Quality Engineering

Publication details, including instructions for authors and subscription information: http://www.tandfonline.com/loi/lqen20

Discussion of “The Statistical Evaluation of Categorical Measurements: ‘Simple Scales, but Treacherous Complexity Underneath’”

Oana Danila [a], R. Jock MacKay [a] & Stefan H. Steiner [a]

[a] Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

Published online: 11 Dec 2013.

Quality Engineering, 26:33–39, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2013.846065

Taylor & Francis
Taylor & Francis Group

Discussion of "The Statistical Evaluation of Categorical Measurements: 'Simple Scales, but Treacherous Complexity Underneath'"

Oana Danila,
R. Jock MacKay,
Stefan H. Steiner

Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

INTRODUCTION

The paper "The Statistical Evaluation of Categorical Measurements: Simple Scales, but Treacherous Complexity Underneath" by Jeroen de Mast, Thomas Akkerhuis, and Tashi Erdmann (2014), which we hereafter refer to as de Mast's paper, addresses the important and challenging task of assessing the performance of categorical measurement systems in manufacturing industry and diagnostic tests in medicine.

The paper raises awareness of the complexity that underlies the apparent simplicity of a categorical/binary measurement system. It emphasizes the fact that most of the past and current methods involve assumptions that oversimplify the reality of the measured process. It gives an excellent account of the use of the kappa statistic, commonly used in psychometrics, medicine, and manufacturing as the measure of appraiser agreement and discusses the serious issues related to its behavior and interpretation. The authors also discuss the statistical modeling of binary measurement systems and emphasize the effect of using overly simplified assumptions on inference. They give relevant examples that are helpful in understanding the issues discussed in the paper.

The authors also draw attention to another crucial aspect related to the assessment of a categorical/binary measurement system. That is, aside from the challenges related to statistical modeling, things are further complicated by the realities of the contexts where these measurement systems are used. For example, when a manufacturing process is high quality and noncomforming products are rare, care has to be taken in designing an assessment study so that enough nonconforming products are included in the study sample. Also, it is common that a gold-standard system is not available or is just partially available for the assessment study, which adds another layer of complexity to the study design and analysis.

In this discussion, we focus on the assessment of a single binary measurement system in a specific context that is commonly found in manufacturing industry. In Sections 4 and 5 of de Mast's paper, the authors address some of the issues encountered in this particular context

Article presented at the First Stu Hunter Research Conference in Heemskerk, Netherlands, March 2013.

Address correspondence to Oana Danila, Business and Industrial Statistics Research Group, Department of Statistics and Actuarial Sciences, University of Waterloo, 200 University Ave. West, Waterloo, ON, Canada N2L 3G1. E-mail: omdanila@uwaterloo.ca

and also propose a new method for analysis. Our goal here is to further clarify some of the concepts and statistical models commonly used in this context, to give an alternative to the method proposed in de Mast's paper, and to contrast the two methods.

We start by describing the context of interest and the concepts and statistical models involved. Next, we introduce our method and compare it to the one proposed in de Mast's paper. We close our discussion with some recommendations for the study design tailored to the context of interest and give general conclusions related to the assessment of a binary measurement system.

CONTEXT OF INTEREST

Binary measurement systems (BMSs) are commonly found in manufacturing industry as well as in medicine where they are known as diagnostic or screening tests. Examples of BMSs include systems for visual inspection, go–no–go gauges, leak tests, etc. One important feature of a BMS is that the measurand or the true quality status X is binary (i.e., a product is either conforming, $X=1$, or nonconforming, $X=0$), although conformance might be defined based on a (usually large) collection of product characteristics, some of them continuous and others categorical. Another very important aspect of the measurand X is that it has to be clearly defined based on some prespecified quality standards. For example, when the characteristics underlying the definition of X are continuous, a certain threshold has to be clearly defined in order to classify a product as conforming or nonconforming. In the case discussed in Section 5 in de Mast's paper, a product is defined as conforming when the width and depth of any surface scratch are less than some specified thresholds. Another example involves the visual inspection of blank credit cards by a human operator who checks the cards for various surface imperfections such as scratches, color bleeding, missing letters, etc. In this case, the quality standard defines a card with any of these defects as nonconforming.

The true quality state X can be determined by a definitive or error-free measurement system, known as the *gold standard.* In practice, it is common that the gold standard is not available for the BMS assessment study because it might be too expensive or destructive. In this discussion, we focus on statistical models and study designs used when a gold standard is not available for the assessment study. However, we want to reemphasize here that though information about the true state X might not be available during the study, X still has to be well defined in order to train the BMS for quality inspection.

