Measurement System Analysis for Binary Inspection: Continuous Versus Dichotomous Measurands

JEROEN DE MAST and TASHI P. ERDMANN

Institute for Business and Industrial Statistics of the University of Amsterdam (IBIS UvA), Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands

WESSEL N. VAN WIERINGEN

Department of Epidemiology and Biostatistics, VUmc University Medical Center & Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV, Amsterdam, The Netherlands

We review methods for assessing the reliability of binary measurements, such as accept/reject inspection in industry. Our framework introduces two factors that are highly relevant in deciding which method to use: (1) whether a reference value (gold standard) can be obtained and (2) whether the underlying measurand is continuous or truly dichotomous. Artificially dichotomizing a continuous measurand, as is commonly done, creates complications that are underappreciated in the literature and in practice. In particular, it introduces an intrinsic reason for the assumption of conditional i.i.d. to be violated. For most methods, this is not crucial provided the samples are random (or at least representative). But, also for most methods, it is, in general, not clear how one can obtain a random sample from the relevant population. The taxonomy presents methods that are generally known in industry, such as nonparametric estimation of false-acceptance and false-rejection probabilities, AIAG's analytic method (logistic regression), latent class modeling, and latent trait modeling. The methods discussed are applied to an example presented in the measurement-system-analysis manual from the automotive industry.

Key Words: Agreement; Binary Measurement; Categorical Data; Conditional Independence; Gauge Capability; Latent Variable Modeling; Pass/Fail Inspection; Repeatability; Reproducibility.

Introduction

In industry, binary measurements abound. Think of visual inspections of products where the outcome can be 'pass' or 'fail', functional tests where the outcome is 'ok' or 'nok', and automated tests where some parts are rejected and others are accepted. Also beyond industry, binary classifications are omnipresent, as in diagnostic tests in medicine (think of a pregnancy test).

Binary measurement aims to classify items as 'accept' or 'reject', or in terms of another dichotomy, in such a way as to reflect a relevant underlying property of the items such as 'good' versus 'defective'. This property that underlies the binary classification is traditionally called the item's 'true value' or 'true state', but, following the terminology in ISO's *Guide to the Expression of Uncertainty in Measurement (GUM)* (International Organization for Standardization (1995)), is better called the 'measurand'.

Binary inspections are, like all measurements, subject to error, especially because they are often based

Dr. De Mast is Associate Professor and Principal Consultant at IBIS UvA. He is a Senior Member of ASQ. His email address is j.demast@uva.nl.

Mr. Erdmann is Consultant and PhD Student at IBIS UvA. His email address is t.p.erdmann@uva.nl.

Dr. Van Wieringen is Assistant Professor at the VUmc and the VU. His email address is wvanwie@few.vu.nl.

on visual or other sensory assessments by humans. A measurement system analysis (MSA), an assessment of the quality and reliability of a measurement procedure, is as important for binary inspections as it is for other types of measurements. The literature describes a multitude of methods for studying the reliability of binary measurements; see, for example, Van Wieringen and Van den Heuvel (2005) for an overview; Boyles (2001), Danila et al.(2008), Van Wieringen and De Mast (2008), and Danila et al. (2010) for recent contributions in quality engineering; and Pepe (2003) for a recent overview in the diagnostic sciences. In this paper, we aim to provide insight into the question of when and how these methods should be applied. We introduce two factors that, in our view, are decisive in designing an MSA study for assessing the reliability of a binary measurement procedure. One factor, the availability of a so-called *gold standard*, is generally recognized. The other factor, whether the measurand is a true dichotomy or rather a continuum, is, as we see it, underappreciated, despite the strong ramifications this distinction has for i.i.d. assumptions and the need for random sampling.

In the next section, we introduce and discuss these two factors and the concept of a *false dichotomy*. The subsequent four sections treat the situations where the measurand is dichotomous or continuous and where a gold standard is available or unavailable. In each of the four settings, we briefly describe methods for experimental design and estimation available in the literature and we discuss potential complications that arise, especially in the case of a false dichotomy. Some of our concerns, as well as a proposal for dealing with false dichotomies, are illustrated from an example taken from the automotive industry's MSA reference manual. We summarize the ramifications of our analyses in a Conclusions section.

General Set-Up

We denote the result of an accept/reject type of measurement as Y, which can be 0 ('reject') or 1 ('accept'). The measurand ('true value') is denoted X, which can be a discrete or a continuous property. Our taxonomy of methods discerns four situations, depending on whether a reference value (gold standard) is available and whether the underlying measurand is continuous or a true dichotomy.

