Coefficients of Individual Agreement: A New Approach to Evaluating Agreement between Two Observers or Methods of Measurement

Michael Haber

Department of Biostatistics

Rollins School of Public Health

Emory University

Collaborating with

Huiman Barnhart, Duke University

Jingjing Gao, Emory University

This research is supported by NIMH grant 1-R01-MH070028

December 2008

Existing Coefficients of Interobserver Agreement

For continuous observations – the concordance correlation coefficient (Lin, 1989)

$$CCC(X,Y) = 1 - \frac{E(X-Y)^2}{E\{(X-Y)^2 \mid X, Y \; independent\}}$$

For categorical observations – kappa (Cohen, 1960)

$$\kappa(X,Y) = \frac{P(X=Y) - P(X=Y \mid X, Y \; independent)}{1 - P(X=Y \mid X, Y \; independent)}$$

Both coefficients are often criticized because of their dependence on the marginal distribution of the variable being observed:

CCC increases when the between-subjects variance increases

Kappa may attain ‘funny’ values when the marginal distributions of X and Y are very skewed or unequal

For example: $\kappa = -0.01$ for the table

98	1
1	0

Though $P(X = Y) = 0.98$!

Is the correction for chance agreement justified?

A ‘starting from scratch’ approach

First, decide upon a disagreement function $G(X,Y)$ to quantify the disagreement between X and Y.

Examples for continuous observations

$G(X,Y) = E(X-Y)^2$ (MSD)

$G(X,Y) = E\,|\,X-Y\,|$ (MAD)

$G(X,Y) = E\{|\,X-Y\,|\,/\,X\}$ (MRD)

Or the ‘robust MSD’ (King, Chinchilli, 2001):

$G(x,y) = (x-y)^2$ when $|\,x-y\,| \le a$,

$G(x,y) = a^2$ when $|\,x-y\,| > a$,

$G(X,Y) = E(G(x,y))$.

For categorical observations:

$G(X,Y) = E(X-Y)^2 = P(X \neq Y)$

How to Scale the Disagreement Function ?

Compare the observed disagreement with the expected disagreement when X and Y are in 'acceptable' agreement.

X and Y are in 'acceptable' agreement if the disagreement function does not change when replacing one of the observers by the other, i.e., if $G(X,Y) \approx G(X,X')$ and $G(X,Y) \approx G(Y,Y')$. Where $G(X,X')$ is the disagreement between two replicated observations made by observer X.

Therefore we need replicated observations made by the same observer on the same subject.

Coefficients of Individual Agreement (CIA's?)

When X is a reference (gold standard) and Y is a 'new' observer, define

$$\psi_G^R = \frac{G(X, X')}{G(X, Y)}$$

When both X and Y are 'new', define

$$\psi_G^N = \frac{[G(X, X') + G(Y, Y')]/2}{G(X, Y)}$$

The type of comparison in ψ^R is used in individual bioequivalence when comparing a new drug to a reference drug.

ψ^N varies between 0 and 1, while ψ^R may exceed 1.

Before using these coefficients we must make sure that $G(X, X')$ (when using ψ^R) or both $G(X, X')$ and $G(Y, Y')$ are ‘reasonably small’.

For ‘reasonably good’ agreement we require $\psi \geq 0.8$, which means that replacing one observer by the other does not increase the within-subject disagreement by more than 25%.

Observations

Suppose that both observers observe the same N study subjects, indexed by $i = 1,...,N$. Let X_{ik} denote the k-th replicated observation of X ($k = 1,...,K_i$) and Y_{il} denote the l-th replicated observation of Y ($l = 1,...,L_i$) on subject i.

In the special case where X and Y are continuous and $G(X,Y) = MSD(X,Y) = E(X-Y)^2$

Cnsider the model:

$$X_{ik} = \mu_{Xi} + e_{Xik},\ E(e_{Xik}) = 0,\ Var(e_{Xik}) = \sigma^2_{eX},$$
$$Cov(\mu_{Xi}, e_{Xik}) = 0;$$
$$Y_{il} = \mu_{Yi} + e_{Yil},\ E(e_{Yil}) = 0,\ Var(e_{Yil}) = \sigma^2_{eY},$$
$$Cov(\mu_{Yi}, e_{Yil}) = 0;$$

X and Y are conditionally independent, given i.

