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

RECENT DEVELOPMENTS IN PRECISION EVALUATION IN NON-QUANTITATIVE MEASUREMENTS

Tomomichi Suzuki, Yusuke Tsutsumi, Ryosuke Hamaguchi and Hironobu Kawamura

Tokyo University of Science, Chiba, Japan, suzuki@ia.noda.tus.ac.jp

Abstract: This paper describes the recent developments in precision evaluation in non-quantitative measurements. Precision evaluation in quantitative measurements is a thoroughly discussed topic and the established methods are in use. Many methods are proposed for qualitative data, but their effectiveness and statistical properties are not so clear. This paper introduces, compares and discusses the methods to evaluate precision for qualitative data.

Keywords: Precision, Repeatability, Reproducibility, Qualitative Data, ISO 5725

1. INTRODUCTION

Precision evaluation in quantitative measurements is a thoroughly discussed topic and the established methods are in use. ISO published ISO 5725 *accuracy* (*trueness and precision*) *of measurement methods and results* which deals with quantitative measurements. Now there is a movement to develop a document for non-quantitative measurements.

Many methods are proposed for qualitative data, but their effectiveness and statistical properties are not so clear. This paper introduces, compares and discusses the methods to evaluate precision for qualitative data.

2. PRECISION FOR QUANTITATIVE DATA

ISO 5725[1] uses two terms trueness and precision to describe the accuracy of a measurement method. Trueness refers to the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value. Precision refers to the closeness of agreement between test results.

ISO 5725 defines the precision using the model shown in equation (1).

$$y = m + B + e \qquad (1)$$

where y is the measurement result, m is the general mean (expectation), B is the laboratory component of bias under repeatability conditions, and e is the random error in every measurement under repeatability conditions. The term B is considered a random variable whose expectation equals zero and whose variance is expressed as σ_L^2. The term e is a random variable whose expectation equals zero and whose variance is expressed as σ_e^2.

ISO 5725[1] also introduce two types of precision, repeatability and reproducibility.

Repeatability is the precision under repeatability conditions. Repeatability conditions are defined as “conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time.” Repeatability indicates the smallest variation for a particular measurement method.

Reproducibility is the precision under reproducibility conditions. Reproducibility conditions are defined as “conditions where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment.” Reproducibility indicates the largest variation for a particular measurement method.

Using the notation shown in equation (1), the repeatability variance σ_r^2 and reproducibility variance σ_R^2 can be expressed as follows.

$$\sigma_r^2 = \sigma_e^2 \qquad (2)$$

$$\sigma_R^2 = \sigma_r^2 + \sigma_e^2 = \sigma_L^2 + \sigma_e^2 \qquad (3)$$

The estimates of repeatability variance and reproducibility variance are calculated from interlaboratory studies or collaborative assessment experiments.

3. PRECISION FOR QUALITATIVE DATA

There are many types of non-quantitative measurements. The type of data vary accordingly, for example, binary data, categorical data, ordinal data, etc. In this paper, the methods to evaluate precision for binary data are considered.

The data format of binary data are shown in table 1. Every laboratory measures the identical test item number of times. The number of laboratories is usually denoted as p, and the number of replication is usually denoted as n. If the study is conducted on multiple items or levels, the data for each item or level can be expressed as table 1. In the binary data case, the value of y_{ik} is either 0 (negative, non-detect, fail, etc.) or 1 (positive, detect, pass, etc.).

The methods to evaluate precision for binary data are presented in the following chapter.

Table 1. Data Format for Binary Data

Laboratory	Run 1	...	Run k	...	Run n
1	y_{11}		y_{1k}		y_{1n}
2	y_{21}		y_{2k}		y_{2n}
:					
:					
i	y_{i1}		y_{ik}		y_{in}
:					
:					
p	y_{p1}		y_{pk}		y_{pn}

4. QULAITATIVE METHODS

This chapter describes the methods proposed to evaluate precision for qualitative data.

