BMC Medical Research Methodology

BioMed Central

Research article

Interval estimation and optimal design for the within-subject coefficient of variation for continuous and binary variables

Mohamed M Shoukri*[1,2], Nasser Elkum[1] and Stephen D Walter[3]

Address: [1]Department of Biostatistics, Epidemiology and Scientific Computing King Faisal Specialist Hospital and Research Centre, P.O. Box 3354, Riyadh 11211, Saudi Arabia, [2]Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada and [3]Department of Epidemiology and Biostatistics, McMaster University Hamilton, Ontario, Canada

Email: Mohamed M Shoukri* - shoukri@kfshrc.edu.sa; Nasser Elkum - nkum@kfshrc.edu.sa; Stephen D Walter - walter@mcmaster.ca

* Corresponding author

Published: 10 May 2006

BMC Medical Research Methodology 2006, **6**:24 doi:10.1186/1471-2288-6-24

Received: 03 July 2005
Accepted: 10 May 2006

This article is available from: http://www.biomedcentral.com/1471-2288/6/24

Abstract

Background: In this paper we propose the use of the within-subject coefficient of variation as an index of a measurement's reliability. For continuous variables and based on its maximum likelihood estimation we derive a variance-stabilizing transformation and discuss confidence interval construction within the framework of a one-way random effects model. We investigate sample size requirements for the within-subject coefficient of variation for continuous and binary variables.

Methods: We investigate the validity of the approximate normal confidence interval by Monte Carlo simulations. In designing a reliability study, a crucial issue is the balance between the number of subjects to be recruited and the number of repeated measurements per subject. We discuss efficiency of estimation and cost considerations for the optimal allocation of the sample resources. The approach is illustrated by an example on Magnetic Resonance Imaging (MRI). We also discuss the issue of sample size estimation for dichotomous responses with two examples.

Results: For the continuous variable we found that the variance stabilizing transformation improves the asymptotic coverage probabilities on the within-subject coefficient of variation for the continuous variable. The maximum like estimation and sample size estimation based on pre-specified width of confidence interval are novel contribution to the literature for the binary variable.

Conclusion: Using the sample size formulas, we hope to help clinical epidemiologists and practicing statisticians to efficiently design reliability studies using the within-subject coefficient of variation, whether the variable of interest is continuous or binary.

Background

Measurement errors can seriously affect statistical analysis and interpretation; it therefore becomes important to assess the magnitude of such errors by calculating a reliability coefficient and assessing its precision. For instance medical diagnosis, clinicians have now become cognizant to the paramount importance of obtaining accurate measurements to ensure safe and efficient delivery of care to their patients. Experiments designed to measure validity and precision of instruments used in biomedical and epidemiological research are ubiquitous. For example, Ashton [1] demonstrated the importance of evaluating the reliability of manual and automated methods for quantifying total white matter lesions burden in multiple sclero-

sis patients. They compared the coefficient of variations of three methods. In oncology, Schwartz et al. [2] used the coefficient of variation to evaluate the repeatability in bi-dimensional computed tomography measurements of three techniques: hand-held calipers on film, electronic calipers on a workstation, and an auto-contour technique on a workstation. Differences between the coefficients of variation were statistically significantly different for the auto-contour technique, compared to the other techniques. The coefficient of variation is often used to compare variables measured on different scales. For example, in social sciences, when the intent is to compare the variability in school performance with the variability of household income, a comparison of standard deviations makes no sense because income and school performance are measured on different scales. The correct comparison may be based on the coefficient of variation because it adjusts for scale. Other applications of the coefficient of variation are given in Tian [3].

Scientists have developed several indices to assess the reliability and reproducibility of quantitative measurements. The intra-class correlation (ICC), the proportion of the between-subject variance to the total variance, has been widely used as an index of measurement reliability. For a comprehensive review on the ICC and its applications, we refer the reader to Fleiss [4], Dunn [5] and Shoukri [6]. One of the criticisms of the ICC is that its value depends on the population from which the study subjects have been obtained, and this may lead to difficulties in comparing results from different studies. Accordingly, Quan and Shih [7] (QS) considered an alternative measure, the Within-Subject Coefficient of Variation (WSCV) as an alternative to the ICC for assessing measurements reproducibility or test-re-test reliability. Because of the requirement that repeated observations are made on each subject, they used the one-way random effects model (REM) as a mechanism to describe the data. Although the use of the WSCV as a measure of reproducibility is long standing, the issue of sample size determination has not been adequately investigated. Sample size estimation is one of the most important issues in the design of any study that uses inferential statistics.

When the ICC is used as the index of reliability, Donner and Eliasziw [8] provided contours of exact power for selected numbers of subjects (k) and numbers of replicates (n). These power results were then used to identify optimal designs that minimize the study costs. Assuming a constant number of replicates per subject, Walter et al. [9] considered an approximation to determine the required number of subjects to achieve fixed levels of power. Bonett [10] calculated the sample size required to achieve a prescribed expected width for the confidence interval on the ICC. Shoukri et al. [11] derived the values of k and n that allocate the sample resources optimally and minimize the variance of the estimated ICC under cost constraints. The cost structure that was considered was general and followed the general guidelines identified by Flynn et al. [12].

