Journal of Statistical Planning and Inference 139 (2009) 396 – 410

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Inferences on the intraclass correlation coefficients in the unbalanced two-way random effects model with interaction☆

Ren-Dao Ye[a,b,*], Song-Gui Wang[b]

[a] School of Finance, Hangzhou Dianzi University, Zhejiang 310018, China
[b] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Article history:
Received 3 October 2007
Received in revised form
8 May 2008
Accepted 9 May 2008
Available online 18 May 2008

MSC:
62F25
62J10
62K99
62P10

Keywords:
Generalized *p*-value
Generalized confidence interval
Modified large-sample approach
Interrater reliability
Intrarater reliability
Coverage probability

ABSTRACT

In this paper, the hypothesis testing and interval estimation for the intraclass correlation coefficients are considered in a two-way random effects model with interaction. Two particular intraclass correlation coefficients are described in a reliability study. The tests and confidence intervals for the intraclass correlation coefficients are developed when the data are unbalanced. One approach is based on the generalized *p*-value and generalized confidence interval, the other is based on the modified large-sample idea. These two approaches simplify to the ones in Gilder et al. [2007. Confidence intervals on intraclass correlation coefficients in a balanced two-factor random design. J. Statist. Plann. Inference 137, 1199–1212] when the data are balanced. Furthermore, some statistical properties of the generalized confidence intervals are investigated. Finally, some simulation results to compare the performance of the modified large-sample approach with that of the generalized approach are reported. The simulation results indicate that the modified large-sample approach performs better than the generalized approach in the coverage probability and expected length of the confidence interval.

1. Introduction

The assessment of reliability of measurement is of great importance in many areas, such as social, behavioral, and medical sciences (Bartko, 1966; Fleiss, 1986; Lin et al., 2002). Unreliable or imprecise measurement may have serious undesirable consequence (Fleiss, 1986). Researchers have become increasingly aware of the observer or rater as a source of measurement error. However, the measurement error can impair the statistical analysis and its interpretation. Hence it is necessary to quantify the amount of measurement error by reliability coefficients such as intraclass correlation coefficient, concordance correlation coefficient, and kappa statistic (Fleiss and Cohen, 1973; Landis and Koch, 1977; Fleiss and Shrout, 1979; Fleiss, 1986; Armitage et al., 1994; Lin et al., 2002; Gilder et al., 2007). In this paper, we focus on two particular intraclass correlation coefficients that measure interrater reliability and intrarater reliability.

Specifically, we consider a reliability study in which each of the a subjects is independently measured by each of the b raters with an unequal number of replicates. It is also assumed that both subjects and raters are randomly selected from their

☆ This work was supported by the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality.

* Corresponding author at: School of Finance, Hangzhou Dianzi University, Zhejiang 310018, China.
E-mail address: yerendao2003@163.com (R.-D. Ye).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.05.004

populations. The corresponding model is an unbalanced two-way random effects model with interaction. The kth measurement on the ith subject by the jth rater may be represented as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, a, \ j = 1, \ldots, b, \ k = 1, \ldots, n_{ij} \geqslant 1, \tag{1.1}$$

where μ is the overall mean, α_i is the random effect due to the ith subject with $\alpha_i \sim N(0, \sigma_\alpha^2)$, β_j is the random effect due to the jth rater with $\beta_j \sim N(0, \sigma_\beta^2)$, γ_{ij} is the effect of the interaction between the ith subject and the jth rater with $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$, and ε_{ijk} is the random error with $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. All the random variables are mutually independent. The parameter space of $(\mu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma_\varepsilon^2)$ is

$$\Omega = \{(\mu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, \sigma_\varepsilon^2) : -\infty < \mu < \infty, \sigma_\alpha^2 \geqslant 0, \sigma_\beta^2 \geqslant 0, \sigma_\gamma^2 \geqslant 0, \sigma_\varepsilon^2 > 0\}.$$

The variance of a single measurement Y_{ijk} is $\mathrm{Var}(Y_{ijk}) = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2$, and the covariance between two measurements Y_{ijk} and $Y_{ij'k'}$ is $\mathrm{Cov}(Y_{ijk}, Y_{ij'k'}) = \sigma_\alpha^2$. Then the intraclass correlation coefficient to measure interrater reliability is

$$\rho_{\mathrm{inter}} = \mathrm{Corr}(Y_{ijk}, Y_{ij'k'}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2}, \tag{1.2}$$

which is defined as the proportion of total variability in observed measurements accounted for by the subject-to-subject variability. It can also be interpreted as the correlation between two randomly selected measurements on a single subject by two different randomly selected raters (Damon and Harvey, 1987; Armitage et al., 1994; Sahai and Ageel, 2000; Gilder et al., 2007).

Similarly, the intraclass correlation coefficient to measure intrarater reliability is

$$\rho_{\mathrm{intra}} = \mathrm{Corr}(Y_{ijk}, Y_{ijk'}) = \frac{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_\varepsilon^2}, \tag{1.3}$$

which is defined as the correlation between two randomly selected measurements on the same subject for a randomly selected rater (Damon and Harvey, 1987; Sahai and Ageel, 2000; Gilder et al., 2007).

In the literature, several general methods for the inferences on the intraclass correlation coefficients have worked with balanced model, which are based on the fact that the traditional decomposition of the total sums of squares yields independent and chi-squared-type sums of squares (Arteaga et al., 1982; Burdick and Graybill, 1988, 1992; Zou and McDermott, 1999; Cappelleri and Ting, 2003; Tian and Cappelleri, 2004; Gilder et al., 2007). However, collecting balanced data in the context of the present problem seems to be unrealistic because of the raters burden, and unavailability of the raters during the sampling period. Moreover, measured data sets are typically small, and hence methods that are applicable for small unbalanced data sets are really warranted.

The main purpose of our paper is to provide two approaches that are readily applicable for the inferences on ρ_{inter} and ρ_{intra} in model (1.1). One approach is based on the concepts of generalized p-value and generalized confidence interval, introduced by Tsui and Weerahandi (1989) and Weerahandi (1993), respectively, the other is based on the modified large-sample idea. The major appeal of these approaches is applicable to small samples. These ideas have turned out to be very satisfactory for obtaining tests and confidence intervals for many complex problems (Zhou and Mathew, 1994; Weerahandi, 1991, 1995, 2004; Roy and Mathew, 2005; Mathew and Webb, 2005; Ho and Weerahandi, 2007; Burdick and Graybill, 1992; Gui et al., 1995; Cappelleri and Ting, 2003; Gilder et al., 2007).

This article is organized as follows. In Section 2, we briefly introduce the concepts of generalized p-value and generalized confidence interval. In Section 3, we outline the generalized approach to construct hypothesis testing and interval estimation for the intraclass correlation coefficients. In Section 4, we outline the modified large-sample approach to construct confidence bounds for the intraclass correlation coefficients. In Section 5, we describe the method of computing coverage probabilities and expected lengths of the generalized confidence interval. In Section 6, we present simulation studies on the coverage probability and expected length of the proposed approaches in different situations.

2. The generalized *p*-value and generalized confidence interval

The setup where generalized p-value and generalized confidence interval are defined is as follows. Let X be a random variable whose distribution depends on a scalar parameter of interest δ and a nuisance parameter ξ, where ξ could be a vector. Suppose we want to test

$$H_0 : \delta \leqslant \delta_0 \quad \text{vs} \quad H_1 : \delta > \delta_0, \tag{2.1}$$