Then:

$$MSD(X,Y) = E(\mu_{Xi} - \mu_{Yi})^2 + \sigma^2_{eX} + \sigma^2_{eY},$$
$$MSD(X,X') = 2\sigma^2_{eX},\ MSD(Y,Y') = 2\sigma^2_{eY}$$

$$\psi^N_{MSD} = \frac{\sigma^2_{eX} + \sigma^2_{eY}}{E(\mu_{Xi} - \mu_{Yi})^2 + \sigma^2_{eX} + \sigma^2_{eY}}$$

$$\psi^R_{MSD} = \frac{2 \cdot \sigma^2_{eX}}{E(\mu_{Xi} - \mu_{Yi})^2 + \sigma^2_{eX} + \sigma^2_{eY}}$$

Estimation

Subject-specific disagreements:

$$G_i(X,Y) = E\{G(X_{ik}, Y_{il}) \mid i\},$$

$$G_i(X,X') = E\{G(X_{ik}, X_{ik'}) \mid i,\ k \neq k'\}$$

$$G_i(Y,Y') = E\{G(Y_{il}, Y_{il'}) \mid i,\ l \neq l'\}$$

They are estimated by:

$$\hat{G}_i(X,Y) = Mean_{k,l}[G(x_{ik}, y_{il})],$$

$$\hat{G}_i(X,X') = Mean_{k<k'}[G(x_{ik}, x_{ik'})^2],$$

$$\hat{G}_i(Y,Y') = Mean_{l<l'}[G(y_{il}, y_{il'})^2]$$

Then the population-level disagreement functions are estimated as the sample means of the estimated subject-specific disagreements:

$$\hat{G}(X,Y) = Mean_i[\hat{G}_i(X,Y)]$$

$$\hat{G}(X,X') = Mean_i[\hat{G}_i(X,X')]$$

$$\hat{G}(Y,Y') = Mean_i[\hat{G}_i(Y,Y')]$$

The estimated ψ's are then obtained by substituting the estimated G's into their definitions:

$$\hat{\psi}_G^R = \frac{\hat{G}(X, X')}{\hat{G}(X, Y)},$$

$$\hat{\psi}_G^N = \frac{[\hat{G}(X, X') + \hat{G}(Y, Y')]/2}{\hat{G}(X, Y)}$$

Standard Errors of Estimated ψ's

Consider ψ^N, whose estimate can be written as:

$$\hat{\psi}_G^N = \frac{[\overline{G}(X,X') + \overline{G}(Y,Y')]/2}{\overline{G}(X,Y)},$$

where each $\overline{G}$ is the sample mean of the subject-specific disagreements.

To simplify the notation let $G^{(1)} = G(X,X')$,

$G^{(2)} = G(Y,Y'),\ \ G^{(3)} = G(X,Y).$

Then write $\hat{\psi}_G^N = A/B$ where

$A = (\overline{G}^{(1)} + \overline{G}^{(2)})/2$ and $B = \overline{G}^{(3)}.$

For $p = 1,2,3$ denote the sample variances:

$S^2(G^{(p)}) = [\sum_i (G_i^{(p)} - \overline{G}^{(p)})^2]/(N-1),$

so that $\hat{Var}(\overline{G}^{(p)}) = S^2(G^{(p)})/N.$

In addition, for $1 \le p < q \le 3$ denote the sample covariance of $G^{(p)}$, $G^{(q)}$ by

$$C(G^{(p)}, G^{(q)}) = [\sum_i (\hat{G}_i^{(p)} - \overline{G}^{(p)})(\hat{G}_i^{(q)} - \overline{G}^{(q)})] / (N-1)$$

so that $\hat{C}ov(\overline{G}^{(p)}, \overline{G}^{(q)}) = C(G^{(p)}, G^{(q)}) / N.$

We wrote $\hat{\psi}_G^N = A / B$ where

$A = (\overline{G}^{(1)} + \overline{G}^{(2)}) / 2$ and $B = \overline{G}^{(3)}$.