In this paper, we derive the optimal allocation for the number of subjects and the number of repeated measurements needed to minimize the variance of the maximum likelihood estimator (MLE) of the WSCV. In Section 2 we present the random effects model, the definition of the WSCV, and the asymptotic distribution of its MLE for continuous data. In Section 3, we use the calculus of optimization to find the optimal combinations (n, k) that minimize the variance of the MLE of WSCV for normally distributed variables. The use of the WSCV for dichotomous data has never been investigated before, and a novel contribution in this paper is the estimation of WSCV for binary outcome measurements, and sample size requirements, with emphasis on the case of two ratings per subject (i.e. $n = 2$). We devote Section 4 to the binary data, and general discussion is presented in Section 5.

Methods

Estimating the WSCV for continuous variables

Assumptions

Consider a random sample of k subjects with n repeated measurements of a continuous variable Y, and denote by Y_{ij} the j^{th} reading made on the i^{th} subject under identical experimental conditions ($i = 1,2,...k$; $j = 1,2,...n$). In a test-retest scenario, and under the assumption of no reader's effect (*i.e.* the readings within a specific subject are exchangeable), Y_{ij} denotes the reading of the j^{th} trial made on the i^{th} subject. A useful model for analyzing such data is given by:

$$Y_{ij} = \mu + s_i + e_{ij} \quad i = 1,2,...k;\ j = 1,2,...n \qquad (1)$$

where μ is the mean of Y_{ij}, the random subject effects s_i are normally distributed with mean 0 and variance σ_s^2 , or $N(0,\ \sigma_s^2)$, the measurement errors e_{ij} are $N(0,\ \sigma_e^2)$, and the s_i and e_{ij}terms are independent. We assume that the subjects are randomly drawn from some population of interest.

Quan and Shih [7] defined the WSCV parameter in the above model as:

$$\theta = \sigma_e/\mu \qquad (2)$$

With model (1), it is assumed that the within subject variance is the same for all subjects.

Maximum likelihood estimator

Under the above set-up, the log-likelihood function has the form

$$l = -\frac{nk}{2}\log 2\pi - \frac{1}{2}k(n-1)\log\sigma_e^2 - \frac{1}{2}k\left[\log\left(\sigma_e^2/n + \sigma_s^2\right)\right] - \frac{\sum_i\sum_j (y_{ij}-\mu)^2}{2\sigma_e^2} + \frac{n^2\sigma_s^2\sum_i(\bar{y}_i-\mu)^2}{2\sigma_e^2\left(\sigma_e^2+n\sigma_s^2\right)}.$$

Define $\sigma^2 = \sigma_s^2 + \sigma_e^2$ and $\rho = \sigma_s^2/(\sigma_s^2 + \sigma_e^2)$, the intra-class correlation coefficient, for which $\sigma_s^2 = \rho\sigma^2$ and $\sigma_e^2 = \sigma^2(1-\rho)$. Because the design is balanced the maximum likelihood estimators (MLE) for μ, σ^2, and ρ are given in closed forms by: $\hat{\mu} = k^{-1}\sum_{i=1}^{k}\bar{y}_i, \hat{\theta} = \sqrt{MSW}/\hat{\mu}$, $\sigma_s^2 = \frac{(k-1)MSB - (k)MSW}{nk}$, and the estimated ICC is

$$\hat{p} = \frac{(k-1)MSB - (k)MSW}{(k-1)MSB + k(n-1)MSW},$$ where

$$MSW = \frac{1}{k(n-1)}\sum_i^k\sum_j^n\left(y_{ij}-\bar{y}_i\right)^2,$$

$$MSB = \frac{n}{k-1}\sum_i^k\left(\bar{y}_i-\hat{\mu}\right)^2$$ are respectively, the within-subject and between subjects mean squares as obtained from the usual one-way ANOVA table, and $\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$. Note that the *MSB* does not exist for $k = 1$, which means that to obtain a sensible estimate of ρ as an index of reliability, the study should include more than one subject.

Results

The asymptotic variance-covariance matrix of the MLE's is obtained by inverting Fisher's information matrix. The large sample variance of $\hat{\theta}$ can be obtained using delta method (see Kendall vol. 1 [13]) and was shown by Quan and Shih [7] to be:

$$\text{var}(\theta) = A(\rho,n,\theta)/k = \frac{\theta^4}{nk}\left[1 + n\frac{\rho}{1-\rho}\right] + \frac{\theta^2}{2k(n-1)} \quad (3)$$

To construct an approximate confidence interval on $\hat{\theta}$, it is assumed that for large k, $\sqrt{k}(\hat{\theta} - \theta)$ follows a normal distribution with mean 0 and variance $A(\rho,n,\theta)$. An approximate $100(1-\alpha)\%$ confidence interval on θ can be given as $\hat{\theta} \pm Z_{\alpha/2}\sqrt{\text{var}(\hat{\theta})}$, where $Z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ cut off point of the standard normal distribution.

Due to the dependence of the variance of $\hat{\theta}$ on the true parameter value θ itself, we found that the asymptotic coverage deviates from its nominal levels for some values of θ. To improve the coverage probability we suggest a variance stabilizing transformation to remove the dependence of var($\hat{\theta}$) on θ.

