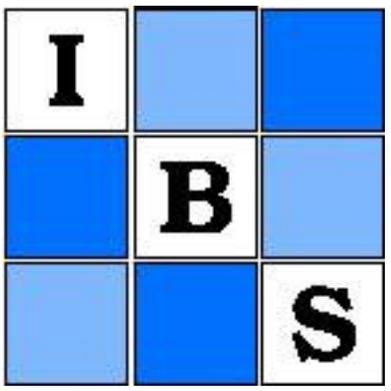

BIOMETRICS 45, 255–268
March 1989

A Concordance Correlation Coefficient to Evaluate Reproducibility

Lawrence I-Kuei Lin

Baxter Healthcare Corporation, Route 120 and Wilson Road,
Round Lake, Illinois 60073, U.S.A.

SUMMARY

A new reproducibility index is developed and studied. This index is the correlation between the two readings that fall on the 45° line through the origin. It is simple to use and possesses desirable properties. The statistical properties of this estimate can be satisfactorily evaluated using an inverse hyperbolic tangent transformation. A Monte Carlo experiment with 5,000 runs was performed to confirm the estimate's validity. An application using actual data is given.

1. Introduction

In an assay validation or an instrument validation process, the reproducibility of the measurements from trial to trial is of interest. Also, when a new assay or instrument is developed, it is of interest to evaluate whether the new assay can reproduce the results based on a traditional gold-standard assay (Westgard and Hunt, 1973; Bauer and Kennedy, 1981). Such validation processes are often evaluated by using the Pearson correlation coefficient, the paired t-test, the least squares analysis of slope (= 1) and intercept (= 0), the coefficient of variation, or the intraclass correlation coefficient. There are drawbacks to all of these, however, in that none alone can fully assess the desired reproducibility characteristics. For example, to evaluate the blood cell counter for hematology analysis in a laboratory, it is desirable to have duplicates of the same blood sample measurement by the counter at different times (usually at most 1 day apart) yield results as close together as possible. If we plot the first measurement against the second measurement of the red blood cell counts for all blood samples available, we would like to see, within a tolerable error, that the measurements fall on a 45° line through the origin (45°). The Pearson correlation coefficient measures a linear relationship but fails to detect any departure from the 45° line (see Figure 1). The paired t-test fails (see Figure 2) to detect poor agreement in pairs of data such as (1, 3), (2, 3), (3, 3), (4, 3), and (5, 3). Combining the above two methods cannot detect poor agreement in pairs of data such as (1, 2.8), (2, 2.9), (3, 3.0), (4, 3.1), and (5, 3.2). The least squares approach fails to detect departure from intercept equal to 0 and slope equal to 1 if data are very scattered (see Figure 3, lower plot). In other words, the more the data are scattered (nonreproducible), the less chance one could reject the hypothesis. The least squares approach can reject a highly reproducible assay due to very small residual error (see Figure 3, upper plot). This is also true if the paired t-test is used (see Figure 2, lower plot). The coefficient of variation and the intraclass correlation coefficient allow duplicate readings to be interchangeable. In other words, these methods consider duplicate readings as replicates (random) rather than two distinct readings. Two

Key words: Accuracy; Asymptotic normality; Concordance correlation coefficient; 45° line through the origin; Precision; Z-transformation.

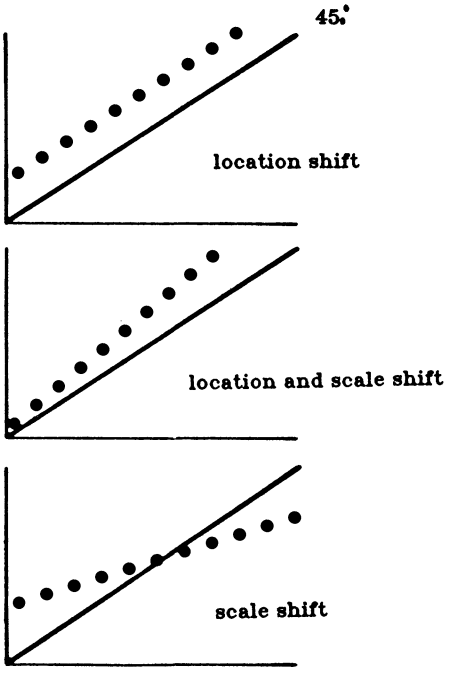

Figure 1. Cases when Pearson correlation coefficient fails to detect nonreproducibility.