Now calculate

$$\hat{V}ar(A) = [S^2(G^{(1)}) + S^2(G^{(2)}) + 2 \cdot C(G^{(1)}, G^{(2)})] / 4N$$

$$\hat{V}ar(B) = S^2(G^{(3)}) / N,$$

$$\hat{C}ov(A, B) = [C(G^{(1)}, G^{(3)}) + C(G^{(2)}, G^{(3)})] / 2N,$$

Finally, substitute these in the approximation for the variance of a ratio:

$$\hat{V}ar(\hat{\psi}^N) = \hat{V}AR(\frac{A}{B}) \approx \frac{A^2}{B^2}[\frac{\hat{V}ar(A)}{A^2} + \frac{\hat{V}ar(B)}{B^2} - \frac{2 \cdot \hat{C}ov(A, B)}{A \cdot B}]$$

Example 1: Systolic Blood Pressures

Systolic blood pressure (SBP) was measured on 85 subjects by two experienced human observers using a sphygmomanometer and by a semi-automatic blood pressure monitor. Three replications were made in quick succession with each of the three methods on each subject (Bland and Altman, 1999).

Observer	Mean	SD	$\hat{\sigma}_W$
Human 1	127.4	30.8	6.2
Human 2	127.3	30.5	6.2
Monitor	143.0	31.8	9.3

Mean and SD are based on the means of the three observations made by each observer

The agreement between the two human observers is excellent, $\hat{\psi}^N_{MSD} = 1.44$

We will focus on the agreement between the first human observer (X) and the monitor (Y). For ψ^R the human observer is considered the reference.

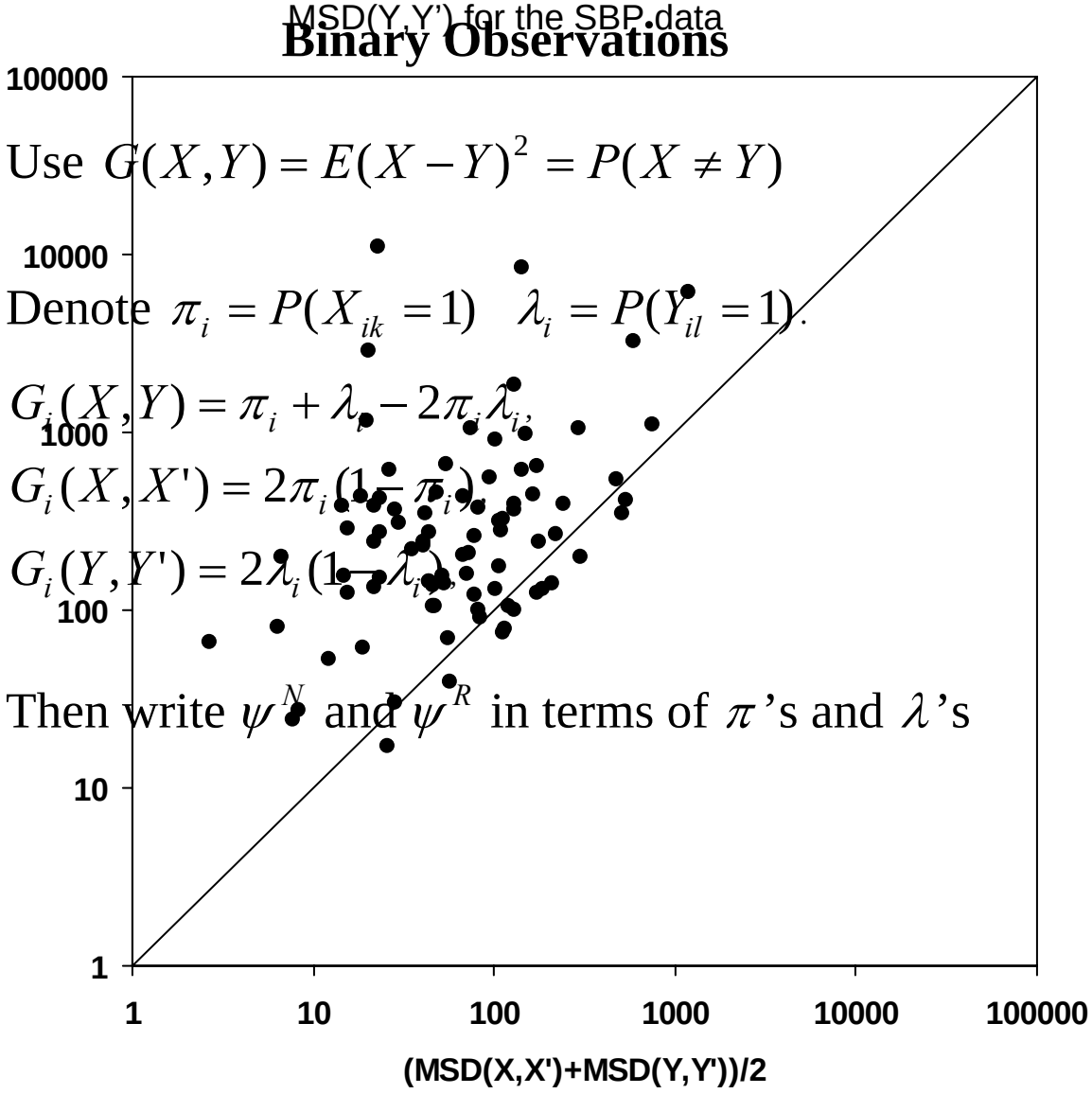
G	(X,X')	(Y,Y')	(X,Y)	ψ_G^N	ψ_G^R
MSD	74.8	166.3	678.6	0.18 (0.11,0.31)* (0.09,0.27)#	0.11 (0.07,0.21)* (0.05,0.17)#
MAD	6.7	9.0	18.4	0.43 (0.35,0.52)*	0.36 (0.28,0.46)*
MRD	0.053		0.156		0.34 (0.27,0.43)*