In the context that we focus on here, the BMS is nondestructive, so repeated measurements on a product are possible. In addition, the BMS is stable during the time of the study and the manufacturing process is under statistical control. The goal of the assessment study is to estimate the two misclassification probabilities; the false acceptance probability $\alpha=\Pr(Y=1\,|\,X=0)$, also known as the customer's risk; and the false rejection probability $\beta=\Pr(Y=0\,|\,X=1)$, also known as the producer's risk, where Y is the result of the BMS inspection.

BASIC LATENT CLASS MODEL

In the case where information about the true state X is not known, statistical models with latent variables are commonly used. The simplest one, called the conditional independence model in Section 4 of de Mast's paper, has been used for a long time in medicine, psychology, and manufacturing (Boyles 2001; Hui and Zhou 1998; Van Wieringen and de Mast 2008) and is now considered overly simplistic by most researchers in these fields (Qu et al. 1996; Torrance-Rynard and Walter 1997; Vacek 1983).

The model, which we call here the *basic latent class*, makes three main assumptions:

1. Homegeneity of misclassication rates: α and β are constant within the populations of nonconforming and conforming products, respectively.
2. Conditional independence: given the true state, two or more repeated measurements on the same product are independent:

$$\Pr(Y_1=y_1, Y_2=y_2\,|\,X=x)=\Pr(Y_1=y_1|\,X=x)\times\Pr(Y_2=y_2\,|\,X=x).$$

3. Measurements on different products are independent.

As a consequence of these assumptions, two measurements on the same product, Y_1 and Y_2, are marginally dependent and the likelihood contribution for a randomly selected product measured k times is a mixture of two binomial distributions. In order to identify all of the parameters of interest, a minimum number (usually three) of repeated measurements on the same product are required.

The homogeneity assumption is the most difficult to justify in practice, although the conditional independence receives the most criticism. Usually, as mentioned in de Mast's paper, when the definition of the measurand X is based on a complex combination of product characteristics, some products are more difficult to correctly classify than others. In the example involving a visual inspection for scratches, products with deeper or longer scratches are more easily (with higher chance) classified as nonconforming than those with barely visible scratches. In addition, there are other product characteristics not related to X, such as the color of the product, that influence the chance of correct classification. Therefore, in most cases, assuming that the misclassification errors are constant represents an unrealistic simplification, and models relying on this assumption lead to serious bias in the estimators of α and β (Albert and Dodd 2004; Danila et al. 2012; de Mast et al. 2011).

The conditional independence assumption as defined above is more of a mathematical construct and has nothing to do with the design of the assessment study. As we can see in Section 4 in de Mast's paper and later in our discussion, conditional independence has to be assumed at a certain level of conditioning, no matter how complicated the model for explaining the variation in misclassification rates is. For example, in Section 4, the authors assume that, when conditioning on the misalignment variable Z and X, repeated measurements on the same product are independent.

RANDOM EFFECTS MODELS

In practice, there might exist many characteristics, $\mathbf{Z} = (Z_1, \ldots, Z_p)$, that influence the chance that, for example, a nonconforming product is accepted. These characteristics might be related to the definition of the measurand X or not and, are not directly measurable during the study and therefore are considered latent. As a consequence, the chance of accepting a nonconforming product varies from product to product, and the same can be true about the chance of rejecting a conforming product. In the surface scratches example, the chance of accepting a product with scratches depends on the length, width, depth, and shape of the scratch and also on the color of the product, lighting conditions, etc.

A general approach that deals with this situation assumes explicitly or implicitly that the pass probability varies within the populations of conforming/nonconforming products and that, conditioning on all latent characteristics, including X, the repeated measurements on one product are independent. As a result, given only X, repeated measurements on a product are not independent as assumed by the basic latent class model.

Several statistical models have been proposed under this approach. One class of models, including the one proposed in Sections 4 and 5 in de Mast's paper (which we will call de Mast's model), assumes that the probability of accepting a product is defined by a function of some identified product characteristics related to the measurand X and some other product or environmental variables. These characteristics are not directly measured during the assessment study (i.e., they are latent). Another class of models assumes a composite effect of all of the latent variables on the probability of accepting (rejecting) a nonconforming (conforming) product, which is captured by linking α_i (β_i) to a random effect specific to the product i ($i = 1, \ldots, n$) (Albert and Dodd 2004; Qu et al. 1996) or by directly considering α_i and β_i as random effects (Albert and Dodd 2004; Danila et al. 2012; Fujisawa and Izumi 2000).

