

Indices of Repeatability, Reproducibility, and Agreement

April 26, 2013

Document History:

- April 16, 2013: added metrics and basic study designs of repeatability, reproducibility, and agreement, along with notations.
- April 26, 2013: added notation for data, estimators for each of the various metrics for repeatability, and provided references.

Repeatability

Each patient in the study undergoes two or more scans in quick succession, under identical conditions (i.e. scans acquired on the exact same device or identical devices, acquisition protocol and parameters held constant, no treatment administered in between scans).

Notation:

- σ^2 : the variability attributed to measurement error, namely the variance of repeat measurements on the same patient (assumed to be constant across all patients in the population of interest).
- ξ_i : the average value of repeat measurements on the i^{th} patient (will vary across patients).
- τ^2 : the between-patient variability, namely the variance of the average measurements across patients (i.e. the variance of the ξ_i .)
- μ : the grand average measurements, namely the average measurement across both patients and repeat measurements between patients (i.e. the mean of the ξ_i .)

Index	Definition and Formula	Range	Interpretation	Comments
Repeatability Coefficient (RC)	$1.96\sqrt{2\sigma^2}$	$[0, \infty)$	The value under which the difference between any two repeat measurements on the same patient acquired under identical conditions should fall with 95% probability.	Not useful for comparing repeatability of measurands of differing units.
Within-Case Coefficient of Variance (wCV)	σ/μ	$[0, \infty)$	Variation in repeat measurements on the same patient relative to typical measurement values.	Only meaningful when measurement values are positive.

Intra-class Correlation Coefficient (ICC)	$\frac{\tau^2}{\tau^2 + \sigma^2}$	[0, 1]	Proportion of total variation in measurements explained by between-patient differences rather than variation in repeat measurements for the same patient.	Not appropriate for comparing repeatability across different patient populations.
---	------------------------------------	--------	---	---

Data: N patients each undergo p scans repeated under identical conditions and using the same data acquisition protocol and parameter values, with no intervening treatment between scans.

- Y_{ij} : the value of the j^{th} measurement for the i^{th} patient.

Inference: Bland and Altman (1986) [[REF: Bland and Altman 1986, “Statistical methods for assessing agreement between two methods of clinical measurement”]] suggest using one-way analysis of variance (ANOVA) to estimate the RC. The ANOVA estimator is $\widehat{RC} = 1.96\sqrt{2\hat{\sigma}^2}$, with

$$\hat{\sigma}^2 = \frac{1}{N(p-1)} \sum_{i=1}^N \sum_{j=1}^p (Y_{ij} - \bar{Y}_{i.})^2$$

where $\bar{Y}_{i.} = \frac{1}{p} \sum_{j=1}^p Y_{ij}$ is the mean of the repeat measurements for the i^{th} patient. An exact 95% confidence interval for the RC is

$$\left(\widehat{RC} \sqrt{\frac{N(p-1)}{\chi_{N(p-1),0.975}^2}}, \widehat{RC} \sqrt{\frac{N(p-1)}{\chi_{N(p-1),0.025}^2}} \right)$$

where $\chi_{N(p-1),0.025}^2$ and $\chi_{N(p-1),0.975}^2$ are the 2.5th and 97.5th percentiles of a chi-square distribution with $N(p-1)$ degrees of freedom. We then may examine whether this confidence interval lies strictly below some pre-determined scientifically relevant threshold.

We can also construct estimators of the wCV and ICC based on one-way ANOVA. For wCV, the estimator is $\widehat{wCV} = \hat{\sigma}/\hat{\mu}$, where $\hat{\sigma}$ is the square root of the estimator of the variance of repeat measurements $\hat{\sigma}^2$ as defined above and

$$\hat{\mu} = \frac{1}{Np} \sum_{i=1}^N \sum_{j=1}^p Y_{ij}$$

No closed-form expression for exact confidence intervals for the wCV exist, but Quan and Shih (1996) derive an approximate confidence interval for when the number of patients N gets large:

$$\widehat{wCV} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{N}} \sqrt{\frac{\sum_{i=1}^N p(\bar{Y}_i - \hat{\mu})^2 / N}{p\hat{\mu}^4} + \frac{1}{2(p-1)\hat{\mu}^2}}$$

[[REF: Quan and Shih 1996, "Assessing reproducibility by the within-subject coefficient of variation with random effects models"]]

For ICC, the estimator is

$$ICC = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}$$

with

$$\hat{\tau} = \frac{1}{p} \left(\frac{1}{N-1} \sum_{i=1}^N p(\bar{Y}_i - \hat{\mu})^2 - \hat{\sigma}^2 \right)$$

An exact 95% confidence interval for the ICC is

$$\left(1 - p \left[\frac{\sum_{i=1}^N p(\bar{Y}_i - \hat{\mu})^2}{\hat{\sigma}^2 F_{N-1, N(p-1), 0.025}} + (p-1) \right]^{-1}, 1 - p \left[\frac{\sum_{i=1}^N p(\bar{Y}_i - \hat{\mu})^2}{\hat{\sigma}^2 F_{N-1, N(p-1), 0.975}} + (p-1) \right]^{-1} \right)$$

where $F_{N-1, N(p-1), 0.025}$ and $F_{N-1, N(p-1), 0.975}$ are the 2.5th and 97.5th percentiles of an F distribution with $N-1$ and $N(p-1)$ degrees of freedom. We then may examine whether this confidence interval lies strictly above some pre-determined scientifically relevant threshold.

