13

Measuring Agreement in Method Comparison Studies – A Review

Pankaj K. Choudhary[1] **and H. N. Nagaraja**[2]

[1]*Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX U.S.A.*
[2]*Department of Statistics, The Ohio State University, Columbus, OH U.S.A.*

Abstract: Assessment of agreement between two or more methods of measurement is of considerable importance in many areas. In particular, in medicine, new methods or devices that are cheaper, easier to use, or less invasive, are routinely developed. Agreement between a new method and a traditional reference or gold standard must be evaluated before the new one is put into practice. Various methodologies have been proposed for this purpose in recent years. We review the literature focussing on the assessment of agreement between two methods, and on the selection of the best when several methods are compared with a reference. A real data set is analyzed to illustrate the various approaches.

Keywords and phrases: Limits of agreement, gold standard, intersection-union tests, intraclass correlation, concordance correlation, selection of the best

13.1 Introduction and General Overview

When multiple methods are available for measuring a variable of interest, one is led to the task of some sort of comparison that depends on the objective of the study. General goals of such studies are [Lewis et al. (1991)]:

(a) *Comparison*: A new method has to be evaluated by comparison with an established standard, often called a *gold standard* or a *reference method.* Neither method may be accurate or precise. The goal is to learn the extent to which the measurements from the two methods agree and understand the nature of their differences. If the methods agree sufficiently well, we can use them interchangeably or use the new one, which is cheaper or more convenient, in place of the gold standard.

(b) *Calibration*: Compare an approximate method with a known accurate and precise method whose measurement error is negligible. The goal is to establish a mathematical relationship between their measurements so that the approximate method can be used as a predictor of the accurate and precise method (and hence of the true measurement).

(c) *Conversion*: Compare two approximate methods that measure two variables in different units that are surrogates for the same underlying quantity. The goal is to interpret the results of one in terms of the other.

The focus of this chapter is on surveying the recent, growing literature on the first topic. We concentrate only on the case when the measurements are continuous. Lin (2003), Lin et al. (2002) and Shoukri (1999) provide brief reviews of this area, but our scope is comprehensive and broader. The problem of assessing agreement in the categorical measurements case has been discussed extensively elsewhere. See the reviews by Kraemer, Periyakoil and Noda (2002), and Banerjee et al. (1999), and the books by Shoukri (2004) and Fleiss (1981). Also, see Cameron (1982) for an introduction to calibration problems, and to Lewis et al. (1991) for conversion problems.

Bland and Altman (1986) who present the *limits of agreement* approach and Lin (1989) who introduces the *concordance correlation coefficient* are the two classical references. The former, a favorite of medical researchers, has over 6100 citations in the Institute for Scientific Information database at the time of writing. When two methods are compared, the data consist of a random sample of paired measurements, (X_{1j}, X_{2j}), $j = 1, \ldots, n$, taken from a bivariate population (X_1, X_2), where X_1 and X_2 arise from the reference and the test method, respectively. The following model is often assumed:

$$X_{ij} = T_j + \beta_i + \epsilon_{ij};\ i = 1, 2,\ j = 1, \ldots, n;\ \text{ where,} \tag{13.1}$$

(a) T_j is the true unobservable measurement for the jth subject, distributed as $N(\mu_T, \sigma_T^2)$;

(b) β_i is the fixed bias of the ith method;

(c) ϵ_{ij} is the random error having $N(0, \sigma_{\epsilon_i}^2)$ distribution, $i = 1, 2$, and

(d) $(T_j, \epsilon_{1j}, \epsilon_{2j})$ are mutually independent for all j.

This model is known as the *Grubbs' model* in the literature when the objective of the experiment is to estimate the bias and precision of the methods [see e.g., Grubbs (1982) and Dunn and Roberts (1999)]. The quantity σ_T^2 is also known as the *between-subject variance* and $\sigma_{\epsilon_i}^2$ as the *within-subject variance* for, or the *measurement error variance* of the ith method. Let $E(X_i) = \mu_i = \mu_T + \beta_i$, $Var(X_i) = \sigma_i^2 = \sigma_T^2 + \sigma_{\epsilon_i}^2$, $i = 1, 2$, and $\rho = \sigma_T^2/(\sigma_1\sigma_2)$. Then

(X_1, X_2) is bivariate normal with means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation ρ. Notice that $\rho = (\rho_1\rho_2)^{1/2}$, where $\rho_i = \sigma_T^2/(\sigma_T^2 + \sigma_{\epsilon_i}^2)$, is the *reliability coefficient* of the ith method [see, e.g., Fleiss (1986)]. Now, define $D = X_2 - X_1$ and $D_j = X_{2j} - X_{1j}$; $j = 1, \ldots, n$. Thus, D is $N(\mu, \sigma^2)$, where $\mu = \mu_2 - \mu_1$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 = \sigma_{\epsilon_1}^2 + \sigma_{\epsilon_2}^2$, and the D_j are a random sample from this population.

In the above setup, the two methods are in *perfect agreement* if all the paired measurements (X_{1j}, X_{2j}) lie on the 45^o line through the origin. It can be characterized by any of the following equivalent conditions:

(A1) $\beta_1 = \beta_2$, $\sigma_{\epsilon_1}^2 = \sigma_{\epsilon_2}^2 = 0$;

(A2) $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$, $\rho = 1$;

(A3) $\mu = 0$, $\sigma^2 = 0$.

Typically, the criteria (A2) and (A3) are used in the literature. Although they are equivalent when there is perfect agreement, quantification of disagreements differ. In applications, point estimators are followed by confidence intervals on relevant parameters that are generally obtained by inverting a test of hypotheses of the form

$$\begin{aligned} H &: \text{The methods lack satisfactory agreement } \textit{vs} \\ K &: \text{The methods have satisfactory agreement.} \end{aligned} \tag{13.2}$$

The advantage of this formulation is that we look for evidence in the data to claim satisfactory agreement. This way the type-I error is actually the error of wrongly concluding satisfactory agreement. This formulation was first proposed by Lin (1992) and is now well established. However, there is no unanimity on what the best formulation of (H, K) is, in terms of parameters of the data. In the terminology of the US Food and Drug Administration, *satisfactory agreement* is also referred to as *substantial equivalence* [see, e.g., Lin, Whipple and Ho (1998)].

If insufficient agreement is inferred, it helps to pay attention to the nature and extent of disagreement. It may happen that X_1 and X_2 are highly correlated and have similar means, but $\sigma_{\epsilon_2}^2$ is smaller than $\sigma_{\epsilon_1}^2$. Then the new method is certainly worthy of adoption. Many times a simple linear calibration of the new method $(X_2' = a + bX_2)$ may be enough for sufficient agreement between X_2' and X_1.

Let $\overline{X}_i$, S_i^2 and S_{12} be the usual unbiased estimators of $E(X_i)$, $Var(X_i)$ and $Cov(X_1, X_2)$. The components of variance in the model (13.1) are estimated as $\hat{\sigma}_T^2 = S_{12}$ and $\hat{\sigma}_{\epsilon_i}^2 = S_i^2 - S_{12}$. These are also known as *Grubbs' estimators* [see Grubbs (1982)]. Thus, the various parameter estimators are: $\hat{\mu}_i = \overline{X}_i$, $\hat{\sigma}_i^2 = S_i^2$, $\hat{\rho} = S_{12}/(S_1S_2)$, $\hat{\mu} = \overline{X}_2 - \overline{X}_1$ and $\hat{\sigma}^2 = S_1^2 + S_2^2 - 2S_{12}$.

A key issue is how much a method (say, the gold standard) agrees with itself, because it limits the amount of agreement that is possible between two methods. Popular terms for this phenomenon of agreement with itself include "reliability", "reproducibility" and "repeatability". But these terms have been used in other settings too [see, e.g., Lin (1989) and Bland and Altman (1999)]. The various measures of this agreement include the intraclass correlation (or the reliability coefficient) computed from one-way models [see, e.g., Fleiss (1986) and Dunn (1989, 1992)], intraclass correlation from two-way models [see e.g., Fleiss (1986)], within-subject variance [see Bland and Altman (1999)], and within-subject coefficient of variation [see Quan and Shih (1996) and the related correspondence]. See also the comments in Hawkins (2002), and Dunn and Roberts (1999). Generally, this issue is addressed in a separate reliability study and will not be discussed here.

This chapter is organized as follows. Sections 13.2 and 13.3 discuss various approaches for assessment of agreement between two methods. In Section 13.4 we provide an illustrative real example. In Section 13.5 we change our focus to the comparison of k (≥ 2) methods with a gold standard. The goal of this comparison is to select the best among k — the one that agrees most with the reference. Some concluding remarks are presented in Section 13.6.

In what follows, $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cdf and pdf, respectively, and $\Phi_2(\cdot,\cdot;\rho)$ is the cdf of a bivariate standard normal distribution with correlation ρ. A χ^2-distribution will be denoted by χ^2_k, and a t-distribution will be denoted by t_k, where k is the degrees of freedom. An F distribution with degrees of freedom l and m will be denoted by $F(l,m)$. The notation H_i (K_i) is used for a null (alternative) hypothesis, with $i = 0, 1, \ldots$.

13.2 Early Approaches

The examples in Altman and Bland (1983) indicate that the correlation

$$\rho = \sigma_T^2/\left((\sigma_T^2 + \sigma_{\epsilon_1}^2)(\sigma_T^2 + \sigma_{\epsilon_2}^2)\right)^{1/2},$$

arising from the model (13.1), was a widely used measure of "agreement" in the medical literature. Further, a statistically significant result of testing $H_0 : \rho = 0$ *vs* $K_0 : \rho \neq 0$ was often taken as evidence of agreement. But this test is generally useless because two methods designed to measure the same quantity will rarely be uncorrelated. Also, ρ is just a measure of strength of linear relationship, not of agreement. It is possible to have $\rho = 1$ even when $\mu_2 - \mu_1 = a$ ($\neq 0$) and $\sigma_2^2/\sigma_1^2 = b$ ($\neq 1$). Altman and Bland (1983) and Bland and Altman (1986, 1990, 1995a) draw attention to the following deficiencies of ρ:

(a) The value of ρ increases as $\sigma_T^2/\sigma_{\epsilon_i}^2$ increases. In practice, investigators try to assess the agreement of methods over a wide range of subjects resulting in large σ_T^2 and ρ values. This property makes ρ unsuitable as a measure of agreement, because how σ_T^2 and $\sigma_{\epsilon_i}^2$ compare is not related to the goal of assessing agreement.