4.1 ISO BASED METHOD

This section describes the method proposed by Wilrich[3][4]. This method is based on the approach of ISO 5725 part 2. Treating the measurement values (zero and one) as the same as ISO 5725 in analysis (ANOVA etc.).

The model adopted by Wilrich is shown below. It is similar to the ISO 5725 model.

$$y_{ij} = \pi + \left(\pi_i - \pi\right) + e_{ij} \qquad (4)$$

where π_i denotes the ratio of obtaining the positive result (value="1") for laboratory number i, and π denotes the overall ratio of obtaining the positive result, which can be shown as

$$\pi = \sum_{i=1}^{p} \pi_i / p \qquad (5).$$

Values of π_i and π are estimated by the following equations.

$$\hat{\pi}_i = \sum_{j=1}^{n} y_{ij} / n \qquad (6)$$

$$\hat{\pi} = \sum_{i=1}^{p} \hat{\pi}_i / p \qquad (7)$$

The repeatability variance, the between-laboratory variance component, and the reproducibility variance are estimated by equations (8) to (10).

$$\hat{\sigma}_r^2 = \left(\frac{n}{n-1}\right) \cdot \frac{1}{p} \sum_{i=1}^{p} \hat{\pi}_i \left(1 - \hat{\pi}_i\right) \qquad (8)$$

$$\hat{\sigma}_L^2 = \max\left(0, \frac{1}{p-1} \sum_{i=1}^{p} \left(\hat{\pi}_i - \hat{\pi}\right)^2 - \left(\frac{1}{n-1}\right) \cdot \frac{1}{p} \sum_{i=1}^{p} \hat{\pi}_i \left(1 - \hat{\pi}_i\right)\right) \qquad (9)$$

$$\hat{\sigma}_R^2 = \hat{\sigma}_r^2 + \hat{\sigma}_L^2 \qquad (10)$$

The basic idea behind the method is ISO 5725-2 and that seems reasonable. However, it is not totally clear that the same procedure is equally effective in the qualitative data. Uhlig et al.[5] point out that the estimates for reproducibility variance is not adequate because the estimator does not take into account the true value of π. It will be influenced by the true value of π. Horie et al.[6]

showed that the same variances as above are obtained when beta-binomial distribution (which is one of natural assumptions) is assumed.

4.2 ACCORDANCE AND CONCORDANCE

This section describes the method proposed by Langton[7]. This method introduces two new concept of precision, accordance and concordance. Accordance is a concept similar to repeatability and concordance is a concept similar to reproducibility.

Accordance is derived by considering the pairs of the measurement results within a laboratory. If the results match, the accordance tends to become higher. All the pairs within the laboratory are compared, and that is done for all the laboratories. Concordance, on the other hand, considers the pairs of the measurement results between laboratories. All the pairs between all the combinations of two laboratories are considered. The larger value of concordance means the results of two different laboratories tends to match more, which can be interpreted as smaller variation between laboratories, hence smaller reproducibility variance.

Accordance for a specific laboratory i is calculated by

$$A_i = \frac{k_i(k_i - 1) + (n - k_i)(n - k_i - 1)}{n(n-1)} \quad (11),$$

where n represents number of runs and k_i number of '1's for laboratory i .

Overall accordance is calculated by

$$A = \frac{1}{p}\sum_{i=1}^{p} A_i \quad (12).$$

Concordance for a specific laboratory i is calculated by

$$C_i = \frac{1}{n^2 L(L-1)}\left\{2\sum_{i=1}^{L} x_i\left(\sum_{i=1}^{L} x_i - nL\right) + nL(nL-1) - A_i nL(n-1)\right\} \quad (13)$$

and overall concordance is calculated by

$$C = \frac{1}{L}\sum_{i=1}^{L} C_i \quad (14).$$

The original proposal is based on the fixed effect case. The estimation of accordance and concordance based on the random effect case is presented by van der Voet et al.[8].

4.3 KAPPA STATISTIC

The method uses the Fleiss' Kappa statistic [9]. The method is also known as Attribute Agreement Analysis [10].

