

A new permutation-based method for the evaluation of the agreement between two observers with replicated binary observations

Michael Haber and Yi Pan

Department of Biostatistics and Bioinformatics
Emory University

ISCB Meeting, August 2009

OUTLINE

- 1 Introduction
 - Why Agreement is important?
- 2 Coefficient of Individual Equivalence
 - Equivalence
 - General Definition
 - CIE for Binary Data
- 3 Simulation Study in Estimation of CIE
 - Data Generation
 - Set Ups
 - Simulation Result
- 4 Hypothesis Test
 - Coefficient of Individual Agreement
 - Hypothesis Test of CIA and CIE
 - Association between CIA and CIE
- 5 Example
 - Background
 - Result of Example
- 6 Conclusion and Future Work

Why Agreement is Important?

- In agreement studies, we are interested in whether the observers can be used interchangeably, or whether a new method that is easy to use can replace an existing standard method that may be expensive or invasive.
- Simple examples :
 - In a blood pressure study (Bland and Altman, 1999), automatic blood pressure machine VS. human observers
 - In a carotid stenosis screening study, MRA-2D, MRA-3D VS. IA

Equivalence

- Why there are so many methods to assess agreement?
 - There is no clear definition of 'good' or 'acceptable' agreement.
 - Hawkins (2002) defined equivalence of two observers (X, Y) as follows: Let $f(x|i)$ and $f(y|i)$ be the conditional distributions of the values of X and Y on subject i . Then X and Y are equivalent if $f(x|i) = f(y|i)$ for all the subjects.
 - From a statistical point of view, when X and Y are equivalent, it does not matter if the next measurement is made by X or by Y .
 - Equivalence can be considered as a definition of acceptable agreement.

CIE for General Case

- $G_i(X, Y)$ denotes the disagreement function between X and Y for subject i . Usually mean squared deviation (MSD) was used to construct $G_i(X, Y)$.
 $G_i(X, Y) = E[X - Y]^2$.
- We're interested in comparing the observed value of $G_i(X, Y)$ to its expected value under equivalence, which will be denoted by G_i^E .

$$CIE = \frac{E_i(G_i^E)}{E_i(G_i(X, Y))}.$$

- Values close to 1 indicate that the observed disagreement is similar to the disagreement that can be expected if the two observers are 'equivalent' in the sense that for each subject, the (conditional) distributions of the values they would report are identical.
- The CIE compares the observed disagreement to the expected under chance disagreement.

Estimation of CIE

Assume we have K replicated observations by X and L replicated observations by Y . K, L may vary across subjects, but here we assume that they are fixed in order to simplify the notation.

Estimation of $E_i G_i(X, Y)$

$$X_i = (X_{i1}, \dots, X_{iK}), Y_i = (Y_{i1}, \dots, Y_{iL})$$

$$\hat{G}_i(X, Y) = G(X_i, Y_i) = \text{mean}_{k,l}[G(X_{ik}, Y_{il})]$$

$$\hat{G}(X, Y) = \text{mean}_i[\hat{G}_i(X, Y)].$$

Estimation of $E_i(G_i^E)$

Consider all the C_K^{K+L} possible assignments of K X 's and L Y 's to the $K + L$ observations made on this subject. Under equivalence, all these assignments are equally likely, and hence the expected value of the disagreement function for this subject is the mean of $\hat{G}_i(X, Y)$ over all C_K^{K+L} assignments.

$$\hat{G}_i^E = \frac{\sum_{j=1}^{C_K^{K+L}} \hat{G}_{ij}(X, Y)}{C_K^{K+L}}.$$

CIE for Binary Data

Estimation of $E_i G_i(X, Y)$

In general, $G_i(X, Y) = E[X - Y]^2$. For binary outcome, assume $P(X_{ik} = 1) = \pi_i$, and $P(Y_{il} = 1) = \lambda_i$. $G_i(X, Y) = P(X_{ik} \neq Y_{il} | i) = P(X_{ik} = 1)P(Y_{il} = 0) + P(Y_{il} = 1)P(X_{ik} = 0) = \pi_i(1 - \lambda_i) + \lambda_i(1 - \pi_i)$. Furthermore, $T_i \sim \text{Bin}(K, \pi_i)$, $U_i \sim \text{Bin}(L, \lambda_i)$. The estimated disagreement for this subject is $\hat{G}_i(X, Y) = [T_i(L - U_i) + (K - T_i)U_i]/KL$.