(b) A consequence of (a) is that, on its own, the value of ρ does not tell us much about agreement. It is possible to have two data sets such that (i) the differences of the paired measurements in one set is identical with those in the other (and hence the two sets exhibit the same degree of agreement), but (ii) the estimate of ρ is quite low in one case and is very high in the other [see also Atkinson and Nevill (1997)].

These concerns also hold for other correlation type measures, namely concordance correlation and intraclass correlation, discussed later in this section.

Other popular, but inappropriate, early approaches include a paired-t test of $H_0 : \mu = 0$, and a test of H_0 : slope = 1, intercept = 0, when the test method is regressed on the reference method. Notice that the paired-t test only assesses whether the methods agree *on average*, not *for every subject*. Further, as Lin (1989) demonstrates with both graphs and real data, both these tests can be misleading. They may reject H_0 if the scatter around the 45^o line is near zero (indicating good agreement), and may fail to reject it if the scatter is very high (indicating poor agreement). In addition, a correct regression approach must account for the fact that even the reference method measures with error. But again, since our goal is to quantify the disagreement between methods, and not how accurately one method can be predicted from the other, we do not discuss regression models with errors in variables or structural equations models here and refer the reader to Kelly (1985), Linnet (1993), Nix and Dunston (1991), and Fuller (1987).

13.2.1 The limits of agreement (LOA) approach

Bland-Altman Plot

The basic idea due to Bland and Altman (1986) is that if a large proportion (such as 95%) of the differences are sufficiently close to zero then the two methods have satisfactory agreement. The process of judging this agreement has two components: (a) 95% LOA, defined by $\hat{\mu} \pm 1.96\hat{\sigma}$, and (b) the plot of mean, $(X_1 + X_2)/2$, versus difference, D, with LOA superimposed. This plot is popularly known as the Bland-Altman plot. Statistical software SAS JMP produces such a plot for matched pair data.

The LOA estimate the set $(\mu - 1.96\sigma, \mu + 1.96\sigma)$. One could declare sufficient agreement if the differences within these limits are not practically (or clinically) important as determined by the investigator specified threshold δ_0 (> 0). Bland

and Altman recommend that this δ_0 be specified in advance and Bland and Altman (1999) observe that its choice "will depend on the use to which the result is put, and is a question of clinical judgement." See also Hawkins (2002).

The uncertainty in the estimation of LOA is accounted for by the approximate 95% confidence intervals (CIs) for the two limits. It is $(\hat{\mu} - 1.96\hat{\sigma}) \pm t_{n-1}(\alpha/2)1.71\hat{\sigma}/n^{1/2}$ for $\mu - 1.96\sigma$ (and similar for the upper limit), where $t_k(\alpha)$ is the upper αth percentile of a t_k distribution.

The Bland-Altman plot is an excellent supplement to the usual scatterplot of the data. It reveals interesting features of the data and also helps in diagnosing departures from the various model assumptions and suggesting remedies. Hawkins (2002) describes how the Bland-Altman plot can be used in conjunction with the standard regression diagnostics, and presents several real examples to illustrate the common model violations and discusses ways to handle them. He says, "the ideal plot resembles the ideal plot of residuals against the fitted values in a regression problem". Some common departures are:

(a) *LOA band not centered at zero or that is wide*: A plot not centered at zero indicates a bias between the methods. If the rest of the features of the plot are close to ideal, the new method can be recalibrated by adding a constant for good agreement with the old. However, if the points do not lie in a narrow band, it suggests that the variability of the differences is not small. This will result in wide LOA, and is a serious problem to resolve.

(b) *Linear trend*: The differences increase (or decrease) with increasing magnitude of the measurements in the range of measurement [see also Bland and Altman (1995b)]. Under the model in (13.1), this implies that the $\sigma^2_{\epsilon_i}$ are different. If $\sigma^2_{\epsilon_2} > \sigma^2_{\epsilon_1}$, a possible resolution is to recalibrate the new method by multiplying with a constant. Otherwise, the new is more precise than the old, and obviously is preferred. As another possible resolution, Bland and Altman (1986, 1999) suggest that a log transformation of the data may make the variabilities comparable. The analysis then proceeds with the log-scale differences, and the results are back-transformed to the ratio scale for interpretation.

(c) *Heteroscedasticity*: If the scatter of differences does not remain the same over the range of measurement, the model assumption that σ_i^2 is not related to μ_i is violated for one or both of the methods. A common violation corresponds to a "right opening megaphone" — an increase in scatter of differences as the magnitude increases. Often, a log transformation of the data corrects this problem. Sometimes a more sophisticated variance stabilizing transformation may be needed.

(d) *Outliers*: Vertical outliers flag the subjects for whom the measurements

differ by an unusually large amount. It is recommended to assess the impact of the outliers on the results.

(e) *Non-linear curvature*: A non-linear recalibration may be needed to correct the problem.

Significance Tests

One can use the paired-t test for $H_0 : \mu_1 = \mu_2$, the Pitman-Morgan test for $H_0 : \sigma_1^2 = \sigma_2^2$, and the Bradley-Blackwood test for $H_0 : \mu_1 = \mu_2,\ \sigma_1^2 = \sigma_2^2$. The last two are described in Krummenauer (1999). Bartko (1994) shows how all of these can be implemented using the standard two-way ANOVA output. These tests (of point null hypotheses) are generally used only to supplement the graphical analysis, and rarely play a prominent role in the assessment of agreement [see also the discussion in Bland and Altman (1999)]. The hypotheses of whether the agreement is satisfactory or not is generally of the form (13.2). Bartko (1994) suggests adding an elliptical tolerance region of the bivariate distribution of the difference and the mean to the Bland-Altman plot, so that the relative magnitudes of the between-subject and the within-subject variations can be amplified.

The normality of differences can be assessed with the usual histogram and normal probability plot.

Confidence Regions

Beyond the plots, interest lies in quantifying disagreement by the estimation of the parameter region $(\mu - 1.96\sigma, \mu + 1.96\sigma)$, and its comparison with the threshold interval $(-\delta_0, \delta_0)$. For this comparison we must use a CI estimate of the region so that the uncertainty in the estimation is also accounted for. The usual LOA approach achieves this by constructing separate two-sided CIs for the two endpoints. However, Lin et al. (1998) argue that instead of the two-sided CIs, we need one-sided CIs, an upper confidence bound (UCB) for $\mu+1.96\sigma$ and a lower confidence bound (LCB) for $\mu-1.96\sigma$, because the interest lies in bounding the region $(\mu - 1.96\sigma, \mu + 1.96\sigma)$. Consequently, we can infer satisfactory agreement if

$$(\hat{\mu} - a_n\hat{\sigma}, \hat{\mu} + a_n\hat{\sigma}) \subset (-\delta_0, \delta_0),$$

where $a_n = 1.96 + 1.71n^{-1/2}t_{n-1}(\alpha)$. Based on this, Lin et al. (1998) derive a simple sample size formula for use in planning method comparison studies. They also indicate that the above rule can be thought of as a large-sample intersection-union test (IUT) [see Casella and Berger (2002, p. 380), for an introduction] of

$$H_1 : \text{Complement of } K_1 \ \textit{vs.}\ K_1 : -\delta_0 < \mu - 1.96\sigma,\ \mu + 1.96\sigma < \delta_0, \quad (13.3)$$

which is of the form (13.2). For these hypotheses, however, Liu and Chow (1997) have given an exact IUT in the context of assessment of individual bioequivalence. This test rejects H_1 if

$$(\hat{\mu} - b_n\hat{\sigma}, \hat{\mu} + b_n\hat{\sigma}) \subset (-\delta_0, \delta_0),$$

where $b_n = n^{-1/2}t_{n-1}\big(\alpha, n^{1/2}z(\frac{\pi_0}{2})\big)$, $z(\frac{\pi_0}{2})$ is the $(\frac{\pi_0}{2})$th upper percentile of a $N(0,1)$ distribution, and $t_k(\alpha, \Delta)$ is the upper αth percentile of a non-central t_k-distribution with non-centrality parameter Δ. As shown in Choudhary and Nagaraja (2004a), the interval $(\hat{\mu} - b_n\hat{\sigma}, \hat{\mu} + b_n\hat{\sigma})$ can also be interpreted as a large-sample two-sided tolerance interval with 0.95 content at confidence level $1 - \alpha$. See, e.g., David and Nagaraja (2003) or Guttman (1988) for an introduction to tolerance intervals. A comparison of these two tests of (13.3) and the associated sample size formulae will be of interest.

The strength of the LOA approach lies in its intuitive appeal and simplicity. Further, since the limits are based on difference, they are not affected by the between-subject variation in the data. This approach was recently generalized by Bland and Altman (1999) to accommodate replicate measurements on each subject from every method.

13.2.2 Intraclass correlation and related measures

Fleiss (1986, Ch. 1) gives an overview of the intraclass correlation coefficient (ICC) as a measure of agreement between k (≥ 2) methods. In our setting, the ICC is defined under the two-way mixed model (13.1) with additional assumptions that $i = 1, \ldots, k$ (≥ 2), $\sigma^2_{\epsilon_i} = \sigma^2_\epsilon$ for all i, and $\sum_{i=1}^k \beta_i = 0$. It serves as an index of agreement among k methods and is given by

$$\rho_I = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_\beta + \sigma^2_\epsilon}, \tag{13.4}$$

where $\sigma^2_\beta = \sum_{i=1}^k \beta_i^2/(k-1)$. This ICC is also known as the inter-method (or inter-rater) reliability. Note that $0 < \rho_I \leq 1$ and equals 1 only when there is perfect agreement among all the k methods.