Variance Stabilizing Transformation (VST)

To improve the estimated coverage proportion, we propose a variance stabilizing transformation g (see Kendall vol.1 page 541 [13]) where, $g = \int(\text{var}(\hat{\theta}))^{-1/2}\, d\theta$. With θ defined as in equation (2), it can be shown that

$$g(\theta) = \sqrt{k(n-1)/2}\log\left[\frac{\left(1+c\theta^2\right)^{1/2}-1}{\left(1+c\theta^2\right)^{1/2}+1}\right]$$ where,

$c = 2(1-\frac{1}{n})(1+n\rho^*)$, and $\rho^* = \rho/(1-\rho)$ Letting

$$f(\theta,n,\rho) = \sqrt{(n-1)/2}\log\left[\frac{\left(1+c\theta^2\right)^{1/2}-1}{\left(1+c\theta^2\right)^{1/2}+1}\right],$$ we may establish assuming the function g is bounded and differentiable, that, $f(\hat{\theta},n,\rho)$ is asymptotically normally distributed with mean $f(\theta,n,\rho)$ and variance $1/k$. Therefore, we can construct $100(1-\alpha)\%$ confidence limits on θ based on the above transformation. The upper and lower $(1-\alpha/2)100\%$ confidence bound on θ are respectively given by:

$$\hat{\theta}_u = \frac{2\exp(\xi_1(2(n-1))^{-1/2})}{c^{1/2}[1-\exp(\xi_1(\frac{2}{n-1})^{1/2})]}$$ and,

$$\hat{\theta}_l = \frac{2\exp(\xi_2(2(n-1))^{-1/2})}{c^{1/2}[1-\exp(\xi_2(\frac{2}{n-1})^{1/2})]}$$

where, $\xi_1 = f(\hat{\theta},n,\rho) + z_{\alpha/2}/\sqrt{k}$, and $\xi_2 = f(\hat{\theta},n,\rho) - z_{\alpha/2}/\sqrt{k}$.

Note that the limits of the interval depend on the unknown value of the intra-class correlation, which can be replaced by its MLE as defined in section 2.1.

To examine the finite sample behavior of the VST based confidence interval estimator, a Monte-Carlo study was conducted under model (1) using the S-Plus program. The values of ρ were, 0.3, 0.4, 0.6, 0.7, and 0.8; $\mu = 10$, and θ

Table 1: Estimated coverage probabilities under the VST. The nominal level is 95%. (θ = 0.04)

k	ρ	*n*		
		2	3	5
12	0.3	0.929	0.934	0.947
	0.4	0.943	0.937	0.930
	0.6	0.943	0.948	0.938
	0.7	0.941	0.943	0.934
	0.8	0.939	0.931	0.941
25	0.3	0.961	0.946	0.949
	0.4	0.942	0.946	0.953
	0.6	0.939	0.935	0.969
	0.7	0.946	0.936	0.948
	0.8	0.936	0.936	0.940
50	0.3	0.956	0.954	0.949
	0.4	0.945	0.939	0.950
	0.6	0.952	0.955	0.934
	0.7	0.953	0.948	0.940
	0.8	0.950	0.935	0.937
75	0.3	0.948	0.948	0.944
	0.4	0.956	0.955	0.955
	0.6	0.952	0.943	0.949
	0.7	0.945	0.944	0.930
	0.8	0.946	0.945	0.946

= 4%, 10% or 20%. Sample size (k) = 12, 25, 50, 75, and number of replicates (n) = 2, 3, and 5. The number of repetitions for each simulation was 1000. Tables 1, 2, 3, and 4, demonstrate the coverage proportion for the 95% nominal level confidence interval on the WCV. The estimated coverage proportions were close to the 95% nominal level.

Table 2: Estimated coverage probabilities under the VST. The nominal level is 95%. (θ = 0.01)

K	ρ	*n*		
		2	3	5
12	0.3	0.943	0.954	0.945
	0.4	0.930	0.934	0.952
	0.6	0.949	0.942	0.938
	0.7	0.920	0.932	0.929
	0.8	0.927	0.926	0.913
25	0.3	0.938	0.951	0.945
	0.4	0.936	0.953	0.955
	0.6	0.926	0.948	0.946
	0.7	0.928	0.939	0.931
	0.8	0.925	0.939	0.915
50	0.3	0.946	0.936	0.946
	0.4	0.957	0.935	0.942
	0.6	0.940	0.955	0.936
	0.7	0.940	0.947	0.935
	0.8	0.947	0.930	0.925
75	0.3	0.941	0.948	0.945
	0.4	0.941	0.947	0.941
	0.6	0.935	0.933	0.941
	0.7	0.932	0.929	0.949
	0.8	0.960	0.936	0.925

Example 1

Accurate and reproducible quantification of brain lesion count and volume in multiple sclerosis (MS) patients using magnetic resonance imaging (MRI) is a vital tool for evaluation of disease progression and patient response to therapy. Current standard methods for obtaining these data are largely manual and subjective and are therefore error-prone and subject to inter-and intra-operator variability. Therefore, there is a need for a rapid automated lesion quantification method. Ashton *et al*. [1] compared manual measurements and an automated data technique known as Geometrically Constrained Region Growth (GEORG) of the brain lesion volume of 3 MS patients, each measured 10 times by a single operator for each method. The data are presented in Table 5.

Based on the guidelines for the levels of reliability provided by Fleiss [4], a value of an ICC above 80% indicates an excellent reliability, and from Table 3 both methods cross this threshold level. However, based on the WSCV values, the manual method is definitely less reproducible than the automated method (the GEORG is 5 times more reproducible than the manual). This example demonstrates the usefulness of the WSCV over the ICC as a measure of reproducibility. Clearly, one should construct a formal test on the significance of the difference between two correlated within-subject coefficients of variation. There are several competing methods to construct such a test (e.g. LRT, Wald, and Score tests) but this issue is quite involved and so we intend to report our findings in a future publication.

