Fleiss and Cuzick treated α as a weighted average of these within-group estimators with weights proportional to $n_i - 1$, while Mak used the unweighted average.

As for β, Fleiss and Cuzick estimated it as $1 - 2\hat{p}(1-\hat{p})$, where $\hat{p} = (\sum x_{i.})/\sum n_i$ and the estimator is proposed as

$$\hat{\rho}_{FC} = 1 - \frac{\sum_i^N x_{1.}(n_i - x_{1.})/n_i}{(\sum_i^N n_i - N)\hat{p}(1-\hat{p})}.$$

The unbiased estimator of β for the ith and jth subjects is shown to be

$$\frac{x_{i.}x_{j.} + (n_i - x_{i.})(n_j - x_{j.})}{n_i n_j}$$

and Mak estimated β as unweighted average a sum of $N(N-1)/2$ between-subjects estimators and the estimator is proposed as

$$\hat{\rho}_M = 1 - \frac{(N-1)\sum_i^N x_{1.}(n_i - x_{i.})/[n_i(n_i-1)]}{\sum_i^N (x_{i.}^2/n_i^2) + \sum_i^N (x_{i.}/n_i)(N - 1 - \sum_i^N (x_i/n_i))}$$

2.4 Relationship between ICC and Chance-corrected Indices

Bartko (1966) proposed three different ANOVA models depending on the rater effects. Fleiss (1975) summarized three models and linked the connection to the chance-corrected indices. Blackman and Koval (1993) summarized the relationship between various kappa indices and intraclass correlation coefficients estimated from ANOVA. We present a brief summary here.

Let

$$\begin{aligned}
SS_b &= [4n_{11}n_{22} + (n_{11} + n_{22})(n_{12} + n_{21})], \\
SS_w &= (n_{12} + n_{21})/2, \\
SS_j &= (n_{12} - n_{21})^2/2N, \\
SS_r &= [4n_{12}n_{21} + (n_{11} + n_{22})(n_{12} + n_{21})]/2N,
\end{aligned}$$

representing between subjects, within subjects, between raters and residual sum of squares respectively each with $N-1$, N, 1 and $N-1$ degrees of freedoms (Blackman and Koval, 1993).

If the potential differences between raters can be ignored, then a one-way random effects model can be used. The appropriate estimate of ICC is shown to be

(Bartko, 1966)

$$R_1 = \frac{MS_b - MS_w}{MS_b + MS_w}$$
$$= \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} + n_{21})^2 + (n_{12} + n_{21})}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21}) - (n_{12} + n_{21})},$$

equivalent to Mak's $\hat{\rho}$. If N is sufficiently large so that the difference between N and $N-1$ is negligible,

$$R_2 = \frac{MS_b - MS_w}{MS_b + MS_w}$$
$$= \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})},$$

which is equivalent to Scott's $\hat{\pi}$ and Bloch and Kraemer's MLE.

When the raters are considered to be a random sample from a large population, one should use a two-way random effects model. The appropriate estimate of ICC is shown to be (Bartko, 1966)

$$R_3 = \frac{MS_b - MS_r/2}{(MS_b + MS_r)/2 + (MS_j + MS_r)/N + MS_r}$$
$$= \frac{\frac{4n_{11}n_{22}+(n_{11}+n_{22})(n_{12}+n_{21})}{4N(N-1)} - \frac{4n_{12}n_{21}+(n_{11}+n_{22})(n_{12}+n_{21})}{4N(N-1)}}{\frac{4n_{12}n_{21}+(n_{11}+n_{22})(n_{12}+n_{21})}{2N(N-1)} + \frac{(n_{12}+n_{21})^2}{2N^2} - \frac{4n_{12}n_{21}+(n_{11}+n_{22})(n_{12}+n_{21})}{2N^2(N-1)}}.$$

Again when N is large enough to ignore the difference between N and $N-1$,

$$R_3 = \frac{\frac{1}{N^2}(n_{11}n_{22} - n_{12}n_{21})}{\frac{1}{2N^2}[4n_{11}n_{22} + (n_{11} + n_{22})(n_{12} + n_{21}) + n_{12}^2 + n_{21}^2] - \frac{1}{2N^3}[4n_{12}n_{21} + (n_{11} + n_{22})(n_{12} + n_{21})]}.$$

If the term of order $1/N$ is ignored,

$$R_3 = \frac{2(n_{11}n_{22} - n_{12}n_{21})}{(n_{11} + n_{12})(n_{12} + n_{22}) + (n_{11} + n_{21})(n_{12} + n_{21})},$$

which is Cohen's $\hat{\kappa}$.

If the raters are fixed, then a two-way mixed effects model should be used. The appropriate estimate of ICC is shown to be (Bartko, 1966)

$$\begin{aligned} R_4 &= \frac{MS_b - MS_r}{MS_b + MS_r} \\ &= \frac{2(n_{11}n_{22} - n_{12}n_{21})}{(n_{11}+n_{12})(n_{21}+n_{22}) + (n_{11}+n_{21})(n_{12}+n_{22})}. \end{aligned}$$

Thus, Scott's π, Cohen's κ and Intraclass kappa can all be interpreted as both chance-corrected agreement measurements and intraclass correlation coefficients. The choice of ANOVA model depends on the assumed rater effects. The one-way effects model assumes no rater difference, implying marginal homogeneity. The two-way effects model allows differences between raters, implying the marginal probabilities can be unequal. When the marginal probabilities are equal, the two-way effects model reduces to the one-way effects model.

Ridout *et al.* (1999) showed the equivalence of ANOVA estimator and Mak's ρ. Fleiss and Cuzick (1979) proved that, when MSB has a divisor N instead of $N-1$, then

$$\hat{\rho} = \frac{\hat{\rho}_{FC}}{1 - \frac{(1-\hat{\rho}_{FC}\sum_i^N (n_i - \sum_i^N n_i/N)^2}{N(N-1)(\sum_i^N n_i/N)^2}}$$

and ANOVA estimator and Fleiss-Cuzick estimator are virtually identical with N large enough.

With the link between chance-corrected agreement indices and ICC, Kraemer (1979) has elucidated the meaning of ICC, discussing its effect on estimation, precision and statistical power. Similar discussions can be found in Lachin (2004).

2.5 Sample Size Estimation

As with many science studies, sample size determination is an essential step to fulfill the study objectives. A study with a too large sample size wastes resources

while a study with a too small sample size cannot answer the study question with reasonable certainty.

One type of approaches to calculate sample size is from a hypothesis testing perspective. A graph of the general relationship between reliability coefficient, statistical power and sample size when detecting the null hypothesis of a specified difference in true scores is presented by Lachin (2004). It shows that the sample size required to maintain the same power increases when the reliability coefficient decreases. The inflation factors are also listed for corresponding reliability coefficient values, e.g., when the reliability equals 0.8, 25% extra sample size is needed, when the reliability is only 0.5, the sample size should be doubled to maintain the desired level of power.

The asymptotic variance for Cohen's kappa (Fleiss *et al.*, 1969) made the sample size calculation feasible. Flack *et al.* (1988) assumed equal marginal rating probabilities and presented a sample size calculation for two raters and multiple categories by maximizing the standard error. The sample size is obtained in order that a test of null hypothesis that kappa is no larger than a certain value has a pre-specified significance level and power. By maximizing the standard error, this approach can also be applied to obtain a sample size that gives a targeted kappa's confidence interval length.

Cantor (1996) extended this approach by allowing unequal marginal probabilities with two raters and two categories. His approaches allow one to calculate sample size requirements for either testing the null hypothesis that estimated kappa equals a certain value, or testing if two kappas estimated from two independent samples are equal with desired significance level and power.

On the base of common correlation model, a goodness- of-fit approach (Donner and Eliasziw, 1992) was applied to facilitate sample size calculation with prespecified power. It follows from that the goodness-of-fit statistic has one degree

of freedom non-central chi-square distribution under the alternative hypothesis $H_1 : \kappa = \kappa_1$ with non-centrality parameter

$$\lambda(1) = N \sum_{i=1}^{3} \frac{[p_i(\kappa_1) - p_i(\kappa_0)]^2}{p_i(\kappa_0)}.$$

If $1 - \beta(1, \lambda, \alpha)$ denotes the power of the goodness-of-fit statistic, then the sample size N can be determined using tables of the non-central chi-square distribution (Haynam *et al.*, 1970) to ensure that power exceeds a pre-specified level. The formula of sample size to conduct a two-sided test with significance level α and power $1 - \beta(1, \lambda, \alpha)$ is shown as

$$N = \lambda(1, 1-\beta, \alpha) \times$$
$$\left\{ \frac{[p(1-p)(\kappa_1 - \kappa_0]^2}{p^2 + p(1-p)\kappa_0} + \frac{2[p(1-p)(\kappa_1 - \kappa_0]^2}{p(1-p)(1-\kappa_0)} + \frac{[p(1-p)(\kappa_1 - \kappa_0]^2}{(1-p)^2 + p(1-p)\kappa_0} \right\}^{-1},$$

where $\lambda(1, 1-\beta, \alpha)$ is the tabulated non-centrality parameter (Haynam *et al.*, 1970). Note that in this one degree of freedom case, the value of $\lambda(1, 1-\beta, \alpha)$ is the same of $(Z_{1-\alpha/2} + Z_{1-\beta})^2$ where $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are the critical values of the standard normal distribution corresponding to α and β.

Altaye *et al.* (2001) generalized the common correlation model and the goodness-of-fit approach to multiple raters by applying the joint probability function first proposed by Bahadur (1961) and later reparameterized to an expression in terms of positively rating probability instead of correlation by George and Bowman (1995). Similar ideas have been adopted to extend the goodness-of-fit sample size formula to multiple raters. More detailed derivation of the model will be introduced in the next chapter, as the common correlation model is the basis on which we derive the variance and sample size formulas.