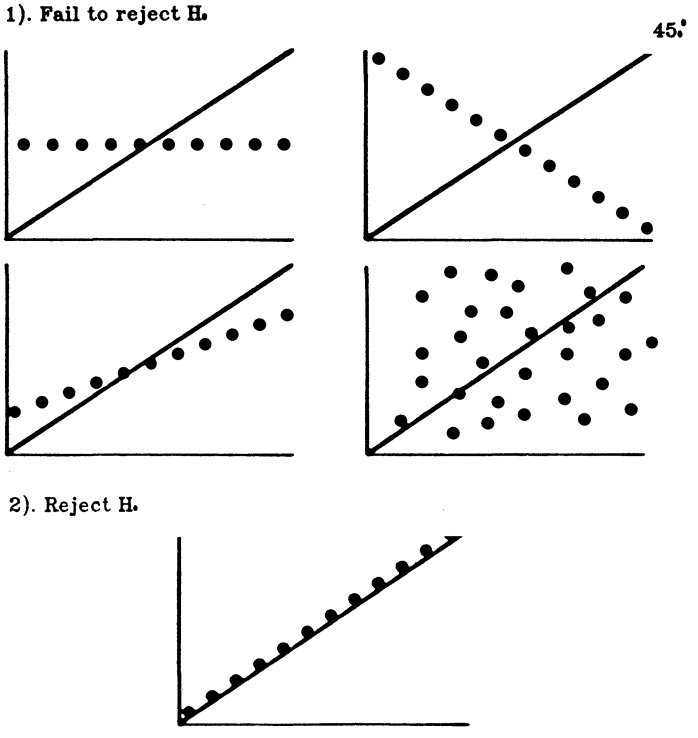

Figure 2. Cases when paired *t*-test can be misleading.
H_0: Means are equal

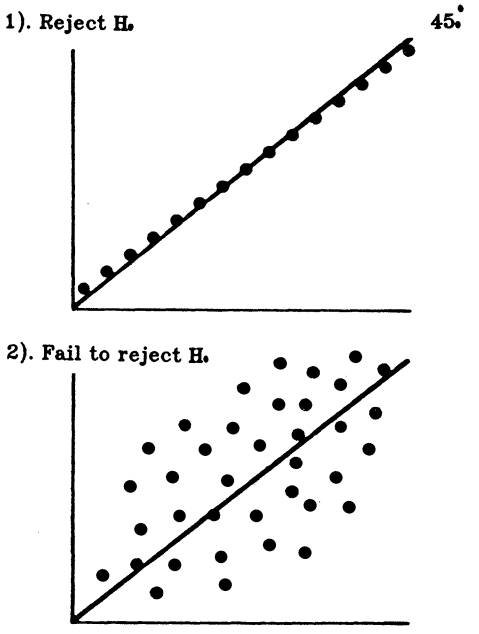

Figure 3. Cases when least squares analysis can be misleading.
H_0: Intercept = 0 and slope = 1

distinct readings would occur in the case of a first reading (earlier) versus a second reading (later) or as a reading of assay A versus a reading of assay B.

This study proposes a desirable reproducibility index, to be called a concordance correlation coefficient, which evaluates the agreement between two readings (from the same sample) by measuring the variation from the 45° line through the origin (the concordance line).

2. The Concordance Correlation Coefficient

Let us assume that pairs of samples (Y_{i1}, Y_{i2}), $i = 1, 2, \ldots, n$, are independently selected from a bivariate population with means μ_1 and μ_2 and covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The degree of concordance between Y_1 and Y_2 can be characterized by the expected value of the squared difference, i.e.,

$$\begin{aligned} \mathrm{E}[(Y_1 - Y_2)^2] &= (\mu_1 - \mu_2)^2 + (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) \\ &= (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2(1 - \rho)\sigma_1\sigma_2, \end{aligned}$$

where ρ is the Pearson correlation coefficient. This also represents the expected squared perpendicular deviation from the 45° line, multiplied by 2.

If each pair, Y_1 and Y_2, in the population are in perfect agreement, $\mathrm{E}[(Y_1 - Y_2)^2]$ would be 0. The following transformation is proposed in order for the value of the index to be

scaled between −1 and 1:

$$\rho_c = 1 - \frac{E[(Y_1 - Y_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}$$

$$= 1 - \frac{\text{Expected squared perpendicular deviation from } 45^\circ_o \text{ line}}{\text{Expected squared perpendicular deviation from } 45^\circ_o \text{ line when } Y_1 \text{ and } Y_2 \text{ are uncorrelated}}$$

or,

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \rho C_b,$$

where

$$C_b = [(v + 1/v + u^2)/2]^{-1},$$

$$v = \sigma_1/\sigma_2 = \text{scale shift},$$

$$u = (\mu_1 - \mu_2)/\sqrt{\sigma_1\sigma_2} = \text{location shift relative to the scale.}$$

Here, $0 < C_b \leq 1$ is a bias correction factor that measures how far the best-fit line deviates from the 45°_o line (measure of accuracy). No deviation from the 45°_o line occurs when $C_b = 1$. The further C_b is from 1, the greater the deviation is from the 45°_o line. The Pearson correlation coefficient ρ measures how far each observation deviated from the best-fit line (measure of precision).

If we let $\beta_1 = (\sigma_1/\sigma_2)\rho$ and $\beta_0 = \mu_1 - \beta_1\mu_2$ represent the regression slope and intercept, respectively, from the conditional distribution of Y_1 given Y_2, then

$$\rho_c = \frac{2\beta_1\sigma_2^2}{(\sigma_1^2 + \sigma_2^2) + [(\beta_0 - 0) + (\beta_1 - 1)\mu_2]^2}.$$

This concordance correlation coefficient, ρ_c, possesses the following characteristics:

(i) $-1 \leq -|\rho| \leq \rho_c \leq |\rho| \leq 1$.
(ii) $\rho_c = 0$ if and only if $\rho = 0$.
(iii) $\rho_c = \rho$ if and only if $\sigma_1 = \sigma_2$ and $\mu_1 = \mu_2$.
(iv) $\rho_c = \pm 1$ if and only if
  (a) $(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2\sigma_1\sigma_2(1 \mp \rho) = 0$, or equivalently,
  (b) $\rho = \pm 1$, $\sigma_1 = \sigma_2$, and $\mu_1 = \mu_2$, or equivalently,
  (c) each pair of readings is in perfect (1) agreement (for example, 1, 1; 2, 2; 3, 3; 4, 4; 5, 5) or in perfect reversed (−1) agreement (for example, 5, 1; 4, 2; 3, 3; 2, 4; 1, 5).