Sample size estimation

In the following development we discuss the second objective of this paper. We assume that the investigator is interested in the number of replicates, n, per subject, so that the variance of the estimate of θ is minimized, given that the total number of measurements is fixed *a priori* at $N = nk$.

Efficiency criterion

For fixed total number of measurements $N = nk$, equation (3) gives:

$$\mathrm{var}(\theta) = \frac{\theta^4}{N}\left[1 + n\frac{\rho}{1-\rho}\right] + \frac{n\theta^2}{2N(n-1)} \qquad (4)$$

The necessary condition for $\mathrm{var}(\hat{\theta})$ to have a unique minimum is that $\partial\mathrm{var}(\hat{\theta})/\partial n = 0$. This, and the additional condition that $\partial^2\mathrm{var}(\hat{\theta})/\partial n^2 > 0$ are both satisfied so long

Table 3: Estimated coverage probabilities under the VST. The nominal level is 95%. (θ = 0.02)

K	ρ	n		
		2	3	5
12	0.3	0.947	0.941	0.945
	0.4	0.936	0.945	0.942
	0.6	0.932	0.939	0.950
	0.7	0.937	0.938	0.935
	0.8	0.934	0.925	0.921
25	0.3	0.951	0.951	0.954
	0.4	0.942	0.943	0.943
	0.6	0.940	0.946	0.934
	0.7	0.940	0.941	0.920
	0.8	0.939	0.934	0.909
50	0.3	0.948	0.947	0.935
	0.4	0.941	0.945	0.948
	0.6	0.939	0.952	0.947
	0.7	0.939	0.944	0.949
	0.8	0.941	0.931	0.908
75	0.3	0.951	0.953	0.956
	0.4	0.948	0.947	0.951
	0.6	0.950	0.945	0.944
	0.7	0.936	0.931	0.931
	0.8	0.937	0.930	0.912

as $0 < \rho < 1$. Differentiating (4) with respect to n, equating to zero and solving for n we obtain

$$n^* = 1 + \sqrt{\frac{(1-\rho)}{2\rho\theta^2}} \quad 0 < \rho < 1. \qquad (5)$$

The required number of subject is thus $k^* = N/n^*$.

Table 4 shows few optimal allocations of (n, k) for $\rho = 0.6$, 0.7, and 0.8, $\theta = 0.1$, 0.2, 0.3 and 0.4, when $N = 24$.

Note that in practice, only integer values of (n, k) are used, and because $N = nk$ is fixed *a priori*, we first round the optimum values of n to the nearest integer; then $k = N/n$ was rounded to the nearest integer. The values of var($\hat{\theta}$) at the rounded optimal allocations for different values of ρ, θ and n showed that the net loss or gain in efficiency due to rounding is negligible. It is clear that to efficiently estimate the WSCV for large values of θ we need smaller number of replicates and larger number of subjects.

Fixed width confidence interval approach

Bonett [10] discussed the issue of sample size requirements that achieve a pre-specified expected width for a confidence interval about ICC. This approach is useful in planning a reliability study in which the focus is on estimation rather than hypothesis testing. He demonstrated that the effect of inaccurate planning value of ICC is more serious in hypothesis testing applications. Shoukri et al. [11] argued that the hypothesis testing approach might not be appropriate while planning a reproducibility study. This is because, in most cases, values of the coefficient under the null and alternative hypotheses may be difficult to specify. An alternative approach is to focus on the width of the CI for θ. Since the approximate width of an $(1 - \alpha)100\%$CI on θ is, $2z_{\alpha/2}\,\mathrm{var}(\hat{\theta})^{1/2}$, an approximate sample size that yields an $(1 - \alpha)100\%$ CI for θ with a desired width w obtains by setting $w = 2z_{\alpha/2}\{\mathrm{var}(\theta)\}^{1/2}$ and then solving for k:

$$k = \frac{4z^2 A(\rho, n, \theta)}{w^2}.$$

We observe that, for fixed n and θ, larger values of ρ require larger number of subjects to satisfy the criterion. As an example, suppose that it is of interest to construct 95% CI on θ with expected width $w = 0.05$, $\rho = 0.3$, and an afforded number of replicates $n = 2$. If the hypothesized value of θ is 0.10, then $k = 31$, and if θ is 0.3 (*i.e.* lower reliability), then $k = 323$.

Table 4: Results for 10 replicates on each of three patient's total lesion burden. Values are given volumes in cubic centimeters.

		Replicates									
Patient	Method*	1	2	3	4	5	6	7	8	9	10
1	M	20	21.2	20.8	20.6	20.2	19.1	21	20.4	19.2	19.2
	G	19.5	19.5	19.6	19.7	19.3	19.1	19.1	19.3	19.2	19.5
2	M	26.8	26.5	22.5	23.1	24.3	24.1	26	26.8	24.9	27.7
	G	22.1	21.9	22	22.1	21.9	21.8	21.7	21.7	21.7	21.8
3	M	9.6	10.5	10.6	9.2	10.4	10.4	10.1	8	10.1	8.9
	G	8.5	8.5	8.3	8.3	8.3	8	8	8	8	8.1

*M = Manual, G = Geometrically constrained region growth.

