Open Journal of Statistics, 2016, 6, 254-273
Published Online April 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.62022

Bias and Mean Square Error of Reliability Estimators under the One and Two Random Effects Models: The Effect of Non-Normality

Mohamed M. Shoukri[1,2], Tusneem Al-Hassan[3], Michael DeNiro[1], Abdelmoneim El Dali[4], Futwan Al-Mohanna[1,2]

[1]Department of Cell Biology, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[2]Al-Faisal University College of Medicine, Riyadh, Saudi Arabia
[3]The Oncology Center, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[4]Department of Biostatistics, Epidemiology and Scientific Computing, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
Email: shoukri@kfshrc.edu.sa

Received 14 February 2016; accepted 23 April 2016; published 26 April 2016

Abstract

The coefficient of reliability is often estimated from a sample that includes few subjects. It is therefore expected that the precision of this estimate would be low. Measures of precision such as bias and variance depend heavily on the assumption of normality, which may not be tenable in practice. Expressions for the bias and variance of the reliability coefficient in the one and two way random effects models using the multivariate Taylor's expansion have been obtained under the assumption of normality of the score (Atenafu *et al.* [1]). In the present paper we derive analytic expressions for the bias and variance, hence the mean square error when the measured responses are not normal under the one-way data layout. Similar expressions are derived in the case of the two-way data layout. We assess the effect of departure from normality on the sample size requirements and on the power of Wald's test on specified hypotheses. We analyze two data sets, and draw comparisons with results obtained via the Bootstrap methods. It was found that the estimated bias and variance based on the bootstrap method are quite close to those obtained by the first order approximation using the Taylor's expansion. This is an indication that for the given data sets the approximations are quite adequate.

Keywords

Rater's Reliability, Random Effects Models, Multivariate Taylor's Expansion, Wald's Confidence Interval, Bootstrap Methods

How to cite this paper: Shoukri, M.M., Al-Hassan, T., DeNiro, M., El Dali, A. and Al-Mohanna, F. (2016) Bias and Mean Square Error of Reliability Estimators under the One and Two Random Effects Models: The Effect of Non-Normality. *Open Journal of Statistics*, **6**, 254-273. http://dx.doi.org/10.4236/ojs.2016.62022

1. Introduction

Statistics is the science of transforming data into information and knowledge. Therefore producing reliable information requires error free data. Measurement errors can seriously affect statistical analysis and interpretation; it therefore becomes important to quantify the magnitude of such errors by calculating what is known as "reliability coefficient" and assessing its statistical properties. The topic of reliability has gained much attention in the literature as evidenced in the books by Dunn [2] [3], and the recent reviews by Shoukri *et al.* [4] and Shoukri [5]. As a general feature of this coefficient, it must distinguish within-subject variation from variation between subjects.

A widely recognized index that possesses this property is the intraclass correlation coefficient (ICC) defined as the proportion of between-subject variation relative to the total variation. In the most frequently adopted design, k subjects are each rated by the same n raters (for inter-rater reliability). A similar approach, however, can also be adopted when a single subject is assessed repeatedly on each of several occasions (test-retest reliability), or when replicates consisting of different occasions are taken on different subjects by a single rater [6]. In each of these cases, and for continuous and categorical assessments, Fisher [7] showed that ρ can be estimated from an appropriate one-way analysis of variance (ANOVA).

There are numerous versions of the intraclass correlation coefficient (ICC) that can give quite different results when applied to the same data. Each form is appropriate for specific situations defined by the experimental design and the conceptual intent of the study. The differences among these forms and their applications were discussed in Shrout and Fleiss [8], and McGraw and Wong [9]. Shrout and Fleiss [8] provided specific guidelines for choosing the appropriate form of the ICC by adopting two linear additive models. The fundamental question is: which appropriate statistical model for the reliability study, may be selected to address the questions of interest.

Much of the work in reliability studies focused on the estimation (point and interval), hypothesis testing, and sample size requirements to achieve certain power [10] and or maximizing the precision of estimation under cost constraint [11].

Only recently the issue of correcting the ICC for bias, under the one-way ANOVA was investigated. Assuming the normality of the distribution of scores, Atenafu *et al.* [1] investigated the issues related to bias correction of the ANOVA estimator of ICC from the one-way layout. The authors investigated the effect of non-normality through Monte-Carlo simulations by generating data from known skewed distributions.

This article has two-fold objectives: First, under the one-way ANOVA, we evaluate the bias, the variance (and hence the mean square error) of the ICC when the assumption of normality is not tenable. We further investigate the effect of non-normality on the sample size requirements to achieve certain levels of power on specific null hypotheses on the reliability parameter for a given level of type I error. Second, under the two-way ANOVA we derive the first order approximation for the bias and the variance of the ICC. This allows for the comparison between the Wald confidence interval and other proposed confidence intervals. We analyze data of two examples using the R package.

The paper is structured as follows: In Section 2 we derive analytic expressions for bias and variance of ICC when the assumption of normality is not satisfied. In Section 3 we extend our approach to the case of two-way data layout. We obtain analytic expressions for bias and variance of ICC, and construct a Wald's type confidence interval. Moreover we evaluate the empirical power using simulations. In Section 4 we introduce two examples and assess the accuracy of the first order approximation for bias and variance using the bootstrap technology for the two-way model. We discuss the results in Section 5.

2. The Effect of Non-Normality on the Bias and Variance of the One-Way ANOVA Estimator of the Reliability Coefficient

In most interrater reliability study, each of a random sample of k subjects is rated independently by n raters. There usually are two situations that are of interest to us:

1) Each subject is rated by asset n different raters, randomly selected from a larger population of raters.

2) A random sample of $n > 1$ raters is selected from a larger population, and each judge rates each subject, that is, each judge rates k subjects. We shall assume that the number of raters is less than the number of subjects.

Conceptually the two situations should produce close estimates of the ICC, but components of variations in the scores should appropriately be specified to avoid misspecification bias. Each of the postulated models speci-

fies the decomposition of a rating made by the *j*th rater on the *i*th subject in terms of various effects. In this paper we consider the decomposition into subject component, rater component and random error component. Depending on the way the study is designed, different assumptions are made about the effects, and the structure of the corresponding ANOVA will be different.

We start by specifying the simplest design used to assess the reliability sets of scores; namely the one-way random effect model. Suppose that we have k subjects and we would like to take n measurements by a single device. How can we assess the consistency of the set of measurements taken from each subject? The one-way model stipulates that:

$$Y_{ij} = \mu + b_i + e_{ij}, \quad i = 1, 2, \cdots, k, \; j = 1, 2, \cdots, n. \tag{1}$$

The Y_{ij} is the jth measurement taken on the ith subject, μ is the bias, b_i is the subject effect, and e_{ij} is a random measurement error, assumed independent of b_i where $b_i \sim N\left(0, \sigma_b^2\right)$ and $e_{ij} \sim N\left(0, \sigma_e^2\right)$. We assume that b_i and e_{ij} are mutually independent random variables. Clearly, $E\left(y_{ij}\right) = \mu$, $\mathrm{Var}\left(y_{ij}\right) = \sigma_b^2 + \sigma_e^2$, and

$$\mathrm{Cov}\left(y_{ij}, y_{il}\right) = \begin{cases} \sigma_b^2 & j \neq l, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore the reliability coefficient, or the intraclass correlation coefficient (ICC) is defined as:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}.$$

The reliability estimate of ρ is obtained once suitable estimate of the components of variance are obtained.

$$\rho = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_e^2} \tag{2}$$

Here, $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$ are the estimates of the corresponding variance components and are obtained either from the maximum likelihood estimation, or the one-way random effects ANOVA (**Table 1**).