This concordance correlation coefficient evaluates the degree to which pairs fall on the 45°_o line. It contains the measurements of accuracy (C_b) and precision (ρ). Any departure from this line would produce $\rho_c < 1$ even if $\rho = 1$.

For n independent pairs of samples, it is natural to use the sample counterparts; i.e., let

$$\hat{\rho}_c = \frac{2S_{12}}{S_1^2 + S_2^2 + (\bar{Y}_1 - \bar{Y}_2)^2},$$

where

$$\bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}, \quad S_j^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_j)^2, \quad j = 1, 2;$$

and

$$S_{12} = \frac{1}{n}\sum_{i=1}^{n}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2).$$

3. Inference

Let us assume that $\hat{\rho}_c$ is the sample concordance correlation coefficient of paired samples from a bivariate normal distribution. Using the transformation theory of functions of asymptotically normal vectors (Serfling, 1980, p. 122), we can show that $\hat{\rho}_c$ is a consistent estimator of ρ_c and has an asymptotic normal distribution with mean ρ_c and variance (see the Appendix)

$$\sigma^2_{\hat{\rho}_c} = \frac{1}{n-2}\left[(1-\rho^2)\rho_c^2(1-\rho_c^2)/\rho^2 + 4\rho_c^3(1-\rho_c)u^2/\rho - 2\rho_c^4u^4/\rho^2\right]. \tag{1}$$

One can improve the normal approximation by using the inverse hyperbolic tangent transformation (or Z-transformation),

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2}\ln\frac{1+\hat{\rho}_c}{1-\hat{\rho}_c}.$$

This yields a better asymptotic normality with mean $Z = \frac{1}{2}\ln[(1+\rho_c)/(1-\rho_c)]$ and variance (see the Appendix)

$$\sigma^2_{\hat{Z}} = \frac{1}{n-2}\left[\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{4\rho_c^3(1-\rho_c)u^2}{\rho(1-\rho_c^2)^2} - \frac{2\rho_c^4u^4}{\rho^2(1-\rho_c^2)^2}\right]. \tag{2}$$

One can replace the parameters of (1) and (2) with their sample counterparts to derive confidence intervals or to do hypothesis testing. Note that use of the Z-transformation approach when assessing the confidence interval for ρ_c not only bounds the value within the open interval $(-1, 1)$, but also provides a more realistic asymmetric interval.

4. Monte Carlo Simulation

To assess the asymptotic normality, a Monte Carlo simulation was performed for five underlying values of ρ_c with samples sizes of $n = 10$, $n = 20$, and $n = 50$. For each of the 15 situations, 5,000 runs were performed. Paired samples were generated from each of the following bivariate normal distributions (five cases), using Statistical Analysis System (SAS) software:

Case 1. Mean (0, 0) and covariance matrix

$$\begin{pmatrix} 1 & .95 \\ .95 & 1 \end{pmatrix},$$

$\rho_c = .95$ with no difference in location and scale parameters.

Case 2. Mean $(-\sqrt{.1}/2, \sqrt{.1}/2)$ and covariance matrix

$$\begin{pmatrix} 1 & .95 \\ .95 & 1 \end{pmatrix},$$

$\rho_c = .905$ with a slight shift in location parameters.

Case 3. Mean $(-\sqrt{.1}/2, \sqrt{.1}/2)$ and covariance matrix

$$\begin{pmatrix} 1.1^2 & .95 \times 1.1 \times .9 \\ .95 \times 1.1 \times .9 & .9^2 \end{pmatrix},$$

$\rho_c = .887$ with slight differences in both location and scale parameters.

Case 4. Mean $(-\sqrt{.1}/2, \sqrt{.1}/2)$ and covariance matrix

$$\begin{pmatrix} .9^2 & .8 \times .9 \times 1.1 \\ .8 \times .9 \times 1.1 & 1.1^2 \end{pmatrix},$$

$\rho_c = .747$ with slight differences in both location and scale parameters, and with a smaller correlation coefficient.

Case 5. Mean $(-\sqrt{.25}/2, \sqrt{.25}/2)$ and covariance matrix

$$\begin{pmatrix} (\frac{4}{3})^2 & .5 \times \frac{4}{3} \times \frac{2}{3} \\ .5 \times \frac{4}{3} \times \frac{2}{3} & (\frac{2}{3})^2 \end{pmatrix},$$

$\rho_c = .360$ with large differences in both location and scale parameters, and with correlation coefficient .5.

In each run, $\hat{\rho}_c$, $\hat{Z}$, and their standard errors based on the sample counterparts of (1) and (2) were calculated. The mean, standard deviation (Std), test statistic value (D), and P-value for normality based on 5,000 runs are reported in Table 1. The test statistic, based on a Kolmogorov goodness-of-fit test, is the greatest distance between the observed cumulative density function and the normal cumulative density function. The asymptotic variances of $\hat{\rho}_c$ were right on the target when sample sizes were 20 or larger. They tended to underestimate in cases 1, 2, and 3 when sample sizes were 10. For this reason, 200 bootstrap (Efron, 1979) samples were used for each run to estimate $\hat{\rho}_c$ and $\hat{Z}$ when $n = 10$. The bootstrap estimates of the standard errors were calculated for each run by taking the standard deviation of those 200 $\hat{\rho}_c$'s and $\hat{Z}$'s. The means of the estimators based on 5,000 runs are reported in Table 2. The "bootstrapped" $S_{\hat{\rho}_c}$ tended to overestimate in cases 1, 2, and 3, and to underestimate in case 5. Based on the value of D in Table 1, the distribution of $\hat{\rho}_c$ was closer to normality when ρ_c was closer to 0, and/or when n became larger.