Table 5: Summary analysis of data in Table 2 and 95% confidence intervals

Method	$\hat{\rho}$	$\hat{\theta}$	95% CI without VST	95% CI with VST
M	0.966	6.5%	(0.034, 0.096)	(0.043, 0.118)
G	0.999	1.2%	(0.006, 0.017)	(0.008, 0.021)

Cost criterion

Funding constraints will often determine the cost of recruiting subjects for a reliability study. Although too small a sample may lead to a study that produces an imprecise estimate of the reproducibility coefficient, too large a sample may result in a waste of resources. Thus, an important decision in a typical reliability study is to balance the cost of recruiting subjects with the need for a precise estimate of the parameter summarizing reliability.

In this section, we determine the combinations (n, k) that minimize the variance of $\hat{\theta}$ subject to cost constraints. Constructing a flexible cost function starts with identifying sampling and overhead costs. The sampling cost depends primarily on the size of the sample and includes costs for data collection, compensation to volunteers, management, and evaluation. On the other hand, overhead costs are independent of sample size. Following Sukhatme et al. [14], we assume that the overall cost function is given as:

$$C = c_0 + kc_1 + nkc_2 \quad (6)$$

where, c_0 is the fixed cost, c_1 the cost of recruiting a single subject, and $c_2$is the cost of making one observation. Using the method of Lagrange multipliers and following Shoukri et al. [11], we write the objective function Ψ in this form

$$\Psi = \text{var}(\hat{\theta}) + \lambda(C - c_0 - kc_1 - nkc_2) \quad (7)$$

where, var($\hat{\theta}$) is given by Equation (3) and λ is the Lagrange multiplier. Differentiating $\Psi$with respect to n, k and λ and equating to zero, we obtain

$$2\theta^2\rho^* n^4 - 4\theta^2\rho^* n^3 - (2\theta^2 r + r - 2\theta^2\rho^* + 1)n^2 + 4\theta^2 rn - 2\theta^2 r = 0 \quad (8)$$

where $r = c_1/c_2$, and $\rho^* = \rho/(1 - \rho)$

Although an explicit solution to (8) is available, the resulting expression is complicated and does not provide any useful insight. The 4^{th} degree polynomial in the left side of (8) has two imaginary roots, one negative and one admissible (positive) root for n. Table 5 summarize the results of the optimization procedure where we provide the optimal n for various values of θ, ρ, and r, noting that:

$$k_{opt} = \frac{(C - c_0)/c_2}{r + n_{opt}}. \quad (9)$$

Table 6: Optimal combinations of (n_{opt}, k_{opt}) which minimize the variance of $\hat{\theta}$ for N = 24.

θ	ρ		
	0.6	0.7	0.8
	(n_{opt}, k_{opt})		
0.1	(6.77, 3.54)	(5.63, 4.26)	(4.54, 5.29)
0.2	(3.89, 6.17)	(3.31, 7.24)	(2.77, 8.67)
0.3	(2.91, 8.23)	(2.54, 9.47)	(2.17, 11.05)
0.4	(2.44, 9.82)	(2.16, 11.12)	(1.88, 12.74)

Results

From Table 7, it is apparent that when $r = c_1/c_2$ increases, the required number of replicates per subject (n) increases, because the cost of making a single observation (c_2) decreases and the cost of recruiting a subject (c_1) increases. When r is fixed, an increase in ρ results in a decline in the required value of n and accordingly an increase in k. An increase in θ also results in a decrease in n. The general conclusion is that it is sensible to decrease the number of items associated with a higher cost, while increasing those with lower cost.

We note that by setting $c_1 = 0$ in Equation (8), we obtain $n_{opt} = 1 + \sqrt{(1-\rho)/2\rho\theta^2}$, as in Equation (5). The situation $c_1 = 0$ is quite plausible, at least approximately if the major cost is in actually making the observations (e.g. expensive equipment, cost of interviews versus free volunteer subjects). This means that a special cost structure is implied by the optimal allocation procedure discussed earlier.

Example 2

To assess the accuracy of Doppler Echocardiography (DE) in determining aortic valve area (AVA) prospective evaluation on patients with aortic stenosis, an investigator wishes to demonstrate a high degree of reliability (ρ = 0.80) in estimating AVA using the "velocity integral method" with a planned value for the WSCV = 0.10. Suppose that the total cost of making the study is fixed at $1600.0. It is assumed that the overhead fixed cost c_0 is

Table 7: Optimal replications (rounded to the nearest integer) of *n* that minimize the variance of $\hat{\theta}$ subject to cost constraints.

			r						
			0.1	**0.5**	**1**	**1.5**	**5**	**10**	**20**
			n_{opt}						
		0.6	62	72	83	92	142	193	266
	0.01	**0.7**	50	58	66	74	114	155	213
		0.8	38	44	52	57	88	118	163
		0.6	16	19	21	24	36	49	67
	0.04	**0.7**	13	15	17	19	29	39	54
		0.8	10	12	14	15	23	30	42
		0.6	7	8	9	10	15	20	28
	0.10	**0.7**	6	7	8	8	12	17	22
		0.8	5	5	6	7	10	13	17
θ	ρ								
		0.6	5	6	7	7	11	14	19
	0.15	**0.7**	4	5	5	6	9	11	15
		0.8	3	4	4	5	7	9	12
		0.6	3	3	4	4	6	8	10
	0.30	**0.7**	3	3	3	4	5	6	9
		0.8	2	2	3	3	4	5	8
		0.6	3	3	3	3	5	6	8
	0.40	**0.7**	2	2	3	3	4	5	7
		0.8	2	2	2	2	3	4	5

absorbed by the hospital. Moreover, we assume that travel cost is $200.0, and the administrative cost using the DE is $200.0 per visit. From Table 5, n_{opt} for $r = 1$, $\rho = 0.8$, and $\theta = 0.10$ is 6. From (9), $k_{opt} = (1600/15)/(1 + 6) = 15$. That is we need 15 patients, with 6 measurements each to minimize $\text{var}(\hat{\theta})$ subject to the given cost.

