XX IMEKO World Congress
Metrology for Green Growth
September 9–14, 2012, Busan, Republic of Korea

A NOTE ON APPLICATION OF MEASUREMENT PRECISION FOR BINARY DATA

R. Hamaguchi, Y. Tsutsumi, H. Kawamura, and T. Suzuki

Tokyo University of Science, Chiba, Japan, j7411623@ed.tus.ac.jp

Abstract: This paper describes the previous works about the precision for binary data, and clarifies about a view of those methods and what kind of feature. Furthermore, propose that the method of estimating precision from ANOVA using the normal approximation by logit transformation. We compare each method, and consider the relation between methods from the result.

Keywords: ISO 5725; Repeatability; Reproducibility; Binary Data

1. INTRODUCTION

As described in introduction of ISO5725-1(1994), *tests performed on presumably identical materials in presumably identical circumstances do not, in general, yield identical results. This is attributed to unavoidable random errors inherent in every measurement procedure; the factors that influence the outcome of a measurement cannot all be completely controlled. In the practical interpretation of measured value, this variability has to be taken into account. For instance the difference between a test result and some specified value may be within the scope of unavoidable random errors*. The ISO, International Organization for Standardization, develops various international standards. Regarding the measurement methods and results, ISO 5725, accuracy (trueness and precision) of measurement methods and results that consists of six parts, and the second part, ISO 5725-2 (1994) is the standard for the basic method to determine repeatability and reproducibility of standard measurement methods, and deals with collaborative interlaboratory experiments to obtain two measures of precision; repeatability and reproducibility. ISO 5725 assumes that measured values are continuous and follow a normal distribution. These assumptions do not hold for qualitative measurements, where the measured values are either 0(negative result) or 1(positive result). Several methodological researches have been carried out for qualitative data based on a variety of assumptions.

Mandel (1997) defined and proposed the method of estimating *repeatability standard deviation* and *reproducibility standard deviation* for a laboratory, but didn't define overall repeatability and reproducibility. Langton et al. (2002) [1] defined *accordance* as the equivalent of repeatability for qualitative data, and *concordance* as the equivalent of reproducibility for qualitative data. *Accordance* and *concordance* are estimated for evaluation of collaborative interlaboratory experiments. Van der Voet and van Raamsdonk (2004) interpreted the method of Langton as fixed effect which represents the observed quantities in terms of interlaboratory that is treated as if those quantities were non-random, and proposed the method that interlaboratory effects might be treated as random effects which is treated as interlaboratory as if those quantities were random. But both Langton and van der Voet didn't give statistical model, and focused on *Accordance* and *concordance*. Wilrich (2006) [2] proposed *repeatability*, Wilrich regarded binary data as continuous quantity, defined repeatability and reproducibility, and proposed estimation method the same calculation method of variance components for continuous quantity. Wieringen et al. (2008) [3] proposed a measurement system for binary data, where the latent class model including some raters is assumed. They estimated the sensitivity and specificity of each rater, using an expectation-maximization (EM) algorithm. Danila et al. (2008) proposed a measurement system for binary data focused on sensitivity and specificity. They focused on misclassification, sensitivity, and specificity because they assumed pass-fail inspection data. Corry et al.(2007) reviewed of evaluation of measurement method of food micro-organisms, and express that evaluation of measurement method for binary data are not hold. Horie et al.(2009)[4] proposed a new method, where a beta-binomial distribution is assumed to estimate the repeatability and reproducibility for binary data, and also clarified mathematical relation between Wilrich method and Langton method. However mathematical relation between Wieringen method and Langton method is not clarified.

Consequently, in this study, these methods including Wieringen method and proposed logit model are considered. Analogy and difference for these methods are also clarified.

