A Formal Test for Binary R&R Measurement Systems

Vahid Partovi Nia* Masoud Asgharian † Samuel Basetto‡

Abstract

Measurement systems studies are an integral part of most quality improvement processes. A desirable measurement system must produce repeatable and reproducible values, such measurement systems are called R&R. Statistical test of R&R, despite its importance, is less studied and no formal statistical test is proposed for widely applied attribute measurement systems. We propose a statistical test of R&R for an attribute gauge allowing implementation of a significance test. The efficiency of the methodology is demonstrated on an example.

Key Words: Attribute gauge R&R, measurement systems analysis, Pearson goodness of fit statistic, Kappa measure of agreement.

1. Introduction

Most of the statistical quality control and improvement techniques are based on a measured value of a product. A measurement system, sometimes referred to as *gauge*, then is involved as a significant part of quality monitoring and improvement. Therefore, it is important to test or check the reliability of the measurement systems or even monitor the underlying measurement system. There often is some uncertainty associated with such measurement. This uncertainty appears because when the same product is measured repeatedly, the result may not be the same. An effective measurement system produces repeatable values, i.e. if the measurement on a product is repeated, the measurement value remains unchanged and equal to the true measurement. Although, in real world the value of the measurement might have slight variability if the experiment is repeated, this variability must not be a function of the trials or the operator who applies the measurement system. Furthermore, if the same operator measures the same product repeatedly, the error in different replication should not have any specific pattern. In other words, it is of interest to check if the variability of the measurement is a pure noise. In statistical sense, this means to test if the variability of the measurement system is independent of the operator and replication. The end result of many continuous measurement is translated to a pass or fail decision eventually, a binary value. Hence, such measurement systems reduce to a classification problem, being well-studied topic in statistical literature. The variability due to different operators or different replications cause misclassification of the units, i.e. some correct units may be assigned to the deffective group and vice-versa.

Before starting the technical development, we first review some basic definitions of technical terms used throughout this paper. *Capability* refers to the precision of the measurement system, i.e. the less the misclassification, the more capability the measurement system has. In a practical measurement system we expect that the capability of the measurement does not depend on who applies it (*reproducibility*), and how many times the

*Corresponding author, GERAD http://gerad.ca and Department of Mathematics and Industrial Engineering, École Polytechnique de Montreal, Montreal, QC, Canada H3T 1J4, vahid.partovinia@polymtl.ca.

†Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada H3A 2K6. masoud@math.mcgill.ca.

‡Department of Mathematics and Industrial Engineering, École Polytechnique de Montreal, Montreal, QC, Canada H3T 1J4, samuel.basetto@polymtl.ca.

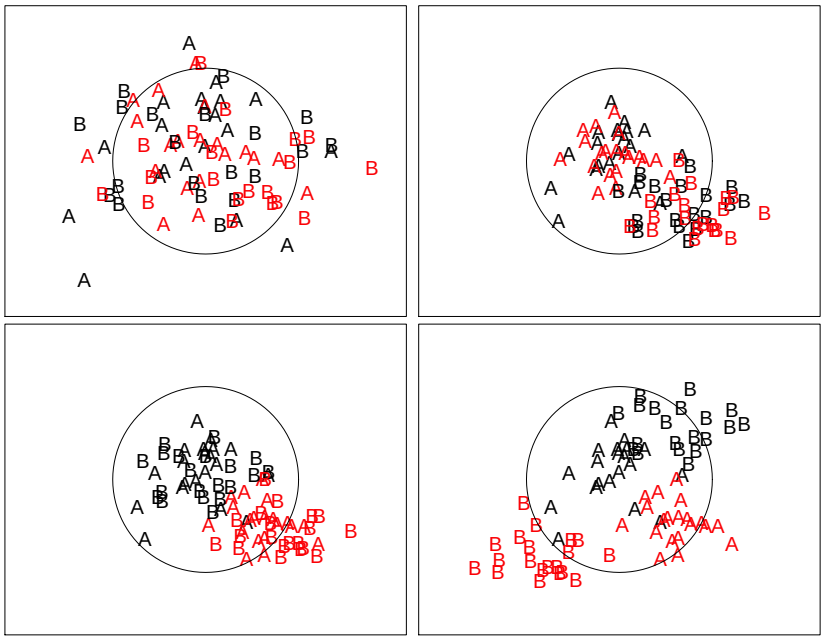

Figure 1: Operator A and B repeat their measurement on 20 parts two times, the first try is coloured red and the second try is coloured black. The tolerance is denoted by the black circle and the true measurement of all parts is the centre of the circle. Example of a measurement system being repeatable and reproducible (top left), being repeatable but not reproducible (top right), being reproducible but not repeatable (bottom left), neither repeatable nor reproducible (bottom right).

measurement systems is used by the same operator (*repeatablity*). For a visual interpretation of these concepts see Figure 1.