Shoukri *et al.* (2004) provided a solution concerning efficient allocation when the product of the number of subjects and the number of raters is fixed. The sample size calculated minimizes the variance of the estimate of intraclass kappa. It has

been recommended that with the acceptable level of reliability coefficient, two or three measurements for each subject is reasonable.

However, it is generally considered that the reliability studies are designed to estimate the level of agreement and results of the studies are often reported in terms of estimates of agreement instead of hypothesis testing. The specification of the alternative hypothesis testing level is difficult in practice. Besides, rejection of the null hypothesis is not informative, since researchers need to know more than the fact that the agreement is not caused by chance. As confidence intervals are considered to be more informative than a single estimate, another type of approach from the precision perspective seems to be consistent with the trend.

Among those approaches, Donner (1999) discovered that after solving confidence intervals from the goodness-of-fit statistic, the interval width depends only on the total number of subjects N, total number of discordant pairs of ratings $n_{12} + n_{21}$, the observed success rate $\hat{p}$ and the probability of coverage $(1 - \alpha)$. One can estimate sample size needed to ensure the estimated confidence interval width is less than a pre-specified value by replacing $\hat{p}$ with anticipated value p, $n_{12} + n_{21}$ replaced with the expected value from the common correlation model for two raters, given by $N\mathrm{p}_2(\kappa_1) = 2Np(1-p)(1-\kappa_1)$.

Another formula also from precision perspective is provided by Donner and Rotondi (2010) for multiple raters. After pre-specifying the targeted agreement level, the lower confidence interval limit and the overall success rate, an iterative procedures can be adopted to determine the minimum sample size so that a one-sided 95% confidence interval has an expected lower bound. However, those existent approaches implicitly achieve a pre-specified precision with 50% chance, which will be shown from the evaluation study in Chapter 4.

The above issue has lead to our proposal in the next chapter of sample size formulas that explicitly incorporates both precision and assurance probability.

Chapter 3

DERIVATION OF SAMPLE SIZE FORMULAS

3.1 Introduction

In the planning of a reliability study, the researcher is interested in how many subjects need to be recruited in order to estimate reliability with reasonable precision. This chapter derives the sample size formulas that explicitly incorporate both the precision and assurance probability. In section 3.2, the common correlation model underlying the estimation of the the ICC is introduced. The point estimate of the ICC, its variance and confidence interval are introduced. In section 3.3, two sample size formulas based on the model introduced in section 3.2 are derived to have desirable pre-specified assurance probability characteristics.

3.2 Common Correlation Model with Multiple Raters

Under the same notations used previously that x_{ij} represents the binary measurement assigned from the jth rater $(j = 1, 2, ..., n)$ to the ith subject $(i = 1, 2, ..., N)$ and $\pi = \mathrm{P}(x_{ij} = 1)$ represents the underlying success rate for all the n raters, a model specifically for correlated binary data was firstly proposed by Bahadur (1961) under the assumption of interchangeability. Note that when this assumption is in doubt, the Cochran's Q-statistic test can be performed to test no rater bias, as illustrated by Fleiss (1973). Letting $X_i = \sum_{j=1}^{N} x_{ij}$ representing the total rating

scores for the ith subject, the joint probability of measurements can be expressed as (Bahadur, 1961)

$$\mathrm{P}(X_i = x_i) =$$
$$\binom{n}{x_i} \pi^{x_i}(1-\pi)^{n-x_i}[1 + \sum_{i<j} \rho Z_i Z_j + \sum_{i<j<k} \rho_3 Z_i Z_j Z_k + \ldots + \rho_n Z_1 Z_2 \ldots Z_n]$$

where

$$Z_i = \frac{x_i - \pi}{[\pi(1-\pi)]^{\frac{1}{2}}},$$
$$\rho_2 = \rho = \mathrm{E}(Z_i Z_j), \ldots, \rho_n = \mathrm{E}(Z_1 Z_2 \ldots Z_n).$$

In interrater agreement studies, the parameter of interest is ρ, which is shown (Fleiss and Cuzick, 1979) to be equal to the ICC obtained from a one-way random ANOVA model applied to dichotomous data. The correlation between any two ratings by the same rater, ρ is defined as

$$\rho = \frac{\mathrm{E}(x_{ij}x_{ik}) - \pi^2}{\pi(1-\pi)} \quad \text{for } j \neq k.$$

Using maximum likelihood estimation method to estimate parameters in the joint probability function could be attempted. However as n increases, the number of parameters needed to be estimated proliferates rapidly and this method becomes challenging.

As shown in Altaye *et al.* (2001) and Donner and Rotondi (2010) later work resorts to a reparameterization presented by George and Bowman (1995). The joint probability of measurements is then expressed in terms of success probabilities rather than correlations:

$$\mathrm{P}(X_i = x_i) = \binom{n}{x_i} \sum_{u=0}^{n-x_i} (-1)^u \binom{n-x_i}{u} \lambda_{x_i+u},$$

where $\lambda_k = \mathrm{P}(x_{i1} = 1, x_{i2} = 1, \ldots, x_{ik} = 1)$, $\lambda_0 = 1$ and the maximum likelihood estimate of λ_k is given by (Altaye *et al.*, 2001)

$$\widehat{\lambda_k} = \frac{\sum_{u=0}^{n-k} \binom{n-u}{k} m_{n-u}}{N\binom{n}{k}}$$

where m_j represents the number of subjects whose total rating scores x_i equals j $(j = 0, 1, \ldots, n)$.

By applying the expression provided by Altaye *et al.* (2001)

$$\lambda_k = \pi^k + \rho(1-\pi)\sum_{j=1}^{k-1} \pi^{k-j},$$

the joint probability of measurements can be written as (Zou and Donner, 2004)

$$\Pr(X = x) = \begin{cases} \rho(1-\pi) + (1-\rho)(1-\pi)^n & x = 0 \\ \binom{n}{x}(1-\rho)\pi^x(1-\pi)^{n-x} & 1 \leqslant x \leqslant n-1 \\ \rho\pi + (1-\rho)\pi^n & x = n \end{cases}$$

and ρ must satisfy

$$\max\left[-\frac{(1-\pi)^n}{(1-\pi)-(1-\pi)^n},\ -\frac{\pi^n}{\pi-\pi^n}\right] \leqslant \rho \leqslant 1.$$

Note that this model has been adopted by Donner *et al.* (1981) in the context of designing cluster randomization trials.

This model is more general than the one by Altaye *et al.* (2001). To see this, consider the case of $n = 3$, where the joint probability of measurements reduce to

$$\begin{aligned} p_0 &= \Pr(X_i = 0) = (1-\pi)^3 + 3\rho\pi(1-\pi)^2 - \rho_3[\pi(1-\pi)]^{\frac{3}{2}} \\ p_1 &= \Pr(X_i = 1) = 3\pi(1-\pi)^2 - 3\rho\pi(1-\pi)(2-3\pi) + 3\rho_3[\pi(1-\pi)]^{\frac{3}{2}} \\ p_2 &= \Pr(X_i = 2) = 3\pi^2(1-\pi)^2 + 3\rho\pi(1-\pi)(1-3\pi) - 3\rho_3[\pi(1-\pi)]^{\frac{3}{2}} \\ p_3 &= \Pr(X_i = 3) = \pi^3 + 3\rho\pi^2(1-\pi) + \rho_3[\pi(1-\pi)]^{\frac{3}{2}} \end{aligned}$$

which can be summarized in Table 3.1. When there are only two raters , it further reduces to the well-known common correlation model for two raters Bloch and Kraemer (1989). This model can be reparameterized to (Donner and Rotondi, 2010)

$$\begin{aligned}
\mathrm{P}(X_i = 0) &= (1-\pi)^3 + \rho\pi[(1-\pi)^2 + (1-\pi)] \\
\mathrm{P}(X_i = 1) &= 3\pi(1-\pi)^2(1-\rho) \\
\mathrm{P}(X_i = 2) &= 3\pi^2(1-\pi)(1-\rho) \\
\mathrm{P}(X_i = 3) &= \pi^3 + \rho\pi(1-\pi^2).
\end{aligned}$$

Table 3.1: Data Layout for Three Raters

Category	Ratings	Frequency	Probability
0	$(0,0,0)$	m_0	p_0
1	$(0,0,1),(0,1,0),(1,0,0)$	m_1	p_1
2	$(0,1,1),(1,1,0),(1,0,1)$	m_2	p_2
3	$(1,1,1)$	m_3	p_3
Total		N	1

Since the observed frequencies m_j $(j = 0, 1, \ldots, n)$ follow a multinomial distribution with parameters $(N, p_j\ (j = 0, 1, \ldots, n))$, the reliability coefficient, ICC, is defined as (Kraemer, 1979)

$$\rho = \frac{\lambda - \pi^2}{\pi(1-\pi)}$$

where

$$\pi = \frac{\sum_{j=1}^{n} j p_j}{n}, \ \ \lambda = \frac{\sum_{j=2}^{n} j(j-1)p_j}{n(n-1)},$$

and the maximum likelihood estimates are shown to be

$$\hat{\rho} = \frac{\hat{\lambda} - \hat{\pi}^2}{\hat{\pi}(1-\hat{\pi})}$$

where

$$\hat{\pi} = \frac{\sum_{j=1}^{n} jm_j}{nN}, \ \hat{\lambda} = \frac{\sum_{j=2}^{n} j(j-1)m_j}{nN(n-1)}.$$

By applying moment generating function and delta method, the variance formula of $\hat{\rho}$ is proposed as (Zou and Donner, 2004)

$$\begin{aligned} &\text{var}(\hat{\rho}) = \frac{1-\rho}{N} \times \\ &\left\{ \frac{2}{n(n-1)} - \left[3 - \frac{1}{\pi(1-\pi)} \right] \rho + \frac{n-1}{n} \left[4 - \frac{1}{\pi(1-\pi)} \right] \rho^2 \right\}. \end{aligned} \tag{3.1}$$