Estimation of $E_i(G_i^E)$

Assume $W_i = T_i^A + U_i^A$ is fixed and T_i^A is hypergeometric, where A is one of the C_K^{K+L} .

$$\hat{G}_i^E = E_H[2(T_i^A)^2 + (L - K - 2W_i)T_i^A + K \cdot W_i]/(K \cdot L),$$

$$E(T_i^A) = W_i K / M,$$

$$E(T_i^A)^2 = [(W_i K \cdot L \cdot (M - W_i) + W_i^2 K^2 (M - 1)]/[M^2 (M - 1)],$$

$$M = K + L.$$

Data Generation

- $T_i \sim N(\mu_T, \sigma_T^2)$. $A_i \sim N(\mu_A, \sigma_A^2)$, $B_i \sim N(\mu_B, \sigma_B^2)$. T , A and B are mutually independent.
- $U_i \sim N(t_i + a_i, \sigma_U^2)$, $V_i \sim N(t_i + b_i, \sigma_V^2)$.
- If $U > C$ then $X = 1$, else then $X = 0$ and if $V > C$ then $Y = 1$, else then $Y = 0$.

$$\pi_i = P(X_i = 1) = P(U_i > C) = 1 - \Phi\left(\frac{C - (t_i + a_i)}{\sigma_U}\right)$$

$$\lambda_i = P(Y_i = 1) = P(V_i > C) = 1 - \Phi\left(\frac{C - (t_i + b_i)}{\sigma_V}\right).$$

Set Ups

- $\mu_T = 138, C = 140, \sigma_U = \sigma_V = 3$.
- $\mu_A = 0, \mu_B = 1, 3, 5$ to accommodate good, moderate and poor agreement, respectively.
- Sample size: $N = 50, 100$ and 200 .
- Replication combinations: $(K, L) = (3, 3), (4, 2)$ and $(1, 2)$.
- Number of Simulation: 1000 times for each scenario.

Table: Estimation of CIE when (K,L)=(3,3) and (4,2)

N	K	L	μ_B	Bias(CIE)	SE^a	SE^b	CP(CIE)
50	3	3	1	0.0039	0.0397	0.0381	0.9090
50	3	3	3	0.0027	0.0375	0.0389	0.9450
50	3	3	5	0.0037	0.0335	0.0341	0.9420
100	3	3	1	0.0025	0.0281	0.0280	0.9270
100	3	3	3	0.0024	0.0277	0.0278	0.9380
100	3	3	5	0.0019	0.0234	0.0242	0.9450
200	3	3	1	0.0014	0.0202	0.0200	0.9340
200	3	3	3	0.0012	0.0201	0.0198	0.9390
200	3	3	5	0.0009	0.0171	0.0172	0.9430
50	4	2	1	0.0034	0.0535	0.0536	0.9230
50	4	2	3	0.0053	0.0519	0.0520	0.9340
50	4	2	5	0.0045	0.0448	0.0462	0.9460
100	4	2	1	0.0020	0.0384	0.0385	0.9410
100	4	2	3	0.0015	0.0362	0.0370	0.9510
100	4	2	5	0.0023	0.0320	0.0325	0.9450
200	4	2	1	0.0005	0.0283	0.0273	0.9400
200	4	2	3	0.0013	0.0266	0.0263	0.9380
200	4	2	5	0.0011	0.0232	0.0230	0.9390