Availability of a Gold Standard

The measurand is often unknowable on principle. But what we may be able to know instead is the result of the application of a higher order, authoritative measurement procedure. This sometimes-available, but usually hypothetical, authoritative result is the item's 'reference value'; in the diagnostic sciences, it is called a 'gold standard'. Although the measurand and the reference value are conceptually not the same, for practical purposes, we take the reference value to play the role of the measurand (meaning that we assume that there is no error in the reference classification). For example, by means of a more thorough analysis or examination, one may establish whether a rejected part is truly defective or whether a woman who obtains a positive result from a pregnancy test is truly pregnant; the result of this higher order analysis is the reference value or gold standard. If a gold standard is unavailable, an assessment of the reliability of binary inspections must treat the measurand as a latent value, and the methods to be discussed for that situation resort to latent variable modeling.

Continuous and Dichotomous Measurands

In some cases, the measurand is dichotomous (that is, $X \in \{0, 1\}$). The proverbial example is a pregnancy test: one is either pregnant or not. Note that the measurand is whether a woman is or is not pregnant; the measurand is not the levels of chemical markers that such tests detect, as these are just the intermediate results and not the ultimate property that the test aims to establish. An industrial example of a dichotomous measurand is in functional tests on lightbulbs—the measurand X is whether the lightbulb is good or defective, while the measurement Y is 'accept' or 'reject'.

In other cases, the measurand is a continuum ($X \in \mathbb{R}$); an item is rejected if the appraiser assesses the measurand to be beyond a certain threshold *USL* (upper specification limit) on this continuum. As an example, consider a visual inspection where products are accepted or rejected based on whether their wrapping is good (meaning that the wrapping should not be too crooked). The underlying, continuous measurand X is the crookedness of the wrapping, while the measurement Y is 'accept' or 'reject'. Note that this measurand is not measured directly; as a matter of fact, it is not even operationally defined nor is there an explicit, quantitative norm for crookedness, and *USL* is, consequently, only given a vague and ambiguous definition, for example, in the form of a photo.

A convenient way of modeling the stochastics of measurement procedures is by means of characteristic curves (which are, actually, only *curves* if the

measurand is continuous),

$$q(x) := P(Y = 0 \mid X = x),$$

and, therefore,

$$P(Y = 1 \mid X = x) = 1 - q(x).$$

If X is dichotomous, then $p = P(X = 0)$ is the defect rate, $q(0)$ is the probability of correct rejection, and $q(1)$ is the false-rejection probability. If X is continuous, then $F_X(x) = P(X \leq x)$, and $q(x)$ is typically an S-curve, such as defined by the logit function,

$$\log\left(\frac{q(x)}{1 - q(x)}\right) = (x - \delta)/\sigma \qquad (1)$$

(see Figure 1). Items with $X > USL$ are defective, while items with $X > \delta$ are likely to be rejected. Thus, the curve's inflection point δ can be interpreted as the threshold that appraisers appear to apply (with $q(\delta) = 0.5$), as opposed to USL, which is the nominal rejection bound. The difference $\delta - USL$ could be interpreted as systematic measurement error; in cases where false acceptance has more serious consequences than false rejection, it could be advantageous to design the inspection procedure to have an inflection point δ strictly below USL. The value σ is a discrimination parameter, larger values corresponding to poorer measurement reliability.

Note: in our discussion, we will ignore appraisers as a factor, as this complication distracts from the points we aim to bring across. Thus, we assume that repeated measurements of an item are done by the same appraiser or that the appraisers are interchangeable (that is, have identical characteristic curves). All the methods to be discussed can be extended to involve characteristic curves q_j for each appraiser j separately. These extensions are typically straightforward and formulas can be found in the provided literature references.

Repeated measurements of an item are in general not independent, their having the same underlying X value inducing correlation. The estimation methods discussed in the next sections assume that, besides X, there are no other properties of the items and no environmental factors that induce dependencies among the measurement results. Inference concerns the infinite sequence of random variables $\{Y_{ij}\}$, with items $i = 1, 2, \ldots$; and repeated measurements $j = 1, 2, \ldots$ (the *population*). During the MSA study, we observe a finite part of this sequence (the *sample*), namely, $\{Y_{ij}\}$ with $i = 1, \ldots, n$ and $j = 1, \ldots, m$, with n the number of items in the sample, and $m \geq 1$ the number of repeated measurements per item. In