Determining the capability of a gauge involves designing a statistical experiment in which several units are measured under controlled conditions. A simple variable gauge study expresses the capability of the gauge as a variance or as a standard deviation of the measurement error. This is modelled as the sum of two variance components, one component for the repeatability and the other for the reproducibility. Therefore an R&R test reduces to a test on variance components. Often a measure of R&R is defined based on the point estimation. An interval estimation of variance components have also been suggested in the literature Burdick and Larsen (1997).

A gauge is variable if the measurement value is continuous and is attribute if the measurement is discrete, such as a pass-fail gauge. An attribute gauge refers to a categorical-outcome measurement system. In most cases the end-product of all measurement systems is acceptance or rejection of a product. Hence, most of the variable measurement systems can be analyzed in the context of the attribute systems. However, the measured continuous value can be used to investigate other properties such as bias, linearity, and stability which is harder to quantify in the context of attribute measurement systems analysis. The focus of this research is a binary outcome measurement system.

Analysis of variable measurement systems was introduced in Montgomery and Runger (1993a) and Montgomery and Runger (1993b) using variance components model. Burdick *et al.* (2003, 2005) provide a comprehensive review on capability analysis for variable measurement systems. Boyles (2001) studies capability analysis for attribute measurement systems. AIAG (2010) provide a methodology for analysis of variable and attribute measurement systems. The method of AIAG (2010) is implemented in major automobile com-

panies and is available in various statistical packages such as STATISTICA, ProMSA, and MINITAB among others. AIAG (2010) suggests an analysis of variance method for analysis of a variable gauge similar to Vardeman and van Valkenburg (1999), and the use of kappa inter-rater agreement (Cohen, 1960) in crosstabs for analysis of a binary measurement system. The method of AIAG (2010), especially in the analysis of attribute systems, involves several serious flaws, although still one of the mostly applied procedures in practice. For instance, the cross-tabulation is affected by the two-by-two crosstab characteristics, and also their suggested method is unable to handle the analysis of attribute multi-class measurement systems. A resolution to the latter problem is proposed by (De Mast and van Wieringen, 2004, 2007). Furthermore, the kappa inter-rater agreement used to characterize the measurement is sensitive to the marginal frequencies of crosstabs, and alternative agreement measures are also proposed in Brennan and Prediger (1981) as well as Gwet (2002). There has been a surge of articles on attribute measurement systems recently, (Bashkansky *et al.*, 2007; De Mast *et al.*, 2011; de Mast and van Wieringen, 2010; Weaver *et al.*, 2012; Lyu and Chen, 2008) generalize the classical methods, (van Wieringen and Van Den Heuvel, 2005) compare the continuous measurement systems analysis with attribute systems, and (Murphy *et al.*, 2009) applies such analysis in practice.

Recently, researchers proposed analysis of the attribute measurement systems using statistical models inspired by variable measurement systems analysis. A fixed-effect nonlinear model was proposed for analysis of attribute measurement systems in Vágó and Kemény (2011a), and a random effect model was suggested in Vágó and Kemény (2011b) when a continuous measurement is additionally available. However, constructing an R&R statistical test based on random effect models is still a difficult task because such a test reduces to test if multiple variance components is zero. Testing variance components with zero is a non-standard testing problem even in linear models, since the null hypothesis lies on the boundary of the parameter space. Therefore, the common likelihood ratio asymptotic approximations, relying on the Taylor approximation of the log likelihood function, are not applicable; for a more detailed discussion see Verbeke and Molenberghs (2003).

We suggest a simple probabilistic approach to model the decision of the operators and to build a statistical test of R&R under mild assumptions. A binary measurement system is the focal point of our study because of its wide applicability. We propose a Pearson goodness of fit statistic that asymptotically follow a chi-square distribution under the R&R hypothesis. This approach provides an asymptotic significance test of R&R. For small sample sizes a finite sample approximation can be considered using parametric or nonparametric bootstrap which we won't discuss in this work further.