Using the notations, $\bar{y}_i = \bar{n}^1 \sum_{j=1}^{n} y_{ij}$, $\bar{y} = (kn)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij} = \frac{1}{k} \sum_{i=1}^{k} \bar{y}_i$, the corrected sums of squares are given as:

$$SSB = n \sum_{i=1}^{k} \left(\bar{y}_i - \bar{y}\right)^2,$$

which is the between subjects sums of squares, and $SSW = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_i\right)^2$, is the within subjects sum of squares. The total sum of squares is thus given by, $SST = SSB + SSW$. Unbiased estimates of the variance components are given by:

$$\hat{\sigma}_e^2 = S_W, \; \hat{\sigma}_b^2 = \frac{S_B - S_W}{n}.$$

Hence the variance components estimator of the ICC is given by:

$$\hat{\rho} = \frac{S_B - S_W}{S_B + (n-1) S_W} \tag{3}$$

(See; Searle *et al*. [12]).

Table 1. The one-way ANOVA table.

S.O.V.	DF	SS	MS	EMS
Between subjects	$k-1$	SSB	$S_B = \frac{SSB}{k-1}$	$\sigma_e^2 + n\sigma_b^2$
Within subjects	$k(n-1)$	SSW	$S_w = \frac{SSW}{k(n-1)}$	σ_e^2
Total	$kn-1$	SST		

With the additional assumptions of normality and independence of $b_i\,(i=1,2,\cdots,k)$ and $e_{ij}\,(j=1,2,\cdots,n)$ we have:

$$S_B \sim \frac{\sigma_e^2}{(1-\rho)(k-1)}\left[1+(n-1)\rho\right]X^2_{k-1}$$

$$S_W \sim \frac{\sigma_e^2}{k(n-1)}X^2_{k(n-1)}.$$

Here X^2_α denotes a chi-square random variable with α degrees of freedom. Using the delta method we can derive the asymptotic bias and variance of $\hat{\rho}$. After some simplifications we can show that:

$$\text{Bias}(\hat{\rho}) = \frac{-2(1-\rho)(1+(n-1)\rho)}{kn^2}\left[n^2\left[1+(n-1)\rho\right]-(1-\rho)\right]. \quad (4)$$

Equation (4) indicates that the estimator of the ICC from the one-way ANOVA is negatively biased for all values of, n, and ρ.

Dropping the assumptions of normality regarding the distributions of b_i and e_{ij} has two consequences:

1) The mean squares S_B and S_W will not have chi-square distributions.

2) The mean squares S_B and S_W are no longer independent, and hence the ratio of the mean squares will not have the usual F-distribution.

Relaxing the assumption of normality both the measures of Kurtosis of b_i and e_{ij} are needed in the calculation of the asymptotic variance of $\hat{\rho}$ and $\hat{\theta}$ and the amount of bias [13] [14].

Let δ_e and δ_b denote respectively the coefficients of kurtosis of e_{ij} and b_i. These quantities are defined as:

$$\delta_b = \left\{E\left(b_i^4\right)/\sigma_b^4\right\}-3$$

$$\delta_e = \left\{E\left(e_{ij}^4\right)/\sigma_e^4\right\}-3.$$

Using results for the balanced one way ANOVA [14] we have: $\text{Var}(S_W)=c_1\sigma_e^4$, $\text{Var}(S_B)=c_2\sigma_e^4$ and $\text{Cov}(S_W,S_B)=c_{12}\sigma_e^4$, where

$$c_1 = \left\{k(n-1)\right\}^{-1}\left[2+\delta_e\frac{n-1}{n}\right]$$

$$c_2 = 2(k)^{-1}\left[\rho^2(1-\rho)^{-2}n^2(1+\delta_b)+2n\rho(1-\rho)^{-1}+\left(1+\frac{\delta_e}{n}\right)\right]$$

$$c_{12} = (kn)^{-1}\delta_e.$$

Using the delta method, the first order approximation for the variance of $\hat{\rho}$ is

$$\text{Var}(\hat{\rho}) = \text{Var}(S_B)\left(\frac{\partial\rho}{\partial S_B}\right)^2 + 2\text{Cov}(S_B,S_W)\left(\frac{\partial\rho}{\partial S_B}\right)\left(\frac{\partial\rho}{\partial S_W}\right)+\text{Var}(S_W)\left(\frac{\partial\rho}{\partial S_W}\right)^2.$$

Simplifying we get:

$$\text{Var}(\hat{\rho}) = \frac{(1-\rho)^2}{n^2}\left\{\left[1+(n-1)\rho\right]^2c_1-2(1-\rho)\left[1+(n-1)\rho\right]c_{12}+(1-\rho)^2c_2\right\}. \quad (4)$$

We shall write the $\text{Var}(\hat{\rho})$ as $\frac{\eta(\rho)}{k}$. Note that under normality $\delta_b=\delta_e=0$, and in this case the variance expression reduces to variance expression given in Donner [15]:

$$\text{Var}(\hat{\rho}) = \frac{2(1-\rho)^2\left[1+(n-1)\rho\right]^2}{kn(n-1)}. \quad (5)$$

Using the Taylor's expansion for the two variables case (see Appendix I) we obtain, to the first order of approximation the asymptotic bias of the ANOVA estimator of the ICC when the assumption of normality is not satisfied as:

$$\begin{aligned}\text{Bias}(\hat{\rho}) = \frac{c_1}{n^2}(n-1)(1-\rho)^2\left[1+(n-1)\rho\right] - \frac{c_2}{n^2}(1-\rho)^3 \\ + \frac{c_{12}}{n^2}(1-\rho)^2\left[n-2(n-1)(1-\rho)\right].\end{aligned} \tag{6}$$

We can then evaluate the mean square error $\text{MSE} = \text{Var}(\hat{\rho}) + \left(\text{Bias}(\hat{\rho})\right)^2$. Expressions (4) and (6) demonstrate the dependence of these quantities on the kurtosis of both the between subject and within variables. To calculate the estimated bias and variance we need not specify the complete distributions of b_i and e_{ij} but good guesses for δ_b and δ_e will suffice.

One question that may be stated is: which component of variation has the largest effect on the bias and variance of the reliability estimate. We answer this question empirically in two different ways. First in **Table 2**, we demonstrate the direct effect of the combinations of the model parameters on the bias. Subsequently, we investigate the effect of departure from normality on the sample size requirements in a typical reliability study and summarize the results in **Table 2**. We see from **Table 2** that for selected values of the parameters combination (n, k, ρ) the smallest bias occurs when $\delta_b = \delta_e = 0$. Larger values δ_b increase the bias of the estimates of ICC and has more adverse effect on the bias than that caused by large values of δ_e. The conclusion here is that, we are worse-off by miss-specifying the distribution of the between subjects effects relative to the error term distribution. We note also from Equation (6) that δ_e is divided by the factor $\{kn\}$ in c_1, c_2, and c_{12}. The implication is that, as the number of subjects increase, the kurtosis of the error term has negligible effect on the bias and variance of the estimated reliability.

Note that the selected values for δ_b and δ_e are not arbitrary as it may seem. For example, if we assume that the error term $\{e\}$ is a random variable that has a mixture of two normal distributions with $p \cdot d \cdot f \quad N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ and a mixing proportions p_1, and $p_2 = 1 - p_1$, we can show in general that

$$E(e) = \mu = \sum_{j=1}^{2} p_j \mu_j$$

Table 2. Effect non-normality on bias one-way ANOVA.

n	k	ρ	δ_e	δ_b	BIAS
20	10	0.20	0.000	6.000	−0.0460
20	20	0.20	6.000	0.000	−0.0019
10	30	0.40	0.000	6.000	−0.0457
10	30	0.40	6.000	3.000	−0.0237
20	50	0.20	0.000	0.000	−0.0015
10	10	0.40	0.000	3.000	−0.0796
10	20	0.40	6.000	3.000	−0.0355
2	30	0.70	6.000	6.000	−0.0650
5	25	0.80	6.000	0.000	−0.0092
5	40	0.80	0.000	6.000	−0.0451
7	126	0.70	0.000	0.000	−0.0024
7	126	0.70	3.000	0.000	−0.0022
7	126	0.70	0.000	3.000	−0.0109
7	117	0.65	3.000	3.000	−0.0115
3	16	0.95	3.000	3.000	−0.0261

$$\tau^2 = \text{var}(e) = \sum_{j=1}^{2} p_j\left(\sigma_j^2 + \mu_j^2\right) - \mu^2 .$$

And a coefficient of kurtosis δ given by:

$$\delta = \frac{1}{\tau^4}\left[\sum_{j=1}^{2} p_j\left[3\sigma_j^4 + 6\left(\mu_j - \mu\right)^2\sigma_j^2 + \left(\mu_j - \mu\right)^4\right]\right].$$

For the case, $p_1 = p_2 = 0.5$, $\mu_1 = 0.5$, $\mu_2 = -0.5$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, the $\mu = 0$, $\tau^2 = 1$ and the kurtosis will be 3.0625. This justifies our choice for the values of δ_b and δ_e.