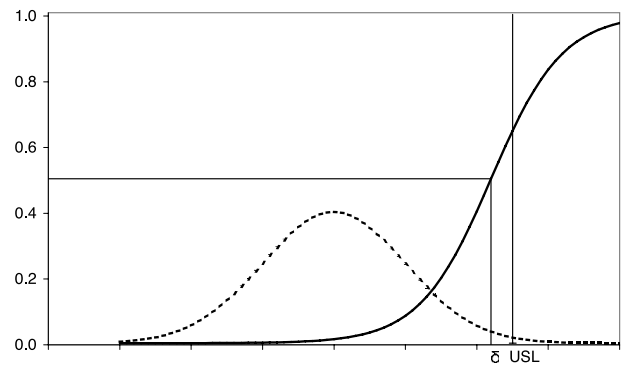

FIGURE 1. Characteristic Curve $q(x) = P(Y = 0 \mid X = x)$ (Solid Curve) and Density $f_X(x) = d\,P(X \leq x)/dx$ (Dashed Curve).

general, to allow extrapolation of sample statistics to inferences on population parameters, we need the $\{Y_{ij}\}$, $i = 1, 2, \ldots$; $j = 1, 2, \ldots$ to be independent and identicaly distributed (i.i.d.) conditional on the measurands X_i (or, in Bayesian terminology, we need the sequence to be exchangeable in Y conditional on X; Lindley and Novick (1981)). As we will see next, this assumption of conditional i.i.d. is easily violated in practice.

False Dichotomies and Conditional I.I.D.

In practice, one often evaluates binary inspections in terms of $q(0)$ and $q(1)$, even if the measurand is continuous. Thus, one treats a continuous measurand as artificially dichotomous by defining a dummy measurand $\widetilde{X}$, which is 1 if $X < USL$ and 0 if $X \geq USL$. We have

$$\widetilde{p} = P(\widetilde{X} = 0) = \int_{USL}^{\infty} f_X(x)dx,$$

and further,

$$\begin{aligned}
\widetilde{q}(0) &= P(Y = 0 \mid \widetilde{X} = 0) \\
&= \int_{USL}^{\infty} q(x) f_X(x)dx \Big/ \int_{USL}^{\infty} f_X(x)dx. \\
\widetilde{q}(1) &= P(Y = 0 \mid \widetilde{X} = 1) \\
&= \int_{-\infty}^{USL} q(x) f_X(x)dx \Big/ \int_{-\infty}^{USL} f_X(x)dx. \qquad (2)
\end{aligned}$$

The $\widetilde{q}(0)$ and $\widetilde{q}(1)$ of the artificial dichotomy are the average $q(x)$ over the relevant intervals of x, weighted by $f_X(x)$. Treating a continuous measurand as dichotomous creates complications, as, in general, it creates an intrinsic reason for the conditional i.i.d. assumption to be violated. For example, repeated inspections of an item i that *are* independent

conditional on a continuous measurand X (that is, $P(Y_{i1} = 0, Y_{i2} = 0 \mid X_i = x) = q^2(x)$) are, in general, *not* independent conditional on the artificially dichotomized $\widetilde{X}_i$:

$$\begin{aligned} &P(Y_{i1} = 0, Y_{i2} = 0 \mid \widetilde{X}_i = 0) \\ &= \frac{\int_{-\infty}^{\infty} P(Y_{i1} = 0, Y_{i2} = 0, \widetilde{X}_i = 0 \mid X_i = x) f_X(x) dx}{P(\widetilde{X}_i = 0)} \\ &= \frac{\int_{USL}^{\infty} P(Y_{i1} = 0, Y_{i2} = 0 \mid X_i = x) f_X(x) dx}{\int_{USL}^{\infty} f_X(x) dx} \\ &= \int_{USL}^{\infty} q^2(x) f_X(x) dx \Big/ \int_{USL}^{\infty} f_X(x) dx, \end{aligned} \tag{3}$$

which is, in general, not equal to $(\int_{USL}^{\infty} q(x) f_X(x) dx / \int_{USL}^{\infty} f_X(x) dx)^2 = \widetilde{q}^2(0)$, unless q is a step function with step at $x = USL$. In words, Y depends not only on $\widetilde{X}$ (*whether* the item is good or defective), but in addition, it depends on X (the degree of goodness or defectiveness). Violations of conditional i.i.d. when a continuous measurand is treated as dichotomous may also concern the identically distributed aspect, especially when the distribution of X values in the sample is not identical to the population distribution. We introduce the phrase *false dichotomies* for such measurands that are artificially dichotomized and whose characteristic curve is not a step function.

Treating a continuous measurand as dichotomous, and thus creating a false dichotomy, introduces an intrinsic reason for conditional i.i.d. not to hold, and this fact has consequences for estimation and sampling.