2. COLLABORATIVE ASSESSMENT EXPERIMENT

2.1. Data format

Table1 shows data format. Suppose π_i is the probability of a measurement value in laboratory i, and probability of π_i is repeated j measurement experiments in interlaboratory i in the measurement result of binary data. Measured values of .each within-laboratory are obtained with the same method on the identical test items in the same

laboratory by the same operator using the same equipment within short intervals of time

Table 1 Data Format

Laboratory number	Repeated number						
	1	2	…	*j*	…	*n*	
1	0	1	…	1	…	0	π_1
2	0	0	…	1	…	1	π_2
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
i	0	0	…	y_{ij}	…	1	π_i
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
L	1	1	…	0	…	1	π_L

2.2 Precision

In ISO 5725, accuracy of a measurement result, measurement method, or measurement system is a general term that involves trueness and precision. Trueness, the closeness of agreement between the average value obtained from a large series of measurement results and an accepted reference value, is usually expressed in terms of bias, which is the difference between expectation of the measurement results and accepted reference value. Precision, the closeness of agreement between independent measurement results obtained under stipulated conditions, is usually expressed in terms of standard deviations of the measurement results.

Generally, when the accuracy of a measurement method is to be repeated, it is known that two measures of accuracy, named repeatability and reproducibility, are required. Repeatability is measurement results under repeatability conditions, where the independent measurement results are obtained using the same method on the identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. Reproducibility is measurement results under reproducibility conditions, where the measurement results are obtained using the same method on identical test items in different laboratories with different operators using different equipment.

In the ISO 5725 series, the basic model for a measurement result y is given by $y = m + B + e$ to estimate accuracy of measurement method. m is general mean (expectation), B is laboratory component of variation (under repeatability conditions), e is random error (under repeatability conditions).

The expectation and variance of B are assumed to be 0, and σ_L^2 , the between-laboratory variance, respectively. Expectation of e is assumed to be 0 and its variance, the within-laboratory variance, is assumed to be equal in all laboratories and is denoted as repeatability variance σ_r^2 . Repeatability standard deviation σ_r and reproducibility standard deviation σ_R are defined as Eqs. $\sigma_r = \sqrt{V(e)}$ and $\sigma_R = \sqrt{V(B) + V(e)}$.

2.3 Gauge R&R

In Gauge R&R, repeatability and reproducibility are difined as one of the indices which show the variation in a measuring method. Repeatability in Gauge R&R represent the variation in measurement when measurement results are obtained with the same method on the identical object by the same operator using the same equipment within short intervals of time. Reproducibility in Gauge R&R represent the variation in measurement when measurement results are obtained with the same method on the identical object by the different operator using the different equipment. These concepts are much the same, but definitions are a little different.

A Evaluation method for qualitative data is known Kappa model. Fleiss's Kappa (1981) [5] is used as AAA (Attribute Agreement Analysis).

3. METHODS

3.1 Langton Methods

S.D. Langton (2002) proposed that *accordance* is defined as qualitative equivalent of repeatability and *concordance* is defined as qualitative equivalent of reproducibility.

'Accordance'

Accordance is the (percentage) ratio that two identical test materials analyzed by the same laboratory under standard repeatability conditions will both be given the same result (i.e. both found positive or both found negative). Accordance is given by eq.1.

$$A_i = \frac{x(x-1) + (n-x)(n-x-1)}{n(n-1)}$$

$$A = \frac{1}{L}\sum_{i=1}^{L} A_i \qquad (1)$$

x: number of success
n: number of trials
L: number of laboratories
A_i: Accordance for laboratory

'Concordance'

Concordance is the percentage ratio that tow identical test materials sent to different laboratories will both be given the same result (i.e. both found positive or both found negative result). Concordance is given by eq.2.

$$C_i = \frac{2x(x-nL) + nL(nL-1) - A_i nL(n-1)}{n^2 L(L-1)}$$

$$C = \frac{1}{L}\sum_{i=1}^{L} C_i \qquad (2)$$

x: number of success
n: number of trials
L: number of laboratories
A_i: Accordance for laboratory
C_i: Concordance for laboratory

3.2 Wieringen Methods

The experimental design for evaluating the R&R of a binary measurement system involves some objects that are judged repeatedly by some raters. Measurement result is 1 (pass) or 0 (fail). Van Wieringen et al (2008) supposed to using Latent class model. Sensitivity shows the probability that evaluate a confirming product as pass. Specificity shows the probability that evaluate defective as fail.

They estimate confirming rate, sensitivity, and specificity by maximum likelihood method. Moreover, they represent repeatability and reproducibility using log likelihood.

3.3 AAA(Attribute Agreement Analysis)

The data include measuring object's factor and appraiser's factor, and the data repeatedly measured by each operator. AAA is the method of analyzing the agreement of within appraiser and between appraiser using Fleiss's kappa statistics. In AAA, "within appraiser's kappa" and "between appraiser's kappa" are estimated respectively by the probability whose evaluation actually corresponds, and probability whose evaluation corresponds by chance.