The ANOVA table for this model is given in Table 13.1 [see also McGraw and Wong (1996)] where $\overline{X}_{i\cdot} = \sum_{j=1}^n X_{ij}/n$, $\overline{X}_{\cdot j} = \sum_{i=1}^k X_{ij}/k$, and $\overline{X}_{\cdot\cdot} = \sum_{i=1}^k \sum_{j=1}^n X_{ij}/(nk)$. The ICC is then estimated by

$$\hat{\rho}_I = \frac{SMS - EMS}{SMS + (k-1)EMS + (k/n)(IMS - EMS)}. \tag{13.5}$$

Its approximate $100(1-\alpha)\%$ LCB [McGraw and Wong (1996)] is

$$\frac{n(SMS - F_{n-1,\nu}(\alpha)\, EMS)}{F_{n-1,\nu}(\alpha)\,[k\, IMS + (kn - k - n)EMS] + n\, SMS},$$

Table 13.1: ANOVA Table for estimating the ICC ρ_I of (13.4)

Source	d.f.	SS	MS	E(MS)
Subjects	$n-1$	$k\sum_{j=1}^{n}(\overline{X}_{\cdot j}-\overline{X}_{\cdot\cdot})^2$	*SMS*	$k\sigma_T^2+\sigma_\epsilon^2$
Methods	$k-1$	$n\sum_{i=1}^{k}(\overline{X}_{i\cdot}-\overline{X}_{\cdot\cdot})^2$	*IMS*	$n\sigma_\beta^2+\sigma_\epsilon^2$
Error	$(n-1)(k-1)$	By Subtraction	*EMS*	σ_ϵ^2
Total	$nk-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\overline{X}_{\cdot\cdot})^2$		

where

$$\nu=\frac{(a\,IMS+b\,EMS)^2}{(k-1)^{-1}(a\,IMS)^2+[(n-1)(k-1)]^{-1}(b\,EMS)^2},$$

$a=(k/n)[\hat{\rho}_I/(1-\hat{\rho}_I)]$, $b=1+(n-1)\,a$, and $F_{l,m}(\alpha)$ is the upper αth percentile of an $F(l,m)$ distribution.

This model can accommodate several methods, but the assumption of equal error variance is generally not reasonable. Further, ρ_I is non-negative but its estimates may be negative, and it is also sensitive to the between-subject variation. If ρ_I is low, it is unclear whether the lack of agreement is due to low between-subject variation and/or high error variation and/or location bias. See also Müller and Büttner (1994) for a critique of this measure.

Assuming X_1 and X_2 have the same means, St. Laurent (1998) has considered the model

$$X_{2j}=X_{1j}+\epsilon_j;\ j=1,\ldots,n,$$

and suggested the associated ICC as a measure of agreement. Here ϵ_j is the measurement error in the test method. It is assumed that the (i) X_{1j} and ϵ_j are independent and identically distributed (i.i.d.) with means μ_1 and zero, variances σ_1^2 and σ_ϵ^2, respectively, and (ii) X_{1j} and ϵ_j are mutually independent. The ICC is then given by

$$\rho_G=\frac{\sigma_1^2}{\sigma_1^2+\sigma_\epsilon^2}. \tag{13.6}$$

which is the squared correlation between X_1 and X_2. Clearly, $\rho_G>0$ and it equals 1 when $\sigma_\epsilon^2=0$. St. Laurent refers to $\rho_G^{1/2}$ as the *gold standard correlation.* When X_{1j} and ϵ_j are normally distributed, the maximum likelihood estimator (MLE) of ρ_G is

$$\hat{\rho}_G=\frac{(n-1)\hat{\sigma}_1^2}{(n-1)\hat{\sigma}_1^2+\sum_{i=1}^{n}D_i^2}$$

and the LCB of ρ_G is given by

$$\frac{F_{n,n-1}(\alpha)}{F_{n,n-1}(\alpha)+(\hat{\rho}_G^{-1}-1)(n-1)/n}.$$

St. Laurent (1998) also develops a large sample theory for the non-normal case. Recently, Harris, Burch, and St. Laurent (2001) have developed a family of estimators of ρ_G that includes the MLE and they indicate that at times other members of the family may be preferable to the MLE in terms of mean-squared error. Being an ICC, ρ_G also suffers from the drawback of being sensitive to the between-subject variation. In addition, the assumption of no bias between the methods may not be justified.

13.2.3 Concordance correlation approach

Lin (1989) proposed a concordance correlation coefficient (CCC) as an index of agreement and defined it as

$$\rho_C = 1 - \frac{E(X_1 - X_2)^2|_\rho}{E(X_1 - X_2)^2|_{\rho=0}} = \frac{2\rho\sigma_1\sigma_2}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2}. \tag{13.7}$$

It represents the expected squared distance of a point (X_1, X_2) from the 45^o line through the origin, scaled to lie between $[-1, 1]$. This distance, $\theta = E(X_1 - X_2)^2$, is also referred to as the mean squared deviation (MSD). Thus ρ_C measures how close the (paired) observations are to the 45^o line. It can also be written as ρC_b, where

$$C_b = 2/(v + 1/v + u^2), v = \sigma_1/\sigma_2, \text{ and } u = (\mu_1 - \mu_2)/(\sigma_1\ \sigma_2)^{1/2}. \tag{13.8}$$

Thus the CCC has two components: (a) ρ, the correlation, which Lin calls the "precision" component, that measures how close the observations are to the best fit line, and (b) $C_b \in (0, 1]$, the "accuracy" component that measures how close the best fit line is to the 45^o line. The CCC has the following properties:

(i) $|\rho_C| \leq |\rho| \leq 1$, (ii) $\rho_C = 0$ iff $\rho = 0$, (iii) $\rho_C = \rho$ iff $\sigma_1 = \sigma_2$, $\mu_1 = \mu_2$, and (iv) $\rho_C = \pm 1$ iff $\rho = \pm 1$, $\sigma_1 = \sigma_2$, and $\mu_1 = \mu_2$.

The estimator of ρ_C, obtained by replacing the population moments with the sample moments, is

$$\hat{\rho}_C = \frac{2\,\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2}{(\hat{\mu}_1 - \hat{\mu}_2)^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}. \tag{13.9}$$

Under bivariate normality for (X_1, X_2), Lin (1989) showed that $\hat{\rho}_C$ is asymptotically normal with mean ρ_C. He suggests the transformation $W = \tanh^{-1}(\hat{\rho}_C)$ for faster convergence; W is asymptotically normal with mean $\tanh^{-1}(\rho_C)$ and variance

$$\sigma_W^2 = \frac{1}{n-2}\left[\frac{\rho_C^2(1-\rho^2)}{\rho^2(1-\rho_C^2)} + \frac{2\rho_C^3(1-\rho_C)u^2}{\rho(1-\rho_C^2)^2} - \frac{\rho_C^4 u^4}{2\rho^2(1-\rho_C^2)^2}\right],$$

where u is given in (13.8). Another advantage of this transformation is that the asymptotic CI for ρ_C is constrained to lie within $[-1, 1]$ and an approximate level $(1-\alpha)$ LCB for ρ_C is

$$\tanh\left(\tanh^{-1}(\hat{\rho}_C) - z(\alpha)\hat{\sigma}_W\right),$$

where $\hat{\sigma}_W$ is obtained by replacing the population moments in σ_W^2 with the corresponding sample moments.

Based on the allowable losses in the "precision" and "accuracy" components described above, Lin (1992) first computes ρ_C^*, which represents the smallest acceptable value of ρ_C that the investigator is willing to consider as evidence of satisfactory agreement. Then he proposes to test

$$H_2 : \rho_C \leq \rho_C^* \ \ vs. \ \ K_2 : \rho_C > \rho_C^*,$$

using the LCB of ρ_C obtained as above. If it exceeds ρ_C^*, one rejects H_2 and infers satisfactory agreement. Lin also gave a sample size formula based on the above test of hypotheses.

In practice, however, an LCB for ρ_C is computed and compared with a cutoff such as 0.75. The practice of accepting satisfactory agreement if the LCB for CCC (or ICC) exceeds 0.75 [see, e.g., Lee, Koh and Ong (1989) and Atkinson and Nevill (1997)] may not be wise. When ρ is close to 1, a location/scale bias may not be reflected well in these measures, and hence they may lead to wrong conclusions. See Section 13.4 for an example.

The CCC formulation has motivated further research. Chinchilli et al. (1996) introduced a weighted version of ρ_C to handle repeated measurements data. King and Chinchilli (2001) have generalized ρ_C to incorporate distance functions other than the squared error, and constructed robust forms of ρ_C. They also demonstrated the relationship between ρ_C and the *kappa statistic*, a measure of agreement for nominal/ordinal measurements, and introduced further extensions [see also Robieson (1999)]. Barnhart, Haber and Song (2002) have constructed an overall CCC, a weighted average of all pairwise CCC's, for measuring agreement between k (> 2) methods. Additionally, Barnhart and Williamson (2001) suggested a generalized estimating equations approach to model ρ_C when covariates are present. Vonesh, Chinchilli and Pu (1996) have used CCC type of index in goodness-of-fit contexts – to evaluate the agreement between observed values and the values predicted by a model and for model selection.

Liao and Lewis (2000) allude to some deficiencies of CCC and suggest a slightly modified index. However, its performance tends to be similar. They also extend it to handle situations when parameters cannot be assumed to remain fixed over the entire range for measurement.

Lin (1989) pointed out that the CCC tends to produce results similar to the ICC. It was finally noted by Nickerson (1997) that, when $k = 2$, the estimate

of the ICC ρ_I defined in (13.5), becomes

$$\hat{\rho}_I = \frac{2\,\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2}{(\hat{\mu}_1 - \hat{\mu}_2)^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}^2/n}.$$

This differs from $\hat{\rho}_C$ in (13.9) in the denominator by $\hat{\sigma}^2/n$. She notes that this difference tends to be small resulting in close values for these coefficients.