Gold Standard Available, Dichotomous Measurand: Nonparametric Estimation

In this and the subsequent sections, we consider each of the four situations defined by our set-up and we describe methods that can be used in each situation and possible complications. The first situation we discuss is where the measurand is dichotomous and a gold standard is available. The distribution of X is characterized by $p = P(X = 0)$, with p the *true defect rate*. Further, $p(1) = P(X = 0 \mid Y = 1)$ and $p(0) = P(X = 0 \mid Y = 0)$. The probabilities of interest are $q(1) = P(Y = 0 \mid X = 1)$ and $q(0) = P(Y = 0 \mid X = 0)$, and we have the rejection rate $q = P(Y = 0)$. The inspection procedure's error rates are given by the false-acceptance probability, $FAP = 1 - q(0)$, and the false-rejection probability, $FRP = q(1)$. Variants of nonparametric estimation of error rates are common in the diagnostic sciences; see, for instance, Pepe (2003, Chapter 2). Also, the AIAG *MSA Manual* (Automotive Industry Action Group, 2003, pp. 128–134) presents approaches akin to the ones discussed in this section.

By expressing *FAP* and *FRP* in terms of $q(x)$, where x is assumed to be 0 or 1, the approaches in this section assume that the measurand is a dichotomy. But they are often applied to cases where the measurand is continuous, thus creating a false dichotomy. An example are credit cards, which, before they are released for use, are inspected for bleeding of colors. If one assesses the quality of this inspection procedure in terms of $q(0)$ and $q(1)$, one treats a continuous measurand (X = degree of bleeding) as dichotomous ($\widetilde{X} = 0$ or 1).

Farnum: Samples of Good and Defective Items

One set-up for an MSA study proposed, for example, in Farnum (1994), is to obtain a sample of n_0 defective items and a sample of n_1 good items. These items are classified by the inspection procedure under study, which gives the totals $m_{0|0}$, $m_{1|0}$, $m_{0|1}$, and $m_{1|1}$ (where, for example, $m_{1|0}$ is the number of items with $Y = 1$ and $X = 0$). Estimation is straightforward from sample proportions: $\hat{q}(0) = m_{0|0}/n_0$ and $\hat{q}(1) = m_{0|1}/n_1$; the *FAP* and *FRP* are derived from these values.

We study what happens if the measurand is a false dichotomy, with $\widetilde{X} = 1$ or 0, depending on whether a continuous property X is smaller or larger than a threshold *USL*. The discussion concerns the expected value of the sample proportion estimators, such as

$$\begin{aligned} \mathrm{E}(\hat{q}(1)) &= \mathrm{E}(m_{0|1}/n_1) \\ &= \int_{-\infty}^{USL} q(x) f_X^s(x) dx \Big/ \int_{-\infty}^{USL} f_X^s(x) dx, \end{aligned} \tag{4}$$

with F_X^s the sampling distribution of X (i.e., the distribution determined by the sampling mechanism). We discuss potential complications, especially for false dichotomies, in a number of scenarios.

Random Sampling

Truly random samples from the subpopulations of defective and good items allow unbiased estimation of *FAP* and *FRP*, even in the case of a false dichotomy. Namely, random samples ensure that, in Equation (4), the distribution F_X^s of X in the sample

equals the population distribution F_X and, therefore, $\mathrm{E}(\hat{q}(1)) = \widetilde{q}(1) = P(Y = 0 \mid \widetilde{X} = 1)$, per Equation (2).

Nonrandom Sampling for True Dichotomies

Also, nonrandom samples allow unbiased estimation, provided $q(x)$ is a step function (step at $x = USL$); that is, provided the measurand is truly dichotomous. Equations (3) and (4) show that, if $q(x)$ is a step function, $\mathrm{E}(\hat{q}(1))$ is independent of the sampling distribution F_X^s, and there is no intrinsic reason for repeated classifications of an item not to be i.i.d. conditional on the measurand.

Nonrandom Sampling for False Dichotomies

But if the dichotomy is false, nonrandom sampling may create a bias, due to the fact that F_X^s may not be identical to F_X, and thus, conditional i.i.d. of the Y_{ij} is violated (or, in Bayesian terminology, Y_{ij} in the sample and population are not exchangeable). This bias can be arbitrarily large, as illustrated from the following two numerical examples. Both examples concern a situation where the population distribution F_X of X values is the standard normal. Suppose, further, that $USL = 2.5$ and that the inspection procedure's characteristic curve is given by Equation (1) with $\delta = USL$ and $\sigma = 0.5$, which gives $FRP = 0.0284$ (from Equation (2)). The first example of a nonrandom sample is inspired by the tendency among some practitioners to sample items guided by the idea of "covering the whole range"; as a result, the distribution F_X^s of X values in the sample might approach a uniform distribution on the interval $[-3, 3]$. Such a sampling approach would give an expected result of $\mathrm{E}(\widehat{FRP}) = 0.0630$, overestimating the FRP by more than a factor of two. Our second example considers a sample consisting mainly of difficult-to-judge parts, which we interpret by taking F_X^s to be a normal distribution with mean 2.5 and standard deviation 0.5. The expected result is $\mathrm{E}(\widehat{FRP}) = 0.325$, overestimating the FRP by nearly a factor of 12.