2. Methodology

Suppose the total of $n = n_0 + n_1$ units of a product are used in an attribute gauge R&R study. The true state of the unit is a binary variable $\theta_u = k, k \in \{0, 1\}$, where $k = 0, 1$ represents, respectively, unacceptable acceptable units. We consider a balanced design, i.e. we assume that the study is run over I operators and all operators independently repeat their measurements exactly J times for each unit. We index units, operators and their replications by u, i, and j, respectively. The measurement result is $Y_{iju} \in \{0, 1\}$, a binary random variable being the measurement of operator i in its jth replication on unit u. The capability of the system is defined as the strength of the measurement system to discover the true state of the unit. We parametrize this quantity by $\pi_{ij}^{(0)} = \Pr(Y_{iju} = 0 \mid \theta_u = 0)$ for the unacceptable units and $\pi_{ij}^{(1)} = \Pr(Y_{iju} = 1 \mid \theta_u = 1)$ for acceptable units. The probabilities $\pi_{ij}^{(0)}$ is the probability that operator i in its jth essay correctly classifies an

unacceptable unit, $\pi_{ij}^{(1)}$ is defined similarly for *acceptable* unit. This notation of capability implies that the capability is a function of the units only through the true state of the unit u, i.e. θ_u. Obviously $1-\pi_{ij}^{(0)}$ and $1-\pi_{ij}^{(1)}$ are the misclassification probabilities for units being unacceptable or acceptable, respectively.

The true state θ_u is available for all units of the designed experiment, but this information is hidden from the operators in all of its trials. AIAG (2010) suggest to quantify the probability of agreement between operator $i \neq i'$

$$\Pr(Y_{iju} = k_1, Y_{i'ju} = k_2), k_1, k_2 \in \{0, 1\},$$

using the Kappa measure of agreement. These probabilities are used to study the R&R property of an attribute measurement system. We propose to build the R&R analysis directly on the capabilities, i.e. on the conditional probability

$$\Pr(Y_{iju} = k_1, Y_{i'ju} = k_2 \mid \theta_u = k). \tag{1}$$

The advantage of such analysis is its simple interpretation and direct statistical inference because under R&R assumption

$$\Pr(Y_{iju} = k_1, Y_{i'j'u} = k_2 \mid \theta_u = k) = \Pr(Y_{iju} = k_1 \mid \theta_u = k) \Pr(Y_{i'j'u} = k_2 \mid \theta_u = k), \tag{2}$$

for an $i \neq i'$ or a $j \neq j'$.

The immediate consequence of (2) is a well-known way of testing an R&R system, a Pearson test. Measurements Y_{iju} alone maybe useless for construction of an interpretable formal test. One can, however, use $Y_{ijk} \mid \theta_u$ since the true state θ_u is known for each unit u. The latter conditional probability $\Pr(Y_{ijk} = k \mid \theta_u)$ is the capability. In contrast, Y_{ijk} is a mere measurement and does not reflect any property of a measurement system.

To build a Pearson test it is required to calculate the observed, and the expected frequencies under the R&R assumption. The observed frequency assumption is the number of correct decisions for unacceptable units which can be written formally as

$$O_{ij}^{(0)} = \sum_{u=1}^{n_0} \mathbb{I}(Y_{iju} = 0 \mid \theta_u = 0).$$

Likewise, for acceptable units

$$O_{ij}^{(1)} = \sum_{u=1}^{n_1} \mathbb{I}(Y_{iju} = 1 \mid \theta_u = 1).$$

in which $\mathbb{I}$ denotes the indicator function.

If the system is R&R, the capability of the measurement is constant among different operators and among their trials for unacceptable units, i.e. $\pi_{ij}^{(0)} = \pi_{i'j'}^{(0)}$ leading to the following expected frequency under the R&R assumption

$$E^{(0)} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{u=1}^{n_k} \mathbb{I}(Y_{iju} = 0 \mid \theta_u = 0).$$

The same analogy applies for acceptable units, i.e.

$$E^{(1)} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{u=1}^{n_k} \mathbb{I}(Y_{iju} = 1 \mid \theta_u = 1).$$

Having the observed and expected frequencies under the R&R assumption allows us to define a formal statistical test using the Pearson statistic

$$V^{(0)} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left\{O_{ij}^{(0)} - E^{(0)}\right\}^2}{E^{(0)}}.$$

The random variable $V^{(0)}$ follows a chi-square distribution with $IJ-1$ degrees of freedom under R&R assumption. This test checks if the system is R&R on unacceptable units.

