Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and Inference 129 (2005) 229–237

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Selecting the instrument closest to a gold standard

Pankaj K. Choudhary[a,*], H.N. Nagaraja[b]

[a] *Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75083-0688, USA*
[b] *Department of Statistics, Ohio State University, Columbus, OH 43210-1247, USA*

Available online 21 August 2004

Abstract

We consider the comparison of two instruments with a gold standard with the goal of finding the best one—the one that agrees most with the gold standard. Using natural log of the mean squared deviation as the measure of agreement, we present a large sample two-stage procedure with good small sample properties. When the differences of the paired measurements are bivariate normal, a first-stage sample of size 15 is adequate for its application. We illustrate the procedure using a dataset from the literature.

Keywords: Agreement; Bivariate normal distribution; Concordance; Multiple comparisons with the best; Selection of the best; Two-stage procedures

1. Introduction

We consider the problem of comparing two instruments (or methods of measurement) with a gold standard in method comparison studies. The goal is to find the instrument that agrees most with the gold standard and we refer to it as the best one. We assume that one measurement per subject is available from every instrument, the measurements are continuous, and have the same unit.

Let the triplet (G, X_1, X_2) denote the measurements on a typical subject by the gold standard, the first and the second instrument, respectively, and D_i be the difference $G - X_i$ $(i = 1, 2)$. We assume that $\mathbf{D} = (D_1, D_2)$ follows a bivariate normal (BVN) distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$, standard deviation $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ and correlation ρ, where $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \rho) \in$

* Corresponding author. Tel.: +972-883-4436.
E-mail addresses: pankaj@utdallas.edu (P.K. Choudhary), hnn@stat.ohio-state.edu (H.N. Nagaraja).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2004.06.049

$\Omega = \{(\boldsymbol{\mu}, \boldsymbol{\sigma}, \rho) : -\infty < \mu_i < \infty, 0 < \sigma_i < \infty, i = 1, 2; |\rho| < 1\}$. Thus, $\mathbf{D}^{(2)} = (D_1^2, D_2^2)$ has mean $\boldsymbol{\theta} = (\theta_1, \theta_2)$, standard deviation $\boldsymbol{\psi} = (\psi_1, \psi_2)$ and correlation γ, where

$$\theta_i = \mu_i^2 + \sigma_i^2, \psi_i^2 = 2\sigma_i^4 + 4\mu_i^2\sigma_i^2 \quad \text{and} \\ \gamma = (\psi_1\psi_2)^{-1}\{2\rho\sigma_1\sigma_2(\rho\sigma_1\sigma_2 + 2\mu_1\mu_2)\}. \tag{1}$$

We take the mean squared deviation (MSD), θ_i, as the measure of agreement between the ith instrument and the gold standard, where smaller θ_i value indicates better agreement. Let [1], [2] be the unknown indices such that $\theta_{[1]} \leqslant \theta_{[2]}$. Thus, the instrument associated with $\theta_{[1]}$ is the best one and our goal is to find it. The role of MSD in measuring agreement between two instruments has been discussed by Lin (2000) who also presented some large-sample results. For various other measures of agreement, see the reviews by Lin (2003) and Lin et al. (2002).

In the ranking and selection literature (see Gupta and Panchapakesan, 1979 for an introduction), Mukhopadhyay and Chou (1984) describe a two-stage procedure for selecting the component of a multivariate normal distribution with the smallest mean, assuming non-negative correlations. But in our case, (a) the BVN assumption for $\mathbf{D}^{(2)}$ is not justified in general—we are assuming it for $\mathbf{D}$, and (b) the covariance matrix of $\mathbf{D}^{(2)}$ is not free of $\boldsymbol{\theta}$, the parameter of interest. So we cannot assume any simplifying structure on this matrix (such as non-negative correlation, equal variances, etc.). Also, the standard multiple comparisons with the best (MCB) techniques (see Chapter 4, Hsu, 1996 for an introduction) cannot be directly employed.

Let

$$\lambda_i = \log(\theta_i), \quad \tau_i^2 = \psi_i^2/\theta_i^2, \quad i = 1, 2, \quad \text{and} \quad \tau_d^2 = \tau_1^2 + \tau_2^2 - 2\gamma\tau_1\tau_2, \tag{2}$$

where ψ_i, θ_i, and γ are given by (1). Also, let $D_{[i]}$ be the difference associated with $\theta_{[i]}$. Hence, $(\mu_{[1]}, \mu_{[2]})$ and $(\sigma_{[1]}, \sigma_{[2]})$ are the mean and standard deviation of $(D_{[1]}, D_{[2]})$, and $(\theta_{[1]}, \theta_{[2]})$ and $(\psi_{[1]}, \psi_{[2]})$ are the mean and standard deviation of $(D_{[1]}^2, D_{[2]}^2)$. Let $\mathbf{D}_j = (D_{1j}, D_{2j})$, $j = 1, 2, \ldots,$ denote a sequence of i.i.d. observations on $\mathbf{D}$. Based on the first m observations, let $\hat{\mu}_i(m)$ and $\hat{\sigma}_i^2(m)$ denote the usual unbiased estimators μ_i and σ_i^2, and let the sample correlation be $\hat{\rho}(m)$. The estimators of functions of these five parameters are constructed by plugging-in their sample counterparts, and are denoted by the usual hat notation. When it is clear from the context, we will suppress the sample size from the notations of the estimators. Finally, let $z(\alpha)$ and $t_k(\alpha)$, respectively denote the upper α-th quantiles of a $N(0, 1)$ distribution and a t-distribution with k degrees of freedom.

In the method comparison literature, St. Laurent (1998), and Hutson et al. (1998) have considered the problem of measuring agreement of two instruments with a gold standard. St. Laurent assumes a random effects model $X_i = G + \varepsilon_i$, where ε_i, $i = 1, 2$, are correlated random variables with zero means and distributed independently of G. This model assumes no bias among the two instruments and the gold standard. St. Laurent takes the intraclass correlation between (X_i, G) as the measure of agreement (see Harris et al., 2001, for an extension of this measure). For selection purposes he employs a nonparametric bootstrap confidence interval (CI) for the difference between the two intraclass correlations. This approach is ad hoc and the no bias assumption cannot in general be justified in practice.