For confidence intervals, the commonly applied Wald method produces upper and lower confidence interval limits as

$$\hat{\rho} \pm Z_{\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\rho})}$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. However, in studies by Donner and Eliasziw (1992) and Altaye *et al.* (2001), it was shown that Wald method performs poorly especially when the sample size is small, say less than 100, or when π and ρ are extreme values, likely the result of forced symmetry to construct limits. A second approach equates the observed Pearson χ^2 statistic to the critical value of $\chi^2_{(1)}$ distribution at the pre-specified significance level and then solved for two admissible roots as the confidence limits. Nevertheless, the iterative procedures are rather time-consuming and complicated. An alternative approach that inverts the modified Wald test (Lee and Tu, 1994; Donner and Zou, 2002) is to solve the cubic equation

$$(\hat{\rho} - \rho)^2 = Z^2_{\alpha/2} \widehat{\text{var}}(\hat{\rho}).$$

By replacing $\hat{\pi}$ for π and $\hat{\rho}$ for ρ to obtain $\widehat{\text{var}}(\hat{\rho})$, the equation above can be rewritten as

$$a\rho^3 + b\rho^2 + c\rho + d = 0$$

where

$$
\begin{aligned}
a =& -\frac{Z^2_{\alpha/2}}{N}\frac{n-1}{n}\left[4-\frac{1}{\pi(1-\pi)}\right] \\
b =& \left\{\frac{Z^2_{\alpha/2}}{N}\left[\frac{n-1}{n}(4-\frac{1}{\pi(1-\pi)})+(3-\frac{1}{\pi(1-\pi)})\right]-1\right\} \\
c =& -\left\{\frac{Z^2_{\alpha/2}}{N}\left[3-\frac{1}{\pi(1-\pi)}+\frac{2}{n(n-1)}\right]-2\hat{\rho}\right\} \\
d =& \frac{Z^2_{\alpha/2}}{N}\frac{2}{n(n-1)}-\hat{\rho}^2
\end{aligned}
$$

and upper and lower $(1-\alpha)\times 100\%$ confidence interval limits can be obtained as the two admissible roots of the above cubic equation. When $a = 0$, it becomes a quadratic equation and confidence interval limits are

$$
\begin{aligned}
\rho_L &= \max(-1, \frac{-c-\sqrt{c^2-4bd}}{2b}), \\
\rho_U &= \min(\frac{-c+\sqrt{c^2-4bd}}{2b}, 1).
\end{aligned}
$$

When $a \neq 0$, the confidence interval limits are the two roots that are within the range of -1 and 1, that is

$$
\begin{aligned}
\rho_L &= \max(-1, -\frac{b}{3a}+2\sqrt{-\beta}\cos\left[\frac{\arccos\frac{\alpha}{\sqrt{-\beta^3}}+2\pi}{3}\right]), \\
\rho_U &= \min(-\frac{b}{3a}+2\sqrt{-\beta}\cos\left[\frac{\arccos\frac{\alpha}{\sqrt{-\beta^3}}-2\pi}{3}\right], 1),
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha &= -\frac{b^3}{27a^3}-\frac{d}{2a}+\frac{bc}{6a^2}, \\
\beta &= \frac{c}{3a}-\frac{b^2}{9a^2}.
\end{aligned}
$$

3.3 Derivation of Sample Size Formulas

When considering the sample size requirement for reliability studies, one's interest often lies in whether the estimated agreement level exceeds a certain threshold ρ_L. Several qualifiers may be used to describe a particular level of agreement. A commonly used set of qualifiers are provided by Landis and Koch (1977), where $\rho < 0$ is labelled 'poor', $\rho \leqslant 0.2$ 'slight', $\rho \leqslant 0.4$ 'fair', $\rho \leqslant 0.6$ 'moderate', $\rho \leqslant 0.8$ 'substantial', and $\rho > 0.8$ 'almost perfect' agreement. The goal here is to derive a formula for sample size N to ensure a pre-specified $1-\beta$ assurance probability, by fixing the lower confidence limit of estimated reliability coefficient $\hat{\rho}_L$ for multiple raters to be no less than a pre-specified value ρ_L. Here the approach follows that of Zou (2012) and starts by writing

$$\begin{aligned} 1-\beta =& \Pr(\hat{\rho}_L \geqslant \rho_L) \\ =& \Pr\left[\hat{\rho} - Z_\alpha \sqrt{\text{var}(\hat{\rho})} \geqslant \rho_L\right]. \end{aligned} \tag{3.2}$$

where $\text{var}(\hat{\rho})$, as derived by Zou and Donner (2004), was given in Equation (3.1).

Let

$$f(\rho) = (1-\rho)\left\{\frac{2}{n(n-1)} - \left[3 - \frac{1}{\pi(1-\pi)}\right]\rho + \frac{n-1}{n}\left[4 - \frac{1}{\pi(1-\pi)}\right]\rho^2\right\},$$

which is the numerator of the variance formula for ρ. The equation (3.1) can be written as

$$\begin{aligned} 1-\beta &= \Pr\left[\hat{\rho} - Z_\alpha\sqrt{f(\rho)/N} \geqslant \rho_L\right] \\ &= \Pr\left[\hat{\rho} \leqslant -\rho_L - Z_\alpha\sqrt{f(\rho_L)/N}\right]. \end{aligned} \tag{3.3}$$

By central limit theorem, we know

$$\hat{\rho} \sim N\left[\rho,\ \frac{f(\rho)}{N}\right]$$

so that equation (3.3) simplifies to

$$1-\beta = \Pr\left[\frac{\hat{\rho}-\rho}{\sqrt{f(\rho)/N}} \leqslant \frac{-\rho-\rho_L - Z_\alpha\sqrt{f(\rho_L)/N}}{\sqrt{f(\rho)/N}}\right].$$

After solving the above equation, the sample size N can be derived as

$$N = \left[\frac{Z_\alpha\sqrt{f(\rho_L)} + Z_\beta\sqrt{f(\rho)}}{\rho-\rho_L}\right]^2 \tag{3.4}$$

where Z_β is the upper β quantile for the standard normal distribution. Also the assurance probability $1-\beta$ can be calculated when given N, ρ_L, π and ρ_0:

$$Z_\beta = \frac{-\rho-\rho_L - Z_\alpha\sqrt{f(\rho_L)/N}}{\sqrt{f(\rho)/N}}.$$

If a minimum level of precision is required by restricting the confidence interval, the minimum sample size can be computed for a two-sided confidence interval of $\hat{\rho}$ such that each tail is no larger than a pre-specified width ω with a pre-specified assurance probability $1-\beta$.

Under the above requirements, the equation can be written as

$$1-\beta = \Pr\left[Z_{\alpha/2}\sqrt{\text{var}(\hat{\rho})} \leqslant \omega\right],$$

which implies

$$1-\beta = \Pr\left[\sqrt{f(\hat{\rho})} \leqslant \frac{\omega\sqrt{N}}{Z_{\alpha/2}}\right].$$

To derive the distribution of $\sqrt{f(\hat{\rho})}$, the delta method is applied as

$$\begin{aligned}\text{var}(\sqrt{f(\hat{\rho})}) &= \left(\sqrt{f(\rho)}\right)'^2 \text{var}(\hat{\rho}) \\ &= \frac{1}{4N}[f'(\rho)]^2\end{aligned}$$

where $f'(\rho)$ is the first-order derivative of $f(\rho)$ with respect to ρ, given by

$$\begin{aligned} f'(\rho) = & -3\frac{n-1}{n}\left[4-\frac{1}{\pi(1-\pi)}\right]\rho^2 \\ & +2\left\{3-\frac{1}{\pi(1-\pi)}+\frac{n-1}{n}\left[4-\frac{1}{\pi(1-\pi)}\right]\right\}\rho \\ & -\frac{2}{n(n-1)}-3+\frac{1}{\pi(1-\pi)}. \end{aligned}$$

Again by central limit theorem,

$$\sqrt{f(\hat{\rho})} \sim N\left(\sqrt{f(\rho)},\ \frac{1}{4N}[f'(\rho)]^2\right).$$

Now solve for N asymptotically:

$$1-\beta = \Pr(Z \leqslant \frac{\omega\sqrt{N}/Z_{\alpha/2}-\sqrt{f(\rho)}}{|f'(\rho)|/2N}),$$

$$N = \left[\frac{\sqrt{f(\rho)}+\sqrt{f(\rho)+2\omega Z_\beta|f'(\rho)|/Z_{\alpha/2}}}{2\omega/Z_{\alpha/2}}\right]^2. \tag{3.5}$$

If ρ, ω, N and π are previously known, the assurance probability $1-\beta$ can be calculated as

$$Z_\beta = \frac{\omega\sqrt{N}/Z_{\alpha/2}-\sqrt{f(\rho)}}{|f'(\rho)|/2\sqrt{N}}.$$

Sample size formulas are now derived to allow requirements on both precision and assurance probability. Equation (3.4) gives the sample size that assures the lower $(1-\alpha)100\%$ one-sided confidence interval limit is no less than a pre-specified value ρ_L with $1-\beta$ assurance probability, and Equation (3.5) gives the sample size that assures the width of $(1-\alpha)100\%$ two-sided confidence interval is no larger than a pre-specified value ω with $1-\beta$ assurance probability. To prove the sample size formulas have satisfactory performance, evaluation studies are carried out in next chapter.

Chapter 4

EVALUATION OF THE FORMULAS

4.1 Introduction

The results in Chapter 3 are derived under asymptotic conditions. Therefore a numerical study was conducted to study performance under practical situations. Both exact evaluation and simulation studies can be used to provide empirical estimation of the sampling distribution of the estimators. Therefore this chapter adopts this technique to evaluate empirical sample size formulas performance.