Kappa statistic, or kappa coefficient is calculated using the equation below.

$$\hat{\kappa} = \frac{\hat{P}_o - \hat{P}_e}{1 - \hat{P}_e} \quad (15)$$

where $\hat{P}_o$ denotes the probability that the results actually matched and $\hat{P}_e$ denotes the probability that the results match by chance. This kappa statistic can be calculated for between appraisers (between laboratories) and also for within appraisers (within laboratories).

4.4 SENSITIVITY AND SPECIFICITY

This section describes the method proposed by sensitivity and specificity, which are also popular in handling the qualitative data. The main characteristic of this method is that it assumes the true value which is either 0 (negative) or 1 (positive). Sensitivity indicates the ability to identify positive results, and specificity indicates the ability to identify negative results. They are calculated by

Sensitivity = TP/(TP+FN) (16)
Specificity = TN/(TN+FP) (17)

where TP stands for number of true positives, TN for number of true negatives, FP for number of false positives, and FN for number of false negatives. These statistics themselves do not indicate the precision similar to repeatability nor reproducibility. Some methods applying sensitivity and specificity are proposed to evaluate the measurement system [11][12].

4.5 POD MODEL

This method[13] introduces a new concept POD (probability of detection) as a measure for precision values. It assumes underlying distribution, namely normal distributions. The confidence intervals are calculated based on the assumptions.

5. CHARACTERIZATION

The methods for evaluating precision in binary measurements can be characterized in many ways. The characteristics or viewpoints to characterize the methods are proposed and describe below.

Data Type: Only binary data are considered in chapter four, but it would be better if the method can be applied to other qualitative data such as general categorical data. Some methods are directly applicable to multinomial data, and that should be considered.

Material Levels: As discussed in chapter three, measurements are replicated for identical items or items that can be considered as identical. The number of those identical items varies between the methods. Especially for Attribute Agreement Analysis and in Wieringen's Method, the number of the identical items is large, namely the items being the products from a production process.

Intermediate Precision: This is defined in ISO 3534-2[14] as precision in conditions where test results or measurement results are obtained with the same method, on identical test/measurement items in the same test or measurement facility, under some different operating condition. Although it is difficult just to discuss the precision measures themselves, the expansion of the methods to accommodate intermediate precisions should be considered.

True Value: This is if the true value is assumed in the method or the model. The true value can be assumed for each item (positive/negative), or in other cases, assumed for the probability of obtaining positive/negative results. Some descriptive statistics do not assume any true value.

Confidence Interval: Usually point estimation is given for estimating the measures for precision. Some methods also give confidence interval for the estimation, which will be useful when considering the precision of the precision.

Compatibility: Some of the concepts are new in qualitative cases. It would be easier to explain the methods if their ideas are similar to the cases for quantitative cases. Therefore the compatibility with the quantitative cases or compatibility with the existing methods is considered.

The methods described in chapter four are characterized based on the proposed viewpoints. The summary of the methods by the proposed characterization are given in Table 2.

Table 2. Summary of the methods for qualitative precision

Method	Data Type	Material Levels	Intermediate Precision	True Value	Confidence Interval	Compatibility
Mandel	binary	single	n/a	no	n/a	good
Wilrich	binary	single	yes	yes (ratio)	yes	good
Langton	binary /multi	single	n/a, possible	no	n/a	no
Sensitivity Specificity	binary	single/ multiple	n/a	yes (item)	n/a	no
Kappa Statistic	binary /multi	single/ multiple	yes	no	n/a	no
Wieringen	binary	multiple	n/a, possible	yes (item)	n/a	not bad
POD model	binary	multiple	n/a	yes	yes	no

6. DISCUSSION

The methods for evaluating precision in binary measurements are examined. The aforementioned methods have many statistical characteristics in common. Some of the statistics have direct mathematical relation between them. The accordance and concordance have close relation with repeatability and reproducibility proposed by Wilrich. The relation between repeatability and accordance is shown below.

$$\frac{1-A}{2} = \hat{\sigma}^2_{r:wilrich} \qquad (18)$$

The relation between reproducibility and concordance is shown below.

$$\frac{1-C}{2} = \hat{\sigma}^2_{R:wilrich} \qquad (19)$$

Therefore, the statistical properties of these statistics are basically the same. There are also close relation between the Kappa Statistics and the statistics proposed by both Wilrich and Langton.