Naive Sampling Without Swapping for False Dichotomies

The fact that, for false dichotomies, the quality of the estimation hinges on the randomness of the samples creates a potentially serious problem for Farnum's set-up, as it is all but clear how such random samples of good and defective items can be obtained. A naive way to do so has one collect random samples from the streams of accepted and rejected items and use the gold standard to single out and remove the falsely accepted and falsely rejected items, thus obtaining samples of n_0 defective and n_1 good items. These samples are then used for the MSA study; that is, they are classified by the inspection procedure under study and the FAP and FRP are estimated from the results. We refer to this sampling scheme as *naive sampling without swapping*. The problem with this procedure is that the subsamples of n_0 and n_1 items, thus obtained, are not representative for the subpopulations of defective and good items in false-dichotomy cases. For example, items with X values close to USL are underrepresented in the sample of defective items, as they have a larger probability of slipping through and therefore a smaller probability of being in the stream of rejects and the assessment of the inspection's FAP will be too optimistic. The Appendix shows, by calculation, that the sampling distribution of X in the subsamples thus obtained is different from the distribution in the subpopulations of defective and good items and, consequently, that the resulting estimators for the FAP and FRP are biased. The Appendix also shows that this bias is modest ($|\mathrm{E}(\widehat{FAP}) - FAP| < 0.035$ and $|\mathrm{E}(\widehat{FRP}) - FRP| < 0.030$) if the inflection point δ of the inspection procedure's characteristic curve is equal to USL. If $\delta \neq USL$, the bias can be arbitrarily large.

Naive Sampling with Swapping for False Dichotomies

A variant of the previous case is to take random samples from the streams of accepted and rejected items, but falsely accepted or rejected items are not removed but added to the other sample. We refer to this sampling plan as *Naive Sampling with Swapping*. Also under this strategy, the estimated FAP and FRP are biased, with similar consequences as for naive sampling without swapping; see the Appendix.

Plan I: Samples of Accepted and Rejected Items

An alternative for Farnum's set-up is to randomly select m_0 items from the stream of rejected items and m_1 accepted items. This is Plan I in Danila et al. (2008). Determining for these items the measurand gives the numbers $m_{0|0}$, $m_{1|0}$, $m_{0|1}$, and $m_{1|1}$. From these, $\hat{p}(0) = m_{0|0}/m_0$ and $\hat{p}(1) = m_{1|0}/m_1$ and, for example,

$$\hat{q}(1) = \frac{\hat{q}(1 - \hat{p}(0))}{\hat{q}(1 - \hat{p}(0)) + (1 - \hat{q})(1 - \hat{p}(1))}, \qquad (5)$$

with $\hat{q}$ an historical estimate of the rejection rate q.

Also for this approach, we study the applicability in the case of false dichotomies. The estimates for $q(0)$ and $q(1)$ are derived from equations of the form of Equation (5). Thus, one needs a good estimate for q, $p(0)$, and $p(1)$. To ensure the latter, the two subsamples of accepted and rejected items must be random, even in the truly dichotomous case, in order that the sample proportions $m_{0|0}/m_0$ and $m_{1|0}/m_1$ are unbiased estimates of $p(0)$ and $p(1)$. Random sampling from the streams of accepted and rejected items will be straightforward in most cases, and we conclude that Plan I is feasible even in the case of a false dichotomy.

Plan II: A Sample from the Total Items Population

A third option (Plan II in Danila et al. (2008)) is to collect a random sample of n items from the study population of items and determine each item's measurand X, which gives the totals n_0 and n_1 of defective and good items, and next apply the classification procedure under study, which gives the totals $m_{0|0}$, $m_{1|0}$, $m_{0|1}$, and $m_{1|1}$. Estimation is done from sample proportions: $\hat{q}(0) = m_{0|0}/n_0$ and $\hat{q}(1) = m_{0|1}/n_1$. If there is additional information about either p or q from historical data, then, in Plan II, the probabilities can be estimated more efficiently using an approach described in Danila et al. (2008).

Equations of the type of Equation (4) tell us that, in the case of a false dichotomy, the sample must be really random to avoid complications as discussed for Farnum's approach. In truly dichotomous cases, where conditional i.i.d. holds, even a nonrandom sample allows good estimation of $q(0)$, $q(1)$, and the *FAP* and *FRP*, but inferences involving the defect rate p are impaired. Whereas a truly random sample is difficult to achieve in Farnum's set-up, it will be mostly straightforward under Plan II. However, a practical problem with Plan II is that the number n_0 of defects in the random sample will be zero or very low in the typical situation where p is close to zero, which makes estimation of $q(0)$ precarious.