The first class of models further assumes some joint distribution for the latent variables $\mathbf{Z}$ and a certain characteristic function $q(\mathbf{Z})$, whereas the second class of models assumes a certain distribution for the random effects. Then, given $\mathbf{Z}$ or the random effects, repeated measurements on a product i are assumed independent. Note that conditional independence has to be assumed at one point in order to build a statistical model for the repeated measurements, with the Waterloo approach the level of conditioning is deeper than with the basic latent class model.

The approach we proposed in Danila et al. (2012), called here the Waterloo approach, assumes that the random effects α_i and β_i are beta-distributed with means μ_α and μ_β, respectively, and also assumes conditional independence given the random effects. Therefore, for a conforming product i, the probability of observing a total number of passes $S_i = s$ out of k total number of measurements, given β_i, is

$$\Pr(S_i = s \mid \beta_i, X_i = 1) = \binom{k}{s}(1-\beta_i)^s \beta_i^{k-s}$$

Similarly, for a nonconforming product i,

$$\Pr(S_i = s \mid \alpha_i, X_i = 0) = \binom{k}{s}\alpha_i^s (1-\alpha_i)^{k-s}$$

The product-specific random effects α_i (β_i) can be interpreted as the proportion of time that product i would pass (fail) the BMS inspection, given that a large number of repeated inspections are conducted. The goal of the assessment study now becomes the estimation of the average error rates μ_α and μ_β, which are the average customer's and producer's risks.

The Waterloo approach and de Mast's model are conceptually similar. That is, they both assume that the misclassification rates vary within the populations of conforming and nonconforming products. However, these models also differ in several aspects:

- de Mast's model specifies the joint distribution of some characteristics $\mathbf{Z} = (Z_1, \ldots, Z_p)$ and links the individual α_i and β_i to these variables through the characteristic function $q(\mathbf{Z})$. This approach requires that all characteristics with a potential effect on individual misclassification rates be identified. When there is only one product characteristic influencing the measurement process as in the misalignment example in Section 4 of de Mast's paper, the model includes two parameters. However, when there are two or more latent variables, the number of model parameters indexing the joint distribution of the latent variables $\mathbf{Z}$, including the correlation parameters, increases dramatically. Note that in the example from Section 5 of de Mast's paper, the latent variables $\mathbf{Z}$ are assumed independent, which might not be realistic in practice. On the other hand, the Waterloo model directly specifies a distribution for the individual misclassification rates and it does not require identification of relevant latent variables. The model just considers the composite effect of all of them and includes five parameters—two mean two variation parameters and one parameter corresponding to the conforming rate.
- As a consequence, in the case where the process influencing the BMS performance can be mainly explained by a small number of product characteristics $\mathbf{Z}$ ($p \leq 2$) as in the misalignment example, de Mast's model might be a better choice than the Waterloo model. However, the Waterloo model is more parsimonious for cases where $p \geq 3$, because it is flexible and easy to understand. Note that distributions other than the beta can be assumed for the random effects α_i and β_i.
- The estimation procedure for de Mast's model can become quite difficult once the number of latent variables becomes large, which is a common case in practice. On the other hand, the estimation of the parameters from the Waterloo model is relatively simple.

As mentioned at the beginning of this section, other statistical models have been proposed for analyzing data from an assessment study when a gold standard is not available. However, work by Albert and Dodd (2004) showed that latent class models are not robust to departures from the assumed distribution of the random effects. These authors demonstrated that the estimators of the average misclassification rates might be severely biased when the underlying model generating the data is very different than the fitted one. In addition, in a realistic study design, it can be quite difficult to choose between several competing models. They cautioned practitioners about blindly relying on estimates of average misclassification rates obtained from studies for which no gold standard information is available.

On the other hand, when the gold standard is available for complete verification of a random sample of products, Danila et al. (2013) showed that the simple moment estimators for the average misclassification rates are unbiased even when the individual rates vary from product to product. Albert and Dodd (2008) also showed that when the gold standard is used only on a random collection of products from the study sample, different latent class models become robust to mispecification of the distribution of the underlying random effects. Albert (2009) provided another solution to the lack of robustness problem and recommends including in the assessment study an imperfect reference test with good performance and known characteristics.