3.4 Proposed methods

In the ANOVA about a population defective fraction, there is a method using the normal approximation by logit transformation. We propose the method of estimating repeatability and reproducibility using the above-mentioned ANOVA.

When the number of times of detection of defective products when it measures n times in a laboratory i (i = 1,...,L) is set to x_i, probability of detection π_i^* can be expressed as follows.

$$\pi_i* = \frac{x_i + 0.5}{n+1} \tag{3}$$

When normal approximation by logit transformation is performed to this, it can express as follows.

$$Logit(\pi_i*) = \ln \frac{\pi_i*}{1-\pi_i*} \sim N\left(\mu_i, \frac{1}{n\pi_i(1-\pi_i)}\right) \tag{4}$$

μ_i is a population mean value and π_i is a population defective fraction.

When it is considered as L_i=$Logit$(π_i^*), repeatability variance can be estimated as follows by the one-way layout ANOVA.

$$\sigma_r^2 = \frac{1}{L}\sum_{i=1}^{L} \frac{1}{n\pi_i(1-\pi_i)} \tag{5}$$

Reproducibility variance can be estimated as follows

$$\sigma_R^2 = \frac{1}{L-1}\left(\sum_{i=1}^{L} L_i^2 - \frac{\left(\sum_{i=1}^{L} L_i\right)^2}{L}\right) \tag{6}$$

If it is defined as the sum of repeatability variance and between laboratory variance being reproducibility variance like ISO5725, between laboratory variance can be estimated as follows.

$$\sigma_L^2 = \sigma_R^2 - \sigma_r^2 \tag{7}$$

4. CONSIDERING THESE METHODS

In this chapter, In order to clarify the relation of each method from the application result of Langton, Wieringen, AAA, and the proposal, the application to the result containing a factor of a measuring object like Wieringen or AAA is considered. That is, the result of the binary to multiple objects is assumed. This assumption is for making it easy to compare directly by applying all the methods to the same data.

Langton and the proposal are usually inapplicable to multiple measuring objects. However, when it applies to one certain object, the precision of the every for each can be calculated. This is performed to all the objects and comparison with Wieringen or AAA is performed by the method of getting the average.

When the whole Probability of detection is assumed to be 0.5, with number of objects n = 200, the number of operators m = 3, with three repetitions, assuming that the number of objects are n = 200, the number of operators are m = 3, with three repetitions, the data which gave variation to the agreement in an operator, and the agreement between operators are used.

The scatter diagram between precision in an application result is shown in figure 1~figure 4. The correlation coefficient matrix between precision in an application result is shown in Table 2. In the table, Wieringen (r) shows the repeatability of Wieringen's method, Wieringen (R) shows the reproducibility of Wieringen's method, logit (r) shows the repeatability of the proposal, and logit (R) shows the reproducibility of the proposal.

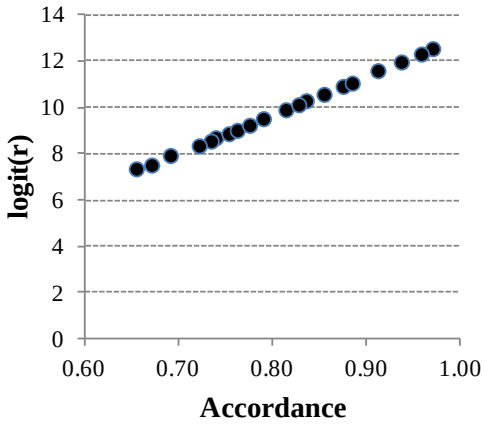

Figure 1 Scatter diagram of Accordance and logit(r)