Estimating the WSCV for dichotomous responses

Assumptions

Consider a random sample of k subjects, each is blindly evaluated n times by the same rater. We assume that all subject responses y_{ij} (where $j = 1, 2, \ldots n$) are dichotomous and are conditionally independent with probabilities $P(y_{ij} = 1) = p_i$ $(i = 1,2,\ldots,k)$ and $p(y_{ij} = 0) = 1 - p_i$. Thus, for fixed p_i, the conditional distribution of the random variable $y_{i\bullet} = \sum_j y_{ij}$ follows binomial distribution with parameters n and p_i. To account for the variation of response probabilities between subjects, as considered by Mak [15], we assume further that the probabilities p_i are independently and identically distributed as a beta distribution, Beta (α,β), with mean $\pi = \alpha/(\alpha + \beta)$ and variance $\pi(1 - \pi)\rho$. Given these assumptions, one can show that the correlation between y_{ij} and y_{il} is in fact ρ. Define $\bar{y}_{i\bullet} = y_{i\bullet}/n$ and $s_i^2 = \sum_j \left(y_{ij} - \bar{y}_{i\bullet}\right)^2/(n - 1)$, $\hat{\mu} = \frac{1}{k}\sum_i \bar{y}_{i\bullet}$, and $s^2 = \frac{1}{k}\sum_i s_i^2$. We therefore estimate the WSCV for binary assessments by:

$$\hat{v} = s/\bar{y}.$$

A case of special interest to clinical epidemiologists is when $n = 2$, or a test re-test reliability study involving two readings per subject. For this case we investigate the sample size issue in the following section.

Results

The special case n = 2

Under the above set-up, the common correlation model (CCM) (see Mak [15], Bloch and Kraemer [16]) provides an appropriate description for the joint distribution of (Y_{ij}, Y_{il}):

$$P_{11} = P(Y_{ij} = 1, Y_{il} = 1) = \pi^2 + \rho\pi(1 - \pi).$$

$$P_{10} = P_{01} = P(Y_{ij} = 1, Y_{il} = 0) = P(Y_{ij} = 0, Y_{il} = 1) = (1 - \rho)\pi(1 - \pi). \quad (10)$$

$$P_{00} = P(Y_{ij} = 0, Y_{il} = 0) = (1 - \pi)^2 + \rho\pi(1 - \pi).$$

The data layout can be summarized as in Table 8.

Table 8: Data layout for a 2 × 2 binary classification

		1st measurement (Y_1)	
		1	0
2nd Measurement Y_2	1	k_{11}	k_{01}
	0	k_{10}	k_{00}

where, $k = k_{11} + k_{01} + k_{10} + k_{11}$.

Since for the i^{th} subject, the mean of two measurements is $\mu_i = \frac{1}{2}(Y_{i1} + Y_{i2})$ when summed up over all subjects we have:

$$\sum_{i=1}^{k} \mu_i = \frac{1}{2}[(k_{11} + k_{10}) + (k_{11} + k_{01})]$$

Therefore, an unbiased estimator of the population mean π is:

$$\hat{\mu} = \frac{1}{2k}[k_{10} + k_{01} + 2k_{11}],$$

and the MSW, is $S^2 = \frac{1}{2k}(k_{10} + k_{01})$. Hence the sample coefficient of variation for binary responses is:

$$\hat{\nu} = \frac{S}{\hat{\mu}} = \frac{\sqrt{2k(k_{10} + k_{01})}}{k_{10} + k_{01} + 2k_{11}}. \qquad (11)$$

Since $E(\hat{\mu}) = \pi$ and $E(SD) = \pi(1 - \pi)(1 - \rho)$ we define $\nu = \sqrt{\pi^{-1}(1-\pi)(1-\rho)}$ as the population coefficient of variation for dichotomous outcome with its MLE given by $\hat{\nu}$. The CCM can be re-parameterized by substituting $\left(1 - \frac{\nu^2}{1-\pi}\pi\right)$ for the reliability coefficient ρ. Applying the delta method, the first order approximation to the variance of $\hat{\nu}$ is shown to be:

$$\text{var}(\hat{\nu}) = k^{-1}(a_1 + a_2 + a_3),$$

Table 9: Data analysis from a mammography study by Powell et al. (1999).