* Percentile-based bootstrap CI's

CI's based on estimated SE's

$MRD(X,Y) = E\,|\,X - Y\,|\,/\,X$, is of interest mainly when X is the reference.

Figure 1a: Comparing MSD(X,Y) with the mean of MSD(X,X') and MSD(Y,Y') for the SBP data



Use $G(X,Y) = E(X-Y)^2 = P(X \neq Y)$

Denote $\pi_i = P(X_{ik} = 1)$ $\lambda_i = P(Y_{il} = 1)$.

$G_i(X,Y) = \pi_i + \lambda_i - 2\pi_i\lambda_i,$

$G_i(X,X') = 2\pi_i(1-\pi_i),$

$G_i(Y,Y') = 2\lambda_i(1-\lambda_i),$

Then write ψ^N and ψ^R in terms of π's and λ's

A Model for Diagnostic Agreement

$X, Y = 1$ for positive Dx, $X, Y = 0$ for negative.

$T = 1$ for ill, $T = 0$ for not ill

$\omega = P(T = 1)$ (prevalence)

$\eta_t = P(X = 1 \mid T = t),\ t = 0,1$

$\theta_t = P(Y = 1 \mid T = t),\ t = 0,1$

η_1, θ_1 = sensitivities of X, Y

η_0, θ_0 = complements of specificities of X, Y

$$G(X,Y) = \omega[\eta_1 + (1-2\eta_1)\theta_1] + (1-\omega)[\eta_0 + (1-2\eta_0)\theta_0]$$