In Section 4.2 of this chapter, the parameter selection, the background parameter settings and criteria to evaluate the performance of the estimators are determined. Then detailed procedures and methods to carry out both exact evaluation and simulation studies are discussed. Section 4.3 reports and discusses the evaluation and simulation results. A conclusion of this chapter is given in Section 4.4.

4.2 Study Design

4.2.1 Parameter Selection and Data Generation

There are various sample size formulas in the literature dealing with the two rater case (Cantor, 1996; Donner and Eliasziw, 1992) and in recent years several formulas have been extended to multiple raters (Altaye *et al.*, 2001; Donner and Rotondi,

2010). Consider a scenario in which a group of n raters, here, assumed to be between two and five raters, each assigns dichotomous measurements to a total of N subjects. The aim is to measure reliability with desired precision, either using a pre-specified lower confidence interval limit ρ_L, or a pre-specified half confidence interval width ω, with the nominal significance level α and the nominal assurance probability $1 - \beta$. The raters have the same tendency given by probability p to assign subjects a positive measurement and within each subject, the true intraclass correlation coefficient ρ represents the correlation between any two measurements. The significance level α is set always equal to 0.05 so that the confidence intervals are either 95% one-sided lower confidence interval or 95% two-sided confidence interval depending on the sample size formulas.

Values for the true reliability coefficient ρ were selected based on the suggestions from Landis and Koch (1977) indicating 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) agreement. Since in many clinical investigations an ICC of 0.6 is expected to be the minimal acceptable level, the values for ρ were chosen to be 0.6, 0.7 and 0.8, and respectively, the values of pre-specified lower confidence interval limit ρ_L were set to be 0.4, 0.6, and 0.6, or alternatively the values of the half pre-specified confidence interval widths were 0.2, 0.1 and 0.2, respectively. Under those conditions, the performances of both the general cases with wide confidence intervals and extreme cases with narrow confidence intervals can be compared.

As for the values of the probability of raters assigning a positive measurement, due to the fact that both the values of p and $1 - p$ result in the same agreement level, only the p values of 0.1, 0.3 and 0.5 are considered.

Also, in this evaluation study, sample sizes calculated from the formulas under 50% assurance probability level are compared to those from Donner (1999) and Donner and Rotondi (2010) to show that their sample size approach implicitly only

incorporate 50% assurance probability. The sample size formulas are evaluated under 80% assurance level as well.

For the cases of two and three raters, an exact evaluation study (Zou *et al.*, 2009) was carried out to assess all the possibilities. For each parameter combination, the required minimum sample size N was calculated at the first step. Then with the total sample size N, every possible allocation to $n + 1$ categories of total measurement scores can be listed, that is, every subject has a total measurement score from 0 (every rater assigned a negative measurement), 1, ..., to n (every rater assigned a positive measurement). Under each allocation, the corresponding confidence interval can be constructed to evaluate the sample size formulas.

For the cases of four and five raters, since an exact evaluation based on all the possibilities is very computationally intensive, a simulation study is conducted instead. Under each parameter combination, with the minimum sample size N obtained from the formulas, a total of 100 random sets of numbers were generated following a multinomial distribution with probabilities obtained from the joint probability density function. Similarly, confidence intervals can be constructed to assess the performance of the sample size formulas. All numerical studies were performed using R Version 3.2.5 (R Core Team, 2016).

4.2.2 Confidence Interval Methods Compared

The traditional confidence intervals calculated by the Wald method is shown to perform poorly (Donner and Eliasziw, 1992; Altaye *et al.*, 2001). An alternative approach to obtain the confidence interval is to invert a modified Wald test (Rao and Mukerjee, 1997) so that the confidence interval limits are two admissible roots of the cubic equation in terms of ρ. In this numerical study, these two approaches are referred as "Wald" and "M. Wald" and their performance in terms of coverages and assurance probabilities were compared.

4.2.3 *Procedures and Evaluation Criteria*

Given each combination of parameters (n, p, ρ, ρ_L or ω, α, β), by applying the sample size formulas proposed in the last chapter, the minimum sample size N can be calculated. With the sample size N, every possible allocation of sample size to $n+1$ categories of total measurement scores can be listed for the evaluation study, or alternatively random sets of numbers can be generated for a simulation study. The numbers of subjects assigned a total measurement score follows a multinomial distribution with probabilities obtained from the joint probability equation introduced in the last chapter

$$\Pr(X=x)=\begin{cases}\rho(1-\pi)+(1-\rho)(1-\pi)^n & x=0\\ \binom{n}{x}(1-\rho)\pi^x(1-\pi)^{n-x} & 1\leqslant x\leqslant n-1\\ \rho\pi+(1-\rho)\pi^n & x=n\end{cases}$$

Then when $n=3$ the probability for every particular condition can be calculated through the joint density

$$\begin{aligned} f(m_0,m_1,m_2,m_3;N,p_0,p_1,p_2,p_3) &= \frac{N!}{m_0!m_1!m_2!m_3!}p_0^{m_0}p_1^{m_1}p_2^{m_2}p_3^{m_3}, \qquad (4.1)\\ m_0+m_1+m_2+m_3 &= N,\\ p_0+p_1+p_2+p_3 &= 1. \end{aligned}$$

Under every condition, $100(1-\alpha)\%$ confidence intervals $(L,\ U)$ can be obtained by using the Wald and modified Wald method.

In both the evaluation and simulation studies, the comparison of generated results to the "true" values provides a measure of performance. This is achieved by assessing the confidence intervals for the selected parameters. The empirical confidence intervals under every condition were obtained to measure how often the true intraclass correlation coefficient ρ could be correctly predicted by an in-

terval based on either Wald or modified Wald methods. This measurement was quantified by calculating coverage probabilities.

For this study, coverage is defined as (Zou *et al.*, 2009)

$$\text{Coverage} = \sum_{i=1}^{S} \frac{N!}{m_{0i}!m_{1i}!\ldots m_{ni}!} p_0^{m_{0i}} p_1^{m_{1i}} \ldots p_n^{m_{ni}} \times I(L < \rho) \times 100$$

for the sample size calculated by a pre-specified 95% one-sided lower confidence interval limit and

$$\text{Coverage} = \sum_{i=1}^{S} \frac{N!}{m_{0i}!m_{1i}!\ldots m_{ni}!} p_0^{m_{0i}} p_1^{m_{1i}} \ldots p_n^{m_{ni}} \times I(L < \rho < U) \times 100$$

for the sample size calculated by a pre-specified 95% two-sided confidence interval width where S equals the number of possible allocations of sample size N to $n+1$ categories of total measurement scores.

For the simulation study, the coverage is defined as the proportion of times, in large number of different data sets generated randomly using the same parameter combination, that the obtained confidence interval contains the true intraclass correlation coefficient ρ.

Empirical coverages approximately equal to the nominal 95% indicate the confidence interval method performs satisfactorily. The criteria used here to assess coverage have also been adopted by many other authors (Zou, 2007, Robey and Barcikowski, 1992): strict criterion (94.5% to 95.5%); moderate criterion (93.75% to 96.25%); liberal criterion (92.5% to 97.5%).

Since the sample size formulas explicitly incorporate the assurance probability β, it is reasonable to evaluate the empirical assurance probability to assess their performance by seeing how often the empirical 95% one-sided lower confidence interval is truly no less than the pre-specified lower confidence interval limit ρ_L, or how often half of the empirical 95% two-sided confidence interval width is no larger than the pre-specified confidence interval width ω.

For the evaluation study, the empirical assurance probability is calculated as

$$\text{Assurance} = \sum_{i=1}^{S} \frac{N!}{m_{0i}!m_{1i}!\ldots m_{ni}!} p_0^{m_{0i}} p_1^{m_{1i}} \ldots p_n^{m_{ni}} \times I(\rho_L < L) \times 100\%$$

for the sample size calculated by a pre-specified a 95% one-sided lower confidence interval limit and

$$\text{Assurance} = \sum_{i=1}^{S} \frac{N!}{m_{0i}!m_{1i}!\ldots m_{ni}!} p_0^{m_{0i}} p_1^{m_{1i}} \ldots p_n^{m_{ni}} \times I((U-L) < 2w) \times 100\%$$

for the sample size calculated by a pre-specified a 95% two-sided confidence interval width. The empirical assurance probabilities are expected to be close to the nominal level if a sample size formula performs well. For the simulation study, the empirical assurance probability is defined as the proportion of times, in repeated data generation, that the obtained lower confidence interval is no less than the pre-specified a 95% one-sided lower confidence interval limit, or the obtained confidence interval width is no larger than the pre-specified a 95% two-sided confidence interval width.

In order to prove validity, the results calculated by using the sample size formulas proposed in the last chapter under 50% assurance probability are compared to the results from Donner (1999) and Donner and Rotondi (2010).

4.3 Discussion of Evaluation Results

4.3.1 Sample Size

Table 4.1 displays the minimum sample sizes required to achieve pre-specified 95% one-sided lower confidence interval limit for different parameter combinations with two raters.

Fewer subjects are generally required in order to achieve a higher agreement level ρ. When the distance between ρ and the threshold ρ_L is a constant 0.2, then

the minimum sample size increases from 125 to 150 when ρ is reduced from 0.8 to 0.6, and p and β are fixed at 0.1 and 0.5 respectively. As for the case where ρ and ρ_L are fixed, the minimum sample sizes all drop as the success rate p increases. Specifically, when p increases from 0.1 to 0.3, the sample sizes are all reduced by more than half for all combinations. It is natural to demand a larger sample size if stronger assurance probability is desired. To maintain the pre-specified lower confidence interval limit, the sample size almost doubles when increasing the assurance probability from 50% to 80%. The requirement for greater assurance probability will create significantly greater costs associated with recruiting more subjects.