As well, non-normality has an effect on the required sample size through its influence on the variance of the estimated reliability coefficient. Suppose that we need to determine the number of subjects to detect the departure from the null hypothesis $H_0 : \rho = \rho_0$ in the direction of the one-sided alternative $H_1 : \rho = \rho_1 > \rho_0$, with type-one error rate α and power $1-\beta$. For fixed n, we have:

$$k = \frac{\left[z_\alpha\sqrt{\eta[\rho_0]}\right] + z_\beta\sqrt{\eta[\rho_1]}^2}{\left(\rho_0 - \rho_1\right)^2}. \tag{7}$$

If we set the Type I error rate at 5% and power at 80%, for given values of ρ_0, ρ_1, n, δ_b and δ_e, the estimated values of k are given in **Table 3**.

In this table we demonstrate the interplay between the effect size ($\rho_0 - \rho_1$), δ_b, δ_e, and the required sample size. Specifically, the two red-colored rows of **Table 3** show that for the same effect size $\left(\rho_0 - \rho_1\right) = \left(0.4 - 0.8\right)$, δ_b = 0, and δ_e = 6, and n = 5, we need to recruit 6 subjects (see the first red row), while for the same range of values of the reliability parameter we need to recruit 21 subjects if δ_b = 6, and δ_e = 0 (the second red row). The worst situation when the two components of variation are far from being normal. This illustrates the impact of the departure from normality of the distribution of between subjects on the sample size requirements.

Table 3. Effect of non-normality on sample size under the one-way ANOVA.

n	ρ_0	ρ_1	δ_e	δ_b	k
20	0.00	0.20	0.000	3.000	20
20	0.00	0.20	3.000	0.000	7
20	0.00	0.20	0.000	0.000	7
10	0.00	0.40	0.000	6.000	17
10	0.00	0.40	6.000	3.000	12
10	0.00	0.40	0.000	3.000	11
10	0.00	0.40	6.000	3.000	12
7	0.70	0.90	0.000	3.000	26
7	0.70	0.90	3.000	0.000	8
7	0.70	0.90	0.000	6.000	38
7	0.64	0.70	3.000	3.000	839
7	0.64	0.80	3.000	3.000	84
7	0.64	0.80	0.000	0.000	21
5	0.70	0.90	3.000	3.000	28
5	**0.40**	**0.80**	**6.000**	**0.000**	**6**
5	**0.40**	**0.80**	**0.000**	**6.000**	**21**
3	0.95	0.98	3.000	3.000	67
3	0.95	0.50	3.000	3.000	14
3	0.95	0.98	0.000	0.000	18

In the previous section we investigated the effect of departure from normality on the bias and the sample size requirements. In the simple case of one-way design the evaluations depend on a number of parameters. In order to extend the one-way model to the more complex model of a two-way layout, we adopt the situation when a random sample of *n* raters is selected from a larger population, and each judge rates each subject, that is, each judge rates *k* subjects. We investigate the issues of bias, mean square error (as a measure of precision of the estimated reliability parameter) and the power of hypothesis testing when the scores are not normally distributed, and when the model generating the data is that of a two-way layout. Although the extension is straight forward, we have to deal with several parameters, some of them are treated as nuisance, and others are considered essential so that we can produce useful results.

3. Bias and Variance of Estimating the Reliability under the Two-Way Random Effects Models

As summarized in the previous section, the sampling theory and formula for the standard error of the reliability estimates rely heavily on the normality assumptions, despite the fact that real data seldom satisfy these assumptions. At best we may expect that normality would be only approximately satisfied, and it does not logically follow, of course, that approximately satisfying the normality requirements guarantees automatic approximation of the actual distribution to the distribution given under normal theory. A similar problem exists for statistical inference in the two-way fixed model ANOVA, though it has been found that the distribution of the ratio of mean squares is quite robust with respect to non-normality under certain conditions [16] [17].

The present model is the two-way mixed model ANOVA, with one observation per cell, and the primary concern is the distribution of a function of the variance component estimates, unlike the fixed model, in which the primary concern is the location parameter estimates. Thus findings in [16] cannot be generalized to reliability theory without thorough investigation. In this section we consider the model:

$$y_{ij} = \mu + b_i + r_j + e_{ij},\ i = 1, 2, \cdots, k;\ j = 1, 2, \cdots, n. \tag{8}$$

The corrected sums of squares as shown in **Table 4** are given as:

$$SSB = n\sum_{i=1}^{k}\left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2,\quad SSR = k\sum_{j=1}^{n}\left(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}\right)^2,\quad SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}\right)^2.$$

The total sum of squares:

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2 = SSB + SSR + SSE,$$

where:

$$\bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}\bar{y}_{\cdot j} = \frac{1}{k}\sum_{i=1}^{k} y_{ij},\quad \bar{y}_{\cdot\cdot} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}.\ \text{Moreover,}\ \ \mathrm{Cov}\left(y_{ij}, y_{il}\right) = \sigma_b^2,\ j \neq l.$$

It is assumed that b_i, r_j, and e_{ij} are mutually stochastically independent with

$$b_i \sim N\left(0, \sigma_b^2\right),\ r_i \sim N\left(0, \sigma_r^2\right),\ e_{ij} \sim N\left(0, \sigma_e^2\right).$$

The variance of Y_{ij} is:

$$\mathrm{var}\left(Y_{ij}\right) = \sigma_b^2 + \sigma_r^2 + \sigma_e^2$$

Table 4. The ANOVA table under the two way layout.

S.O.V.	DF	SS	MS	EMS
Between subjects	$k-1$	*SSB*	*SB*	$\sigma_e^2 + n\sigma_b^2$
Within subjects	$n-1$	*SSR*	*SR*	$\sigma_e^2 + k\sigma_r^2$
Error	$(k-1)(n-1)$	*SSE*	*SE*	σ_e^2
Total	$nk-1$	*TSS*		

and the covariance between two measurements on the same subject, taken by the *j*th and *j*'th raters, is

$$\operatorname{cov}\left(Y_{ij}, Y_{ij'}\right) = \sigma_b^2 .$$

Under the above set-up, we have:

$$\frac{(k-1)SB}{\sigma_e^2 + n\sigma_b^2} \sim X^2_{(k-1)}, \quad \frac{(n-1)SR}{\sigma_e^2 + k\sigma_r^2} \sim X^2_{(n-1)} \quad \text{and} \quad \frac{(n-1)(k-1)SE}{\sigma_e^2} \sim X^2_{(n-1)} .$$

Hence, under this model, the appropriate intraclass correlation to measure interrater reliability becomes:

$$\rho_2 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_r^2 + \sigma_e^2} . \tag{9}$$

Under the assumption of normality the second, third central moments of mean squares *SB*, *SR*, and *SE* are given **Table 5**.