There are two related issues with ICC, CCC and other such measures that have drawn much criticism in the literature [see, e.g., Müller and Büttner (1994), Bland and Altman (1995b), Atkinson and Nevill (1997,) and Lin and Chinchilli (1997)]. Firstly, these measures depend on the between-subject variation. Thus deficiencies of correlation that were listed on Page 218 continue to hold for CCC.

Secondly, although CCC is a convenient single index for agreement that lies in $[-1, 1]$ and combines components of a systematic bias, a difference in variabilities and a low correlation, there lies its major weakness. If a lack of satisfactory agreement is concluded using CCC, the reason is unclear. To deal with this issue Lin and Torbeck (1998) and Lin et al. (2002) suggest supplementing the CCC with CIs for its ρ ("precision") and C_b ("accuracy") components.

Recently, Lin et al. (2002) compared the power properties of the test (of agreement) based on CCC with those based on total deviation index [Lin (2000)] and the coverage probability, and found that the CCC based test has inferior power properties. These terms and tests are introduced next.

13.3 Recent Developments

13.3.1 Approaches based on percentiles and coverage probability

The LOA approach involved the following steps:

(a) Specify a threshold interval $(-\delta_0, \delta_0)$ such that the differences in this interval are practically equivalent to zero.

(b) Quantify the observed disagreement by estimating the range in which a pre-specified large proportion of differences (say π_0) are expected to lie.

(c) If the estimated range in (b) is contained in $(-\delta_0, \delta_0)$, declare satisfactory agreement.

In practice, the LOA approach takes $\pi_0 = 0.95$ and uses a CI estimate of $(\mu - 1.96\sigma, \mu + 1.96\sigma)$, the *centrally symmetric* region of 95% probability content. Often times δ_0 is not explicitly specified and the evaluation of agreement

proceeds by assessing whether the estimated range contains any clinically important differences. We saw in Section 13.2.1 that this approach can be thought of as testing the hypotheses (13.3), which can also be expressed as

$$H_1 : \pi_c \leq \pi_0 \ \ vs. \ \ K_1 : \pi_c > \pi_0, \tag{13.10}$$

where π_c is the proportion of centrally symmetric region contained in $(-\delta_0, \delta_0)$.

Lin (2000) and Lin et al. (2002) consider a slightly liberal variation on the above theme: instead of asking for π_0 proportion of *central* differences to lie in $(-\delta_0, \delta_0)$ for satisfactory agreement, they only ask for π_0 proportion of differences to lie in $(-\delta_0, \delta_0)$. There are now two theoretically equivalent ways to proceed.

Total Deviation Index (TDI)

For a specified π_0, consider $q(\pi_0)$, the π_0th percentile of $|D|$, as the measure of agreement and assess satisfactory agreement by testing

$$H_3 : q(\pi_0) \geq \delta_0 \ \ \text{vs.} \ \ K_3 : q(\pi_0) < \delta_0. \tag{13.11}$$

Lin (2000) proposes this approach and calls $q(\pi_0)$ the *total deviation index*.

Since D is $N(\mu, \sigma^2)$ and $Pr\big(|D| < q(\pi_0)\big) = \pi_0$. $q(\pi_0)$ can be written as

$$q(\pi_0) = \sigma\big(\chi_1^2(1 - \pi_0, \mu^2/\sigma^2)\big)^{1/2}, \tag{13.12}$$

where $\chi_1^2(\alpha, \Delta)$ is the upper αth percentile of a non-central χ_1^2-distribution with non-centrality parameter Δ. Lin (2000) argues that the inference based on the estimate of $q(\pi_0)$ in (13.12) is intractable. So he approximates it by

$$q^*(\pi_0) = \big((\mu^2 + \sigma^2)\chi_1^2(1 - \pi_0, 0)\big)^{1/2} = (\mu^2 + \sigma^2)^{1/2} z((1 - \pi_0)/2), \tag{13.13}$$

and for assessing agreement, he modifies the hypotheses (13.11) to

$$H_3^* : q^*(\pi_0) \geq \delta_0 \ \ \text{vs} \ \ K_3^* : q^*(\pi_0) < \delta_0. \tag{13.14}$$

He suggests estimating the MSD $\theta = (\mu^2 + \sigma^2) = E(D^2)$ by $\hat{\theta} = \sum_{i=1}^n D_i^2/(n-1)$ and performing inference using the large-sample $N(0, 1)$ distribution of $\big(\log(\hat{\theta}) - \log(\theta)\big)/\hat{\tau}$, where

$$\hat{\tau}^2 = 2(1 - \overline{D}^4/\hat{\theta}^2)/(n - 1) \tag{13.15}$$

is the estimated asymptotic variance of $\log(\hat{\theta})$. Thus the estimate of $q^*(\pi_0)$ and its approximate $100(1 - \alpha)\%$ UCB respectively become

$$\hat{q}^*(\pi_0) = \hat{\theta}^{1/2} z((1 - \pi_0)/2) \text{ and } \hat{q}^*(\pi_0) \exp\big[\tfrac{1}{2} z(\alpha)\hat{\tau}\big].$$

A test for (13.14) rejects H_3^* if this UCB is less than δ_0. Lin also gives a sample size formula associated with (13.11).

The approximation $q^*(\pi_0)$ for $q(\pi_0)$ will be good only when μ^2/σ^2 is small and Lin gives a range of values of π_0 and μ^2/σ^2 where this approximation can be considered reasonable. Further, the above test of (13.14) has asymptotic level α and is consistent. This leads to the undesirable property that in the limiting case where $n \to \infty$, with probability 1, there will be some regions where the agreement is satisfactory ($q(\pi_0) < \delta_0$), but the test will conclude otherwise (i.e., $q(\pi_0) \geq \delta_0$) and vice versa. Such regions depend on whether the approximation $q^*(\pi_0)$ is conservative (i.e., $q^*(\pi_0) > q(\pi_0)$) or liberal (i.e., $q^*(\pi_0) < q(\pi_0)$).

Coverage Probability Approach

This approach of Lin et al. (2002) takes the *coverage probability* (CP) $\pi = Pr(|D| < \delta_0)$ of the threshold interval $(-\delta_0, \delta_0)$ as the measure of agreement and tests

$$H_4 : \pi \leq \pi_0 \quad \text{vs.} \quad K_4 : \pi > \pi_0. \tag{13.16}$$

These hypotheses are equivalent to (13.11) for specified (δ_0, π_0).

Lin et al. (2002) estimate π as

$$\hat{\pi} = \Phi\big((\delta_0 - \hat{\mu})/\tilde{\sigma}\big) - \Phi\big((-\delta_0 - \mu)/\tilde{\sigma}\big),$$

where $\tilde{\sigma}^2 = (n-1)\hat{\sigma}^2/(n-3)$ and suggest performing inference through the large sample normality of $(\hat{\lambda} - \lambda)/\hat{\psi}$, where $\lambda = \log(\pi/(1-\pi))$, $\hat{\lambda} = \log(\hat{\pi}/(1-\hat{\pi}))$, and

$$\hat{\psi}^2 = \frac{1}{(n-3)\hat{\pi}^2(1-\hat{\pi})^2}\left\{\left[\phi\big((\delta_0 - \hat{\mu})/\tilde{\sigma}\big) - \phi\big((-\delta_0 - \hat{\mu})/\tilde{\sigma}\big)\right]^2 + \frac{1}{2}\left[\phi\big((\delta_0 - \hat{\mu})/\tilde{\sigma}\big)\big((\delta_0 - \hat{\mu})/\tilde{\sigma}\big) - \phi\big((-\delta_0 - \hat{\mu})/\tilde{\sigma}\big)\big((-\delta_0 - \hat{\mu})/\tilde{\sigma}\big)\right]^2\right\}$$

is the estimated asymptotic variance of $\hat{\lambda}$. Thus, in particular, the approximate $100(1-\alpha)\%$ LCB for π becomes

$$e^{\hat{\lambda}^-}/(1 + e^{\hat{\lambda}^-})$$

where $\hat{\lambda}^- = \hat{\lambda} - z(\alpha)\hat{\psi}$. When this LCB exceeds the cutoff π_0, H_4 of (13.16) is rejected.

Recent simulations by Choudhary and Nagaraja (2004a) reveal that this test is overly conservative for moderate sample sizes. When $30 \leq n \leq 50$ and $0.80 \leq \pi_0 \leq 0.95$, the empirical type-I error rate of this test is about 3% or less for a nominal 5% level. However, using the MLE of σ^2 in place of $\tilde{\sigma}^2$ greatly improves its performance.

Lin et al. (2002) provide a sample size formula for (13.16) and note that TDI and CP approaches yield similar powers.

Since the hypotheses (13.11) and (13.16) are equivalent, it is natural to expect that their tests should lead to the same conclusions when both are applied

to the same data set. However, this is clearly not the case for the proposals given above. This issue is addressed in Choudhary and Nagaraja (2004a) where an exact level α test of (13.16) (or equivalently (13.11)) is given and a good, simple approximation to its critical value is presented.

In practice, it is easy to specify π_0 (generally, a number between 0.80 and 0.95) but we have seen in Section 13.2 that choosing δ_0 is relatively difficult and depends on the purpose. In this sense, it can be argued that the formulation (13.11) is better than (13.16) as the former produces an UCB, $\hat{q}^+(\pi_0)$, for $q(\pi_0)$ for a specified π_0. The interval $(-\hat{q}^+(\pi_0), \hat{q}^+(\pi_0))$ then can be used in the same way as the LOA without a specified δ_0. However, its value must be explicitly specified in advance for the formulation (13.16).