The asymptotic standard error of $\hat{Z}$ using sample counterparts of (2) was very close to the true standard error in each of the 15 situations in this study. The bootstrap estimates tended to underestimate when $n = 10$ and were very close to the value of the square root of expression (2) (using sample counterparts) multiplied by $\sqrt{(n-2)/n}$. In Table 1, the distribution of $\hat{Z}$ was much closer to normality than that of $\hat{\rho}_c$ since the D value of $\hat{Z}$ was much smaller in each of the 15 situations. The normality hypothesis was not rejected, even with 5,000 runs, except for case 4, $n = 20$ and case 5, $n = 10$. Even for these two situations, the distributions were sufficiently close to normality for practical purposes. The largest D value of $\hat{Z}$, among the 15 situations considered, was .017.

The asymptotic normality of $\hat{\rho}_c$ and $\hat{Z}$ for data from nonnormal distributions was also examined. Additional Monte Carlo simulations were performed, using SAS software, for data from the uniform and Poisson distributions. Paired samples were generated for cases 1, 3, and 5 from the uniform distribution after standardization. Standardization refers to uniform variates on the interval (0, 1) minus .5, multiplied by $\sqrt{12}$. Paired samples were also generated for the same cases from the Poisson distribution after standardization. Standardization here refers to Poisson variates with mean 9 minus 9, divided by 3. The sample sizes used were $n = 10$ and $n = 50$. The results are provided in Table 3 for the uniform distribution and in Table 4 for the Poisson distribution.

Table 1
Mean, standard deviation (Std), and test of normality based on simulation of 5,000 runs generated from normal distribution

	$n = 10$				$n = 20$				$n = 50$			
			Normality				Normality				Normality	
	Mean	Std	D	P-value	Mean	Std	D	P-value	Mean	Std	D	P-value
Case 1. $\rho_c = .950$; $Z = 1.832$												
$\hat{\rho}_c$	.937	.051	.136	<.01	.942	.027	.092	<.01	.947	.015	.052	<.01
$S_{\hat{\rho}_c}$	.044				.026				.015			
$\hat{Z}$	1.782	.344	.011	.119	1.807	.231	.008	>.15	1.822	.143	.007	>.15
$S_{\hat{Z}}$	.341				.233				.144			
Case 2. $\rho_c = .905$; $Z = 1.499$												
$\hat{\rho}_c$	.874	.078	.111	<.01	.891	.045	.078	<.01	.900	.025	.050	<.01
$S_{\hat{\rho}_c}$	.069				.044				.026			
$\hat{Z}$	1.434	.312	.012	.070	1.468	.212	.010	>.15	1.484	.130	.008	>.15
$S_{\hat{Z}}$	.310				.218				.137			
Case 3. $\rho_c = .887$; $Z = 1.408$												
$\hat{\rho}_c$	.855	.080	.101	<.01	.873	.047	.076	.01	.882	.027	.039	<.01
$S_{\hat{\rho}_c}$	.075				.048				.029			
$\hat{Z}$	1.344	.286	.007	>.15	1.378	.193	.011	>.15	1.397	.119	.008	>.15
$S_{\hat{Z}}$	.293				.205				.129			
Case 4. $\rho_c = .747$; $Z = .966$												
$\hat{\rho}_c$	.699	.153	.082	<.01	.723	.099	.056	<.01	.738	.060	.038	<.01
$S_{\hat{\rho}_c}$	.149				.101				.062			
$\hat{Z}$	.929	.307	.011	.139	.945	.210	.013	.044	.958	.131	.008	>.15
$S_{\hat{Z}}$	.310				.218				.137			
Case 5. $\rho_c = .360$; $Z = .377$												
$\hat{\rho}_c$	.323	.200	.021	<.01	.338	.136	.020	<.01	.353	.088	.014	<.01
$S_{\hat{\rho}_c}$	.195				.140				.089			
$\hat{Z}$	.352	.236	.017	<.01	.360	.158	.008	>.15	.372	.101	.011	.147
$S_{\hat{Z}}$	.231				.162				.102			

Table 2
Actual and bootstrapped estimation of $\sigma_{\hat{\rho}_c}$ and $\sigma_{\hat{Z}}$ ($n = 10$) based on 5,000 runs

Case		Bootstrapped	Actual
1	$S_{\hat{\rho}_c}$	.063	.051
	$S_{\hat{Z}}$	.314	.344
2	$S_{\hat{\rho}_c}$	.087	.078
	$S_{\hat{Z}}$	.284	.312
3	$S_{\hat{\rho}_c}$	.090	.080
	$S_{\hat{Z}}$	.268	.286
4	$S_{\hat{\rho}_c}$	.151	.153
	$S_{\hat{Z}}$	.280	.307
5	$S_{\hat{\rho}_c}$	.172	.200
	$S_{\hat{Z}}$	.207	.236

Table 3
Mean, standard deviation (Std), *and test of normality based on simulation of* 5,000 *runs from the uniform distribution*