Gold Standard Unavailable, Dichotomous Measurand: Latent Class Modeling

In this second situation, the measurand is assumed dichotomous and a gold standard unavailable; X, therefore, is treated as a latent class. One needs a sample of n items from the study population. Van Wieringen and De Mast (2008) find that the standard error of the estimates is minimized by taking a balanced sample, in which the numbers n_1 and n_0 of good and defective items are about equal.

In the gold-standard-available situation, each item is usually measured once (which results in a single Y-value per item and an X-value). Here, in the gold-standard-unavailable situation, it seems unavoidable that each item is measured at least twice (and to ensure identifiability of the model, in the case of a single appraiser, one needs at least three repeats; Van Wieringen (2005)); as a result, one has multiple Y-values for each item, associated with a single unobserved X-value.

The parameters $q(0)$ and $q(1)$ can be estimated from the measurements Y (treating the X as a latent class) by maximizing the likelihood function. Van Wieringen and De Mast (2008) use an EM algorithm to achieve this, under the assumption of conditional i.i.d.; see also Hui and Walter (1980), Boyles (2001), and Beavers et al. (in press). From the results, the error rates *FAP* and *FRP* can be derived. Danila et al. (2010) study the effectiveness of a number of more complex set-ups, exploiting additional information about the rejection rate.

We study potential complications. Without a gold standard, it is difficult to obtain a sample with a sufficiently large number of defective items. In practice, one will sample from the streams of accepted and rejected items, but even in the latter, the percentage of good items is large if p is low,

$$\begin{aligned} &P(X = 1 \mid Y = 0) \\ &\quad = q(1)(1-p)/(q(1)(1-p) + q(0)p). \end{aligned} \tag{6}$$

For example, if $p = 0.01$, $q(0) = 0.95$, and $q(1) = 0.05$, the percentage of good items in the stream of rejects is 84%. As a consequence, an approximately balanced sample is difficult to obtain in practice and, instead, samples may contain just a few defective items; the resulting standard error in the estimation of $q(0)$ and *FAP* will be large. Note that, in this light, it may be a good idea to take a sample only from the stream of rejected items, as suggested by Danila et al. (2010).

Provided one manages to obtain a sample with sufficient defective items, latent class modeling is effective if the measurand is a true dichotomy and conditional i.i.d. holds. However, in case of a false dichotomy, randomization becomes of crucial importance. The whole sample need not be a random sample of items but, to avoid biased estimates, the sub-

samples of n_1 good and n_0 defective items must be random samples from their respective subpopulations (Van Wieringen and De Mast (2008)). Without a gold standard, such random samples are quite difficult to achieve in practice. The naive way to do so, namely, to select m_0 and m_1 items from the streams of rejected and accepted items, results in a sample in which the difficult-to-judge items with X close to the inflection point δ are underrepresented. As a consequence, latent class modeling comes across similar problems as the ones discussed for Farnum's set-up in the previous section; in fact, the situation is comparable to the naive sampling with swapping scenario.

Pepe (2003, pp. 203–205) raises some concerns against the use of latent class models in a context similar to ours, as follows:

1. Latent class modeling may encourage users to study classifications for which the measurand is not well (in Pepe's context: clinically) defined.
2. Validity of the conditional independence assumption cannot be tested.
3. The complex estimation procedure makes it difficult for practitioners to recognize how factors and disturbances affect results.

Acknowledging the legitimacy of item 1 in scientific use and of item 3 in practical use, we think our framework can bring nuance to the second concern. As explained above, latent class modeling is generally problematic if X is continuous (as, in that case, conditional i.i.d. will typically be violated per Equation (3)). However, if X is truly dichotomous, there is no intrinsic reason for conditional i.i.d. to be violated and careful experimental design may enable the assumption to be fulfilled.

Gold-Standard-Available, Continuous Measurand: Logistic Regression

False dichotomies bring about complications for assessing the reliability of binary measurements, as demonstrated in the previous sections. Some of these complications can be handled by careful, random sampling and, in the gold-standard-available situation, especially Plan I is a viable option. An alternative approach is not to artificially dichotomize a continuous measurand, but to treat it as continuous. This and the next section outline some approaches for the gold-standard-available and gold-standard-unavailable situations, respectively.