Define, $\theta_1 = \sigma_b^2 / \sigma_e^2$, and $\theta_2 = \sigma_r^2 / \sigma_e^2$. The ICC is thus written as:

$$\rho_2 = \frac{\theta_1}{1 + \theta_1 + \theta_2} . \tag{10}$$

The variance components estimator of the ICC, using the mean squares given in **Table 4** is:

$$\hat{\rho}_2 = \frac{x_1 - x_3}{x_1 + d_2 x_2 + d_3 x_3} \tag{11}$$

where $x_1 \equiv SB, x_2 \equiv SR$, and $x_3 \equiv SE$, $d_2 = n/k$, and $d_3 = (k-1)(n-1)/k$.

Note that from (10) we can write, $\theta_1 = \dfrac{\rho_2(1+\theta_2)}{(1-\rho_2)}$. This means that the model will have one parameter of interest, ρ_2, and a nuisance parameter θ_2 which will be treated as fixed. Using the Taylor's series expansion as a function of three variables, under the assumption of normality, we use the delta method and the information in **Table 4** to obtain expressions for the variance and bias of $\hat{\rho}_2$ and these are given in Appendix III. We noted from our calculations of the bias, variance and the MSE that:

1) For fixed values of θ_2, *k*, and *n*, the MSE values decrease as the values of ρ_2 increase. For example, when *k* = 20, and *n* = 5, θ_2 = 0.1, and ρ_2 = 0.7, the MSE = 0.007. While for the same set of parameters values, but ρ_2 = 0.9, the MSE = 0.001. This means that the precision of the ICC estimator increases near its upper boundary. This in fact is the situation in reliability studies where high values of the ICC estimator are expected;

2) For fixed values of ρ_2, *k*, and *n*, the MSE values increase (decrease in precision) as the values of θ_2 increase. For example, when *k* = 20, and *n* = 5, θ_2 = 0.1, and ρ_2 = 0.7, the MSE = 0.007. While for the same set of parameters values, with θ_2 = 0.5, the MSE = 0.009. Furthermore, when *k* = 20, and *n* = 5, θ_2 = 0.9, and ρ_2 = 0.7, the MSE = 0.011. The implication is that when the between rater's variance relative to the error variance ratio increases, one should expect a loss in precision of the estimate of ICC.

Table 5. Higher moments of the corrected mean squares under the assumption of normality.

M.S.	μ_2	μ_3	μ_4
SB	$\dfrac{2\sigma_e^4(1+n\theta_1)^2}{k-1}$	$\dfrac{8\sigma_e^6(1+n\theta_1)^3}{(k-1)^2}$	$\dfrac{12\sigma_e^8(1+n\theta_1)^4(k+3)}{(k-1)^3}$
SR	$\dfrac{2\sigma_e^4(1+k\theta_1)^2}{n-1}$	$\dfrac{8\sigma_e^6(1+k\theta_2)^3}{(n-1)^2}$	$\dfrac{12\sigma_e^8(1+k\theta_2)^4(n+3)}{(n-1)^3}$
SE	$2\sigma_e^4/(k-1)(n-1)$	$\dfrac{8\sigma_e^6}{(k-1)^2(n-1)^2}$	$\dfrac{12\sigma_e^8((k-1)(n-1)+4)}{(k-1)^3(n-1)^3}$

3.1. Variances and Covariance under Non-Normality

Note that in this case we need an additional coefficient of kurtosis which we denote by δ_r for the random effect representing raters. In his seminal paper Tukey [18] obtained the variance of the variance estimates under various ANOVA models by employing "polykays". We modified Tukey's results to fit our two-way random effect model. After some algebra we can show that:

$$\operatorname{var}(SB) = \frac{\sigma_e^4}{k}\left[\frac{\delta_b + 2}{\theta_2^2} + \frac{4}{n\theta_2} + \frac{2}{n(n-1)}\right]$$

$$\operatorname{var}(SR) = \sigma_e^4\left[\left(\frac{\delta_r}{n} + \frac{2}{n-1}\right)\frac{\theta_1^2}{\theta_2^2} + \frac{4\theta_1}{k(n-1)\theta_2} + \frac{2}{k(k-1)(n-1)}\right]$$

$$\operatorname{var}(SE) = \frac{\sigma_e^4}{k(n-1)}\left[2 + \frac{n-1}{n}\delta_e\right]$$

$$\operatorname{cov}(SB, SR) = \frac{2}{k(k-1)n(n-1)}\sigma_e^4$$

$$\operatorname{cov}(SB, SE) = \frac{-2}{kn(n-1)}\sigma_e^4$$

$$\operatorname{cov}(SR, SE) = \frac{-2}{k(k-1)(n-1)}\sigma_e^4 .$$

To investigate the effect non-normality, we again use the Taylor's expansion to derive expression for the variance bias, and the mean square error of $\hat{\rho}_2$. See Appendix VI.

To explore the effect of non-normality on the mean square error of $\hat{\rho}_2$, we consider four scenarios, and the results are summarized in **Table 6**, **Table 7**, **Table 8**, and **Table 9**. We fixed the number of subjects, the number raters, the values of the ICC and the nuisance θ_2, but we varied the kurtoses of the variance components parameters. In **Table 6** we set $\delta_b = 3$, $\delta_e = 0$, $\delta_r = 0$, in both the variance and bias terms (Appendix IV) a scenario indicating that the between subjects component of variation is not normally distributed, in **Table 7** we set $\delta_b = 0$, $\delta_r = 0$, $\delta_e = 3$, a scenario indicating that only the error term is not normally distributed, in **Table 8** we set $\delta_b = 0$, $\delta_e = 0$, $\delta_r = 3$, indicating that the between raters component of variation is not normally distributed. **Table 9** summarizes the results when the three kurtosis parameters are set to zero (the normal case). The comparison among the four tables is restricted to variation in the MSE values (the last column in **Table 6**, **Table 7**, **Table 8**, and **Table 9**). We conclude that:

1) When the number of subjects is fairly large ($k > 25$) and for all the parameters values, the kurtosis parameter of any of the components has minor or negligible effect on the MSE.

2) For small values of ρ_2 (<0.4) the bias is positive, and is negative for larger values. The variation in the nuisance parameter θ_2 does affect the bias, the variance and ultimately the MSE. For small values of ρ_2, the MSE decreases with increasing values of θ_2, however, for large values of ρ_2, the MSE increases with increasing values of θ_2.

3) On comparing the MSE values in **Table 6**, **Table 7** and **Table 8** to **Table 9**, we find that a non-zero kurtosis δ_r produces the highest MSE as compared to non-zero δ_b. As we indicated the MSE values are smaller in the case of a non-zero kurtosis of the error term.

4. Data Analysis

In this section we analyze two data sets. Using the large sample bias and variance given in Appendix IV, we construct a Wald's type large sample confidence:

$$\hat{\rho}_2 \pm z_{1-\alpha/2}\sqrt{\operatorname{var}(\hat{\rho}_2)} \tag{12}$$

Because of the lack of exact expressions, and to assess the accuracy of the proposed approximations, we

Table 6. $\delta_b = 3$, $\delta_e = 0$, $\delta_r = 0$.

k	n	ρ_2	θ_2	Bias	Variance	MSE
5	2	0.4	0.01	0.01922	0.15451	0.15488
5	2	0.4	0.02	0.01050	0.14530	0.14541
5	2	0.4	0.10	−0.02814	0.09669	0.09748
5	2	0.8	0.01	−0.07690	0.02164	0.02755
5	2	0.8	0.02	−0.01411	0.02808	0.02828
5	2	0.8	0.10	−0.08968	0.02332	0.03136
10	2	0.4	0.01	0.00454	0.07236	0.07238
10	2	0.4	0.02	−0.00291	0.06695	0.04374
10	2	0.4	0.10	−0.02943	0.04288	0.06101
25	3	0.7	0.01	0.01340	0.00309	0.00327
25	3	0.7	0.02	−0.01761	0.00387	0.00418
25	3	0.7	0.10	−0.02033	0.00736	0.00777
25	3	0.8	0.01	−0.01232	0.00160	0.00175
25	3	0.8	0.02	−0.01704	0.00258	0.00287
25	3	0.8	0.10	−0.02059	0.00705	0.00747
100	7	0.8	0.01	−0.00429	0.00061	0.00062
100	7	0.8	0.02	−0.00517	0.00111	0.00114
100	7	0.8	0.10	−0.00259	0.00121	0.00122