Table 4.2 presents the minimum sample sizes required under the same conditions except the number of raters $n = 3$. Increasing from two raters to three raters will lead to a reduction of the sample size by approximately 30%. In practice, recruiting one extra experienced and well-trained rater can be challenging. The optimal combination of the number of subjects N and raters n required to minimize the variance of the intraclass correlation coefficient for both continuous and binary outcomes is discussed by Shoukri *et al.*, (2004). With an minimum agreement level of 0.6 in many clinical investigations, two or three raters are a safe recommendation.

Also it is worth noticing in Table 4.1 that when the requirement on the precision is strict, (i.e., $\rho = 0.7$, $\rho_L = 0.6$ $1 - \beta = 0.8$,) and the two raters' measurement success probability p is as low as 0.1, the sample size can be impractically large (e.g., 1,058). One alternative way to maintain the same precision is to hire an extra rater, as can be seen in Table 4.2 where the corresponding sample size is comparatively low (e.g., 801).

Tables 4.3 and 4.4 presents the sample sizes needed when there are four or five raters available. The reduction in sample size with one extra rater is modest. For example a total of 32 subjects are needed when $\rho = 0.6$, $\rho_L = 0.4$, $p = 0.3$ and

$\beta = 0.5$. When it comes to five raters, 29 subjects are still needed assumed other parameters are fixed. Therefore, it is not very efficient to hire four or five raters for reliability studies.

Similar patterns in sample size requirements are found in Tables 4.5 to 4.8 for sample sizes that achieve a pre-specified 95% confidence interval width. Increasing from two to three raters affords approximately 30% smaller sample size. Additional raters are inefficient given the modest reduction of sample size.

To address the validity of the proposed approach, the results are compared to those from Donner (1999) and Donner and Rotondi (2010) in Tables 4.9 to Table 4.11. The results from two approaches are close to each other under most parameter combinations. For some extreme cases, e.g., p is low and precision requirement is high, there is a difference between two approaches. But generally, the validity of the sample size formulas under 50% assurance probability can be shown.

4.3.2 Coverage

Empirical coverage is a key factor when assessing confidence interval methods. It is expected to be approximately equal to the nominal coverage of 95%. Over coverage indicates the methodology is too conservative and leads to a loss of statistical power with high type II error. Similarly under coverage indicates the methodology is too liberal with high type I error.

Evaluation results in Table 4.1 to Table 4.8 indicate that empirical coverages provided by confidence intervals constructed through the modified Wald method are consistently close to their nominal 95% level. Under most parameter combinations, the coverage provided by the modified Wald method is met with at least the moderate criterion of 93.75% to 96.25% and within this range they are never below 94.5%, which shows that the method almost always produces coverages higher

than the nominal level. However on occasion the modified Wald method provides coverages that are too conservative. For example for three raters and 50% assurance probability, when $\rho = 0.8$, $\rho_L = 0.6$ and $p = 0.5$, the coverage is 98.54% and there are several cases in which the coverages are above 96.25%. The coverages produced by sample sizes achieving a pre-specified 95% two-sided confidence interval widths are more stable. Almost every coverage is within the interval of 93.75% and 96.25% except for the case where $n = 2$, $\rho = 0.8$, $w = 0.2$, $p = 0.5$ and $\beta = 0.5$, where the coverage is very liberal 93.21%.

In contrast, the Wald method provides rather unsatisfying coverage. The method seldom reaches the nominal 95% level and mostly provide coverages that fall within the range of 90% to 94%, an under coverage that can be considered very liberal. In multiple cases they are even less than 90%. When $n = 3$, $\rho = 0.8$, $w = 0.2$, $p = 0.5$ and $\beta = 0.5$, the coverage is only 85.98%, very distant from the nominal level. Even when the sample size is more than 100, the under coverage is still severe.

4.3.3 Assurance Probability

Empirical assurance probability assesses how often the sample size formulas meet the requirement on precision. Being close to the nominal level is an optimal property. Under the condition of 50% nominal assurance probability, the empirical assurance probabilities provided by the modified Wald method are relatively stable and close to the nominal level, while those provided by the Wald method are more heterogeneous within the range of 40% to 70% and rarely close to the nominal level. For the modified Wald method, occasionally the empirical assurance probability outlies to, for example, 69.28%, but in most cases it deviates mildly from the nominal level.

However, for the condition of 80% nominal assurance probability, the results are more erratic for both methods. With the sample sizes achieving a pre-specified

95% one-sided lower confidence interval limit, the empirical assurance probabilities are still reasonable especially for the modified Wald method which consistently produces assurance probabilities slightly under the nominal level and lies within the range of 75% to 80%. However the Wald method can give empirical assurance probability in a much wider range, indicating the method is very unstable. The empirical assurance probability can be as high as 87.93% when $n = 2$, $\rho = 0.8$, $\rho_L = 0.6$ and $p = 0.5$, and as low as 77.34% when $n = 5$, $\rho = 0.6$, $\rho_L = 0.4$ and $p = 0.5$.

When looking at the results with sample sizes that achieve a pre-specified 95% two-sided confidence interval width, the empirical assurance probability can reach a comparatively wider range for both methods. It is worth noticing that in the evaluation study, the modified Wald method almost always gives empirical assurance probabilities higher than those given by the Wald method, which consistently gives empirical assurance probabilities lower than the nominal level 80%. Thus the modified Wald method performs better in achieving a nominal assurance probability. But in a few cases especially when $p = 0.5$, the empirical assurance probability can be exceptionally high, indicating the sample sizes calculated from the formulas are too conservative.

4.4 Conclusion

Evaluation results indicate that sample sizes calculated with 50% nominal assurance probability from the derived formulas in the last chapter are consistent with pre-specified parameters. These results prove that the previously published sample size formulas implicitly guarantee only 50% probability that the results meet the required estimation precision. The proposed formulas in the last chapter explicitly incorporate both the precision and assurance probability so that based on a specific study objective, sample sizes can be easily calculated with desired and

affordable precision and assurance probability.

In summary, the sample size formulas perform reasonably well in evaluation studies. The modified Wald method is recommended to construct the confidence interval as approval to the original Wald method since it generally maintains the nominal coverage level and assurance probability.

Table 4.1: Minimum sample size to achieve pre-specified 95% one-sided lower confidence interval limit evaluated with empirical percentage coverage and assurance probability for two raters

			50% Assurance Probability					80% Assurance Probability				
			M. Wald			Wald		M. Wald			Wald	
ρ	ρ_L	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.6	0.1	125	96.66	50.07	88.37	65.38	239	96.16	78.08	91.08	84.25
		0.3	52	96.61	48.75	88.93	65.05	100	96.31	78.55	90.85	85.83
		0.5	44	94.89	53.08	86.13	68.95	83	97.16	79.09	92.70	87.03
0.6	0.4	0.1	150	95.91	49.89	92.86	54.63	321	95.55	77.82	93.61	79.70
		0.3	67	95.87	48.75	93.03	56.59	141	95.68	78.93	93.33	81.82
		0.5	57	95.74	50.12	91.64	55.25	119	95.85	79.52	93.23	82.83
0.7	0.6	0.1	497	95.53	49.98	93.55	56.18	1058	95.42	78.96	93.89	81.91
		0.3	208	95.68	50.02	93.31	57.10	442	95.50	79.34	93.91	82.52
		0.5	174	95.35	51.22	92.66	54.27	368	95.79	78.38	94.42	82.10

†CV: empirical coverage percentage based on exact evaluation. ‡AS: empirical assurance probability based on exact evaluation, defined as percentage of times that lower one-sided 95% confidence interval is no less than ρ_L.

Table 4.2: Minimum sample size to achieve pre-specified 95% one-sided lower confidence interval limit evaluated with empirical percentage coverage and assurance probability for three raters

			50% Assurance Probability					80% Assurance Probability				
			M. Wald		Wald			M. Wald		Wald		
ρ	ρ_L	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.6	0.1	95	96.86	50.47	89.56	64.89	180	96.35	76.02	91.07	83.36
		0.3	33	96.62	49.22	88.71	62.88	64	96.30	77.25	91.33	84.16
		0.5	26	98.54	42.93	91.85	63.32	50	95.67	77.90	89.64	85.73
0.6	0.4	0.1	115	96.31	48.72	92.78	54.18	246	95.74	76.49	93.49	78.50
		0.3	39	96.12	47.63	92.85	52.30	84	95.75	76.99	93.77	79.20
		0.5	30	97.11	44.41	92.95	53.60	65	95.58	77.96	93.94	78.59
0.7	0.6	0.1	379	95.80	49.67	93.28	56.48	801	95.52	78.07	93.89	81.36
		0.3	131	95.76	49.40	93.59	55.22	280	95.53	78.69	94.06	81.59
		0.5	102	96.13	46.25	94.19	55.05	218	95.94	79.92	94.30	81.01

†CV: empirical coverage percentage based on exact evaluation. ‡AS: empirical assurance probability based on exact evaluation, defined as percentage of times that lower one-sided 95% confidence interval is no less than ρ_L.

Table 4.3: Minimum sample size to achieve pre-specified 95% one-sided lower confidence interval limit evaluated with empirical percentage coverage and assurance probability for four raters based on 5,000 simulation runs

			50% Assurance Probability					80% Assurance Probability				
			M. Wald			Wald		M. Wald			Wald	
ρ	ρ_L	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.6	0.1	85	96.78	50.34	89.90	65.92	158	96.60	74.86	91.02	83.34
		0.3	28	96.78	48.52	89.14	63.28	54	95.98	76.22	91.20	82.92
		0.5	21	98.06	48.28	90.44	61.52	42	95.80	78.80	90.52	84.40
0.6	0.4	0.1	104	95.84	49.34	92.46	54.22	222	96.10	75.88	93.58	78.46
		0.3	32	96.32	48.14	93.24	52.00	70	96.18	76.16	94.20	77.46
		0.5	23	95.98	45.14	92.98	48.36	51	95.94	76.74	93.96	77.76
0.7	0.6	0.1	337	95.96	49.42	93.26	56.86	710	95.62	78.04	94.00	81.70
		0.3	111	95.44	49.76	93.24	55.04	237	95.40	79.20	93.82	81.76
		0.5	83	95.64	47.76	94.32	53.56	180	95.32	79.16	94.02	81.26

†CV: empirical coverage percentage based on simulation. ‡AS: empirical assurance probability based on simulation, defined as percentage of times that lower one-sided 95% confidence interval is no less than ρ_L.