A similar statistic can be constructed to test the R&R assumption over acceptable units

$$V^{(1)} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left\{O_{ij}^{(1)} - E^{(1)}\right\}^2}{E^{(1)}},$$

which again follows a chi-square distribution with $IJ-1$ degrees of freedom.

This approach helps us to test the R&R hypothesis on unacceptable and acceptable units separately. If one is interested to test this hypothesis overall, the sum of the two Pearson statistic can be used, i.e.

$$V = V^{(0)} + V^{(1)}.$$

The overall test statistic V follows a chi-square distribution with $2IJ-2$ degrees of freedom under the R&R assumption for both unacceptable and acceptable units. The decomposition of V into two parts allows us to refine the inference and diagnose the problem over unacceptable and acceptable parts separately if the test is rejected.

3. Application

This section applies the proposed methodology based on the Pearson statistic. We examine the sample attribute data of (AIAG, 2010, p. 134). An experiment is run using three appraisers, say appraiser A, appraiser B, and appraiser C, over 50 different units among which 32 are acceptable and the remaining 18 are not. The true nature of the measurement is continuous. All units are measured independently with a variable measurement system to keep this continuous measurement as a reference for the correct decision.

Kappa	A	B	C
A		0.86	0.78
B	0.86		0.79
C	0.78	0.79	

Table 1: The Kappa measure of agreement between the three appraisers.

Kappa	A	B	C
True state (θ_u)	0.88	0.92	0.77

Table 2: The Kappa measure of agreement of each appraiser with the true state of parts.

A fast way of classification of units is to ask an expert to pass or fail each unit. Such decision can be made by visual inspection or using some other discrete measurement system such as *go no-go gage*. The main issue is then the possibility of replacing the time consuming continuous measurement system with another faster and cheaper attribute measurement

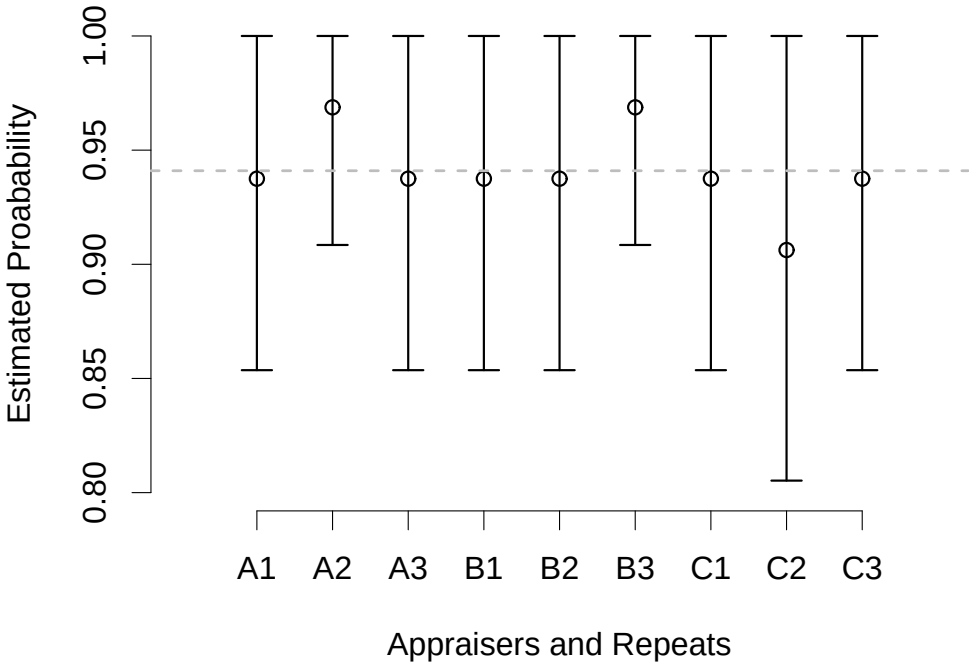

Figure 2: The estimated probabilities of correct decision for appraiser A, B, and C in their three trails over correct units. The confidence intervals are constructed using the normal approximation. The dashed gray line shows the average of these probabilities.

system. In the following we analyze the data to see if the pass-fail inspection gives an R&R measurement system.