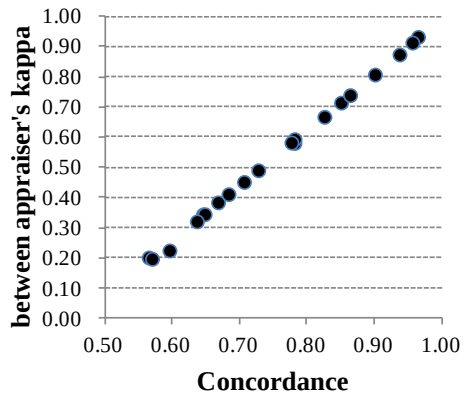

Figure 2 Scatter diagram of Concordance and between appraiser's kappa

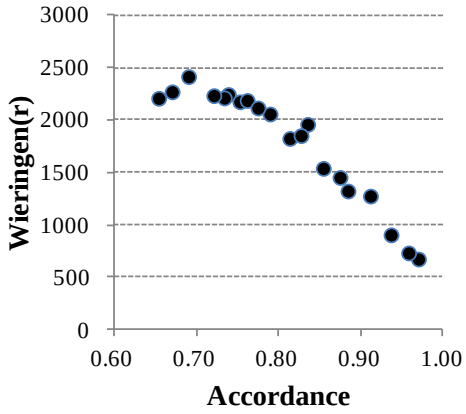

Figure 3 Scatter diagram of Accordance and Wieringen(r)

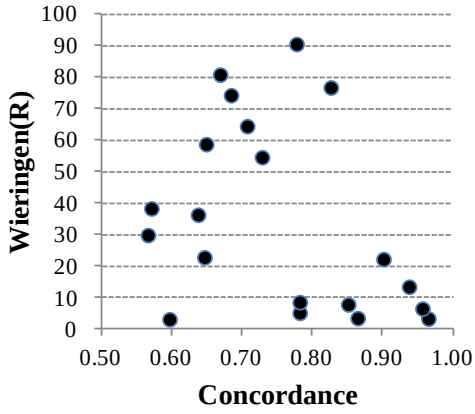

Figure 4 Scatter diagram of Concordance and Wieringen(R)

5. DISCUSSION

Wieringen and AAA which are differ from Langton and proposal, were previously considered measurement error for measuring object. However, AAA is not considered measurement error for measuring object from application results. AAA is not only similar Langton method, but also proposed repeatability is similar Langton's repeatability from results. There are some sort of linear relations between these methods. Proposed between-laboratory correlates with Langton's concordance than proposed reproducibility.

On the other hands, Wieringen's approach differs from other methods. Wieringen considered relationship between measured value for operator and true value for measuring object. Other methods considered error of measured value and agreement.

6. CONCLUSION

In this study, the features of previous studies for precision for binary data were clarified. We proposed logit model and also applied precision for binary data. Each accuracy were compared in various conditions. Consequently, Wieringen is different position and different goal and also another approach.

It is important to consider precision for binary data, Therefore there is needs for standardization. These results would be feedback to international standard.

7. REFERENCES

[1] Langton, S. D., Chevennement, R., Nagelkerke, N., Lombard, B. (2002), Analysing collaborative trials for qualitative microbiological methods: accordance and concordance

[2] Wilrich, P.-Th. (2006), The determination of precision of measurement methods with qualitative results by interlaboratory experiments, Presented to WG2 "Statistics" of ISO/TC34/SC9 "Food Products - Microbiology"

[3] Van Wieringen, W. N., and De Mast, J., (2008), Measurement System Analysis for Binary Data, Technometrics, Vol.50, NO.4, 468-478

[4] Koji Horie, Yusuke Tsutsumi, Yukio Takao, Tomomichi Suzuki (2008), Calculation of Repeatability and Reproducibility for qualitative data", The 6th ANQ 2008

[5] J. L. Fleiss (1981). Statistical Methods for Rates and Propotions, 2nd edition, John Wiley &Sons.

Table 2 correlation coefficient matrix

	Accordance	Concordance	Wieringen(r)	Wieringen(R)	within kappa	between kappa	logit(r)	logit(R)
Accordance	1.00000							
Concordance	0.99251	1.00000						
Wieringen(r)	-0.94430	-0.95709	1.00000					
Wieringen(R)	-0.31822	-0.35499	0.44857	1.00000				
within kappa	0.99990	0.99331	-0.94653	-0.32788	1.00000			
between kappa	0.99552	0.99962	-0.95545	-0.34715	0.99611	1.00000		
logit(r)	0.99973	0.99292	-0.94684	-0.33076	0.99978	0.99577	1.00000	
logit(R)	0.99829	0.98447	-0.93604	-0.31494	0.99796	0.98890	0.99828	1.00000