	$n = 10$				$n = 50$			
			Normality				Normality	
	Mean	Std	D	P-Value	Mean	Std	D	P-Value
	Case 1. $\rho_c = .950$; $Z = 1.832$							
$\hat{\rho}_c$	.930	.052	.132	<.01	.947	.015	.052	<.01
$S_{\hat{\rho}_c}$	.046				.015			
$\hat{Z}$	1.768	.352	.010	>.15	1.821	.142	.007	>.15
$S_{\hat{Z}}$	.350				.145			
	Case 3. $\rho_c = .887$; $Z = 1.408$							
$\hat{\rho}_c$	.857	.081	.102	<.01	.883	.026	.042	<.01
$S_{\hat{\rho}_c}$	.076				.029			
$\hat{Z}$	1.349	.283	.013	.054	1.400	.115	.011	>.15
$S_{\hat{Z}}$	.299				.130			
	Case 5. $\rho_c = .360$; $Z = .377$							
$\hat{\rho}_c$	.315	.200	.030	<.01	.352	.085	.018	<.01
$S_{\hat{\rho}_c}$	.199				.089			
$\hat{Z}$	.342	.231	.017	<.01	.371	.098	.008	>.15
$S_{\hat{Z}}$	.233				.102			

Table 4
Mean, standard deviation (Std), *and test of normality based on simulation of* 5,000 *runs from the Poisson distribution*

	$n = 10$				$n = 50$			
			Normality				Normality	
	Mean	Std	D	P-Value	Mean	Std	D	P-Value
	Case 1. $\rho_c = .950$; $Z = 1.832$							
$\hat{\rho}_c$	.932	.050	.129	<.01	.947	.015	.059	<.01
$S_{\hat{\rho}_c}$	.044				.015			
$\hat{Z}$	1.779	.344	.006	>.15	1.821	.145	.009	>.15
$S_{\hat{Z}}$	.341				.144			
	Case 3. $\rho_c = .887$; $Z = 1.408$							
$\hat{\rho}_c$	.856	.080	.107	<.01	.882	.026	.048	<.01
$S_{\hat{\rho}_c}$	.074				.028			
$\hat{Z}$	1.346	.279	.013	.037	1.397	.116	.010	>.15
$S_{\hat{Z}}$	.288				.129			
	Case 5. $\rho_c = .360$; $Z = .377$							
$\hat{\rho}_c$	.315	.196	.024	<.01	.353	.085	.017	<.01
$S_{\hat{\rho}_c}$	.193				.088			
$\hat{Z}$	.342	.229	.018	<.01	.372	.098	.006	>.15
$S_{\hat{Z}}$	.227				.102			

The asymptotic normality of $\hat{\rho}_c$ and $\hat{Z}$ for samples from the uniform (short-tailed, symmetric) and Poisson (long-tailed, asymmetric to the right) distributions is very similar to that for samples from the normal distribution. These results demonstrate that $\hat{Z}$ is robust for samples from uniform and Poisson distributions. A very encouraging result is that it is robust even with a sample size of 10.

5. Examples

Two examples are considered that investigate the following two questions:

(1) Can a "Portable $ave" machine (actual name withheld) reproduce a gold-standard machine in measuring total bilirubin in blood?
(2) Can an *in-vitro* assay for screening the toxicity of biomaterials reproduce from trial to trial? [This example uses part of the data from Johnson et al. (1985). The two assays in this example are the best and the worst in terms of reproducibility, respectively, in that data set.]

To study the first question, blood samples were taken from 10 animals, one sample per animal. Each sample was measured for bilirubin level by three operators using a Portable $ave machine and a gold-standard Hitachi machine. The salesman of the Portable $ave machine claimed that it can reproduce the results measured by the Hitachi machine, and that "anyone on your staff can run tests." Therefore, this study was conducted at Baxter by three operators. One was a well-trained medical technician in the laboratory (denoted by M.T.); one was a staff nurse who represented the targeted personnel to operate this machine with minimal training (denoted by Nurse); and one was a temporary summer trainee with no training who was the son of a vice president of the laboratory (denoted by S-VP).

The results are plotted in Figure 4. This plot indicates that the Portable $ave machine is reproducible if it is performed by well-trained personnel (symbol M), but not by personnel with minimum or no training (symbols N and S). Table 5 presents the statistical outcome using a variety of different analyses. The first two rows present $\hat{\rho}_c$ and a 95% confidence interval for ρ_c for each operator. These figures clearly reflect high reproducibility performed by the M.T. (.995), mediocre reproducibility by the S-VP (.838), and poor reproducibility by the nurse (.624). For more detailed information, the next three rows show excellent

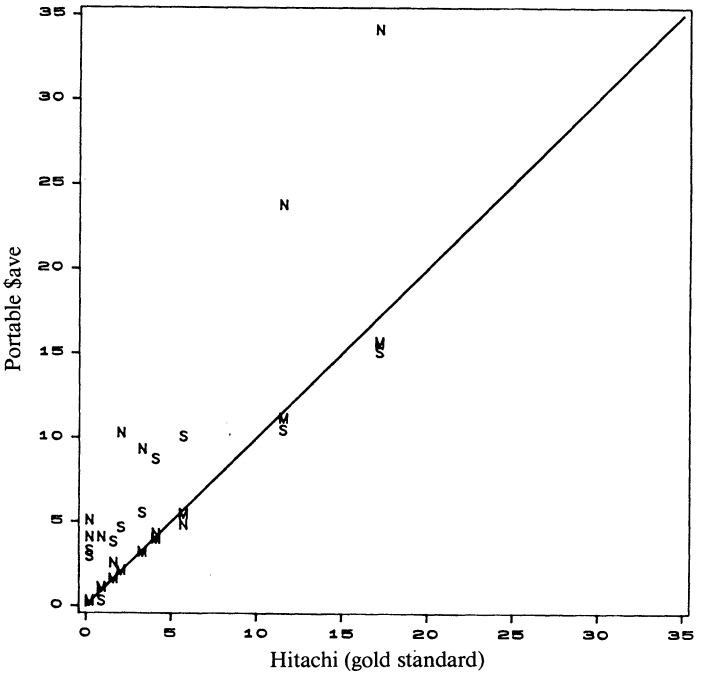

Figure 4. Portable $ave versus Hitachi by three operators for measuring bilirubin in blood. M: M.T.; N: Nurse; S: S-VP