Table 7. $\delta_b = 0$, $\delta_r = 0$, $\delta_e = 3$.

k	n	ρ_2	θ_2	Bias_RO2	Var_RO2	MSE
5	2	0.4	0.01	0.10490	0.12162	0.13262
5	2	0.4	0.02	0.09409	0.11278	0.12164
5	2	0.4	0.10	0.04242	0.07346	0.07526
5	2	0.8	0.01	−0.02304	0.01569	0.01622
5	2	0.8	0.02	−0.02898	0.01602	0.01686
5	2	0.8	0.10	−0.03936	0.03181	0.03336
10	2	0.4	0.01	0.04597	0.06213	0.06424
10	2	0.4	0.02	0.03874	0.05626	0.05776
10	2	0.4	0.10	0.01018	0.03497	0.03508
25	3	0.7	0.01	−0.00422	0.00230	0.00232
25	3	0.7	0.02	−0.00597	0.00259	0.00263
25	3	0.7	0.10	−0.00593	0.00554	0.00557
25	3	0.8	0.01	−0.00449	0.00109	0.00111
25	3	0.8	0.02	−0.00630	0.00161	0.00165
25	3	0.8	0.10	−0.00612	0.00541	0.00545
100	7	0.8	0.01	−0.00165	0.00031	0.00031
100	7	0.8	0.02	−0.00195	0.00055	0.00056
100	7	0.8	0.10	−0.00064	0.00076	0.00076

Table 8. $\delta_b = 0$, $\delta_e = 0$, $\delta_r = 3$.

k	n	ρ_2	θ_2	Bias	Variance	MSE
5	2	0.4	0.01	0.34706	0.27031	0.39076
5	2	0.4	0.02	0.31681	0.24705	0.34742
5	2	0.4	0.10	0.16019	0.13180	0.15747
5	2	0.8	0.01	−0.00454	0.03182	0.03184
5	2	0.8	0.02	−0.01411	0.02808	0.02828
5	2	0.8	0.10	−0.04406	0.01852	0.02046
10	2	0.4	0.01	0.15600	0.14456	0.16890
10	2	0.4	0.02	0.13748	0.12826	0.14716
10	2	0.4	0.10	0.05347	0.05815	0.06101
25	3	0.7	0.01	0.00107	0.00704	0.00705
25	3	0.7	0.02	−0.00193	0.00596	0.00596
25	3	0.7	0.10	−0.00709	0.00370	0.00375
25	3	0.8	0.01	−0.00269	0.00270	0.00271
25	3	0.8	0.02	−0.00511	0.00253	0.00256
25	3	0.8	0.10	−0.00805	0.00302	0.00308
100	7	0.8	0.01	−0.00118	0.00071	0.00071
100	7	0.8	0.02	−0.00174	0.00070	0.00071
100	7	0.8	0.10	−0.00102	0.00050	0.00050

Table 9. Bias and MSE of the ICC estimator under the normality assumption.

k	n	ρ_2	θ_2	Bias	Variance	MSE
5	2	0.4	0.01	0.10483	0.12149	0.13248
5	2	0.4	0.02	0.09384	0.11232	0.12112
5	2	0.4	0.10	0.03829	0.06607	0.06754
5	2	0.8	0.01	−0.02320	0.01535	0.01589
5	2	0.8	0.02	−0.02956	0.01481	0.01569
5	2	0.8	0.10	−0.04731	0.01619	0.01843
10	2	0.4	0.01	0.04591	0.06201	0.06412
10	2	0.4	0.02	0.03851	0.05584	0.05732
10	2	0.4	0.10	0.00682	0.02916	0.02920
25	3	0.7	0.01	−0.00429	0.00221	0.00222
25	3	0.7	0.02	−0.00619	0.00229	0.00232
25	3	0.7	0.10	−0.00800	0.00311	0.00317
25	3	0.8	0.01	−0.00455	0.00099	0.00101
25	3	0.8	0.02	−0.00653	0.00128	0.00133
25	3	0.8	0.10	−0.00825	0.00289	0.00295
100	7	0.8	0.01	−0.00168	0.00027	0.00027
100	7	0.8	0.02	−0.00205	0.00046	0.00047
100	7	0.8	0.10	−0.00103	0.00049	0.00049

compare the results to the distribution free bootstrap techniques. Moreover we compare the approximate Wald's confidence interval, in terms of width, with other proposed approximate confidence intervals proposed in the literature. To be specific we shall compare our results with:

1) Fleiss and Shrout [19] approximate confidence interval for which they applied Satterthwaite's two-moment approximation to arrive at an approximate $100(1-\alpha)$ per cent one-sided upper bound (U) and one-sided lower bound (L).

2) Capelleri JC, Ting N [20] modified the moments approximation initially proposed by Zou and McDermott [21] to obtain more accurate coverage probabilities.

3) BC_a or bootstrap confidence intervals, known as the "bias corrected-bias accelerated" confidence intervals [22] [23].

We analyze the two data sets using the R-packages [24] [25] (**PSY**), and (**bootstrap**). We provide the R-code for the first data set only since it is the same code needed for the second data set.

Example 1: Agreement among pathologists (see **Figure 1**).

Landis and Koch [26] evaluated the agreement among seven who classified most involved lesion of the uterine cervix. Pathologists were asked to categorize 117 using a score in the range (1 - 4): Category 1: negative, category 2: atypical squamous hyperplasia, category 3: carcinoma in situ; category 4: squamous carcinoma with early stromal invasion; category 5: invasive carcinoma. We ignored the categorical nature of the data and estimated the ICC under the model given in (8).

R-CODE

```
x<-Slide_2
x$Slide<-NULL
x$Serial<-NULL
head(x)
library(psy)
library(boot)
icc.c1<-function (data)
{
   score <- as.matrix(na.omit(data))
   n <- dim(score)[1]
   p <- dim(score)[2]
   data2 <- matrix(ncol = 3, nrow = p * n)
   attr(score, "dim") <- c(p * n, 1)
   data2[, 1] <- score
   subject <- as.factor(rep(1:n, p))
   rater <- as.factor(rep(1:p, each = n))
   data2[, 2] <- subject
   data2[, 3] <- rater
   ms <- anova(lm(score ~ subject + rater))[[3]]
   names(ms) <- NULL
   v.s <- (ms[1] - ms[3])/p
   v.r <- (ms[2] - ms[3])/n
   res <- ms[3]
   icc.a <- v.s/(v.s + v.r + res)
   #icc.c <- v.s/(v.s + res)
   return(icc.a)
}
icc.c1(x)
case.fun<-function(d,i)
  icc.c1(d[i,])
icc_boot
icc_boot<-boot(x,case.fun,R=1000)
icc_boot
est<-icc_boot$t
```

m<-mean(est)
var(est)
sd(est)
plot(icc_boot,qdist = "norm")
boot.ci(icc_boot, type = "BCa")
Results:
Bootstrap Statistics: original bias std. error
$\hat{\rho}_2 = t = 0.6475486 \quad -0.00447119 \quad 0.04080831$
boot.ci(icc_boot, type = "all")
Intervals:

Level	Normal	Basic
95%	(0.5720, 0.7320)	(0.5716, 0.7333)
Level	Percentile	BCa
95%	(0.5618, 0.7235)	(0.5749, 0.7320)

The ANOVA estimator $\hat{\rho}_2 = 0.647$ has a standard error = 0.037; which is comparable to the Bootstrap standard error = 0.041. Moreover, the analytical bias in $\hat{\rho}_2$ is −0.002, while the Bootstrap bias is −0.004.

Comments:

It is clear that our proposed Wald's 95% confidence interval is quite close to the BC_a, and this is attributed to the large sample size. Moreover, we can see from the histogram, and the q-q plot of the bootstrap samples of t^*, as shown in **Figure 1**, which is in fact $\hat{\rho}_2$, that it has a symmetric normal distribution even though we are certain that the original scores are not normal.