STUDY DESIGN RECOMMENDATIONS

Next, we discuss some important issues related to the design of a BMS assessment study. Some of these issues are also addressed in Section 4 of de Mast's paper, but here we take the opportunity to emphasize some aspects and add some relevant recommendations.

The standard protocol for conducting an assessment study for a BMS when a gold standard is not available involves the following steps:

- Select a random sample of n products from the manufacturing process.
- Measure each sampled product k times with the BMS.
- Record the total number of passes for each product $s_1, \ldots, s_n$.

If we assume that the misclassification probabilities are constant within the population of conforming and nonconforming products, we can estimate all of the parameters given by the basic latent class model when $k \geq 3$. If we assume a random effects model such as the Waterloo one, all parameters are identifiable when $k \geq 5$.

In practice, it is common that a study is conducted as a regular assessment of a BMS that has been in use for routine inspection. Therefore, large collections of products previously accepted or rejected by the BMS are available for the study. In addition, the BMS tracks the number of products passed over time and therefore prior (baseline) information about the pass rate of the studied BMS is available. Additionally, nowadays, most of the manufacturing processes are high quality and high volume. These common features of the manufacturing process and the BMS, together with the ones mentioned in Section 2, represent challenges in designing a BMS assessment study but also offer additional information that can be turned into solutions when used cleverly.

For example, when the manufacturing process is high quality it yields very few nonconforming products. A random sample from the process might not contain enough nonconforming products to estimate the average customer's risk with good precision unless the sample size is very large. However, a conditional selection where products are randomly selected from the collections of previously rejected products reduces the risk of having too few nonconforming products. It is expected that with such a selection protocol we will better estimate the average customer's risk, but the precision of the average producer's risk will decrease. Danila et al. (2010, 2012) showed that when the study data are supplemented with baseline information about the pass rate, the precision of all estimators is improved when compared to that given by a standard plan with the same study sample size.

As an example, we give some results from a simulation study where data were generated from a beta-distributed random effects model using two selection plans. Figure 1 gives the ratio of sample standard deviations for the average customer's ($\hat{\mu}_\alpha$) and producer's ($\hat{\mu}_\beta$) risks and of the conforming rate $\hat{\pi}_C = \Pr(X = 1)$ given by a beta-distributed random effects model when we compare a standard protocol (SP) vs. a conditional selection (CS) with products selected from the stream of previously failed. For both plans, the study sample includes $n = 1{,}000$ products, each measured $k = 10$ times. In addition, the study data from the CS plan are augmented with prior information about the pass rate based on 10,000 baseline products. For each selection plan, we conducted 1,000 simulation runs for each combination of parameter values (μ_α, $\mu_\beta = 0.02$, 0.1, conforming rate $\pi_C = 0.85$, 0.95), when the variability

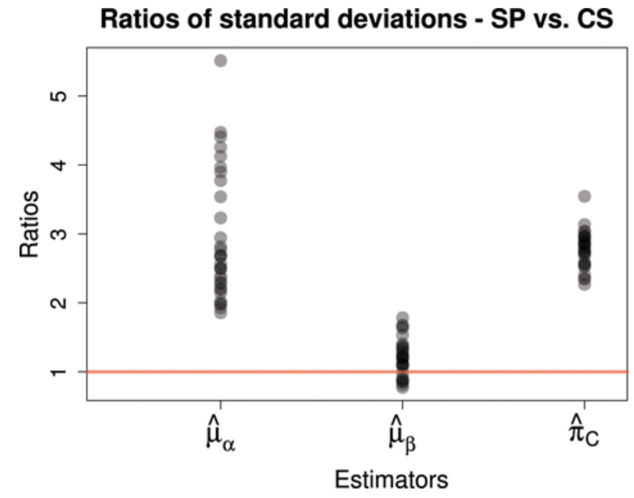

FIGURE 1 ***sd*(*SP*)/*sd*(*CS*); 10, 000 baseline products; $n = 1{,}000$; $k = 10$. (Color figure available online.)**

of product-specific misclassification rates is moderate and low. We note that the gain in precision for $\hat{\mu}_\alpha$ and $\hat{\pi}_C$ is substantial when the CS plan with prior information about the pass rate is used. The estimator of μ_β has better precision for the CS in most cases, although the difference is not as large as for the other two estimators.