Reproducibility

Each patient in the study undergoes two or more scans, but the conditions under which each scan occurs are allowed to vary (e.g. scans taken on different days or using different devices, acquisition protocols, or parameters). No treatment is administered in between scans.

Notation:

- σ^2 : the variability attributed to measurement error, namely the variance of repeat measurements on the same patient had the patient undergone repeat scans under identical conditions (assumed to be constant across all patients in the population of interest).
- ν^2 : the variability attributed to differing conditions, namely the variance of repeat measurements on the same patient assuming each scan occurs under different conditions. The exact formula for ν will vary depending on the study design and the number of conditions that are allowed to vary.
- ξ_i : the average value of repeat measurements on the i^{th} patient (will vary across patients).
- τ^2 : the between-patient variability, namely the variance of the average measurements across patients (i.e. the standard deviation of the ξ_i .)
- μ : the grand average measurements, namely the average measurement across both patients and repeat measurements between patients (i.e. the mean of the ξ_i .)

Index	Definition and Formula	Range	Interpretation	Comments
Reproducibility Coefficient (RDC)	$1.96\sqrt{2\sigma^2 + \nu^2}$	$[0, \infty)$	The value under which the difference between any two repeat measurements on the same patient should fall with 95% probability.	Not useful for comparing reproducibility of measurands of differing units.
Intra-class Correlation Coefficient (ICC)	$\frac{\tau^2}{\tau^2 + \sigma^2 + \nu^2}$	$[0, 1]$	Proportion of total variation in measurements explained by between-patient differences.	Not appropriate for comparing reproducibility across different patient populations.

Data: N patients each undergo one scan using each of the p different acquisition protocols, at each of the p pre-specified time points, or using each of the p devices, with no intervening treatment between scans.

- Y_{ij} : the value of the j^{th} measurement for the i^{th} patient.

Inference: [[Graybill and Wang 1958, "Confidence intervals on non-negative linear combinations of variances]] for RDC.

Agreement

Each patient in the study undergoes both the investigational assay and the standard assay; for the time being, we assume each patient undergoes each assay only once. No treatment is administered in between scans.

NOTE: bias is a special case of agreement where the standard assay is a measurement of the ground truth.

Notation:

- Y : investigational assay measurement.
- X : standard assay measurement.
- σ^2 : the between-patient variability associated with the investigational assay, namely the variance of investigational assay measurements across patients.
- ρ : the correlation between measurements from the investigational and standard assays.
- τ^2 : the between-patient variability associated with the standard assay, namely the variance of standard assay measurements across patients.
- μ : the average investigational assay measurement across patients.
- ν : the average standard assay measurement across patients.
- δ : the average difference between investigational and standard assay measurements across patients.

- v^2 : the variance of the difference between investigation and standard assay measurements across patients.

Index	Definition and Formula	Range	Interpretation	Comments
Mean Squared Deviation (MSD)	$E[(Y - X)^2]$ or $(\mu - \nu)^2 + \sigma^2 + \tau^2 - 2\rho\sigma\tau$	$[0, \infty)$	Mean of the squared difference between investigational and standard assay measurements.	
Bland-Altman Limits of Agreement (LOA)	$\delta \pm 1.96 \left(\frac{N+1}{N}\right)v$	$[0, 1]$	Range of values in which the difference between investigational and standard assay measurements should fall with 95% probability.	If N is small (e.g. below 30), replace 1.96 with the 2.5 th percentile of the t distribution with $n - 1$ degrees of freedom.
Coverage Probability (CP)	$P(Y - X < \epsilon)$ for some prescribed ϵ .	$[0, 1]$	The probability that the absolute difference between investigational and standard assay measurements is less than ϵ .	
Total Deviation Index (TDI)	ϵ such that $P(Y - X < \epsilon)$ equals 0.95.	$[0, \infty)$	A value below which the absolute difference between investigational and standard assay measurements fall with 95% probability.	
Intra-class Correlation Coefficient (ICC)	$\frac{\sigma^2}{\sigma^2 + v^2}$	$[0, 1]$	Proportion of total variation in measurements explained by between-patient differences.	Not appropriate for comparing agreement across different patient populations.
Concordance Correlation Coefficient (CCC)	$\frac{2\rho\sigma\tau}{\sigma^2 + \tau^2 + (\mu - \nu)^2}$	$[0, 1]$	Deviation from (X, Y) to 45° line ($X = Y$), correcting for irreducible deviation between assay measurements (i.e. deviation if assay measurements are uncorrelated).	Not appropriate for comparing agreement across different patient populations.

Correlation	ρ	[0, 1]	The correlation between the investigational and standard assay measurements.	Good for when the standard and investigational assay measurements are on different scales.
Concordance (ROC-type Index)	$P(Y > Y' X > X')$ where X and Y are measurements associated with one randomly selected patient and X' and Y' are those associated with another randomly selected patient.	[0, 1]	The probability that, given any two randomly selected patients, the one with the higher standard assay measurement will also have the higher investigational assay measurement.	Good for when the standard and investigational assay measurements are on different scales.