Table 4.4: Minimum sample size to achieve pre-specified 95% one-sided lower confidence interval limit evaluated with empirical percentage coverage and assurance probability for five raters based on 5,000 simulation runs

			50% Assurance Probability					80% Assurance Probability			
			M. Wald		Wald			M. Wald		Wald	
ρ	ρ_L	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.6	0.1	79	97.36	49.30	90.18	67.50	147	96.86	73.62	91.72	83.30
		0.3	26	97.10	49.80	89.12	62.98	50	96.18	76.24	91.24	83.20
		0.5	19	98.38	47.04	88.20	57.58	38	95.92	76.96	91.50	81.86
0.6	0.4	0.1	99	96.10	47.82	92.78	53.80	210	95.98	75.64	93.98	78.64
		0.3	29	96.10	48.40	93.24	51.70	63	95.82	76.66	94.02	77.78
		0.5	21	96.70	46.28	93.60	48.78	46	95.76	77.10	94.06	77.34
0.7	0.6	0.1	316	96.08	50.46	93.42	57.68	663	95.52	77.70	94.08	81.24
		0.3	102	95.32	49.68	93.28	55.40	218	95.24	77.78	94.14	80.72
		0.5	76	95.90	48.76	94.12	53.96	165	95.62	78.30	94.60	80.80

†CV: empirical coverage percentage based on simulation. ‡AS: empirical assurance probability based on simulation, defined as percentage of times that lower one-sided 95% confidence interval is no less than ρ_L.

Table 4.5: Minimum sample size to achieve pre-specified 95% two-sided confidence interval widths evaluated with empirical percentage coverage and assurance probability for two raters

				50% Assurance Probability					80% Assurance Probability			
				M. Wald		Wald			M. Wald		Wald	
ρ	ω	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.2	0.1	101	94.87	50.35	89.48	52.49	137	95.03	77.94	91.38	73.74
		0.3	42	94.83	44.66	88.41	52.03	57	95.12	79.16	90.54	79.10
		0.5	35	93.21	48.00	87.15	50.54	47	95.13	79.57	94.92	79.59
0.6	0.2	0.1	177	94.95	69.78	93.39	50.26	201	94.90	85.67	93.30	69.68
		0.3	74	94.95	59.23	93.64	48.71	85	94.88	85.34	93.60	76.09
		0.5	62	94.61	55.96	93.32	49.30	72	95.15	88.17	94.88	80.06
0.7	0.1	0.1	569	94.88	52.88	94.54	50.50	633	94.89	75.47	94.48	72.35
		0.3	236	94.90	51.02	94.48	50.08	263	94.98	78.81	94.45	78.05
		0.5	196	95.21	49.79	95.08	49.79	219	95.17	80.87	94.92	80.16

†CV: empirical coverage percentage based on exact evaluation. ‡AS: empirical assurance probability based on exact evaluation, defined as percentage of times that half two-sided 95% confidence interval width is no larger than ω.

Table 4.6: Minimum sample size to achieve pre-specified 95% two-sided confidence interval widths evaluated with empirical percentage coverage and assurance probability for three raters

				50% Assurance Probability					80% Assurance Probability			
				M. Wald		Wald			M. Wald		Wald	
ρ	ω	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.2	0.1	73	95.02	46.07	90.50	50.12	101	94.75	71.27	91.73	72.61
		0.3	28	94.78	51.30	90.05	51.06	37	94.97	78.71	90.93	76.38
		0.5	22	95.45	49.99	85.98	48.87	29	95.39	85.33	93.55	77.77
0.6	0.2	0.1	135	94.76	65.95	92.94	51.79	155	94.79	81.85	93.47	68.09
		0.3	47	94.87	69.28	93.54	52.36	53	94.93	86.31	93.78	72.11
		0.5	36	94.91	67.89	92.91	40.82	41	94.85	97.83	93.55	81.65
0.7	0.1	0.1	426	94.89	51.90	94.46	50.59	478	94.97	72.31	94.57	70.20
		0.3	152	94.97	53.69	94.48	49.97	168	94.98	76.96	94.50	73.47
		0.5	120	94.56	54.79	93.89	49.46	131	94.58	83.83	93.99	77.72

†CV: empirical coverage percentage based on exact evaluation. ‡AS: empirical assurance probability based on exact evaluation, defined as percentage of times that half two-sided 95% confidence interval width is no larger than ω.

Table 4.7: Minimum sample size to achieve pre-specified 95% two-sided confidence interval widths evaluated with empirical percentage coverage and assurance probability for four raters based on 5,000 simulation runs

				50% Assurance Probability					80% Assurance Probability			
				M. Wald		Wald			M. Wald		Wald	
ρ	ω	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.2	0.1	63	94.28	44.92	90.64	51.94	88	94.74	68.46	92.54	71.34
		0.3	24	94.06	54.32	90.84	50.86	32	95.38	80.08	92.12	75.78
		0.5	19	94.44	52.94	87.64	47.82	25	95.04	89.72	90.72	80.06
0.6	0.2	0.1	120	94.40	62.18	92.92	48.84	139	94.82	79.52	93.84	66.30
		0.3	40	94.40	72.12	93.58	53.68	44	94.56	84.96	93.00	69.84
		0.5	30	94.20	88.04	93.28	44.04	33	94.88	97.72	93.90	81.60
0.7	0.1	0.1	373	95.22	51.18	94.82	50.78	421	95.28	70.48	94.92	69.34
		0.3	130	95.30	55.44	95.14	50.34	143	94.20	76.96	93.72	73.12
		0.5	100	94.60	55.86	94.06	45.82	110	95.16	88.48	94.98	81.04

†CV: empirical coverage percentage based on simulation. ‡AS: empirical assurance probability based on simulation, defined as percentage of times that half two-sided 95% confidence interval width is no larger than ω.

Table 4.8: Minimum sample size to achieve pre-specified 95% two-sided confidence interval widths evaluated with empirical percentage coverage and assurance probability for five raters based on 5,000 simulation runs

				50% Assurance Probability					80% Assurance Probability			
				M. Wald		Wald			M. Wald		Wald	
ρ	ω	p	N	CV†	AS‡	CV	AS	N	CV	AS	CV	AS
0.8	0.2	0.1	57	93.52	40.08	90.84	50.82	81	94.42	65.74	92.34	70.96
		0.3	22	94.36	53.92	90.30	52.68	29	94.56	77.70	91.70	74.36
		0.5	18	93.52	55.38	88.20	47.66	23	95.68	90.36	92.54	77.88
0.6	0.2	0.1	113	94.68	62.70	93.40	51.78	132	94.90	80.42	93.86	68.90
		0.3	37	95.40	75.30	93.84	57.80	41	95.14	85.04	94.04	70.38
		0.5	27	95.46	89.62	94.10	38.50	30	95.18	97.70	93.52	83.52
0.7	0.1	0.1	345	94.88	51.08	94.44	51.10	392	94.60	71.06	93.92	70.34
		0.3	120	95.06	55.20	94.94	50.86	132	94.98	76.28	94.28	72.28
		0.5	93	95.50	60.36	94.96	47.66	101	94.72	89.82	94.24	80.34

†CV: empirical coverage percentage based on simulation. ‡AS: empirical assurance probability based on simulation, defined as percentage of times that half two-sided 95% confidence interval width is no larger than ω.

Table 4.9: Comparison of sample sizes to achieve pre-specified 95% one-sided lower confidence interval limit under 50% assurance probability for two raters using Equation (3.4) and the goodness-of-fit (Donner, 1999)

			Sample Size	
ρ_0	ρ_L	p	Eqn(3.4)	GOF
0.8	0.6	0.1	125	116
		0.3	52	52
		0.5	44	44
0.6	0.4	0.1	150	140
		0.3	67	66
		0.5	57	57
0.7	0.6	0.1	497	462
		0.3	208	205
		0.5	174	174

Table 4.10: Comparison of sample sizes to achieve pre-specified 95% two-sided confidence interval widths under 50% assurance probability for two raters using Equation (3.5) and goodness-of-fit (Donner, 1999)

			Sample Size	
ρ	ω	p	$Eqn(3.5)$	GOF
0.8	0.6	0.1	101	97
		0.3	42	43
		0.5	35	37
0.6	0.4	0.1	177	150
		0.3	74	70
		0.5	62	60
0.7	0.6	0.1	569	643
		0.3	236	287
		0.5	196	245

Table 4.11: Comparison of sample sizes to achieve pre-specified 95% one-sided lower confidence interval limit under 50% assurance probability for three raters using Equation (3.4) and the goodness-of-fit (Donner and Rotondi, 2010)

			Sample Size	
ρ	ρ_L	p	Eqn(3.4)	GOF
		0.1	95	78
0.8	0.6	0.3	33	31
		0.5	26	26
		0.1	115	94
0.6	0.4	0.3	39	37
		0.5	30	39
		0.1	379	311
0.7	0.6	0.3	131	124
		0.5	102	102

Chapter 5

ILLUSTRATIVE EXAMPLES

Chapters 3 and 4 have derived and evaluated sample size formulas for designed reliability studies with binary measurements. This chapter presents illustrative examples using data from real trials to demonstrate the use of these formulas. Section 5.1 considers an example arising from a study involving six pathologists reading biopsy specimens from patients suspected to be affected by Crohn's disease (Rogel *et al.*, 1998). A second example is provided in Section 5.2, in which a study was conducted to assess the reliability of four expert clinicians in distinguishing between preparatory grief and depression on dying patients (Kraemer *et al.*, 2002). The common feature of both these studies is that they were carried out to measure reliability using multiple expert raters. They also help to show that sample size calculation can be simple when using the derived formulas, even when the number of raters is large.