AIAG's *MSA Manual* (AIAG, 2003, pp. 135–140) describes a method known as *analytic method*. It prescribes selecting n items such that their measurands $X_1, \ldots, X_n$ are more-or-less equidistant. Each item is to be classified a number m_i of times (with $m_{0|X_i}$ the resulting number of rejects). AIAG gives detailed guidelines for selecting these n items, including the requirements that items 1 and n should be selected extreme enough to ensure that $m_{0|X_1} = 0$ and $m_{0|X_n} = m_i$.

AIAG (2003) suggests assuming a normal ogive as the characteristic curve, $q(x) = \Phi((x-\delta)/\sigma)$, but alternatively, one could take the more traditional logit link function given in Equation (1), which is a traditional logistic regression model (with slope σ^{-1} and intercept $-\delta/\sigma$). For each (known) X_i, we have the corresponding observed proportion $\hat{q}(X_i) = m_{0|X_i}/m_i$ as an estimate of $q(X_i)$. From the observed proportions, δ and σ can be estimated. AIAG recommends plotting the $\hat{q}(X_i)$ against the X_i in a normal probability plot and fitting a straight line. The more conventional way to estimate δ and σ is by maximum likelihood, as is standard in logistic regression.

AIAG (2003, pp. 136) defines the systematic measurement error as $b = \delta - USL$. Further, AIAG suggests expressing reliability as the width of a 99% interval, namely $(\sigma\Phi^{-1}(0.995)+\delta-\sigma\Phi^{-1}(0.005)-\delta)/c$, with c an adjustment constant ($c = 1.08$ if $m_i = 20$ for all i). This mirrors AIAG's guidelines for numerical MSA studies, where measurement reliability is expressed in terms of the length of a 99% prediction interval $5.15\sigma_m$, with σ_m the measurement spread of the numerical gauge.

Alternatively, one may compute similar metrics as in the previous section, such as

$$FAP = \int_{x=USL}^{\infty} (1-q(x))f_X(x)dx \Big/ \int_{x=USL}^{\infty} f_X(x)dx$$

and

$$FRP = \int_{x=-\infty}^{USL} q(x)f_X(x)dx \Big/ \int_{x=-\infty}^{USL} f_X(x)dx.$$

For $q(x)$, we have the ogive determined by $\hat{\delta}$ and $\hat{\sigma}$. The parameters of the probability distribution function F_X of X can be estimated separately by taking a random sample of items and fitting a probability distribution to the X values.

Gold Standard Unavailable, Continuous Measurand: Latent Trait Modeling

In the last situation to be discussed, the measurand is continuous and a gold standard unavailable; X, therefore, is treated as a latent trait. Where

logistic regression is an alternative to nonparametric estimation in the case of false dichotomies, latent trait modeling is the corresponding alternative for latent class modeling. The experimental design is, as for all gold-standard-unavailable methods, one in which each of n randomly selected items is measured two or more times. The characteristic curves are S-curves, similar to the logistic regression model in Equation (1). The difference with logistic regression, is that the X values are unobservable and they are treated as a latent variable. This type of model is standard in the wide and advanced field of item response theory (IRT; see Embretson and Reise (2000) for a recent introduction). Also, for the distribution F_X, one assumes a parametric model, such as $X \sim N(\mu_X, \sigma_X^2)$. Note that the origin and scale of the latent X-continuum are arbitrary and one typically sets them by fixing $\mu_X = 0$ and $\sigma_X = 1$, in which case, $F_X = \Phi$.

The estimation problem is complex and is typically approached using an EM algorithm to compute maximum likelihood estimates. The parameters σ, δ, and the parameters of F_X are estimated simultaneously. An exposition of these algorithms is beyond the scope of this paper, but the reader is referred to the IRT literature (with Embretson and Reise (2000), a recent overview).

Because the x-axis has an arbitrary scale and the X-values are treated as unobservable and dimensionless, one cannot determine, in latent trait modeling, the *FAP* and *FRP* because *USL* is an undefinable parameter. Instead, De Mast and Van Wieringen (2010) propose probabilities of inconsistent ordering, which are the probabilities that an appraiser's classification is inconsistent with his or her own rejection bound δ,

$$\begin{aligned}\pi(1) &= P(Y = 0 \mid X < \delta) \\ &= \frac{\int_{-\infty}^{\delta} (1 - q(x)) f_X(x) dx}{\int_{-\infty}^{\delta} f_X(x) dx}, \\ 1 - \pi(0) &= P(Y = 1 \mid X \geq \delta) \\ &= \frac{\int_{\delta}^{\infty} q(x) f_X(x) dx}{\int_{\delta}^{\infty} f_X(x) dx}, \end{aligned} \tag{7}$$

where *FAP* and *FRP* express both the systematic component of measurement error (that is, $\delta - USL$) and the random component (the degree to which classifications randomly deviate from an appraiser's own δ), these $\pi(1)$ and $1 - \pi(0)$ express the random component only.