The hypotheses (13.16) are often used for the assessment of individual bioequivalence where X_1 and X_2 are respectively the measures of effectiveness of a reference drug and a test drug [see, e.g., Anderson and Hauck (1990) and Wang and Hwang (2001)]. There is also a connection between them and the ones used in statistical quality control. In *acceptance sampling* with two-sided specification limits, D refers to the quality characteristic of an item. If D falls in the specified interval (l, u), the item is conforming, and is non-conforming otherwise. The parameter of interest here is the lot quality as measured by $1 - \pi$ and the decision to accept or reject a lot of items is based on the test of

$$H_4' : 1 - \pi \leq 1 - \pi_0 \quad \text{vs.} \quad K_4' : 1 - \pi > 1 - \pi_0,$$

where $1 - \pi_0$ is the specified *acceptable quality level.* Rejection of H_4' amounts to rejecting the lot. Notice that these hypotheses can be obtained by interchanging H_4 and K_4 of (13.16) while retaining the equality sign in the null. See Hamilton and Lesperance (1995) for a discussion of various tests of these hypotheses.

13.3.2 Approaches based on the intersection-union principle

All the formal approaches discussed so far use a single measure of agreement. Thus, when a lack of satisfactory agreement is inferred, we would not know the cause or extent of disagreement without additional investigation. Choudhary and Nagaraja (2004b) resolve this issue by giving two formulations of the hypotheses (13.2) that preserve the information on all causal indicators. The first tests

$$H_5 : \{|\mu| \geq \delta_\mu\} \cup \{\sigma \geq \delta_\sigma\} \quad \text{vs.} \quad K_5 : \{|\mu| < \delta_\mu\} \cap \{\sigma < \delta_\sigma\}, \tag{13.17}$$

where δ_μ and δ_σ reflect the extent of bias and variability in D, respectively, that the practitioner can tolerate and pre-specify. The second, more detailed formulation tests

$$\begin{aligned} H_6 &: \{|\mu_2 - \mu_1| \geq \delta_\mu\} \cup \{\sigma_2/\sigma_1 \leq \delta_1 \text{ or } \sigma_2/\sigma_1 \geq \delta_2\} \cup \{\rho \leq \delta_\rho\} \quad \text{vs.} \\ K_6 &: \{|\mu_2 - \mu_1| < \delta_\mu\} \cap \{\delta_1 < \sigma_2/\sigma_1 < \delta_2\} \cap \{\rho > \delta_\rho\}, \end{aligned} \tag{13.18}$$

where the pre-specified $\delta_\rho \in (0,1)$ is large, and $0 < \delta_1 < 1 < \delta_2$. Usually $\delta_1 = 1/\delta_2$ is taken so that they are symmetric about zero on the log-scale.

The hypotheses (13.17) and (13.18) are tested using the intersection-union principle [see Casella and Berger (2002, p. 380)], and the tests are inverted to give the CIs that quantify the extent of disagreement on individual indicators. The individual $100(1-\alpha)\%$ CIs for μ and σ associated with (13.17) are, respectively,

$$\left\{\min\left(0,\, \hat{\mu} - t_{n-1}(\alpha) n^{-1/2} \hat{\sigma}\right) \le \mu \le \max\left(0,\, \hat{\mu} + t_{n-1}(\alpha) n^{-1/2} \hat{\sigma}\right)\right\} \qquad (13.19)$$

and

$$\left\{0 < \sigma \le (n-1)^{1/2}\, \hat{\sigma} / \chi_{n-1}(1-\alpha)\right\}.$$

The $100(1-\alpha)\%$ CI for $\mu_2 - \mu_1$ associated with (13.18) is the same as (13.19) above. For σ_2/σ_1 this CI is given by

$$\left\{\, \min(1, \Delta^-) \le \sigma_2/\sigma_1 \le \max(1, \Delta^+)\right\},$$

where

$$\Delta^- = \frac{\hat{\sigma}_2}{\hat{\sigma}_1} \frac{\sqrt{1 - t_1^2 \hat{\rho}^2} - t_1 \sqrt{1 - \hat{\rho}^2}}{\sqrt{1 - t_1^2}},\ \Delta^+ = \frac{\hat{\sigma}_2}{\hat{\sigma}_1} \frac{\sqrt{1 - t_1^2 \hat{\rho}^2} + t_1 \sqrt{1 - \hat{\rho}^2}}{\sqrt{1 - t_1^2}},$$

and $t_1 = \left(n - 2 + t_{n-2}^2(\alpha)\right)^{-1/2} t_{n-2}(\alpha)$. Finally, for ρ one could use the interval suggested by Fisher's z-transformation:

$$\left\{\rho \ge \tanh\left(\tanh^{-1}(\hat{\rho}) - z(\alpha)/(n-3)^{1/2}\right)\right\}.$$

A practical strategy for soliciting the various thresholds from the investigators is described in Choudhary and Nagaraja (2004c). They also illustrate how these CIs can be used in the assessment of agreement if the pre-specification of the thresholds is difficult. A simple sample size formula associated with the test of (13.17) is also given there.

Tests based on the IU principle tend to be conservative. Although the above overall IUT's have size α, it is attained in the limit as the variability diminishes to zero. But in practice, our major goal is to quantify the disagreement in a meaningful way. In this regard IU principle plays a natural role by producing CIs for individual disagreement indicators that are informative and easy to interpret.

Table 13.2 summarizes the basic features of the various approaches we have discussed thus far.

The hypotheses (13.17) and (13.18) resemble the ones used in bioequivalence studies [see, e.g., Berger and Hsu (1996)]. But for average bioequivalence, only the mean responses of the test and the reference drugs must be equivalent, and

Table 13.2: Summary of various approaches for assessing agreement

Approach	**Measure(s) of agreement**	**Remarks**
Limits of agreement	$(\mu - 1.96\sigma, \mu + 1.96\sigma)$	Easy to interpret; most popular
Concordance correlation	$\frac{2\rho\sigma_1\sigma_2}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2}$	May be hard to interpret; sensitive to between-subject variation
ICC	$\rho_I = \sigma_T^2/(\sigma_T^2 + \sigma_\beta^2 + \sigma_\epsilon^2)$; $\rho_G = \sigma_1^2/(\sigma_1^2 + \sigma_\epsilon^2)$	$\rho_I \approx$ CCC; ρ_G requires a reference and assumes no bias; properties similar to CCC
Total deviation index(π_0)	$q^*(\pi_0)$ where $Pr(\|D\| \leq q^*(\pi_0)) \approx \pi_0$	Easy to interpret
Coverage probability (δ_0)	$Pr(\|D\| \leq \delta_0)$	Specifying δ_0 may be hard
IUT based on D	(μ, σ)	Identifies sources and extent of disagreement
IUT based on (X_1, X_2)	$(\mu_2 - \mu_1, \sigma_2/\sigma_1, \rho)$	Similar to IUT based on D, but more informative

for population bioequivalence, the marginal distributions of the responses must be equivalent. In contrast, here we also need the correlation to be close to 1.

A natural measure of agreement, already seen, is the MSD, $\theta = E(D^2) = (\mu^2+\sigma^2)$. The smaller the θ is, the better the agreement is between the methods. As noted earlier, Lin (2000) suggests estimating θ as $\hat{\theta} = \sum_{i=1}^{n} D_i^2/(n-1)$ and using the asymptotic normality of $\log(\hat{\theta})$ for inference. Hence, an approximate $100(1-\alpha)\%$ UCB for θ becomes $\hat{\theta}\exp\{z(\alpha)\hat{\tau}\}$ where $\hat{\tau}$ is given in (13.15). However, the practical utility of θ is limited by the fact that it is hard to interpret. Indeed, as Lin (1989, 2000) points out, the CCC and TDI approaches are attempts to translate this θ into more easily interpreted measures. But, θ (or equivalently $\log(\theta)$) has been the measure of choice in the selection of the instrument that agrees most with a reference. This is partly because the comparison of several methods in terms of this measure is more mathematically tractable. The problem of selection is the topic of Section 13.5.

13.4 An Example

We now illustrate the various approaches summarized in Table 13.2 using the plasma volume data from Bland and Altman (1999). The variable is measured as a percentage of expected values of normal individuals. Two sets of normal values, one due to Hurley (X_1) and the other due to Nadler (X_2) are being compared. We will take the Hurley method in the role of reference for the purpose of illustration. Figure 13.1 gives the scatter plot and the Bland-Altman plot. The solid line in the scatter plot represents the line of equality and the broken lines in the Bland-Altman plot represent the 95% LOA. We see that the two methods are highly correlated and the Nadler method consistently gives higher measurements. Most of the differences lie between 5% to 15%, are centered at around 10%, and increase as the magnitude of measurements increase.

Normal probability plots and formal tests indicate excellent normal fit for the differences and a reasonable bivariate normal fit for (X_1, X_2). Further, there is no evidence for heteroscedasticity or serious outliers in the mean versus difference plot. So we may assume that the model (13.1) holds along with normality. The positive trend observed in the mean versus difference plot is due to the higher variability of the Nadler method.