5.1 Reliability Study of Pathologists Evaluating Biopsy Specimens from Patients with Crohn's Disease

In the study by Rogel *et al.* (1998), six pathologists were evaluated for reliability based on reading a set of 68 intestinal biopsy specimens collected from patients suspected of having Crohn's disease. Crohn's disease is an inflammatory bowel disease and it may affect any part of gastrointestinal tract. Confirmation of diag-

nosis requires evidence of disease based on imaging modalities (Rogel *et al.*, 1998).

The study design consisted of a calibration and evaluation phase. In the calibration phase, six pathologists were asked to independently evaluate a set of biopsies. Any discrepancies were then resolved by discussion in order to reach a consensus rubric to assess typical lesions. Following calibration, the pathologists blindly reviewed 68 new intestinal biopsy specimens in the evaluation phase. Specimens were scored on the presence (1) or absence (0) of lesions. Three of the 23 lesions were retained: epithelioid granuloma (EG), diminution of mucosecretion (DM), and focal infiltrate (FI). The frequency table of the assigned status for the three lesions from each pathologist is listed in Table 5.1. Each subject is given a score ranging from zero indicating perfect agreement on the absence of a lesion when all raters agree on its absence, to a score of one when only one pathologist declared a lesion, up to a maximum score of six indicating perfect agreement on its presence.

For the detection of "epithelioid granuloma (EG)" in the 68 biopsy specimens, the numbers of subjects receiving total scores from 0 (every rater assigned 0 (absence)) to 6 (every rater assigned 1 (presence)) are 30, 4, 5, 6, 3, 5, 15 respectively. The estimated probability of raters assigning presence to EG can be obtained as 0.39. Under the assumption of marginal homogeneity, the estimated intraclass correlation coefficient is 0.66, which is very close to the maximum likelihood estimate of agreement parameter 0.63 from the homogeneous pairwise agreement model in Rogel *et al.* (1998). This indicates substantial agreement among six pathologists.

For the detection of "diminution of mucosecretion (DM)", 29, 8, 5, 6, 10, 9 and 1 subjects respectively were assigned a total score from 0 to 6. The estimated marginal probability is 0.31. The agreement level is relatively low, with $\hat{\rho} = 0.41$ just reaching the moderate agreement level. As for the "focal infiltrate (FI)", $\hat{\rho} = 0.38$ is even poorer, and the estimated marginal probability is 0.35, with 22, 17, 5, 6, 15, 3 and 6 subjects respectively assigned with a total scores from 0 to

Table 5.1: Observed frequencies for six pathologists assigning status (1: presence, 0: absence) for epithelioid granuloma (EG), diminution of mucosecretion (DM), and focal infiltrate (FI). Zero frequencies for all three lesions are not reported.

Raters									Raters								
1	2	3	4	5	6	EG	DM	FI	1	2	3	4	5	6	EG	DM	FI
1	1	1	1	1	1	15	1	6	0	1	1	1	1	0	0	2	0
1	1	1	1	1	0	0	2	1	0	1	1	1	0	1	0	1	0
1	1	1	1	0	1	2	2	0	0	1	1	0	1	1	1	0	0
1	1	1	1	0	0	0	2	0	0	1	1	0	0	1	1	0	0
1	1	1	0	1	1	2	0	0	0	1	1	0	0	0	1	0	0
1	1	1	0	0	1	1	0	6	0	1	0	1	1	1	0	3	0
1	1	0	1	1	1	0	1	0	0	1	0	1	0	1	0	1	0
1	1	0	1	0	1	0	1	0	0	1	0	1	0	0	0	1	0
1	1	0	1	0	0	0	1	1	0	1	0	0	0	0	0	1	0
1	1	0	0	1	1	1	0	0	0	0	1	1	1	0	0	2	0
1	0	1	1	1	1	0	2	2	0	0	1	1	0	1	0	0	2
1	0	1	1	1	0	0	0	1	0	0	1	1	0	0	0	2	1
1	0	1	1	0	1	0	0	7	0	0	1	0	1	1	3	0	0
1	0	1	1	0	0	1	1	0	0	0	1	0	0	1	2	0	1
1	0	1	0	0	1	0	0	1	0	0	1	0	0	0	2	0	2
1	0	1	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0
1	0	0	1	1	1	0	1	1	0	0	0	1	0	1	0	1	2
1	0	0	1	1	0	1	1	0	0	0	0	1	0	0	0	6	1
1	0	0	1	0	1	0	0	2	0	0	0	0	1	0	2	0	0
1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	1	8
1	0	0	0	0	0	0	0	6	0	0	0	0	0	0	30	29	22
0	1	1	1	1	1	1	2	0									

6.

Table 5.2: Sample size for the reliability study of six pathologists assigning status (1: presence, 0: absence) for epithelioid granuloma (EG), diminution of mucosecretion (DM), and focal infiltrate (FI) with specific requirements on 95% lower one-sided confidence interval limit ρ_L, half 95% two-sided confidence interval width w and assurance probability $1-\beta$.

		EG				DM				FI	
		Assurance Pr				Assurance Pr				Assurance Pr	
ρ	ρ_L	50%	80%	ρ	ρ_L	50%	80%	ρ	ρ_L	50%	80%
0.6	0.5	85	189	0.4	0.3	99	230	0.4	0.3	87	203
	0.4	22	48		0.2	21	52		0.2	18	45
	w				w				w		
	0.1	79	114		0.1	105	147		0.1	93	128
	0.2	20	37		0.2	27	47		0.2	24	40
0.7	0.6	79	171	0.5	0.4	105	239	0.5	0.4	93	212
	0.5	22	45		0.3	25	58		0.3	22	51
	w				w				w		
	0.1	68	104		0.1	104	149		0.1	93	130
	0.2	17	35		0.2	26	48		0.2	24	41

Table 5.2 shows the sample size calculated from the sample size formulas for the reliability of six pathologists assigning status of EG, DM and FI to 68 biopsy specimens under different requirements for either 95% lower one-sided confidence interval limit ρ_L or 95% two-sided confidence interval width ω, and assurance probability $1-\beta$. For example, for the detection of EG, if study has a target value of $\rho = 0.6$ and requires $\hat{\rho}_L$ to be no less than 0.5 with 80% assurance probability, the sample size should be at least 189. If the estimated confidence interval width should be within 0.2 with 80% assurance probability, the sample size required re-

duces to 114. For the detection of DM, when the anticipated value of ρ is 0.4, if the study has lowered the requirement that the estimated lower confidence interval limit above 0.2 with 80% assurance probability is already acceptable, only 52 specimens should be collected. Or if the estimated confidence interval width should be no larger than 0.4 with 80% assurance probability, a sample size of 47 is sufficient. For the detection of FI, when the target value of ρ is 0.4, the sample size that assures the estimated lower confidence interval limit no less than 0.3 with 80% assurance probability is relatively large, 203. However if the study can accept this requirement with only half chance, the sample size decreases immediately to less than half, 87.

5.2 Reliability Study of Clinicians Distinguishing Dying Patients with Grief from Depression

In this section, the study considered was conducted by Kraemer *et al.* (2002) to distinguish preparatory grief from depression in dying adult patients. Virtually all patients experience 'preparatory grief' when facing impending death. These feelings are considered to be a sign of positively coping with the dying process, usually exist for a variable period, and will diminish over time. Some patients might also experience negative feelings such as hopelessness and worthlessness and desire for relief from early death. Since these feelings diminish the quality of the dying process, those patients are diagnosed with depression and can be effectively treated (Periyakoil *et al.*, 2005). Depression has become one of the most common psychiatric illnesses in terminally ill patients (McDaniel *et al.*, 1995). The study was conducted to assess the extent to which four expert clinicians could reliably distinguish between these two stages.

Sixty-nine subjects were sampled and four expert clinicians were asked to classify each subject as more indicative of preparatory grief (1) or depression (0). The

numbers of subjects who were assigned with a total of s, $s = 0, 1, 2, 3, 4$ categorizations of preparatory grief are 14, 12, 6, 8 and 29, which results in $\hat{p} = 0.59$ and $\hat{\rho} = 0.58$. The almost substantial reliability level is in agreement with the results presented by the study (Kraemer *et al.*, 2002).

Table 5.3: Sample size for the reliability study of four expert clinicians distinguishing 69 subjects having preparatory grief (1) and depression (0) with specific requirements on 95% lower one-sided confidence interval limit ρ_L, 95% two-sided confidence interval width w and assurance probability $1 - \beta$.

		Assurance Pr				Assurance Pr				Assurance Pr	
ρ	ρ_L	50%	80%	ρ	ρ_L	50%	80%	ρ	ρ_L	50%	80%
0.5	0.4	99	223	0.6	0.5	96	213	0.7	0.6	88	190
	0.3	24	55		0.4	25	54		0.5	24	51
	w				w				w		
	0.1	96	128		0.1	88	122		0.1	74	111
	0.2	24	40		0.2	22	39		0.2	19	36

Table 5.3 presents the sample size needed with different requirements on either the 95% lower one-sided confidence interval limit ρ_L or the 95% two-sided confidence interval width ω, and assurance probability $1 - \beta$. If the study has a target $\rho = 0.7$ and requires the estimated lower confidence interval limit to be no less than substantial agreement level with 80% assurance probability, $N = 190$ subjects should be recruited. On the other hand if the estimated confidence interval width should be within 0.2 with 80% assurance probability, a sample size of 111 is sufficient.

5.3 Summary

This chapter gives two examples that show the derived sample size formulas are convenient to use. Once a reliability study determines the requirement on precision, either a lower confidence interval limit or a confidence interval width, and assurance probability, the corresponding sample size can be easily calculated. Researchers can also construct a table of sample sizes calculated using different sets of parameters so that they can plausibly estimate what precision level and assurance probability to expect with affordable costs in recruiting subjects.