Example 2: Grading of retinopathy (see **Figure 2**).

Retinopathy score of retinal whole mounts was performed using fluorescent microscopy, and images were acquired using a digital camera. The extent of retinal neovascularization (NV) was estimated by implementing a specific retinal NV scoring system. In brief, the entire retina was outlined to distinguish the total retinal area of each eye. Then, the images were given threshold to emphasize only the FITC-perfused vessels. This permitted the measurement of total blood vessel area of each retina and the percentage of each retina that is engrossed with blood vessels. The scoring system was based on selecting several criteria, these are; 1) the size of the central avascular area, 2) blood vessel tuft formation, 3) extra-retinal neovascularization and 4) presence of blood vessel tortuosity. For the purposes of this model, we divided the retina into three areas: zone 1, the inner circumferential third of the retina around the optic disc; zone 2, the middle third of the retina; and zone 3, the outer third of the retina. The extent of disease was specified by clock hours or distance around the retina (number of twelfths similar to a clock). The scoring was performed in a masked fashion, by employing fluorescence microscopy, evaluating and scoring each retina in a blinded manner by three observers. The minimum score according to this method is 0, and the maximum score is 13. Maximal vaso-proliferation in this mouse model has previously been reported to occur from P17 to P21. The average retinopathy score for each animal was used for statistical analysis [27].

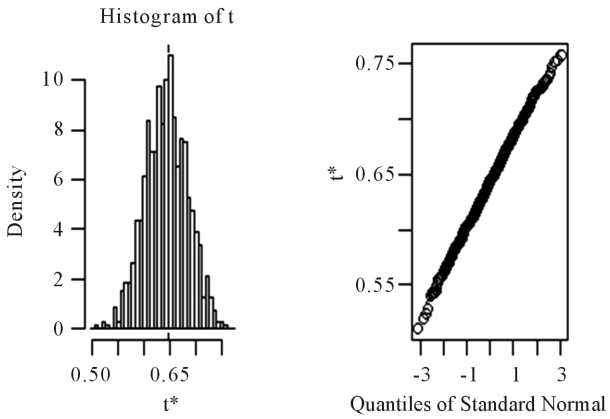

Figure 1. Graphs of histogram and q-q plot of the 1000 bootstrap sample for example 1.

Bootstrap Statistics: original bias std. error
$\hat{\rho}_2 = t_1^* = 0.9580814 \ -0.003262422 \ 0.01569344$

The ANOVA estimator is 0.958, with standard error = 0.019. The analytic bias is −0.006, but the Bootstrap bias is −0.005.

Comments:

Although the sample here is much smaller (k = 16) the Wald's 95% confidence interval is quite close to the corresponding bootstrap confidence interval. But the distribution of $\hat{\rho}_2 = t_1^*$ is far from being normally distributed as can be seen from the histogram (skewed to the left) and the curved q-q plot in **Figure 2**.

For example 1, **Table 10** gives the ANOVA results, and **Table 11** gives the 95% CI for the 6 methods. For example 2, **Table 12** gives the ANOVA results, and **Table 13** gives the 95% CI for the 6 methods.

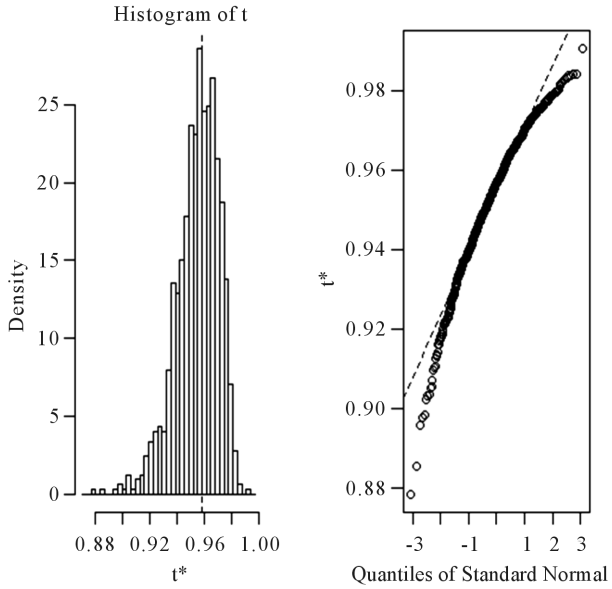

Figure 2. Graphs of histogram and q-q plot of the 1000 bootstrap sample for example 2.

Table 10. ANOVA table for the first data set.

S.O.V.	d. f	s.o.s	M.S.
Slides	116	613.04	5.285
Raters	6	77.72	12.954
Error	969	195.42	0.281

Table 11. Comparing 95% CI by alternative methods.

Method	95% confidence units
FS	(0.600, 0.740)
Modified Z.M.	(0.414, 0.784)
Wald's	(0.574, 0.719)
Bootstrap normal	(0.572, 0.732)
Bootstrap BCa	(0.575, 0.732)
Bootstrap percentile	(0.562, 0.724)

Table 12. ANOVA table for data of example 2.

S.O.V.	d.f	s.o.s	M.S.
Subject	15	1066	71.067
Raters	2	0.542	0.271
Error	30	32	1.071

Table 13. Comparing 95% confidence intervals by different methods for data of example 2.

Method	95% confidence units
FS	(0.891, 0.989)
Modified ZM	(0.907, 0.984)
Wald's	(0.918, 0.994)
Bootstrap normal	(0.905, 0.999)
Bootstrap BCa	(0.908, 0.999)
Bootstrap percentile	(0.901, 0.974)

5. Discussion

The use of the intraclass correlation to assess the reliability of judgments made by the observers is ubiquitous in medical research. Unreliable measurements can eventually affect the diagnosis of diseases and hence expose patients to undesired health risks. Several examples regarding the applications can be found in [2] [3] [5].

1) In the one-way we observe that from the tables that the effect of normality on the variance and the bias of the estimated ICC depend on the kurtosis of the distributions of the components of variation. We noted that the kurtosis of the between subjects component of variation is multiplied by the factor k^{-1}, kurtosis of the between rater's component of variation is multiplied by a factor n^{-1} and that of the error component multiplied by a factor $\left(k\left(n-1\right)\right)^{-1}$. Therefore the effect of non-normality of the between raters component will dominate that of the subjects component would dominate the effect of non-normality of the error (within subjects) components. In general, non-zero kurtosis of the observed distribution of measurements substantially affects the sampling distribution, the bias, and standard error of reliability estimates. For the two-way model, there is an interaction between the values of ρ and the nuisance parameter (ratio of between subjects variance to the error variance).

2) The magnitude of ρ has an effect on the sampling distribution of the reliability estimate. The larger the value of ρ the smaller the mean square error of its estimator. This observation is relevant within the framework of reliability studies since we are interested in producing large values of the estimate.

6. Summary

From the above discussion we summarize our conclusions in three points:

Firstly, the effect of non-normality of the scores distribution is not tangible for $k > 25$ (fairly large), where k is the number subjects. Secondly, large value of reliability, which is our main concern, has smaller variance, and hence a Wald's confidence interval on the population ICC would be acceptable. This conclusion is based on the empirical evidence; as we have shown that the Wald's interval is almost identical in length to the interval based on the bootstrap methods. Thirdly, in the design stage of a reliability study a reasonable guess of the kurtosis is required to satisfy, for given effect size and type I error rate, the sample size requirements to achieve a certain power on specific hypotheses.

We should also note that the estimation problem in the two-way model can be more complex particularly when the raters are required to repeat the assessment of each patient, in order to reduce bias and increase precision of the reliability estimator. In this case the sample size question can be: what is the number of subjects, the

number of raters, and the number of repeats per subjects to achieve certain power level. In such case one has to seriously consider the cost of replications in the early stages of designing reliability study.