Rater	k_{00}	$k_{10} + k_{01}$	k_{11}	$\hat{\rho}$	$\hat{\nu}$	$SE(\hat{\nu})$	95% C.I.
DI	9	5	44	0.73	26%	0.061	(0.14,0.38)
FS	9	7	42	0.65	31%	0.064	(0.19,0.43)

Table 10: Data Layout for the CCM

Category	Ratings	Frequency	Probability
1	(1,1)	n_1	$P_1(c) = \pi(1 - \upsilon^2\pi)$
2	(1,0) or (0,1)	n_2	$P_2(c) = 2\pi^2\upsilon^2$
3	(0,0)	n_3	$P_3(c) = 1 - \pi - \pi^2\upsilon^2$
Total		k	1

where $a_1 = \upsilon^2(1 - \upsilon^2\pi)(1 - \pi + \upsilon^2\pi^2)/\pi$,

$a_2 = (1 - 2\pi\upsilon^2)^2(1 - 2\pi^2\upsilon^2)/8\pi^2$, and

$a_3 = \upsilon^2(1 - 2\upsilon^2\pi)(1 - \upsilon^2\pi)$. (12)

We suggest an approximate $(1 - \alpha)100\%$ confidence interval as $\hat{\nu} \pm z_{\alpha/2}\sqrt{\text{var}(\hat{\nu})}$.

Example 3

To illustrate the methodology discussed in this section, we use data from an investigation of mammography by Powell et al. [17] concerning the equivalence of film-screen (FS) and digital images (DI). Two readings were made on the presence/absence (1/0) of malignancy by each rater on the same set of $k = 58$ patients. The data and the results of the analysis are summarized in Table 9. Both methods seem to have the same levels of reliability in terms of ICC and WSCV. We note that the 95% confidence interval is somewhat relatively wide, and this may be due to the fact that the sample size is not large enough.

Note that if the observed frequencies in the sample of k subjects are given as in Table 10, we can write a simpler estimator of the WSCV as $\hat{\nu} = \sqrt{2kn_2}/(n_2 + 2n_1)$. To construct an estimate of the confidence interval on υ, the MLE of ρ, $\hat{\rho} = 1 - \frac{n_2}{2k\hat{\pi}(1-\hat{\pi})}$ and $\hat{\nu}$ should be substituted in equation (12) where, from Donner and Eliasziw [18] $\hat{\pi} = \frac{n_1 + 2n_2}{2k}$.

Sample size estimation

Methods

There has been increasing attention given recently to estimation of sample size using a confidence interval rather than a significance testing approach (e.g. Gardner and Altman [19]). This is consistent with recent arguments made by many authors, including Goodman and Berlin [20] who state that "confidence intervals should play an important role when setting sample size" and that " the size of a confidence interval can be predicted in the planning stages of an experiment and this can be a great help

in understanding the implications of different sample size choices".

For comparative interest we also present sample size requirements needed to test $H_0 : \upsilon = \upsilon_0$, where υ_0 is some hypothesized value of the WSCV υ.

Fixed width confidence interval (CI) on υ

Following the approach described in Section 3.2, the approximate width of an $(1 - \alpha)100\%$ CI on υ is, $2z_{\alpha/2}\{\text{var}(\hat{v})\}^{1/2}$. An approximate sample size that yields an $(1 - \alpha)100\%$ CI for υ with a desired width w obtains by setting $w = 2z_{\alpha/2}\{\text{var}(\hat{v})\}^{1/2}$ and then solving for k:

$$k = 4\, z^2_{\alpha/2}\,(a_1 + a_2 + a_3)/w^2. \quad (13)$$

Hypothesis testing procedure

Donner and Eliasziw (DE) [18] developed the Goodness-of-fit (GOF) to efficiently construct a confidence interval and to estimate the sample size required to test a specific hypothesis on intra-class kappa value. Here we use the GOF to estimate the sample size needed for ensuring enrollment of a sufficient number of subjects in a reproducibility study. This follows from the observation that, to test the null hypothesis, $H_0 : \upsilon = \upsilon_0$ then:

$$\chi^2_G = \sum_{l=1}^{3} \frac{[n_l - kP_l(v_0)]^2}{kP_l(v_0)} \quad (14)$$

has a non-central chi-square distribution with one degree of freedom under the alternative hypothesis $H_1 : \upsilon = \upsilon_1$ with non-centrality parameter

$$\eta = k\sum_{l=1}^{3} \frac{[P_l(v_1) - P_l(v_0)]^2}{P_l(v_0)}. \quad (15)$$

Following DE it can be shown that the sample size needed to conduct a two-sided test with significance level α and power 1 - β is:

$$k = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\pi^2(v_1^2 - v_0^2)^2}[(1 - v_0^2\pi)^{-1} + 2v_0^{-2} + \pi^2(1 - \pi - \pi^2 v_0^2)^{-1}]^{-1} \quad (16)$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the critical values of the standard normal distribution corresponding to α and β.

As an example suppose it is of interest to test H_0: $\upsilon = 0.04$ versus H_1: $\upsilon = 0.1$, where υ_0 corresponds to high reliability. To ensure with 80 per sent probability a significant result at $\alpha = 5\%$ and $\pi = 0.30$ when $\upsilon_1 = 0.10$, we compute the required number of subjects from the above equation as $k = 986$ and when $\pi = 0.50$, $k = 355$. For the sake of comparison to the fixed width CI procedure, suppose it is of interest to construct 95% CI on υ with expected width $w = 0.10$. If the hypothesized values of υ is 0.10 and $\pi = 0.30$, then from (13) $k = 1100$, and if $\pi = 0.5$, then $k = 400$.

