Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 1199–1212
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Confidence intervals on intraclass correlation coefficients in a balanced two-factor random design

Kye Gilder[a,*], Naitee Ting[b], Lili Tian[c], Joseph C. Cappelleri[d], R. Choudary Hanumara[e]

[a]*Biogen Idec Inc., San Diego, CA 92122, USA*
[b]*Pfizer Inc., Global Research & Development, New London, CT 06320, USA*
[c]*Department of Biostatistics, School of Public Health and Health Professions, Buffalo, NY 14214-3000, USA*
[d]*Pfizer Inc., Global Research & Development, Groton, CT 06340, USA*
[e]*Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881, USA*

Received 19 November 2003; accepted 25 March 2006
Available online 11 May 2006

Abstract

A modified large-sample (MLS) approach and a generalized confidence interval (GCI) approach are proposed for constructing confidence intervals for intraclass correlation coefficients. Two particular intraclass correlation coefficients are considered in a reliability study. Both subjects and raters are assumed to be random effects in a balanced two-factor design, which includes subject-by-rater interaction. Computer simulation is used to compare the coverage probabilities of the proposed MLS approach (GiTTCH) and GCI approaches with the Leiva and Graybill [1986. Confidence intervals for variance components in the balanced two-way model with interaction. Comm. Statist. Simulation Comput. 15, 301–322] method. The competing approaches are illustrated with data from a gauge repeatability and reproducibility study. The GiTTCH method maintains at least the stated confidence level for interrater reliability. For intrarater reliability, the coverage is accurate in several circumstances but can be liberal in some circumstances. The GCI approach provides reasonable coverage for lower confidence bounds on interrater reliability, but its corresponding upper bounds are too liberal. Regarding intrarater reliability, the GCI approach is not recommended because the lower bound coverage is liberal. Comparing the overall performance of the three methods across a wide array of scenarios, the proposed modified large-sample approach (GiTTCH) provides the most accurate coverage for both interrater and intrarater reliability.

MSC: 62F25; 62J10; 62K99; 62P10

Keywords: Modified large-sample approach; Generalized confidence interval; Reliability study; Interrater reliability; Intrarater reliability; Variance components

1. Introduction

Since the introduction of the intraclass correlation coefficient by R. A. Fisher, its use as a measure of reliability has received much attention (Bartko, 1966). Its usefulness and application in the social, behavioral, and medical sciences has been clearly demonstrated (Bartko, 1966; Lin et al., 2002; Fleiss, 1986). Researchers have become increasingly aware of the observer or rater as a source of measurement error. Because measurement error can impair the statistical

* Corresponding author. Tel.: +1 858 4015676; fax: +1 858 4018003.
E-mail address: Kye.Gilder@biogenidec.com (K. Gilder).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2006.03.002

analysis and its interpretation, it is important to quantify the amount of measurement error by use of some form of a reliability index such as a kappa statistic, concordance correlation coefficient, or an intraclass correlation coefficient (Lin et al., 2002; Fleiss, 1986; Fleiss and Cohen, 1973; Landis and Koch, 1977; Fleiss and Shrout, 1979; Armitage et al., 1994).

The intraclass correlation coefficient is based upon standard analysis of variance (ANOVA) models and the estimation of variance components. Although assumptions of normality for these models may not be warranted in certain cases, the ANOVA procedure is generally robust and permits the estimation of the appropriate variance components and the intraclass correlation coefficients (Landis and Koch, 1977; Searle, 1971). Typically, intraclass correlation coefficients are the ratio of variance components of interest to total variance. There are numerous forms of intraclass correlation, each appropriate for different models and objectives (Fleiss and Cohen, 1973; Müller and Büttner, 1994; St. Laurent, 1998). The appropriate version is dictated by the specific situation defined by the experimental design and conceptual intent of the reliability study (Fleiss and Shrout, 1979). This paper concentrates on two particular intraclass correlation coefficients that measure interrater reliability and intrarater reliability.

For the balanced two-factor random design with only one observation per subject–rater cell, in which the interaction term cannot be separated from the error term, confidence intervals for interrater reliability have been presented by several authors including Fleiss and Shrout (1978), Zou and McDermott (1999), Cappelleri and Ting (2003), Arteaga et al. (1982), Burdick and Graybill (1988, 1992) and Tian and Cappelleri (2004). For the balanced two-factor random design with multiple replicates per subject–rater combination (cell), which affords assessment of subject-by-rater interaction, confidence intervals for interrater reliability have been presented using a modified large-sample (MLS) approach by Leiva and Graybill (1986) and Burdick and Graybill (1992). For general discussion of intraclass correlation coefficients, Adamec and Burdick (2003) propose a Satterthwaite approach and Hamada and Weerahandi (2000) propose a generalized confidence interval (GCI) approach. These investigations, however, do not provide confidence intervals for intrarater reliability.

This paper, therefore, focuses on applying a type of MLS approach to obtain approximate one-sided and two-sided confidence intervals for two particular forms of intraclass correlation: interrater reliability and intrarater reliability. Both reliabilities are derived from a study in which we assess the balanced two-factor random effects design having multiple replicates per cell. In addition, the proposed MLS and GCI approach are compared to the Leiva and Graybill (LG) method based on lower bound coverage, upper bound coverage, two-sided confidence interval coverage, and confidence interval width.

Section 2 describes the two intraclass correlation coefficients. Section 3 explicates the derivation of the confidence bounds for the intraclass correlation coefficients using the proposed modified large-sample approach. Section 4 introduces the proposed generalized variables for interval estimation for the particular intraclass correlation coefficients. Section 5 covers the existing Leiva and Graybill method. Section 6 explains the computer simulation study, and Section 7 contains the simulation results. Section 8 illustrates the competing procedures with an example. Section 9 provides concluding remarks.

2. Interrater and intrarater reliability

We consider a reliability study in which each of the I subjects is independently measured by each of the J raters a total of K times ($K > 1$). In addition, it is assumed that both subjects and raters are randomly selected from their respective populations.

The appropriate statistical model is the two-factor crossed random effect with interaction term. The kth measurement on the ith subject by the jth rater is represented as

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk},$$

$$i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K, \quad K > 1, \tag{1}$$

where μ is the overall mean, A_i is the effect of the ith subject, B_j is the effect of the jth rater, $(AB)_{ij}$ is the effect of the interaction between the ith subject and the jth rater, and ε_{ijk} is the unexplained error. The terms A_i, B_j, $(AB)_{ij}$, and ε_{ijk} are assumed to be jointly independent normal random variables each with a mean of zero and variances given by σ_A^2, σ_B^2, σ_{AB}^2, and σ_ε^2, respectively. The expected mean squares for model (1) is shown in Table 1.