^aStandard errors based on simulations of CIE

^bMean of estimated standard errors calculated from formulas of CIE

Table: Estimation of CIE when K=1, L=2

N	K	L	μ_B	Bias(CIE)	SE^a	SE^b	CP(CIE)
50	1	2	1	0.0102	0.0825	0.0810	0.935
50	1	2	3	0.0111	0.0733	0.0708	0.936
50	1	2	5	0.0081	0.0600	0.0585	0.947
100	1	2	1	0.0098	0.0595	0.0575	0.932
100	1	2	3	0.0080	0.0504	0.0497	0.938
100	1	2	5	0.0053	0.0402	0.0409	0.945
200	1	2	1	0.0077	0.0415	0.0406	0.947
200	1	2	3	0.0073	0.0368	0.0351	0.938
200	1	2	5	0.0048	0.0290	0.0289	0.949

^aStandard errors based on simulations of CIE

^bMean of estimated standard errors calculated from formulas of CIE

Coefficient of Individual Agreement

- Coefficient of individual agreement for assessing agreement for both quantitative and categorical measurements was proposed by Barnhart et al.(2007).
- The coefficients of individual agreement with a specific disagreement function are defined as:

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

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

where $G(X, X')$ is the disagreement between two replicated measurements by observer X and $G(Y, Y')$ is defined analogously for observer Y . For the definition of ψ_G^R we assume that observer X is the reference.

Hypothesis Test of CIA and CIE

- $H_{01}: CIA = 1$ $H_{11}: CIA \leq 1$
- $H_{02}: CIE = 1$ $H_{12}: CIE \leq 1$

H_{01} and H_{02} are equivalent

By the definition of individual equivalence, $\pi_i = \lambda_i$ is sufficient and necessary condition of $CIE=1$. Let's focus on CIA. To simplify the notation, let's denote $\pi_i = \pi$ and $\lambda_i = \lambda$.

$$CIA = \frac{(G(X, X') + G(Y, Y'))/2}{G(X, Y)}$$

where $G(X, X') = 2\pi(1 - \pi)$, $G(Y, Y') = 2\lambda(1 - \lambda)$ and $G(X, Y) = \pi + \lambda - 2\pi\lambda$. If $CIA=1$, which means

$$\pi(1 - \pi) + \lambda(1 - \lambda) = \pi + \lambda + 2\pi\lambda.$$

Then,

$$\pi - \pi^2 + \lambda - \lambda^2 = \pi + \lambda + 2\pi\lambda.$$

$$\pi^2 + \lambda^2 - 2\pi\lambda = 0.$$

Therefore, $(\pi - \lambda)^2 = 0$ which is equivalent to $\pi = \lambda$. On the other hand, if $\pi = \lambda$, $CIA=1$.

Simulation Study in Hypothesis Test

Table: Estimated Power of CIA and CIE

N	K	L	μ_B	Power(CIA)	Power(CIE)
50	3	3	1	0.172	0.172
50	3	3	3	0.769	0.769
50	3	3	5	1	1
100	3	3	1	0.271	0.271
100	3	3	3	0.952	0.952
100	3	3	5	1	1
200	3	3	1	0.472	0.472
200	3	3	3	0.999	0.999
200	3	3	5	1	1
50	4	2	1	0.134	0.168
50	4	2	3	0.628	0.675
50	4	2	5	0.991	0.982
100	4	2	1	0.22	0.284
100	4	2	3	0.886	0.918
100	4	2	5	1	1
200	4	2	1	0.364	0.472
200	4	2	3	0.992	0.998
200	4	2	5	1	1

Table: Power of CIE when $K=1$, $L=2$

N	K	L	μ_b	Power(CIE)
50	1	2	1	0.143
50	1	2	3	0.345
50	1	2	5	0.686
100	1	2	1	0.169
100	1	2	3	0.511
100	1	2	5	0.903
200	1	2	1	0.209
200	1	2	3	0.724
200	1	2	5	0.995

Association between CIA and CIE test statistics when $K=L$

- Through the simulations, size and power were observed to be identical for CIA and CIE when $K=L=3$. Furthermore, it was true whenever K and L are equal. Therefore, the association between \widehat{CIA} and \widehat{CIE} was investigated analytically when $K=L$.
- Also, from simulations, $\widehat{CIA} - 1$ and $\widehat{CIE} - 1$ were observed to have a linear association. To be more specific, $\widehat{CIE} - 1$ is proportional to $\widehat{CIA} - 1$ when the replications numbers of two observers are the same. Similarly, $SE(\widehat{CIE})$ is also proportional to $SE(\widehat{CIA})$.