Discussion

The ICC has been traditionally used to assess the reliability of a measurement. QS considered the WSCV as an alternative measure of reproducibility for continuous scale measurements. It should be emphasized that our investigation has not allowed for forms of systematic error (e.g. measurement, or trend that is unaccounted for in the model). A reviewer of this paper indicated that this is beyond our scope. In this paper we have dealt with the issue of sample size estimation of the WSCV from continuous and binary scale measurements focusing on random measurement error, in the conventional way that reliability is usually discussed.

As in any reliability study, a crucial decision that a researcher faces in the design stage is the determination of the number of subjects, k and the number of measurements per subject, n. We have discussed two alternative statistical techniques to determine an optimal allocation. When we have prior knowledge of what constitutes an acceptable level of reproducibility, a hypothesis testing approach may be used. We used this approach in the case of binary outcome variable, following the GOF approach proposed by DE. The application of the GOF was straightforward because the number of replicates n = 2 was fixed. However, there are situations, when appropriate values of the reliability coefficient under the null and alternative hypotheses may be difficult to specify. An alternative to hypotheses testing is the efficient allocation of the sample, and the guidelines provided in this article for the continuous scale measurements allow selection of the pair (n, k) that maximizes the precision of the estimated coefficient under cost constrains. We note that cost implications, for dichotomous assessments, are quite important particularly when n is larger than two, which we intend to report on in a future paper.

Finally it is noted that in practice, the optimal allocation must be integer values, and the net loss/gain in precision as a result of rounding the values the values of (n, k) was negligible. Ideally one should adopt one of the available optimization algorithms, often referred to as integer programming models. These models are suited for the optimal allocations problems since the main concern was to find the best solution(s) in a well-defined discrete space.

Conclusion

The WSCV is a useful index measure of measurements reliability. Investigators may design reliability studies using either efficiency or cost considerations. For continuous

measurements, optimal allocation of the sample may be achieved with as few as two replications per subject. For dichotomous data, when each subject is measured twice, investigators may use, either fixed length confidence interval, or power considerations is estimating the sample size. Both methods produce comparable results.

Competing interests

The author(s) declare that they have no competing interests.

The authors contributed equally to this work

Acknowledgements

Drs. M. Shoukri and N. ElKum acknowledge the support by the Research Centre of The King Faisal Specialist Hospital. Dr. Walter acknowledges the support by NSERC Canada.

References

1. Ashton E, Takahashi C, Berg M, Goodman A, Totterman S, Ekholm S: **Accuracy and reproducibility of manual and semi-automated quantification of MS lesions in MRI.** Technical Report, Department of Radiology, University of Rochester Medical Center, Rochester, NY; 2003.
2. Schwartz L, Ginsberg M, DeCorato D: **Evaluation of Tumor Measurements in oncology: Use of film-based and electronic techniques.** *Journal of Clinical Oncology* 2000, **18(10):**2179-2184.
3. Tian L: **Inference on the common coefficient of variation.** *Statistics in Medicine* 2005, **24:**2213-2220.
4. Fleiss J: **Design and analysis of clinical experiments.** Wiley & Sons, New York; 1986.
5. Dunn G: **Design and analysis of reliability Studies.** Oxford University Press, New York; 1989.
6. Shoukri MM: **Measures of inter-observer agreement.** Chapman & Hall/CRC Press. Boca Raton, Florida; 2004.
7. Quan H, Shih WJ: **Assessing reproducibility by the within-subject coefficient of variation with random effects models.** *Biometrics* 1996, **52:**1195-1203.
8. Donner A, Eliasziw M: **Sample size requirements for reliability studies.** *Statistics in Medicine* 1987, **6:**441-448.
9. Walter D, Eliasziw M, Donner A: **Sample size and optimal design for reliability studies.** *Statistics in Medicine* 1998, **17:**101-110.
10. Bonett DG: **Sample size requirements for estimating intraclass correlations with desired precision.** *Statistics in Medicine* 2002, **21:**1331-1335.
11. Shoukri M, Asyali M, Walter S: **Issues of cost and efficiency in the design of reliability studies.** *Biometrics* 2003, **59:**1107-1112.
12. Flynn N, Whitely E, Peters T: **Recruitment strategy in a cluster randomized trial: Cost implications.** *Statistics in Medicine* 2002, **21:**397-405.
13. Kendall M, Stuart A: **The advanced theory of statistics.** *Volume I.* London: Griffin; 1986.
14. Sukhatme P, Sukhatme B, Sukhatme S, Asok C: **Sampling theory of surveys with applications.** Ames, IA: Iowa State University Press; 1984.
15. Mak TK: **Analyzing the intra-class correlation for dichotomous variables.** *Applied Statistics* 1988, **37:**344-352.
16. Bloch DA, Kraemer HC: **2 × 2 kappa coefficients: measures of agreement or association.** *Biometrics* 1989, **45:**269-287.
17. Powell KA, Obouchowski NA, Chilote WA, Barry MM, Ganocik SN, Cardenso G: **File-screen versus digitized mammography: assessment of clinical equivalence.** *American Journal of Roentgenology* 1999, **173:**889-894.
18. Donner A, Eliasziw M: **A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance testing and sample size estimation.** *Statistics in Medicine* 1992, **11:**1551-1519.
19. Gardner M, Altman D: **Confidence intervals rather than P-values: estimation rather than hypothesis testing.** *British Medical Journal* 1986, **292:**746-750.
20. Goodman S, Berlin J: **The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results.** *Annals of Internal Medicine* 1994, **121:**200-206.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/6/24/prepub