$$G(X,X') = 2\omega\eta_1(1-\eta_1) + 2(1-\omega)\eta_0(1-\eta_0)$$

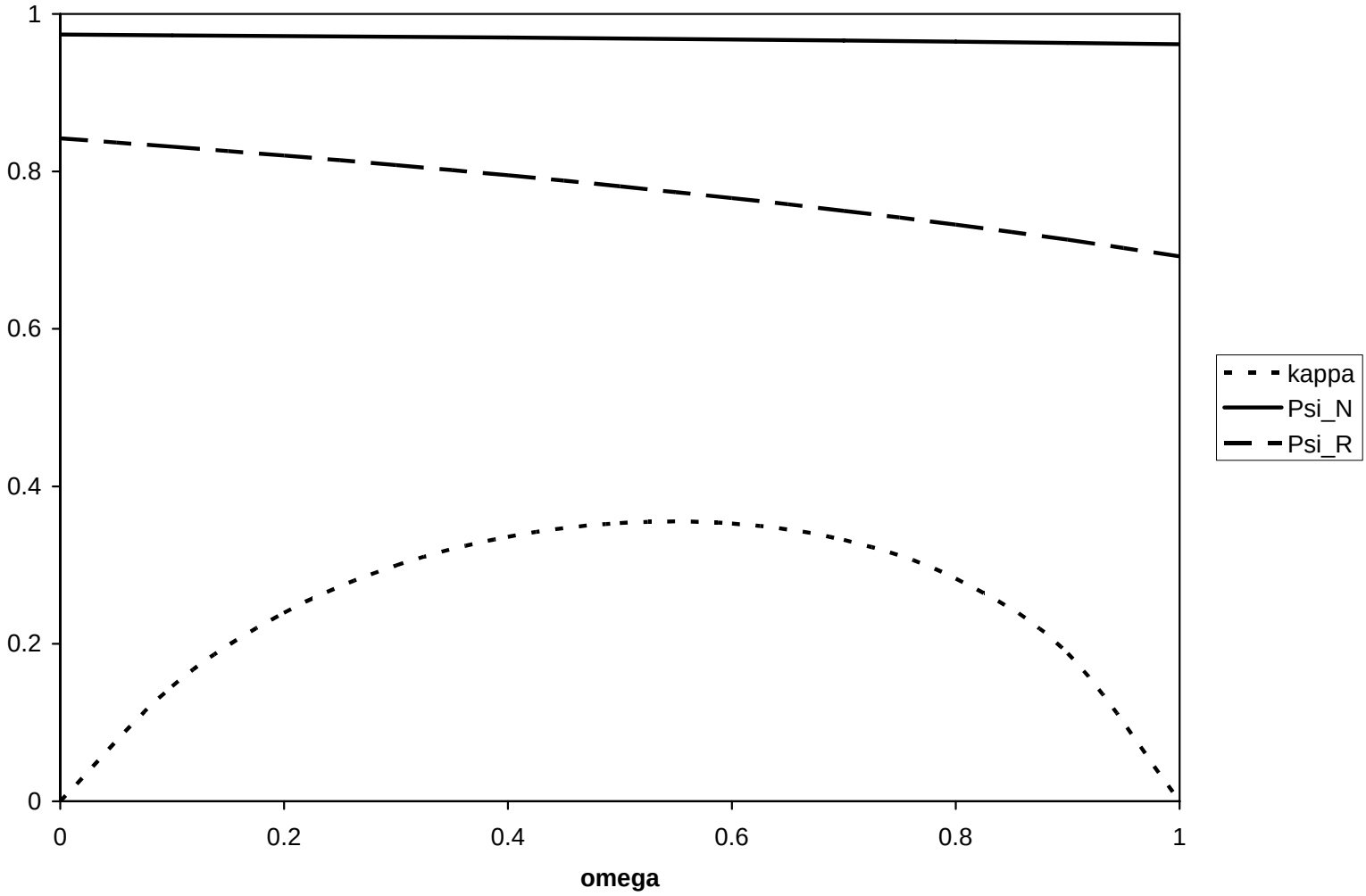
$$G(Y,Y') = 2\omega\theta_1(1-\theta_1) + 2(1-\omega)\theta_0(1-\theta_0)$$

Consider a fixed ‘good’ reference observer X with $\eta_1 > 0.5,\ \eta_0 < 0.5.$

Then ψ^R is an increasing function of the sensitivity and specificity of Y.

ψ^N, ψ^R and κ as functions of prevalence (ω)

η_1= 0.9, η_0=0.2, θ_1=0.8, θ_0=0.3



Estimation (General Case)

$$\hat{G}_i(X,Y) = \hat{\pi}_i + \hat{\lambda}_i - 2\hat{\pi}_i\hat{\lambda}_i$$

$$\hat{G}_i(X,X') = 2K_i\hat{\pi}_i(1-\hat{\pi}_i)/(K_i-1)$$

$$\hat{G}_i(Y,Y') = 2L_i\hat{\lambda}_i(1-\hat{\lambda}_i)/(L_i-1)$$

Where $\hat{\pi}_i$ and $\hat{\lambda}_i$ are the proportions of $X_{ik}=1$ and $Y_{il}=1$, respectively.

Example 2: Diagnosis of Breast Cancer from Mammograms

150 female patients underwent a mammography at the Yale-New Haven Hospital in 1987. Each of ten radiologists read each patient's mammogram and classified it into one of four diagnosis categories:
(1) normal,
(2) abnormal - probably benign,
(3) abnormal - intermediate, or
(4) abnormal - suggestive of cancer.