From statistical point of view, the methods are broadly divided into two groups. The first group consists of Wilrich's method and Langton's method. Although the former assumes a model and the latter does not, they are based on the measurement of one identical sample and do not assume a true value for the sample. The second group consists of Kappa statistic and sensitivity/specificity methods. They assume a true value for each item and evaluate agreements and differences between measurements including the true value.

7. CONCLUSIONS

Many existing methods to evaluate precision for binary data are investigated. In order to characterize the methods, the characteristics and the viewpoints are proposed. Several viewpoints are proposed that can be used to describe the methods for qualitative precision. The existing methods are then summarized based on the proposed viewpoints. The result will be used to clarify the similarities and dissimilarities among the methods.

Also, as we have demonstrated, many close relations became clear between the precision measures adopted by the existing methods. Therefore it can be concluded that many of the existing methods have the common ideas. The model assumptions seem quite different between the methods, but the underlying ideas are not that different as the mathematical calculations are similar. Yet the differences do exist and the methods that should be adopted should change according to users' circumstances.

REFERENCES

[1] ISO 5725, Accuracy (trueness and precision) of measurement methods and result – Part 1 :General principles and definitions, ISO, 1994.

[2] ISO 5725, Accuracy (trueness and precision) of measurement methods and result – Part 2 : Basic methods for the determination of repeatability and reproducibility of a standard measurement methods, ISO, 1994.

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

[4] P.-Th. Wilrich, "The determination of precision of qualitative measurement methods by interlaboratory experiments", Accreditation and Quality Assurance, vol. 15, pp. 439-444, 2010.

[5] S. Uhlig, L. Niewohner, and P. Gowik, "Can the usual validation standard series for quantitative methods, ISO 5725, be also applied for qualitative methods?" Accreditation and Quality Assurance, vol. 16, pp. 533-537, 2011.

[6] K. Horie, Y. Tsutsumi, and T. Suzuki, "Calculation of Repeatability and Reproducibility for Qualitative Data", in Proc. of the 6th ANQ Congress, CDROM, 2008.

[7] S. D. Langton, R. Chevennement, N. Nagelkerke, and B. Lombard, "Analysing collaborative trials for qualitative microbiological methods: accordance and concordance", International Journal of Food Microbiology vol. 79 , pp. 175-181, 2002

[8] H. van der Voet and L. W. D. van Rassmsdonk, "Estimation of accordance and concordance in inter-laboratory trials of analytical methods with qualitative results", International Journal of Food Microbiology, vol. 95, pp. 231-234, 2004.

[9] J. L. Fleiss, "Statistical Methods for Rates and Proportions" 2nd edition, John Wiley &Sons, 1981.

[10] ISO/TR 14468, "Selected illustrations of attribute agreement analysis", ISO, 2010.

[11] O. Danila, S. H. Steiner, and R. J. MacKay, "Assessing a Binary Measurement System", Journal of Quality Technology, vol. 40, pp. 310-318, 2008.

[12] W. N. van Wieringen, J. De Mast, "Measurement System Analysis for Binary Data", Technometrics, vol. 50, pp. 468-478, 2008.

[13] P. Wehling, R. A. LaBudde, S. L. Brunelle, and M. T. Nelson, "Probability of Detection (POD) as a statistical model for the validation of qualitative methods", Journal of AOAC International, vol. 94, pp. 335-347, 2011.

[14] ISO 3534, Statistics – Vocabulary and symbols – Part 2 : Applied Statistics, ISO, 2006.