The measurements observed by the two methods fall in $(52.9, 133.2)$ and the differences $X_2 - X_1$ range from 2.5 to 17.40 with middle 50% in between 7.75 and 10.40. Since 7.75 and 10.40 respectively constitute about 10% and 13% of the measurement scale, these differences seem to be too high for the methods to have satisfactory agreement. This graphical analysis also indicates

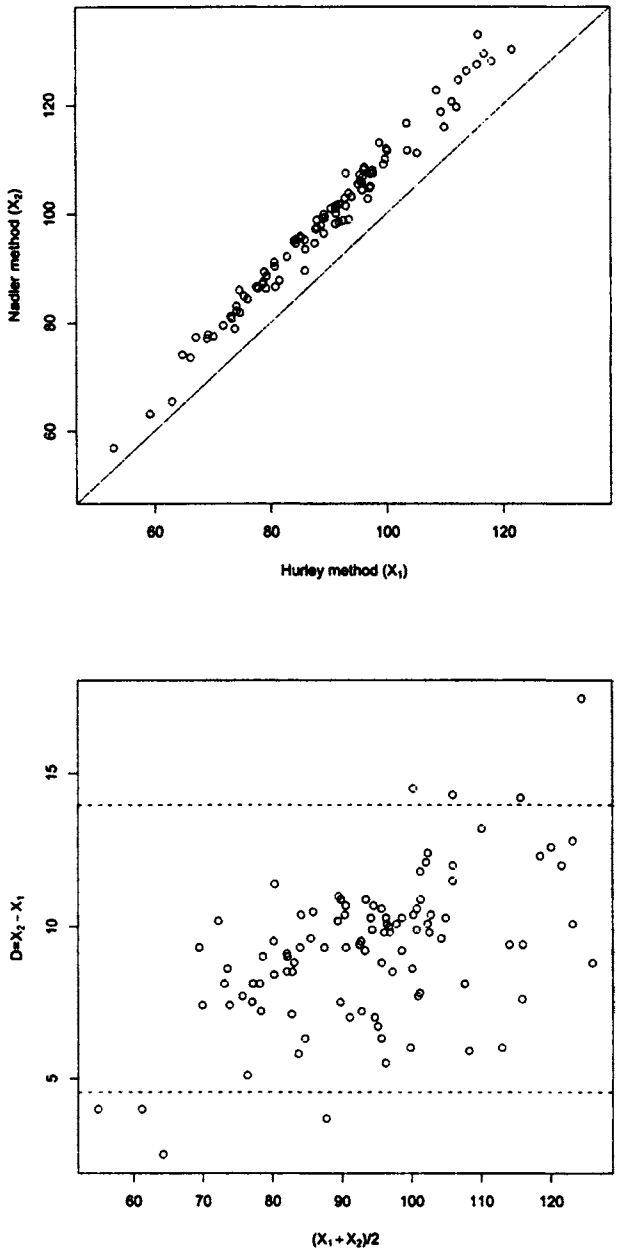

Figure 13.1: The scatter plot and the mean versus difference plot (or the Bland-Altman plot) of the plasma volume measurements

that the higher mean and variance of the Nadler method are the main sources of disagreement.

The various parameter estimates for these data are the following:

$$(\hat{\mu}, \hat{\sigma}) = (9.26, 2.40);\ (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\rho}) = (89.24, 98.50, 13.89, 15.18, 0.99).$$

Table 13.3 presents the estimates of various measures of agreement summarized in Table 13.2 and their CIs when $(\delta_0, \pi_0) = (5, 0.95)$. All the approaches, except perhaps the CCC, confirm that there is substantial disagreement between the methods. The CCC may actually indicate satisfactory agreement since its LCB 0.78 is more than 0.75, a cutoff sometimes suggested for good agreement [see, e.g., Atkinson and Nevill (1997)]. The estimate of CCC is 0.82, which agrees with the estimate of ICC ρ_I (defined in (13.4), not shown in the table) up to two decimal places. This supports Nickerson's (1997) claim that these two quantities tend to be similar in practice. The LCB of ICC ρ_G is small.

From the LOA CIs, we infer that the interval in which the middle 95% of the differences are expected to lie can be as wide as (3.87, 14.66). The TDI approach says that the 95th percentile of the absolute differences can be as large as 19.69, and the CP approach says that the proportion of differences in the threshold interval $(-5, 5)$ can be as low as 2%. The CI for $\mu_2 - \mu_1$ from the two IUTs imply that μ_2 (the mean of Nadler method) is higher because the lower limit of the CI is zero, and this difference can be as high as 9.66. The difference based IUT also suggests that if the methods were considered equivalent, then the differences from $N(9.67, 2.74^2)$ must be considered acceptable. Note that the middle 95% of this distribution is between $9.67 \pm 1.96 \times 2.74 = (4.3, 15.0)$. From the IUT based on (X_1, X_2), we can also infer that σ_2 is higher than σ_1 and the ratio can be up to 1.12, and that the correlation ρ can only be as low as 0.99.

Our investigation thus far indicates that $\rho \approx 1$, $\mu_2 - \mu_1 \approx 9.50$ and $\sigma_2/\sigma_1 \approx 1.10$. Hence, a linear calibration of the Nadler method value (X_2) as $X_2' = aX_2 + b$ may make the methods agree well in the sense that $\mu_2' - \mu_1 \approx 0$ and $\sigma_2'/\sigma_1 \approx 1$ while their correlation remains close to one. Here $\mu_2' = E(X_2')$ and $\sigma_2' = SD(X_2')$. To find a and b, note that

$$\mu_2' - \mu_1 = a\mu_2 + b - \mu_1 \approx 9.50a + b - (1-a)\mu_1;\ \ \sigma_2'/\sigma_1 = a\sigma_2/\sigma_1 \approx 1.10a.$$

This suggests that $a = 1/1.10$ and since $\mu_1 \approx 90$, $b = (90(1.10-1) - 9.50)/1.10 = -0.45 \approx -0.50$. All of the resulting measures of agreement for the recalibrated Nadler method $(X_2' = X_2/1.10 - 0.50)$ now show evidence of excellent agreement. This calibration, however, needs to be validated using an independent experiment.

Table 13.3: Estimates of various measures of agreement and their 95% CIs

Measure	Estimate	95% CI
95% LOA, $\mu \pm 1.96\sigma$	(4.55, 13.97)	$\mu - 1.96\sigma$: (3.87, ∞) $\mu + 1.96\sigma$: ($-\infty$, 14.66)
CCC, ρ_C	0.82	(0.78, 1.00)
ICC, ρ_G	0.68	(0.60, 1.00)
TDI, $q^*(0.95)$	18.84	(0, 19.69)
CP, $\pi(5)$	0.04	(0.02, 1.00)
IUT based on D, (μ, σ)	(9.26, 2.40)	μ: (0, 9.66) σ : (0, 2.73)
IUT based on (X_1, X_2), $(\mu_2 - \mu_1, \sigma_2/\sigma_1, \rho)$	(9.26,1.09, 0.99)	$\mu_2 - \mu_1$: (0, 9.66) σ_2/σ_1: (1.00, 1.12) ρ: (0.99, 1.00)

13.5 Selection Problems in Measuring Agreement

We now discuss a class of method comparison problems where $k\,(\geq 2)$ methods are compared with the gold standard with the goal of either

(a) selecting the best, i.e., the one that agrees most with the gold standard, or

(b) selecting the best when there is evidence that the best agrees sufficiently well with the gold standard.

In the literature, the problem (a) has been discussed in St. Laurent (1998), Hutson, Wilson and Geiser (1998), and Choudhary and Nagaraja (2004d), and the problem (b) in Choudhary and Nagaraja (2004c). We summarize them only for the $k = 2$ case.

Let G and W_i, $i = 1, 2$, represent measurements on a subject by the gold standard and the ith method. Also let D_i be the difference $W_i - G$. We assume that the vector $\mathbf{D} = (D_1, D_2)$ follows a bivariate normal (BVN) distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and non-singular covariance matrix $\Sigma = (\sigma_{ij})_{2\times 2}$. Thus, the vector of squared differences $\mathbf{D}^{(2)} = (D_1^2, D_2^2)$ follows a continuous bivariate distribution with mean $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and non-singular covariance matrix $\Gamma = (\gamma_{ij})_{2\times 2}$, where $\theta_i = \mu_i^2 + \sigma_{ii}$ and $\gamma_{ij} = 2\sigma_{ij}(\sigma_{ij} + 2\mu_i\mu_j)$, $i, j = 1, 2$. Finally, let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) = (\log(\theta_1), \log(\theta_2))$, and define

$$\Psi = (\psi_{ij})_{2\times 2} = \begin{pmatrix} \frac{\gamma_{11}}{\theta_1^2} & \frac{\gamma_{11}}{\theta_1^2} - \frac{\gamma_{12}}{\theta_1\theta_2} \\ \frac{\gamma_{11}}{\theta_1^2} - \frac{\gamma_{12}}{\theta_1\theta_2} & \frac{\gamma_{11}}{\theta_1^2} - 2\frac{\gamma_{12}}{\theta_1\theta_2} + \frac{\gamma_{22}}{\theta_2^2} \end{pmatrix}. \tag{13.20}$$

We take θ_i (or equivalently λ_i) as the measure of agreement between the ith instrument and the gold standard with preference for its smaller values. Thus the goal is to select the component of $\mathbf{D}^{(2)}$ having the smallest mean. In the literature on ranking and selection [see, e.g., Gupta and Panchapakesan (1979), for an excellent introduction], the problem of selecting the component having the smallest (or the largest) mean has been discussed in Mukhopadhyay and Chou (1984) for a k-variate normal population when all the correlations are non-negative. However in our case: (a) the multivariate normal assumption for $\mathbf{D}^{(2)}$ is generally not reasonable — 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 structure for this matrix and hence the standard techniques of multiple comparisons with the best (MCB) [see Hsu (1996, Ch. 4) for an introduction] cannot be directly employed.

We assume each subject is measured only once by the three methods and suppose $\mathbf{D}_l = (D_{1l}, D_{2l})$, $l = 1, 2, \ldots$, is a sequence of i.i.d. observations on $\mathbf{D}$. Let $\hat{\boldsymbol{\mu}}_m = (\hat{\mu}_{1;m}, \hat{\mu}_{2;m})$ and $\hat{\Sigma}_m = (\hat{\sigma}_{ij;m})_{2\times 2}$ denote the usual unbiased estimators of $\boldsymbol{\mu}$ and Σ based on the first m observations on $\mathbf{D}$ (unbiasedness is not necessary for the validity of the large-sample result; it only requires that the estimators be consistent). The estimators $\hat{\boldsymbol{\theta}}_m$, $\hat{\Gamma}_m$, $\hat{\boldsymbol{\lambda}}_m$ and $\hat{\Psi}_m$ are then constructed by plugging-in $\hat{\boldsymbol{\mu}}_m$ and $\hat{\Sigma}_m$ for their population counterparts. We will omit the sample size from the notation of estimators when it is clear from the context.