Four months later the same films were reviewed again, in a random order, by the same radiologists. We consider the two evaluations as replications.

We considered a radiologist's rating as 'positive' only if the mammogram was classified as abnormal and suggestive of cancer.

Each of the study participants was followed up for three years, and then a definitive diagnosis was made. 27 of the 150 patients (18%) had breast cancer.

Proportions of positive ratings, sensitivity and specificity for each radiologist.

Radiologist	Proportion rated positive	Sensitivity	Specificity
A	0.208	0.815	0.927
B	0.140	0.630	0.967
C	0.077	0.333	0.980
D	0.223	0.778	0.898
E	0.180	0.704	0.935
F	0.160	0.722	0.963
G	0.177	0.574	0.911
H	0.107	0.500	0.980
I	0.280	0.796	0.833
J	0.240	0.685	0.858

The total of sensitivity and specificity was highest for radiologist A.

Therefore we illustrate the new coefficients by estimating the agreement between radiologist A and each of the remaining nine radiologists. Radiologist A was considered the reference in estimating ψ^{R}.

Table 2: Estimates of agreement coefficients for nine pairs of radiologists

Radiologists	$\hat{G}(X,X')$	$\hat{G}(Y,Y')$	$\hat{G}(X,Y)$	$\hat{\psi}^N$	$\hat{\psi}^R$	$\hat{\kappa}$
(A, B)	0.040	0.093	0.103	0.645	0.387	0.642
(A, C)	0.040	0.060	0.140	0.357	0.286	0.444
(A, D)	0.040	0.113	0.110	0.697	0.364	0.674
(A, E)	0.040	0.080	0.093	0.643	0.429	0.701
(A, F)	0.040	0.067	0.070	0.762	0.571	0.767
(A, G)	0.040	0.100	0.127	0.553	0.316	0.592
(A, H)	0.040	0.080	0.123	0.486	0.324	0.542
(A, I)	0.040	0.173	0.143	0.744	0.279	0.614
(A, J)	0.040	0.133	0.140	0.619	0.286	0.597

None of the other 9 radiologists has an acceptable agreement with radiologist A when the latter is the reference. The upper 95% CI for ψ^R exceeds 0.8 for radiologists E and F.

Extension to Multiple Observers

J observers $Y_1, Y_2, ..., Y_J$

$$\psi^N = \frac{Mean_{1\le j\le J}[G(Y_j, Y_j^{'})]}{Mean_{1\le j<j'\le J}[G(Y_j, Y_{j'})]}$$

Where $G(Y_j, Y_j^{'})$ is the disagreement between two replicated observations made by observer j.

For ψ^R, the observer Y_J is considered as the reference.

$$\psi^R = \frac{G(Y_J, Y_J^{'})}{Mean_{1\le j\le J-1} G(Y_j, Y_J)}.$$

Summary of New Approach

Using within-subjects rather than between-subjects variability for scaling

Minimal assumptions

Simple methods of estimation and inference

Requires replications

Future Research

Repeated observations corresponding to specific conditions or time points.

Nominal and ordinal observations

Modeling the disagreement function $G(X,Y)$ or the coefficients in term of subject-specific, observer-specific and observation-specific covariates

Random Observers

Multivariate agreement

References

Barnhart HX, Haber M, Lin LI (2007). An overview on assessing agreement with continuous measurements.
Journal of Biopharmaceutical Statistics **17**:529-569.

Barnhart HX, Kosinski AS, Haber M. (2007). Assessing individual agreement. *Journal of Biopharmaceutical Statistics* **17**:697-719.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**:37-46.

Haber M, Barnhart HX. (2006) Coefficients of agreement for fixed observers. *Statistical Methods in Medical Research* **15**:255-271.

Haber M., Barnhart HX. (2008) A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Statistical Methods in Medical Research* **17**:151-171.

Haber M, Gao J, Barnhart HX. (2007) Assessing agreement in studies involving replicated binary observations. *Journal of Biopharmaceutical Statistics* **17**:757-766.

King TS, Chinchilli VM. (2001) Robust estimators of the concordance correlation coefficient.
J Biopharmaceutical Statistics **11**:83-105.

Lin L. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45:**255-68.