Table 5
Statistical analyses to evaluate the reproducibility of "Portable $ave" machine by 3 operators

Method[a]	M.T.	Nurse	S-VP
$\hat{\rho}_c$	.995	.624	.838
95% CI	(.988, .998)	(.292, .822)	(.520, .952)
$\hat{u}$	−.041	.768	.377
$\hat{v}$	.916	1.878	.805
r	.999	.936	.917
P_{pt}	.227	.012	.037
P_{ls}	<.0001	.002	.017
CV	7.98%	72.8%	35.8%
r_i	.995	.617	.848

[a] $\hat{\rho}_c = r[(\hat{v} + 1/\hat{v} + \hat{u}^2)/2]^{-1}$.
95% CI: confidence interval for ρ_c using Z-transformation.
$\hat{u} = (\bar{Y}_1 - \bar{Y}_2)/\sqrt{S_1 S_2}$.
$\hat{v} = S_1/S_2$.
r = Pearson correlation coefficient.
P_{pt} = P-value using paired t-test.
P_{ls} = P-value using least squares analysis.
CV = Coefficient of variation.
r_i = Intraclass correlation coefficient.

precision (r = .999), with minimum location shift ($\hat{u}$ = −.041) and scale shift ($\hat{v}$ = .916) by the M.T. Also shown in these three rows are moderate precision (r = .936), with poor location shift ($\hat{u}$ = .768) and scale shift ($\hat{v}$ = 1.878) by the nurse. The laboratory refused to purchase this machine because it worked well only when it was used by well-trained personnel, contradicting the claim made by the salesman.

If one chose to use the paired t-test (P_{pt} in Table 5), luckily one would draw the correct conclusions in this example. If one chose to use the least squares analysis (P_{ls} in Table 5), the hypothesis of zero intercept and unit slope by the M.T. would be strongly rejected (P < .0001), much more so than that by the nurse (P = .002) and by the S-VP (P = .017), due to a near-zero residual error. The least squares analysis results contradict intuition. The smaller the residual error (more precision), the more likely one would reject zero intercept and unit slope. On the other hand, the larger the residual error (less precision), the less likely one would reject.

The coefficient of variation (CV in Table 5) works well in this example, although one might decide that a CV of 7.98% (by the M.T.) is not acceptable (greater than 5%). The intraclass correlation coefficient (r_i in Table 5) works very well and is nearly identical to $\hat{\rho}_c$ in this example. This coefficient will give results similar to those for $\hat{\rho}_c$ most of the time. However, it will give a negative value when the paired readings are uncorrelated. It cannot distinguish bias from imprecision, which can be characterized by $\hat{u}$, $\hat{v}$, and r when $\hat{\rho}_c$ is used.

In the second example, 10 materials, varying from nontoxic to highly toxic, were evaluated by two biochemical *in-vitro* assays. The two assays were cellular adenosine triphosphate activity using cell line 76 (ATP-76), and cellular adhesion using cell line 74 (CLA-74). These two assays are commonly used in screening the toxicity of materials for use in medical devices. The results are expressed as percent cell function. The lower the value, the higher the toxicity. Two independent trials, approximately 1 week apart, were performed. The purpose was to assess the reproducibility between trials of each assay for material screening. The data are plotted in Figure 5.

Figure 5 shows that ATP-76 had much better agreement than CLA-74 between the two trials. Table 6 presents the statistical outcome. $\hat{\rho}_c$ clearly demonstrates good trial-to-trial

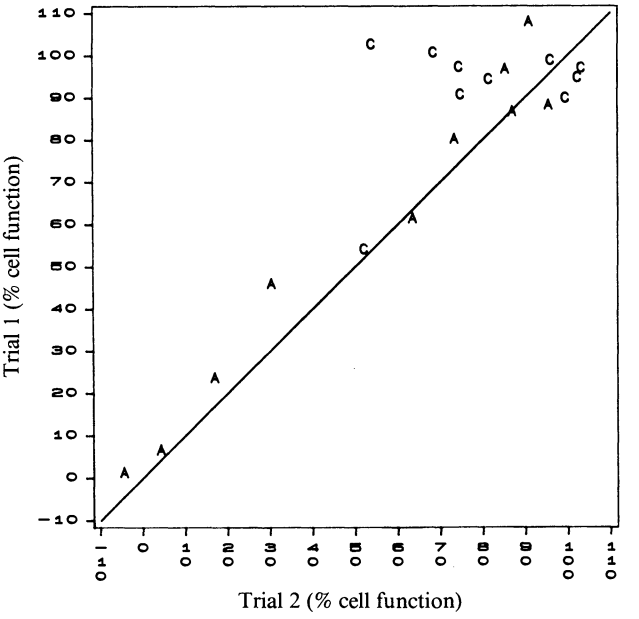

Figure 5. Trial-to-trial reproducibility of two *in-vitro* assays. A: ATP-76; C: CLA-74

Table 6
Statistical analyses to evaluate the trial-to-trial reproducbility by ATP-76 and CLA-74 in-vitro assays

Method[a]	ATP-76	CLA-74
$\hat{\rho}_c$	.969	.283
95% CI	(.874, .992)	(−.245, .681)
$\hat{u}$	.153	.743
$\hat{v}$	.998	.723
r	.980	.376
P_{pt}	.049	.087
P_{ls}	.156	.151
CV	11.46%	17.57%
r_i	.971	.250