13.5.1 Selection of the best

St. Laurent (1998) assumes a random effects model $W_i = G + \epsilon_i$ for the measurements, where $\epsilon_i, i = 1, 2$, are correlated random variables with zero means and distributed independently of G. This model assumes that the two instruments and the gold standard have the same means. Hence the equal agreement is equivalent to the equality of variances of ϵ_1 and ϵ_2. For the inference he uses a nonparametric bootstrap CI for the difference of the ICC's of (W_1, G) and (W_2, G). Recall from (13.6) that the ICC for (W_i, G) is $Var(G)/(Var(G) + Var(\epsilon_i))$. This approach is ad hoc and the equality of means assumption cannot always be justified. In addition, we have seen that the ICC is hard to interpret.

Hutson et al. (1998) consider a large sample $100(1-\alpha)\%$ CI for $\theta_1/(\theta_1+\theta_2)$ constructed from a sample of size n. They infer the first (second) instrument as the true best if the upper (lower) bound of this CI is less (greater) than 0.5, and remain indecisive if the interval contains 0.5. They estimate the mean vector and the covariance matrix of $\mathbf{D}^{(2)}$ with their sample counterparts. When bivariate normality for $\mathbf{D}$ is assumed, the large sample $100(1-\alpha)\%$ level CI for

$\theta_1/(\theta_1+\theta_2)$ becomes

$$\frac{\hat{\theta}_1}{(\hat{\theta}_1+\hat{\theta}_2)} \pm \frac{z(\alpha/2)}{n^{1/2}(\hat{\theta}_1+\hat{\theta}_2)^2}\left(\hat{\gamma}_{11}\hat{\theta}_2^2 - 2\hat{\gamma}_{12}\hat{\theta}_1\hat{\theta}_2 + \hat{\gamma}_{22}\hat{\theta}_1^2\right)^{1/2}.$$

Using Monte Carlo simulations, Choudhary (2002) shows that this CI has a substantial under-coverage. Instead, they recommend $\lambda_1 - \lambda_2$, where $\lambda_i = \log(\theta_i)$, as the metric for comparison since the asymptotic procedures based on the estimators of $\lambda_1 - \lambda_2$ tend to do well for samples with sizes as low as 15. For the purpose of selection, they suggest two approximate CI for $\lambda_1 - \lambda_2$ both of which have asymptotic level $1-\alpha$:

$$\left[\hat{\lambda}_1 - \hat{\lambda}_2 - n^{-1/2}t_{n-1}(\alpha/2)\hat{\psi}_{22}^{1/2},\ \hat{\lambda}_1 - \hat{\lambda}_2 + n^{-1/2}t_{n-1}(\alpha/2)\hat{\psi}_{22}^{1/2}\right],$$
$$\left[\min\{0, \hat{\lambda}_1 - \hat{\lambda}_2 - n^{-1/2}t_{n-1}(\alpha)\hat{\psi}_{22}^{1/2}\},\ \max\{0, \hat{\lambda}_1 - \hat{\lambda}_2 + n^{-1/2}t_{n-1}(\alpha)\hat{\psi}_{22}^{1/2}\}\right].$$

Here $\hat{\psi}_{22}$ is the estimate of ψ_{22} in (13.20). Using the first interval, one infers the first instrument to be the best if its upper (lower) bound is negative (positive), and is indecisive otherwise. The second interval is a constrained (to contain zero) MCB CI for $\lambda_1 - \lambda_2$ [see Hsu (1996, ch. 4)]. If the possibility that $\lambda_1 = \lambda_2$ is ruled out, one infers the first (second) instrument as the best if the upper (lower) bound of this interval equals zero. But when this procedure identifies one instrument to be the best, it does not give a negative upper bound on how much better it is when compared to the unselected. On the other hand, the unconstrained interval allows such an inference. However, the advantage of sacrificing this information is that the constrained interval identifies an instrument to be the best more frequently than the unconstrained one at the same asymptotic level. This sharper inference is desirable for us. Further, assuming $\lambda_1 \neq \lambda_2$ is reasonable from a practical viewpoint as it amounts to assuming that the two instruments do not agree *equally* with the gold standard.

In practice, single-stage CI procedures such as the above may fail to distinguish between two instruments. This difficulty can be avoided by using a two-stage procedure. However, for this, the investigator has to pre-specify a threshold δ (>0) such that whenever $|\lambda_1 - \lambda_2| < \delta$, the two instruments are considered *practically equivalent*, and then the *correct* selection is not important. This δ is also known as the *indifference-zone* in the terminology of ranking and selection procedures [see, e.g., Gupta and Panchapakesan (1979)]. Consider the following two-stage procedure:

Stage 1: Select a random sample of size m, compute $\hat{\psi}_{22;m}$, and define

$$N_m = \max\left\{\left\lceil t_{m-1}^2(\alpha)\,\hat{\psi}_{22;m}\,\delta^{-2}\right\rceil, m\right\}, \tag{13.21}$$

where $\hat{\psi}_{22;m}$ is the estimate of ψ_{22} in (13.20).

Stage 2: Take $N_m - m$ additional i.i.d. observations if $N_m > m$ and compute the estimates $\hat{\lambda}_{i;N_m}, i = 1, 2$, using the complete two-stage sample. Then select the instrument that produces the smaller estimate as the best.

Choudhary and Nagaraja (2004d) show that when m is large and $|\lambda_1 - \lambda_2| \geq \delta$, the probability of correct selection with this procedure is approximately $1-\alpha$. When the differences are bivariate normal, $m = 15$ is a reasonable choice for the first-stage sample size.

13.5.2 Assessment of agreement and selection of the best

Above we focussed on the issue of selection of the best when two instruments are compared with a gold standard. However, for a practitioner, knowing the best instrument is unhelpful unless it also agrees sufficiently closely with the gold standard. Often times this information is not available in advance. Assuming $\lambda_1 \neq \lambda_2$, Choudhary and Nagaraja (2004c) address this problem by developing a two-stage procedure that first determines whether the best has satisfactory agreement with the gold standard through a test of

$$H_7 : \lambda_{[1]} \geq \lambda_0 \quad \text{vs} \quad K_7 : \lambda_{[1]} < \lambda_0$$

before proceeding to its selection. Here [1] is the unknown label of the true best among the two instruments and λ_0 is a user-specified cutoff such that $\{\lambda_i < \lambda_0\}$ is the region of satisfactory agreement.

Let δ be a threshold for practical equivalence of $\lambda_{[1]}$ with λ_0, and of λ_1 with λ_2, in the sense that whenever $|\lambda_{[1]} - \lambda_0| < \delta$ or $|\lambda_1 - \lambda_2| < \delta$ the distinction between the two quantities is not important from practical considerations. Finally, define

$$\hat{\lambda}_{(1)} = (\hat{\lambda}_1 - \hat{\lambda}_2) I(\hat{\lambda}_1 \leq \hat{\lambda}_2) + \hat{\lambda}_2, \; \hat{\psi}_{(11)} = \left(\frac{\hat{\gamma}_{11}}{\hat{\theta}_1^2} - \frac{\hat{\gamma}_{22}}{\hat{\theta}_2^2} \right) I(\hat{\lambda}_1 \leq \hat{\lambda}_2) + \frac{\hat{\gamma}_{22}}{\hat{\theta}_2^2},$$

where $I(A)$ is the indicator function of event A, and take

$$\hat{\nu} = \left(\hat{\psi}_{(11)} - \hat{\gamma}_{12}/(\hat{\theta}_1\hat{\theta}_2)\right) / \left(\hat{\psi}_{(11)}\hat{\psi}_{22}\right)^{1/2}.$$

For pre-specified α, β $(0 < \alpha < 1 - \beta < 1)$, λ_0, δ (> 0) and m (≥ 2); Choudhary and Nagaraja (2004c) propose the following two-stage procedure:

Stage 1: Take a random sample of size m and compute the estimates $\hat{\boldsymbol{\mu}}_m$ and $\hat{\Sigma}_m$. Use them to compute $\hat{\psi}_{22;m}$, and $\hat{\psi}_{(11);m}$ and $\hat{\nu}_m$ defined above. Then solve the equation

$$\Phi_2 \left(\frac{L_m^{1/2}\delta}{\hat{\psi}_{(11);m}^{1/2}} - z(\alpha), \frac{L_m^{1/2}\delta}{\hat{\psi}_{22;m}^{1/2}} ; \hat{\nu}_m \right) = 1 - \beta$$

for L_m, and define $N_m = \max\{\lceil L_m \rceil, m\}$ as the second-stage sample size, where $\lceil L_m \rceil$ denotes the smallest integer $\geq L_m$.

Stage 2: Take $N_m - m$ additional i.i.d. observations if $N_m > m$. Compute $\hat{\lambda}_{(1):N_m}$ and $\hat{\psi}_{(11):N_m}$ using the entire sample. Reject H_7 when $\hat{\lambda}_{(1):N_m} + N_m^{-1/2} z(\alpha) \hat{\psi}^{1/2}_{(11):N_m} \leq \lambda_0$. Further, when H_7 is rejected, infer the instrument that produces $\hat{\lambda}_{(1):N_m}$ as the best.

This procedure has the property that, when m is large,

$$Pr(\text{reject } H_7, \text{ correct selection}) \approx 1 - \beta \text{ or more,}$$

whenever $\lambda_0 - \lambda_{[1]} \geq \delta$ and $|\lambda_1 - \lambda_2| \geq \delta$. Using simulation studies to verify the small-sample properties of this procedure, Choudhary and Nagaraja (2004c) suggest $m = 15$ to be a reasonable choice for the first-stage sample size.

13.6 Concluding Remarks

For the assessment of agreement, we assumed that both the measurements are random. To handle the case when the reference measurements are fixed (non-random), some of the procedure presented here has been adapted by Lin et al. (2002). Further, the approaches discussed here are generally not robust to deviations from normality or outliers. When this assumption is a suspect, none of them are valid, but a simple nonparametric sign test can be used [see Bland and Altman (1999)]. Further, this sign test can be inverted to give a nonparametric UCB for $q(\pi_0)$.