Chapter 6

DISCUSSION

As statistical methods for analyzing reliability studies have developed rapidly in recent decades, it is crucial to consider study designs that drive these analyses. Proper number of subjects (N) and number of raters (n) are of great significance in interpreting results from these studies. In this thesis, closed-form sample size formulas are derived focusing on confidence interval estimation from reliability studies with binary outcomes. These formulas have the advantages of ensuring the reliability studies achieve the pre-specified precision with desired assurance probability.

Before the study begins, the precision thresholds may be chosen somewhat arbitrarily, but the guidelines from Landis and Koch (1977) can give helpful insights. Also, the anticipated ρ_0 could be drastically differ from what is observed upon completion of the study (Donner and Rotondi, 2010). This can have unavoidable impact on achieving the desired precision. Thus a detailed sensitivity analysis is recommended to detect if minor variations in π and ρ_0 leads to extreme changes in the resulted sample size. The precision approaches to calculate the sample size can additionally provide insight into estimating ρ that can be reasonably achieved in the planning stage of the studies.

It is useful to know that $\rho = 0$ indicates either that the heterogeneity of the subjects selected from the population is not well detected by raters, or that the pop-

ulation is indeed homogeneous. Thus, in a homogeneous population, it is very difficult to achieve a high reliability level no matter if the measurement is binary or not. This should not be considered as a flaw, or a paradox in ICC, since this phenomenon only reflects the fact that when raters all tend to give same measurements (p is close to 1 or 0), it is difficult to distinguish subjects from a homogeneous population. In such a population, "noise" quickly overwhelms the "signal" (Kraemer *et al.*, 2002).

Evaluation studies have proven these formulas perform accurately and stably. When using the modified Wald method to construct confidence intervals, those formulas provide empirical confidence interval coverages and assurance probabilities close to the nominal levels even with extreme values.

More importantly, these formulas are easy to use, as shown in the examples. The resulted sample sizes can be used as a test for feasibility during the planning stage of reliability studies. They reveal the information on how many subjects a reliability study should be recruited in order to achieve the pre-specified precision with desired assurance probability. More restrict requirements on precision are always accompanied by increased sample size.

In order to reduce complexity in constructing the model, the assumption of no rater bias is assumed, if every rater possesses the same underlying probability π to assign positive ratings to subjects. This assumption is most appropriate when the emphasis is placed at the measurement process itself in reliability studies, rather than the potential differences among raters Landis and Koch (1977). Even when some differences existed, the consequent effects will be averaged out in both the sample size estimation and further study analysis. However, substantial differences can mislead the estimation procedures. If it is unclear the differences are homogeneous, a test for no rater bias is provided by the famous Cochran's Q-statistic, which is illustrated by Fleiss (1973).

Therefore, in a reliability study context assuming no rater bias, the sample size calculations based on the intraclass kappa using precision approaches are recommended for multiple raters and binary outcomes. The resulting sample size can provide additional insight into the interpretation of a reliability study especially in the planning stages.

BIBLIOGRAPHY

Altaye, M., Donner, A. and Klar, N. (2001). Inference procedures for assessing interobserver agreement among multiple raters. *Biometrics* **57** (2), 584–588.

Armitage, P., Blendis, L. and Smyllie, H. (1966). The measurement of observer disagreement in the recording of signs. *Statistics in Medicine* **129** (1), 98–109.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. *Studies in Item Analysis and Prediction* **6**, 158–168.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* **19** (1), 3–11.

Blackman, N. J. and Koval, J. J. (1993). Estimating rater agreement in 2 x 2 tables: correction for chance and intraclass correlation. *Applied Psychological Measurement* **17** (3), 211–223.

Blackman, N. J.-M. and Koval, J. J. (2000). Interval estimation for cohen's kappa as a measure of agreement. *Statistics in Medicine* **19** (5), 723–741.

Bloch, D. A. and Kraemer, H. C. (1989). 2 x 2 kappa coefficients: measures of agreement or association. *Biometrics* **45** (1), 269–287.

Cantor, A. B. (1996). Sample-size calculations for cohen's kappa. *Psychological Methods* **1** (2), 150–153.

Carey, G. and Gottesman, I. I. (1978). Reliability and validity in binary ratings: areas of common misunderstanding in diagnosis and symptom ratings. *Archives of General Psychiatry* **35** (12), 1454–1459.

Cicchetti, D. V. and Feinstein, A. R. (1990). High agreement but low kappa: ii. resolving the paradoxes. *Journal of Clinical Epidemiology* **43** (6), 551–558.

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

de Vet, H. C., Terwee, C. B., Knol, D. L. and Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology* **59** (10), 1033–1039.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26** (3), 297–302.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* **54** (1), 67–82.

Donner, A. (1999). Sample size requirements for interval estimation of the intraclass kappa statistic. *Communications in Statistics-Simulation and Computation* **28** (2), 415–429.

Donner, A., Birkett, N. and Buck, C. (1981). Randomization by cluster sample size requirements and analysis. *American Journal of Epidemiology* **114** (6), 906–914.

Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* **11** (11), 1511–1519.

Donner, A. and Rotondi, M. A. (2010). Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. *The International Journal of Biostatistics* **6** (1), Article 31, DOI: 10.2202/1557–4679.1275.

Donner, A. and Zou, G. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics* **58** (1), 209–215.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: i. the problems of two paradoxes. *Journal of Clinical Epidemiology* **43** (6), 543–549.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd.

Flack, V. F., Afifi, A., Lachenbruch, P. and Schouten, H. (1988). Sample size determinations for the two rater kappa statistic. *Psychometrika* **53** (3), 321–325.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76** (5), 378–382.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*. J. Wiley & Sons.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* **31** (3), 651–659.

Fleiss, J. L., Cohen, J. and Everitt, B. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* **72** (5), 323–327.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgments: unequal numbers of judges per subject. *Applied Psychological Measurement* **3** (4), 537–542.

George, E. O. and Bowman, D. (1995). A full likelihood procedure for analysing exchangable binary data. *Biometrics* **51** (2), 512–523.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of American Statistical Association* **49**, 732–64.

Goodman, L. A. and Kruskal, W. H. (1979). *Measures of Association for Cross Classifications*. Springer.

Haynam, G., Govindarajulu, Z. and Leone, F. C. (1970). Tables of the cumulative non-central chi-square distribution. *Selected Tables in Mathematical Statistics* **1**, 1–78.

ICH, E. (1998). Guideline: statistical principles for clinical trials, eu: cpmp. Technical report ICH/363/96, FDA: Federal Register.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* **44** (4), 461–472.

Kraemer, H. C., Periyakoil, V. S. and Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine* **21** (14), 2109–2129.

Lachin, J. M. (2004). The role of measurement reliability in clinical trials. *Clinical Trials* **1** (6), 553–566.

Landis, J. R. and Koch, G. G. (1975). A review of statistical methods in the analysis of data arising from observer reliability studies (part i). *Statistica Neerlandica* **29** (3), 101–123.

Landis, J. R. and Koch, G. G. (1977). A one-way components of variance model for categorical data. *Biometrics* **33** (4), 671–679.

Lee, J. J. and Tu, Z. N. (1994). A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes. *Journal of Computational and Graphical Statistics* **3** (3), 301–321.

Mak, T. K. (1988). Analysing intraclass correlation for dichotomous variables. *Applied Statistics* **37** (3), 344–352.

Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C. and Fenton, M. (2000). Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. *The British Journal of Psychiatry* **176** (3), 249–252.

McDaniel, J. S., Musselman, D. L., Porter, M. R., Reed, D. A. and Nemeroff, C. B. (1995). Depression in patients with cancer: diagnosis, biology, and treatment. *Archives of General Psychiatry* **52** (2), 89–99.

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

Periyakoil, V. S., Kraemer, H. C., Noda, A., Moos, R., Hallenbeck, J., Webster, M. and Yesavage, J. A. (2005). The development and initial validation of the terminally ill grief or depression scale (tigds). *International Journal of Methods in Psychiatric Research* **14** (4), 203–212.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Rao, C. R. and Mukerjee, R. (1997). Comparison of LR, score, and Wald tests in a non-iid setting. *Journal of Multivariate Analysis* **60** (1), 99–110.

Ridout, M. S., Demetrio, C. G. and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* **55** (1), 137–148.

Rogel, A., Boelle, P. and Mary, J. (1998). Global and partial agreement among several observers. *Statistics in Medicine* **17** (4), 489–501.

Rogot, E. and Goldberg, I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Diseases* **19** (9), 991–1006.

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* **19** (3), 321–325.

Shoukri, M., Asyali, M. and Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research* **13** (4), 251–271.

Shoukri, M. M. (2010). *Measures of Interobserver Agreement and Reliability*. CRC Press.

Shoukri, M. M. and Donner, A. (2009). Bivariate modeling of interobserver agreement coefficients. *Statistics in Medicine* **28** (3), 430–440.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* **86** (2), 420–428.

Shrout, P. E., Spitzer, R. L. and Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry* **44** (2), 172–177.

Vach, W. (2005). The dependence of cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology* **58** (7), 655–661.

Zou, G. (2012). Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. *Statistics in Medicine* **31** (29), 3972–3981.

Zou, G. and Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics* **60** (3), 807–811.

Zou, G., Huang, W. and Zhang, X. (2009). A note on confidence interval estimation for a linear function of binomial proportions. *Computational Statistics & Data Analysis* **53** (4), 1080–1085.

Curriculum Vitae

Name:	Mengxiao Xu
Post-secondary Education and Degree:	The University of Western Ontario London, Ontario, Canada 2014-2016 M.Sc.
Related Work Experience	Research and Teaching Assistant The University of Western Ontario 2014-2016