[a] $\hat{\rho}_c = r[(\hat{v} + 1/\hat{v} + \hat{u}^2)/2]^{-1}$.
95% CI: 95% confidence interval for ρ_c using Z-transformation.
$\hat{u} = (\bar{Y}_1 - \bar{Y}_2)/\sqrt{S_1 S_2}$.
$\hat{v} = S_1/S_2$.
r = Pearson correlation coefficient.
P_{pt} = P-value using paired t-test.
P_{ls} = P-value using least squares analysis.
CV = Coefficient of variation.
r_i = Intraclass correlation coefficient.

reproducibility for ATP-76 ($\hat{\rho}_c$ = .969) and poor reproducibility for CLA-74 ($\hat{\rho}_c$ = .283 with a 95% CI containing 0). The paired t-test marginally rejected the reproducibility of ATP-76 (P = .049) while it failed to reject the reproducibility of CLA-74 (P = .087). The least squares analysis rejects neither. Due to a higher mean value of CLA-74 compared with that for ATP-76, the coefficient of variation fails to reflect the poorer reproducibility of CLA-74 (17.6%). The intraclass correlation coefficient works well in this example.

6. Conclusion

The concordance correlation coefficient, which is used to evaluate the agreement between paired readings, has desirable characteristics. It is simple to use. Its estimate using the sample counterparts is consistent and has asymptotic normality for bivariate normal data. However, its statistical properties (consistency and asymptotic normality) can be much improved by using the inverse hyperbolic tangent transformation (Z-transformation). It is also robust against samples from the uniform and Poisson distributions even with small sample sizes.

7. Future Studies

This index can be generalized to evaluate agreements among more than two readings. The multiple-reading counterpart is

$$\rho_c = \frac{2 \sum_{i=1}^{p} \sum_{j=1}^{p} \sigma_{ij}}{\sum_{i=1}^{p} \sum_{j=1}^{p} (\mu_i - \mu_j)^2 + 2 \sum_{i=1}^{p} \sigma_i^2}, \quad p > 2.$$

This index may also be a good statistic for use in goodness-of-fit tests. For example, to test for normality, one can measure the agreement between the cumulative density function versus the cumulative normal density function through this index, rather than taking the maximum deviation, as in the Kolmogorov test.

This index can also be applied when Y_1 is random and Y_2 is fixed, in which case the standard error has a simpler structure. The index can be used to characterize agreement between the observed measurements and the theoretical (expected) values. These possibilities are under investigation.

Acknowledgements

I would like to express my sincere gratitude to the editor and referees for their valuable comments, to Laurene Strauch for her assistance in preparing this manuscript, and to my fellow statisticians at Baxter Healthcare Corp. for their support and comments.

Résumé

On présente et on étudie un nouvel indice de reproductibilité. Cet indice est la corrélation entre les deux lectures qui tombent sur la première bissectrice (45 degrés). It est simple à utiliser et a les propriétés souhaitées. Les propriétés statistiques de cette estimation peuvent être évaluées de façon satisfaisante en utilisant la transformation de l'arctangente hyperbolique. On a fait une simulation de Monte Carlo avec 5,000 tirages pour confirmer la validité de l'estimation. On donne une application utilisant des données réelles.

References

Bauer, S. and Kennedy, J. W. (1981). Applied statistics for the clinical laboratory: II. Within-run imprecision. *The Journal of Clinical Laboratory Automation* **1,** 197–201.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7,** 1–26.

Johnson, H. J., Northup, S. J., Seagraves, P. A., Atallah, M., Garvin, P. J., Lin, L., and Darby, T. D. (1985). Biocompatibility test procedure for materials evaluation *in vitro*. II. Objective methods of toxicity assessment. *Journal of Biomedical Materials Research* **19,** 489–508.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Westgard, J. O. and Hunt, M. R. (1973). Use and interpretation of common statistical tests on method-comparison studies. *Clinical Chemistry* **19,** 49–57.

Received February 1986; *revised June and September* 1987.

APPENDIX

Let $(Y_{11}, Y_{12}), \ldots, (Y_{n1}, Y_{n2})$ be independent observations on a bivariate normal distribution. The concordance correlation coefficient is

$$\rho_c = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}.$$

The Z-transformation of ρ_c is $Z = \frac{1}{2}\ln[(1 + \rho_c)/(1 - \rho_c)] = \tanh^{-1}\rho_c$.

The sample analogues are

$$\hat{\rho}_c = \frac{2S_{12}}{S_1^2 + S_2^2 + (\bar{Y}_1 - \bar{Y}_2)^2}$$

and

$$\hat{Z} = \tanh^{-1}\hat{\rho}_c.$$

We first consider the asymptotic normality of $\hat{Z}$. It will then be straightforward to obtain the asymptotic normality of $\hat{\rho}_c$ by simply applying the hyperbolic tangent transformation of $\hat{Z}$. The asymptotic normality of $\hat{Z}$ is algebraically less cumbersome. The transformation may be expressed as $\hat{Z} = g(\mathbf{v})$, where

$$\mathbf{v} = (v_1, v_2, v_3, v_4, v_5) = \left(\bar{Y}_1, \bar{Y}_2, \frac{1}{n}\sum Y_{i1}^2, \frac{1}{n}\sum Y_{i2}^2, \frac{1}{n}\sum Y_{i1}Y_{i2}\right)$$

and

$$g(v_1, v_2, v_3, v_4, v_5) = \frac{1}{4}\ln\left[1 + \frac{4(v_5 - v_1v_2)}{v_3 + v_4 - 2v_5}\right].$$