First we explore the data with the classical analysis using the Kappa measure of agreement. The Kappa measure of agreement between appraisers are reported in Table 1. AIAG (2010) concludes that there is a good agreement of appraisers across themselves and with the true state of the units. Appraisers also show a reasonable amount of agreement with the true state of the units, see Table 2. The Kappa measure does not provide a test and it only measures the amount of agreement between two raters in a two by two crosstab. Our new proposed methodology gives a formal statistical test. By applying the chi-square statistic on the data we obtain the following result. The computed chi-square statistic over the data gives $V = 3.00$. Using the fact that V χ^2_{16} the p-value is > 0.99 and therefore the R&R assumption is confirmed. This result agrees with the analysis of AIAG (2010). Figure 2 illustrates the capabilities with their approximate normal confidence interval for different appraisers over their different replications. The result of the test agrees with the visual inspection over the estimated capabilities of Figure 2 as well.

References

AIAG (2010) *Measurement System Analysis; Reference Manual*. Fourth edition. Automotive Industry Action Group.

Bashkansky, E., Dror, S., Ravid, R. and Grabov, P. (2007) Effectiveness of a product quality

classifier. *Quality Engineering* **19**(3), 235–244.

Boyles, R. (2001) Gauge capability for pass-fail inspection. *Technometrics* **43**(2), 223–229.

Brennan, R. and Prediger, D. (1981) Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement* **41**(3), 687–699.

Burdick, R., Borror, C. and Montgomery, D. (2003) A review of methods for measurement systems capability analysis. *Journal of Quality Technology* **35**(4), 342–354.

Burdick, R. and Larsen, G. (1997) Confidence intervals on measures of variability in r&r studies. *Journal of Quality Technology* **29**(3), 261–273.

Burdick, R. K., Borror, C. M. and Montgomery, D. C. (2005) *Design and Analysis of R & R Studies*. SIAM.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**(1), 37–46.

De Mast, J., Erdmann, T. P. and Van Wieringen, W. N. (2011) Measurement system analysis for binary inspection: Continuous versus dichotomous measurands. *Journal of Quality Technology* **43**(2), 99–112.

De Mast, J. and van Wieringen, W. (2004) Measurement system analysis for bounded ordinal data. *Quality and Reliability Engineering International* **20**(5), 383–395.

De Mast, J. and van Wieringen, W. N. (2007) Measurement system analysis for categorical data: Agreement and kappa type indices. *Journal of Quality Technology* **39**(3), 191–202.

Gwet, K. (2002) Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-rater Reliability Assessment* **1**(6), 1–6.

Lyu, J. and Chen, M.-N. (2008) Gauge capability studies for attribute data. *Quality and Reliability Engineering International* **24**(1), 71–82.

de Mast, J. and van Wieringen, W. N. (2010) Modeling and evaluating repeatability and reproducibility of ordinal classifications. *Technometrics* **52**(1), 94–106.

Montgomery, D. and Runger, G. (1993a) Gauge capability and designed experiments. part i: basic methods. *Quality Engineering* **6**(1), 115–135.

Montgomery, D. and Runger, G. (1993b) Gauge capability analysis and designed experiments. part ii: experimental design models and variance component estimation. *Quality Engineering* **6**(2), 289–305.

Murphy, S. A., Moeller, S. E., Page, J. R., Cerqua, J. and Boarman, M. (2009) Leveraging measurement system analysis (msa) to improve library assessment: the attribute gage r&r. *College & Research Libraries* **70**(6), 568–577.

Vágó, E. and Kemény, S. (2011a) A model-based approach for attribute gauge analysis. *Quality and Reliability Engineering International* .

Vágó, E. and Kemény, S. (2011b) Random effects model for attribute gauge r&r. *Quality and Reliability Engineering International* .

Vardeman, S. and van Valkenburg, E. S. (1999) Two-way random-effects analyses and gauge r&r studies. *Technometrics* **41**(3), 202–211.

Verbeke, G. and Molenberghs, G. (2003) The use of score tests for inference on variance components. *Biometrics* **59**(2), 254–262.

Weaver, B. P., Hamada, M. S., Vardeman, S. B. and Wilson, A. G. (2012) A bayesian approach to the analysis of gauge r&r data. *Quality Engineering* **24**(4), 486–500.

van Wieringen, W. N. and Van Den Heuvel, E. R. (2005) A comparison of methods for the evaluation of binary measurement systems. *Quality Engineering* **17**(4), 495–507.