Our discussion assumed the simple model (13.1) for the measurements. It can be easily extended to handle replicate measurements. More complicated models such as those including method-subject interactions, covariates or the effect of time on the repeated measurements may be called for as well. See Dunn and Roberts (1999), Bland and Altman (1999), and Chinchilli et al. (1996), for some such models. The CCC approach has been generalized in several directions. Similar extensions of the difference based approaches will be of interest.

Acknowledgments. The second author's research was supported by NIH (NCRR) grant # M01 RR00034 awarded to The Ohio State University.

References

1. Altman, D. G. and Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies, *The Statistician*, **32**, 307–317.

2. Anderson, S. and Hauck, W. W. (1990). Consideration of individual bioequivalence, *Journal of Pharmacokinetics and Biopharmaceutics*, **18**, 259–274.

3. Atkinson, G. and Nevill, A. (1997). Comment on the use of concordance correlation to assess the agreement between two variables, *Biometrics*, **53**, 775–777.

4. Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures, *The Canadian Journal of Statistics*, **27**, 3–23.

5. Barnhart, H. X., Haber, M., and Song, J. L. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers, *Biometrics*, **58**, 1020–1027.

6. Barnhart, H. X. and Williamson, J. M. (2001). Modeling concordance correlation via GEE to evaluate reproducibiltiy, *Biometrics*, **57**, 931–940.

7. Bartko, J. J. (1994). Measures of agreement: A single procedure, *Statistics in Medicine*, **13**, 737–745.

8. Berger, R. L. and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets, *Statistical Science*, **11**, 283–319.

9. Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet*, **i**, 307–310.

10. Bland, J. M. and Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement, *Computers in Biology and Medicine*, **20**, 337–340.

11. Bland, J. M. and Altman, D. G. (1995a). Comparing two methods of clinical measurement: A personal history, *International Journal of Epidemiology*, **24**, S7–S14.

12. Bland, J. M. and Altman, D. G. (1995b). Comparing methods of measurement: Why plotting difference against standard method is misleading, *Lancet*, **346**, 1085–1087.

13. Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies, *Statistical Methods in Medical Research*, **8**, 135–160.

14. Cameron, J. M. (1982). Calibration, In *Encyclopedia of Statistical Sciences*, **1**, John Wiley & Sons, New York, pp. 346–351.

15. Casella, G. and Berger, R. (2002) *Statistical Inference*, 2nd edition, Duxbury Press, Pacific Grove, CA.

16. Chinchilli, V. M., Martel, J. K., Kumanyika, S., and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs, *Biometrics*, **52**, 341–353.

17. Choudhary, P. K. (2002). Assessment of Agreement and Selection of the Best Instrument in Method Comparison Studies, *Ph.D. Dissertation*, The Ohio State University, Columbus, OH.

18. Choudhary, P. K. and Nagaraja, H. N. (2004a). Tests for assessment of agreement using probability criteria, *Submitted for publication.*

19. Choudhary, P. K. and Nagaraja, H. N. (2004b). Assessment of agreement using intersection-union principle, *Biometrical Journal* (to appear).

20. Choudhary, P. K. and Nagaraja, H. N. (2004c). A two-stage procedure for selection and assessment of agreement of the best with a gold standard, *Sequential Analysis* (to appear).

21. Choudhary, P. K. and Nagaraja, H. N. (2004d). Selecting the instrument closest to a gold standard, *Journal of Statistical Planning and Inference* (to appear).

22. David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, Third edition, John Wiley & Sons, New York.

23. Dunn, G. (1989). *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors*, Oxford University Press, New York.

24. Dunn, G. (1992). Design and analysis of reliability studies, *Statistical Methods in Medical Research*, **1**, 123–157.

25. Dunn, G. and Roberts, C. (1999). Modelling method comparison data, *Statistical Methods in Medical Research*, **8**, 161–179.

26. Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, John Wiley & Sons, New York.

27. Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments*, John Wiley & Sons, New York.

28. Fuller, W. A. (1987). *Measurement Error Models*, John Wiley & Sons, New York.

29. Guttman, I. (1988). Statistical tolerance regions, *Encyclopedia of Statistical Sciences*, **9**, pp. 272–287, John Wiley & Sons, New York.

30. Grubbs, F. E. (1982). Grubbs' estimators, In *Encyclopedia of Statistical Sciences*, **2**, pp. 542–549, John Wiley & Sons, New York.

31. Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures - Theory and Methodology of Selecting and Ranking Populations*, John Wiley, New York. Republished by SIAM, Philadelphia, 2002.

32. Hamilton, D. C. and Lesperance, M. L. (1995). A comparison of methods for univariate and multivariate acceptance sampling by variables, *Technometrics*, **37**, 329–339.

33. Harris, I. R., Burch, B. D. and St. Laurent, R. T. (2001). A blended estimator for measure of agreement with a gold standard, *Journal of Agricultural, Biological, and Environmental Statistics*, **6**, 326–339.

34. Hawkins, D. M. (2002). Diagnostics for conformity of paired quantitative measurements, *Statistics in Medicine*, **21**, 1913–1935.

35. Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*, Chapman & Hall/CRC, Boca Raton, FL.

36. Hutson, A. D., Wilson, D. C., and Geiser, E. A. (1998). Measuring relative agreement: Echocardiographer versus computer, *Journal of Agricultural, Biological, and Environmental Statistics*, **3**, 163–174.

37. Kelly, G. E. (1985). Use of structural equations model in assessing the reliability of a new measurement technique, *Applied Statistics*, **34**, 258–263.

38. King, T. S. and Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data, *Statistics in Medicine*, **20**, 2131 - 2147.

39. Kraemer, H. C., Periyakoil, V. S., and Noda, A. (2002). Kappa coefficients in medical research, *Statistics in Medicine*, **21**, 2109–2129.

40. Krummenauer, F. (1999). Intraindividual scale comparison in clinical diagnostic methods: A review of elementary methods, *Biometrical Journal*, **41**, 917–929.

41. Lee, J., Koh, D., and Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable, *Computers in Biology and Medicine*, **19**, 61–70.

42. Lewis, P. A., Jones, P. W., Polak, J. W., and Tillotson, H. T. (1991). The problem of conversion in method comparison studies, *Applied Statistics*, **40**, 105–112.

43. Liao, J. and Lewis, J. (2000). An agreement curve, Presented at the Joint Statistical Meetings, Indianapolis, IN.

44. Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility, *Biometrics*, **45**, 255–268. Corrections: 2000, **56**, 324–325.

45. Lin, L. I. (1992). Assay validation using the concordance correlation coefficient, *Biometrics*, **48**, 599–604.

46. Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence, *Statistics in Medicine*, **19**, 255–270.

47. Lin, L. I. (2003). Measuring agreement. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd edition, pp. 561–567, Marcel Dekker, New York.

48. Lin, L. I. and Chinchilli, V. (1997). Rejoinder to the letter to the editor from Atkinson and Nevill, *Biometrics*, **53**, 777–778.

49. Lin, L. I., Hedayat, A. S., Sinha, B. and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools, *Journal of the American Statistical Association*, **97**, 257–270.

50. Lin, L. I. and Torbeck, L. D. (1998). Coefficient of accuracy and concordance correlation coefficient: New statistics for method comparison, *PDA Journal of Pharmaceutical Science and Technology*, **52**, 55–59.

51. Lin, S. C., Whipple, D. M., and Ho, C. S. (1998). Evaluation of statistical equivalence using limits of agreement and associated sample size calculation, *Communications in Statistics—Theory and Methods*, **27**, 1419–1432.

52. Linnet, K. (1993). Evaluation of regression procedures for method comparison studies, *Clinical Chemistry*, **39**, 424–432.

53. Liu, J.-P. and Chow, S.-C. (1997). A two one-sided tests procedure for assessment of individual bioequivalence, *Journal of Biopharmaceutical Statistics*, **7**, 49–61.

54. McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients, *Psychological Methods*, **1**, 30–46.

55. Mukhopadhyay, N. and Chou, W.-S. (1984). On selecting the best component of a multivariate normal population, *Sequential Analysis*, **3**, 1–22.

56. Müller, R. and Büttner, P. (1994). A critical discussion of intraclass correlation coefficients, *Statistics in Medicine*, **13**, 2465–2476.

57. Nickerson, C. A. (1997). Comment on "A concordance correlation coefficient to evaluate reproducibility", *Biometrics*, **53**, 1503–1507.

58. Nix, A. B. J. and Dunston, F. D. J. (1991). Maximum likelihood techniques applied to method comparison studies, *Statistics in Medicine*, **10**, 981–988.

59. Quan, H. and Shih, W. J. (1996). Assessing reproducibility by the within-subject coefficient of variation with random effects models, *Biometrics*, **52**, 1195–1203. Correspondence, *Biometrics*, **56**, 301–302.

60. Robieson, W. Z. (1999). On the weighted kappa and concordance correlation coefficient, *Ph.D. Dissertation*, University of Illinois at Chicago, IL.

61. Shoukri, M. M. (1999). Measurement of Agreement, In *Encyclopedia of Biostatistics*, **1**, pp. 103–117, John Wiley & Sons, New York.

62. Shoukri, M. M. (2004). *Measures of Interobserver Agreement*, Chapman & Hall/CRC, Boca Raton, FL.

63. St. Laurent, R. T. (1998). Evaluating agreement with a gold standard in method comparison studies, *Biometrics*, **54**, 537–545.

64. Vonesh, E. F., Chinchilli, V. P., and Pu, K. W. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models, *Bometrics*, **52**, 572-587.

65. Wang, W. and Hwang, J. T. G. (2001). A nearly unbiased test for individual bioequivalence problems using probability criteria, *Journal of Statistical Planning and Inference*, **99**, 41–58.