The vector $\mathbf{v}$ is expressed as functions of sample moments and has asymptotic 5-variate normality with mean $\mathrm{E}(\mathbf{v}) = (\mu_1, \mu_2, \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, \sigma_{12} + \mu_1\mu_2)$ and variance $n^{-1}\,\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = \{W_{ij}\}_{5\times 5},$$

$$W_{11} = \sigma_1^2, \quad W_{12} = W_{21} = \sigma_{12}, \quad W_{22} = \sigma_2^2,$$

$$W_{13} = W_{31} = 2\mu_1\sigma_1^2, \quad W_{23} = W_{32} = 2\mu_1\sigma_{12}, \quad W_{33} = 2\sigma_1^4 + 4\sigma_1^2\mu_1^2,$$

$$W_{14} = W_{41} = 2\mu_2\sigma_{12}, \quad W_{24} = W_{42} = 2\mu_2\sigma_2^2,$$

$$W_{34} = W_{43} = 2\sigma_{12}^2 + 4\mu_1\mu_2\sigma_{12}, \quad W_{44} = 2\sigma_2^4 + 4\sigma_2^2\mu_2^2,$$

$$W_{15} = W_{51} = \mu_2\sigma_1^2 + \mu_1\sigma_{12}, \quad W_{25} = W_{52} = \mu_1\sigma_2^2 + \mu_2\sigma_{12},$$

$$W_{35} = W_{53} = 2\sigma_{12}\mu_1^2 + 2\sigma_{12}\sigma_1^2 + 2\mu_1\mu_2\sigma_1^2,$$

$$W_{45} = W_{54} = 2\sigma_{12}\mu_2^2 + 2\sigma_{12}\sigma_2^2 + 2\mu_1\mu_2\sigma_2^2,$$

$$W_{55} = \sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_{12}^2 + 2\mu_1\mu_2\sigma_{12},$$

which is the covariance matrix of $(Y_1, Y_2, Y_1^2, Y_2^2, Y_1Y_2)$. It follows from the theory on functions of asymptotically normal vectors (Serfling, 1980, Corollary 3.3) that $\hat{Z}$ is asymptotically normal with mean Z and variance $n^{-1}\mathbf{d}\boldsymbol{\Sigma}\mathbf{d}'$, where

$$\mathbf{d} = \left(\left.\frac{\partial g}{\partial v_1}\right|_{\mathbf{v}=\mathrm{E}(\mathbf{v})}, \ldots, \left.\frac{\partial g}{\partial v_5}\right|_{\mathbf{v}=\mathrm{E}(\mathbf{v})}\right).$$

The elements of **d** are

$$d_1 = \left.\frac{\partial g}{\partial v_1}\right|_{\mathbf{v}=\mathrm{E}(\mathbf{v})} = -\frac{\mu_2}{\sigma_1^2 + \sigma_2^2 + 2\sigma_{12} + (\mu_1 - \mu_2)^2};$$

$$d_2 = \left.\frac{\partial g}{\partial v_2}\right|_{\mathbf{v}=\mathrm{E}(\mathbf{v})} = -\frac{\mu_1}{\sigma_1^2 + \sigma_2^2 + 2\sigma_{12} + (\mu_1 - \mu_2)^2};$$

$$d_3 = d_4 = \left.\frac{\partial g}{\partial v_3}\right|_{\mathbf{v}=\mathrm{E}(\mathbf{v})} = \left.\frac{\partial g}{\partial v_4}\right|_{\mathbf{v}=\mathrm{E}(\mathbf{v})}$$

$$= \frac{-\sigma_{12}}{[\sigma_1^2 + \sigma_2^2 + 2\sigma_{12} + (\mu_1 - \mu_2)^2][\sigma_1^2 + \sigma_2^2 - 2\sigma_{12} + (\mu_1 - \mu_2)^2]};$$

$$d_5 = \left.\frac{\partial g}{\partial v_5}\right|_{\mathbf{v}=\mathrm{E}(\mathbf{v})}$$

$$= \frac{(\sigma_1^2 + \sigma_2^2) + (\mu_1 - \mu_2)^2}{[\sigma_1^2 + \sigma_2^2 + 2\sigma_{12} + (\mu_1 - \mu_2)^2][\sigma_1^2 + \sigma_2^2 - 2\sigma_{12} + (\mu_1 - \mu_2)^2]}.$$

After straightforward but tedious algebraic calculation, it can be shown that the variance of $\hat{Z}$ is

$$\sigma_{\hat{Z}}^2 = \frac{1}{n}\mathbf{d}\boldsymbol{\Sigma}\mathbf{d}'$$

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

One may substitute in the above expressions sample counterparts that are consistent estimates. One may replace the n with $n - 2$ for less bias, as evident in the Monte Carlo study in this paper. The asymptotic normality of $\hat{\rho}_c$ can be obtained by letting

$$\hat{\rho}_c = \tanh(\hat{Z}).$$

From the theory on functions of asymptotically normal statistics (Serfling, 1980, Theorem 3.1), $\hat{\rho}_c$ is asymptotically normal with mean ρ_c and variance

$$(1 - \rho_c^2)^2\sigma_{\hat{Z}}^2.$$

This asymptotic variance depends on the parameter ρ_c much more than $\sigma_{\hat{Z}}^2$ does. Like Fisher's Z-transformation on the sample correlation coefficient, the $\hat{Z}$ approaches normality much more rapidly than the $\hat{\rho}_c$, as confirmed by the Monte Carlo study in this paper.