$$\frac{\hat{G}_i^E - \hat{G}_i(X, Y)}{\frac{(\hat{G}_i(X, X') + \hat{G}_i(Y, Y'))}{2} - \hat{G}_i(X, Y)} = \frac{K - 1}{2K - 1}.$$

This linear association has been verified by the simulation result.

- When $K = L = 2$, this proportion is $\frac{1}{3}$ which is 0.333.
- When $K = L = 3$, this proportion is $\frac{1}{2}$ which is 0.4.
- When $K = L = 4$, this proportion is $\frac{1}{3}$ which is 0.428.

Motivating Example: Mammography Study

- The study was conducted to determine the validity of diagnosis of breast cancer based on mammograms.
- 150 female patients underwent a mammography at the Yale-New Haven Hospital in 1987.
- 10 radiologists read each patient's mammogram twice. Following each reading, the radiologist classified the mammogram into one of four diagnostic categories: (1) normal, (2) abnormal - probably benign, (3) abnormal - intermediate or (4) abnormal - suggestive of cancer.
- We considered the two evaluations as replications. Also, we dichotomized the measurement: (4) = "positive"; (1)-(3) = "negative".
- Since the total of sensitivity and specificity was highest for radiologist A, we illustrated the new coefficients by estimating the agreement between radiologist A and each of the remaining nine radiologists.

Estimated Results Comparing for Nine Pairs of Radiologists

Table: Estimated Results Comparing for Nine Pairs of Radiologists

Raters	CIA	SE(CIA)	95% CI(CIA)	CIE	SE(CIE)	95% CI(CIE)	Test Statistic	P Value
(A,B)	0.645	0.141	(0.369, 0.921)	0.882	0.047	(0.79,0.974)	-2.520	0.006
(A,C)	0.357	0.093	(0.175, 0.540)	0.786	0.031	(0.725,0.847)	-6.900	< 0.0001
(A,D)	0.697	0.125	(0.453, 0.941)	0.899	0.042	(0.818,0.980)	-2.432	0.008
(A,E)	0.643	0.141	(0.367, 0.918)	0.881	0.047	(0.789,0.973)	-2.541	0.006
(A,F)	0.762	0.146	(0.476, 1.000)	0.921	0.049	(0.825,1.106)	-1.635	0.051
(A,G)	0.541	0.110	(0.325, 0.756)	0.847	0.037	(0.775,0.919)	-4.178	< 0.0001
(A,H)	0.486	0.108	(0.274, 0.699)	0.823	0.034	(0.757,0.890)	-4.738	< 0.0001
(A,I)	0.738	0.111	(0.520, 0.956)	0.913	0.037	(0.840,0.985)	-2.351	0.009
(A,J)	0.619	0.109	(0.405, 0.833)	0.873	0.036	(0.802,0.944)	-3.496	0.0002

Conclusion

- No exact definition or null hypothesis exist for good agreement. But individual equivalence is defined as $f(x|i) = f(y|i)$ for all the subjects by Hawkins (2002) and can be viewed as a definition of acceptable agreement.
- In CIE, G^E is 'chance disagreement', which is very different from 'chance agreement' – agreement under independence– used in kappa and CCC.
- Unlike the CIA, the CIE is not based on comparing the between disagreement to the within disagreement. Therefore the CIE is valid even if the within disagreement is unacceptable.
- CIE can be used when one of the observers is a fixed gold standard (which does not require replicated readings), as long as there are replicated observations by the other observer. ψ_R cannot be used in this case.

Future Work

- Generalization to continuous observations.
- Generalization to multiple observers.
- Indices for data with repeated measurements, censoring, outliers, and covariates.
- Indices for multivariate data.
- Sample size calculation for design of agreement study.