CONCLUSIONS

"The Statistical Evaluation of Categorical Measurements: Simple Scales, but Treacherous Complexity Underneath" is a laudable initiative in identifying issues specific to the current statistical methodology for the assessment of categorical/binary measurement systems. The authors warn researchers and practitioners about the possible complexity behind a simple categorical/binary measurement output, and they recommend avoiding models that make overly simplistic assumptions.

The paper also discusses statistical methods for the case where the true quality status of products is not known during the assessment study of a binary measurement system and proposes a new latent class model. This model is a good choice when the underlying process that influences the properties of the measurement system is driven by a small number of identified product or environmental characteristics. However, the model can get quite complicated when this number is large or when there are other unidentified latent variables. We recommend an alternative approach that models the variation of the misclassification rates directly, therefore including a composite effect of all possible latent variables on these rates. This alternative model is generally more parsimonious and involves a simpler estimation procedure.

Based on previous research from the medical field, we further conclude that it is generally difficult to assess the performance of a binary measurement system when no information about the true quality state of the products is available, in which case, the analysis is based on latent class models. The properties of the estimators given by latent class models are usually sensitive to the assumptions regarding the distribution of the product-specific misclassification rates. Partial verification by the gold standard or using a high-performance reference test with known characteristics are two possible solutions to the lack of robustness problem.

In terms of designing an assessment study, we recommend sampling the products used in the assessment study wisely and incorporating any relevant prior information related to the measurement system or the manufacturing process into the analysis. For example, sampling products from the stream of previously rejected and using prior information about the pass rate can substantially improve the precision of the estimators for the (average) customer's and producer's risks.

ABOUT THE AUTHORS

Oana Danila is a postdoctoral fellow in the Department of Statistics and Actuarial Science at the University of Waterloo. She received her M.S. and Ph.D. degrees in statistics from the University of Waterloo. Her research work focuses on the assessment of measurement systems in manufacturing industry and of diagnostic tests in medicine. Oana also worked as a research assistant professor in the Department of Applied Health Sciences at the University of Waterloo and did an internship with the National Cancer Institute, the Clinical Trials Group, in Kingston, Ontario, Canada.

R. Jock MacKay is a retired associate professor in the Statistics and Actuarial Science Department and past director of the Institute for Improvement of Quality and Productivity at the University of Waterloo. He is also an active consultant who has worked with organizations from a wide range of industries, including automotive, telecommunications, aerospace, government, and more.

Stefan H. Steiner is Professor in the Department of Statistics and Actuarial Science as well as the Director of the Business and Industrial Statistics Research Group at the University of Waterloo. He holds a Ph.D. in business administration (management science/systems) from McMaster University. His primary research interests include quality improvement, statistical process control, experimental design, and measurement system assessment. He is a Fellow of the American Society for Quality.

References

Albert, P. S. (2009). Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Statistics in Medicine*, 28(5):780–797.

Albert, P. S., Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60:427–435.

Albert, P. S., Dodd, L. E. (2008). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association*, 103(481):61–73.

Boyles, R. A. (2001). Gage capability for pass-fail inspection. *Technometrics*, 43:223–229.

Danila, O., Steiner, S. H., MacKay, R. J. (2010). Assessment of a binary measurement system in current use. *Journal of Quality Technology*, 42(2):152–164.

Danila, O., Steiner, S. H., MacKay, R. J. (2012). Assessing a binary measurement system with varying misclassification rates using a latent class random effects model. *Journal of Quality Technology*, 44(3):179–191.

Danila, O., Steiner, S. H., MacKay, R. J. (2013). Assessing a binary measurement system with varying misclassification rates when a gold standard is available. *Technometrics*, 55:335–345.

de Mast, J., Erdmann, T. P., 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., Akkerhuis, T., Erdmann, T. (2014). The statistical evaluation of categorical measurements: ''Simple Scales, but Treacherous Complexity Underneath.'' *Quality Engineering*, 26:16–32.

Fujisawa, H., Izumi, S. (2000). Inference about the misclassification probabilities from repeated binary responses. *Biometrics*, 56: 706–711.

Hui, S. L., Zhou, X. H. (1998). Evaluation of diagnostics tests without gold standards. *Statistical Methods for Medical Research*, 7:354–370.

Qu, Y., Tan, M., Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52:797–810.

Torrance-Rynard, V. L., Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, 16:2157–2175.

Vacek, P. (1983). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41:959–968.

Van Wieringen, W. N., de Mast, J. (2008). Measurement system analysis for binary data. *Technometrics*, 50:468–478.