References

[1] Atenafu, E.G., Hamid, J.S., To, T., Willan, A., Feldman, B. and Beyene, J. (2012) Bias-Corrected Estimator for the Intraclass Correlation Coefficient in the Balanced One-Way Random Effects Model. *BMC Medical Research Methodology*, **12**, 126. http://dx.doi.org/10.1186/1471-2288-12-126

[2] Dunn, G. (1992) Design and Analysis of Reliability Studies. *Statistical Methods in Medical Research*, **1**, 123-157. http://dx.doi.org/10.1177/096228029200100202

[3] Dunn, G. (2004) Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors. 2nd Edition, Oxford University Press, New York.

[4] Shoukri, M.M., Asyali, M.H. and Walter, S.W. (2003) Issues of Cost and Efficiency in the Design of Reliability Studies. *Biometrics*, **59**, 1109-1114. http://dx.doi.org/10.1111/j.0006-341X.2003.00127.x

[5] Shoukri, M.M. (2010) Measures of Interobserver Agreement and Reliability. 2nd Edition, Chapman & Hall/CRC Biostatistics Series. http://dx.doi.org/10.1201/b10433

[6] Haggard, E.A. (1958) Intraclass Correlation and the Analysis of Variance. Dryden Press, New York.

[7] Fisher, R.A. (1958) Statistical Methods for Research Workers. Hafner, New York.

[8] Shrout, P.E. and Fleiss, J.L. (1979) Intraclass Correlations: Use in Assessing Rater Reliability. *Psychological Bulletin*, **86**, 420-428. http://dx.doi.org/10.1037/0033-2909.86.2.420

[9] McGraw, K.O. and Wong, S.P. (1996) Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, **1**, 30-46. http://dx.doi.org/10.1037/1082-989X.1.1.30

[10] Walter, S.D., Eliasziw, M. and Donner, A. (1998) Sample Size and Optimal Designs for Reliability Studies. *Statistics in Medicine*, **17**, 101-110. http://dx.doi.org/10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E

[11] Shoukri, M.M., Asyali, M.H. and Donner, A. (2004) Sample Size Requirements for the Design of Reliability Study: Review and Results. *Statistical Methods in Medical Research*, **13**, 251-271.

[12] Searle, S.R., Casella, G. and McCulloch, C.E. (1992) Variance Components. John Wiley & Sons, New York.

[13] Hammersley, J.M. (1949) The Unbiased Estimate and Standard Error of the Intraclass Variance. *Metron*, **15**, 189-205.

[14] Shoukri, M.M., Tracy, D.S. and Mian, I.U.H. (1990) The Effect of Kurtosis in Estimation of the Parameters of the One-Way Random Effects Model from Familial Data. *Computational Statistics and Data Analysis*, **10**, 339-345. http://dx.doi.org/10.1016/0167-9473(90)90016-B

[15] Donner, A. (1986) A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. *International Statistical Review*, **54**, 67-82.

[16] Bonneau, C.A. (1960) The Effect of Violation of Assumptions Underlying the t-Test. *Psychological Bulletin*, **57**, 49-64. http://dx.doi.org/10.1037/h0041412

[17] Scheffe, H. (1959) The Analysis of Variance. Wiley, New York.

[18] Tukey, J.W. (1956) Variance of Variance Components. I. Balanced Designs. *Annals of Mathematical Statistics*, **27**, 722-736. http://dx.doi.org/10.1214/aoms/1177728179

[19] Fleiss, J.L. and Shrout, P.E. (1978) Approximate Interval Estimation for a Certain Class of Intraclass Correlation Coefficient. *Psychometrika*, **43**, 259-262. http://dx.doi.org/10.1007/BF02293867

[20] Capelleri, J.C. and Ting, N. (2003) A Modified Large Sample Approach to Approximate Interval Estimation for a Particular Intraclass Correlation Coefficient. *Statistics in Medicine*, **22**, 1861-1877. http://dx.doi.org/10.1002/sim.1402

[21] Zou, K.H. and McDermott, M.P. (1999) Higher-Moment Approaches to Approximate Interval Estimation for a Certain Intraclass Correlation Coefficient. *Statistics in Medicine*, **18**, 2051-2061. http://dx.doi.org/10.1002/(SICI)1097-0258(19990815)18:15<2051::AID-SIM162>3.0.CO;2-P

[22] Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton.

[23] Davison, A.C. and Hinkley, D.V. (1997) Bootstrap Methods and Their Application. In: *Cambridge Series in Statistical and Probabilistic Mathematics*, No. 1, Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/cbo9780511802843

[24] https://cran.r-project.org/package=psy

[25] http://www.rdocumentation.org/packages/psy

[26] Landis, J.R. and Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**, 150-174. http://dx.doi.org/10.2307/2529310

[27] DeNiro, M., Al-Halafi, A., Al-Mohanna, F.H., Alsamadi, O. and Al-Mohanna, F.A. (2010) Pleiotropic of YC-1 Selectively Inhibit Pathological Retinal Neovascularization and Promote Physiological Revascularization in a Mouse Model of Oxygen-Induced Retinopathy. *Molecular Pharmacology*, **77**, 348-367. http://dx.doi.org/10.1124/mol.109.061366

Appendices

Appendix I: The Taylor's expansion of a function of several variables.

Let $X=(x_1,x_2,\cdots,x_c)$ be a random vector with $\mu=E(X)=(\mu_1,\mu_2,\cdots,\mu_c)$ be a real valued continuously differentiable function of X.

The Taylor series expansion of $Q(\cdot)$ around $\underline{\mu}$ is given by:

$$\begin{aligned}Q(x_1,x_2,\cdots,x_c)&=Q(\mu_1,\mu_2,\cdots,\mu_c)+\sum_{j=1}^{c}\frac{\partial Q}{\partial x_j}(x_j-\mu_j)\\&+\frac{1}{2!}\sum_{j=1}^{c}\sum_{k=1}^{c}\frac{\partial^2 Q}{\partial x_j\partial x_k}(x_j-\mu_j)(x_k-\mu_k)\\&+\frac{1}{3!}\sum_{j=1}^{c}\sum_{k=1}^{c}\sum_{l=1}^{c}\frac{\partial^3 Q}{\partial x_j\partial x_k\partial x_l}(x_j-\mu_j)(x_k-\mu_k)(x_l-\mu_l)+\cdots.\end{aligned}$$

Appendix II: Analytic formula of the bias of the estimated intraclass correlation under the one-way random effects model.

1) The non-normal case:

$$\begin{aligned}\text{Bias}(\hat{\rho})&=\frac{1}{2!}\left[\text{Var}(S_B)\frac{\dot{\partial}^2\hat{\rho}}{\partial S_B^2}+2\text{Cov}(S_B,S_W)\frac{\dot{\partial}^2 f}{\partial S_B\partial S_W}+\text{Var}(S_W)\frac{\dot{\partial}^2\hat{\rho}}{\partial S_W^2}\right]\\&=\frac{c_1}{n^2}(n-1)(1-\rho)^2\left[1+(n-1)\rho\right]-\frac{c_2}{n^2}(1-\rho)^3\\&+\frac{c_{12}}{n^2}(1-\rho)^2\left[n-2(n-1)(1-\rho)\right].\end{aligned}$$

2) Under normality $(\delta_b=\delta_e=0)$

$$c_{12}=0,\ c_1=\frac{2}{k(n-1)},\ c_2=\frac{2}{k(1-\rho)^2}\left[1+(n-1)\rho\right]^2.\ \text{Hence}$$

$$\text{Bias}(\hat{\rho})=\frac{-2(1-\rho)(1+(n-1)\rho)}{kn^2}\left[n^2\left[1+(n-1)\rho\right]-(1-\rho)\right].$$

Appendix III: Analytic expression for variance and bias of the estimated intraclass correlation in the two-way random effects model:

1) Variance Expression under normality:

$$\text{Var}(\hat{\rho}_2)=\text{Var}(x_1)\left(\frac{\dot{\partial}\rho_2}{\partial x_1}\right)^2+\text{Var}(x_2)\left(\frac{\dot{\partial}\rho_2}{\partial x_2}\right)^2+\text{Var}(x_3)\left(\frac{\dot{\partial}\rho_2}{\partial x_3}\right)^2\equiv T_1+T_2+T_3.$$

It would be most informative to express the bias and variance in terms of ρ (the primary parameter of interest) and simplify the expressions accordingly:

$$1+\theta n_1=\frac{(1-\rho_2)+n\rho_2(1+\theta_2)}{1-\rho_2}.$$

We can now write variance expression as

$$T_1=\frac{2(1-\rho_2)^2(1+A)^2\left[1-\rho_2+n\rho_2(1+\theta_2)\right]^2}{(k-1)B^4}$$

$$T_2=\frac{2n^2\rho_2^2d_2^2(1+k\theta_2)^2(1+\theta_2)^2(1-\rho_2)^2}{(n-1)B^4}$$

$$T_3 = \frac{2(1-\rho_2)^2\left[(1+d_3)(1-\rho_2+n\rho_2(1+\theta_2))+d_2(1-\rho_2)(1+k\theta_2)\right]^2}{(k-1)(n-1)B^4}$$

where, $A \equiv d_2(1+k\theta_2)+d_3$, and $B \equiv (1-\rho_2)(1+A)+n\rho_2(1+\theta_2)$.

2) Bias in Estimating $\hat{\rho}_2$ under normality.

A first order approximation of the bias under the assumption of normality is given by:

$$E(\hat{\rho}_2-\rho_2) \simeq \frac{1}{2!}\left[\mathrm{Var}(x_1)\frac{\dot{\partial}^2 f}{\partial x_1^2}+\mathrm{Var}(x_2)\frac{\dot{\partial}^2 f}{\partial x_2^2}+\mathrm{Var}(x_3)\frac{\dot{\partial} f}{\partial x_3^2}\right] = [u_1+u_2+u_3].$$

The terms of the bias expression are:

$$u_1 = \frac{-2(1-\rho_2)(1+A)\left[1-\rho_2+n\rho_2(1+\theta_2)\right]^2}{(k-1)B^3}$$

$$u_2 = \frac{2n\rho_2 d_2^2(1-\rho_2)^2(1+\theta_2)(1+k\theta_2)^2}{(n-1)B^3}$$

and

$$u_3 = \frac{2d_3(1-\rho_2)^2\left[(1-\rho_2)(1-A)+n\rho_2(1+\theta_2)(1+d_3)\right]}{(k-1)(n-1)B^3}.$$

Appendix IV: variance and bias of estimated intraclass correlation under non-normality.

Variance:

$$\begin{aligned}\mathrm{var}(\hat{\rho}_2) &= \left(\frac{\dot{\partial}\hat{\rho}_2}{\partial S_b}\right)^2 \mathrm{var}(S_b)+\left(\frac{\dot{\partial}\hat{\rho}_2}{\partial S_r}\right)^2 \mathrm{var}(S_r)+\left(\frac{\partial\hat{\rho}_2}{\partial S_w}\right)^2 \mathrm{var}(S_w)\\ &\quad +2\,\mathrm{cov}(S_b,S_r)\left(\frac{\dot{\partial}\rho_2}{\partial S_b}\right)\left(\frac{\dot{\partial}\rho_2}{\partial S_r}\right)+2\,\mathrm{cov}(S_b,S_w)\left(\frac{\dot{\partial}\rho_2}{\partial S_b}\right)\left(\frac{\dot{\partial}\rho}{\partial S_w}\right)\\ &\quad +2\,\mathrm{cov}(S_r,S_w)\left(\frac{\dot{\partial}\rho}{\partial S_r}\right)\left(\frac{\dot{\partial}\rho}{\partial S_w}\right)\\ &= T_1^*+T_2^*+T_3^*+\left[T_{12}^*+T_{13}^*+T_{23}^*\right]\end{aligned}$$

where

$$T_1^* = \frac{2C_1}{k\theta_2^2}\left[d_2(1+k\theta_2)+(1+d_3)\right]^2\left[1+\frac{\delta_b}{2}+\frac{2\theta_2}{nk}+\frac{\theta_2^2}{n(n-1)}\right]$$

$$T_2^* = \frac{2d_2^2n^2\theta_1^4}{(n-1)\theta_2^2}C_1\left[1+\frac{n-1}{2n}\delta_r+\frac{2\theta_2}{k\theta_1}+\frac{\theta_2^2}{k(k-1)\theta_1^2}\right]$$

$$T_3^* = \frac{\left(2+\frac{n-1}{n}\delta_e\right)}{k(n-1)}C_1\left[1+n\theta_1+d_2(1+k\theta_2)+d_3(1+n\theta_1)\right]^2$$

$$T_{12}^* = \frac{4d_2n\theta_1}{k(k-1)n(n-1)}C_1\left[(1+d_3)+d_2(1+k\theta_2)\right]$$

$$T_{13}^* = \frac{4\left[d_2(1+k\theta_2)+(1+d_3)\right]C_1}{kn(n-1)}\left[1+n\theta_1+d_2(1+k\theta_2)+d_3(1+n\theta_1)\right]$$

$$T_{23}^{*}=\frac{-4nd_2\theta_1}{k(k-1)(n-1)}C_1\left[1+n\theta_1+d_2(1+k\theta_2)+d_3(1+n\theta_1)\right]$$

where $C_1=\left[1+d_2(1+k\theta_2)+n\theta_1+d_3\right]^{-4}$.

Bias:

The delta method is used to derive the variance of $\hat{\rho}_2$

$$\text{Bias}=\frac{1}{2!}\left[u_1^{*}+u_2^{*}+u_3^{*}+2\left\{u_{12}^{*}+u_{13}^{*}+u_{23}^{*}\right\}\right]$$

$$u_1^{*}=\frac{-4}{k\theta_2^2}D_1\left[1+\frac{\delta_b}{2}+\frac{2\theta_2}{nk}+\frac{\theta_2^2}{n(n-1)}\right]\left[1+d_3+d_2(1+k\theta_2)\right]$$

where $D_1=\left[1+d_3+n\theta_1+d_2(1+k\theta_2)\right]^{-3}$

$$u_2^{*}=2\theta_1d_2^2D_1\left[\delta_r+\frac{2n\theta_1^2}{(n-1)\theta_2^2}+\frac{4n\theta_1}{k(n-1)\theta_2}+\frac{2n}{k(k-1)(n-1)}\right]$$

$$u_3^{*}=\frac{2d_3D_1}{k(n-1)}\left[1+n\theta_1+d_2(1+k\theta_2)+d_3(1+n\theta_1)\right]\left[2+\frac{n-1}{n}\delta_e\right]$$

$$u_{12}^{*}=\frac{2d_2E_1}{k(k-1)n(n-1)}\left[\frac{1+n\theta_2}{\theta_2}-d_2\frac{\theta_1(1+k\theta_2/\theta_1)}{\theta_2}-(2+d_3)\right]$$

$$E_1=\left[\frac{1+n\theta_2}{\theta_2}+d_2\frac{\theta_1(1+k\theta_2/\theta_1)}{\theta_2}+d_3\right]^{-3}$$

$$u_{13}^{*}=-\frac{2E_1}{kn(n-1)}\left[(1+d_3)\left(\frac{1+n\theta_2}{\theta_2}\right)+d_2(1-d_3)\frac{\theta_1}{\theta_2}\left(1+k\frac{\theta_2}{\theta_1}\right)-d_3(1+d_3)\right]$$

$$u_{23}^{*}=\frac{-2E_1}{k(k-1)(n-1)}\left[(1+d_3)d_2\left(\frac{1+n\theta_2}{\theta_2}\right)+\frac{\theta_1}{\theta_2}\left(1+k\frac{\theta_2}{\theta_1}\right)d_2^2\right].$$