Like latent class modeling, latent trait modeling also has some unresolved difficulties. A random sample of items ensures consistent estimates for $\pi(0)$ and $\pi(1)$, but may contain too few defective items and items in the steep part of $q(x)$ for precise estimation. A nonrandom sample, perhaps including more items with larger X values, still allows estimation of the characteristic curve $q(x)$, but the distribution F_X may be misestimated. In logistic regression, this could be solved by estimating the parameters of F_X from a second, random sample, but in latent trait analysis, this is not possible because the scale of the X continuum would be different in the two analyses, and therefore, the fitted $q(x)$ and $F_X(x)$ would be fitted on different x-scales. Also, it is difficult to interpret the fitted characteristic curve in tangible terms, as the x-axis is abstract and dimensionless.

Example: Reliability of a Go/No-Go Gauge

We illustrate and discuss the various methods on the basis of an example taken from the AIAG manual (AIAG, 2003, pp. 125 ff.). The measurand X is continuous and parts are considered 'good' if X is between $LSL = 0.450$ and $USL = 0.545$ and 'defective' otherwise. One could treat the case as artificially dichotomous by defining $\widetilde{X} = 1$ if $0.450 \leq X \leq 0.545$ and $\widetilde{X} = 0$ otherwise. For normal inspection, the reference values X are not available and neither are the $\widetilde{X}$. Instead, the appraisers use a go/no-go gauge that returns 'accept' ($Y = 1$) or 'reject' ($Y = 0$). The aim of the study is to establish the quality of this go/no-go gauge.

The data set gives the results of an experiment in which 50 parts have been gauged three times by each of three appraisers, A, B, and C (giving 9 Y values per part). In addition, the data set gives the 50 parts' X values and the corresponding $\widetilde{X}$ values, so, for the sake of the MSA experiment, a gold standard is available.

Treating the Measurand as Dichotomous: Nonparametric Estimation and Latent Class Analysis

Our first analysis approach is a nonparametric estimation of *FAP* and *FRP*, thus treating the measurand as dichotomous, and taking the $\widetilde{X}$ values as reference values. The 50 parts are claimed to be a random sample from the parts population (AIAG, 2003, p. 126), so the sampling plan is similar to Plan II of Danila et al. Out of 50 parts, 16 are nonconforming,

giving an estimated defect rate of $\hat{p} = 16/50 = 0.32$. Note that a sample of 50 parts is rather small for estimating a defect rate if X is treated as a dichotomy; a 95% confidence interval on p is $[0.21, 0.46]$, which is rather wide. Such a small sample size, when Plan II is used, will typically also create the problem that there are no or just very few defectives in the sample, but the fairly high defect rate here ensures that even in this rather small sample there are sufficient defective parts. Each part was measured 9 times, so altogether 450 measurements were made (302 times 'accept', 148 times 'reject'). This gives an estimated rejection rate of $\hat{q} = 0.329$. The probability of rejecting a defective item is estimated as $\hat{q}(0) = 0.917$ and the probability of rejecting a good item as $\hat{q}(1) = 0.052$ (estimated from sample proportions), giving the following error rates:

- False-acceptance probability $\widehat{FAP} = 0.083$.
- False-rejection probability $\widehat{FRP} = 0.052$.

These error rates could also be calculated for each appraiser separately.

This analysis treats the measurand as dichotomous, but it is in fact continuous and, thus, we are dealing with a false dichotomy. As a consequence, q, $q(0)$, and $q(1)$ are estimated well only if the sample of parts is representative and, in fact, there is some evidence that refutes AIAG's claim that the sample is random. Namely, AIAG states that the process's performance is $P_p = P_{pk} = 0.50$ (where $P_p = (USL - LSL)/6\sigma_X$), suggesting that X has a mean of $\mu_X = 0.4975$ and a standard deviation of $\sigma_X = 0.032$ and that the defect rate is $p = 0.13$. However, the sample defect rate of $\hat{p} = 0.32$ is significantly different (p-value < 0.001) from 0.13, and also the sample's standard deviation $\hat{\sigma}_X = 0.045$ is significantly different from the process standard deviation $\sigma_X = 0.032$ (p-value < 0.001 based on a chi-square test). We conclude that the given sample, being not representative and given that the dichotomy is false, is not suited for this analysis. The miss rates (FAP) and false-alarm rates (FRP) per appraiser as given by AIAG (2003, p. 132) may therefore be misestimated.

If the 50 $\widetilde{X}$ values had not been available for the MSA study, one would have had to resort to a latent class analysis. Using the algorithm described in Van Wieringen and De Mast (2008), the model parameters are estimated as $\hat{p} = 0.361$, $\hat{q}(0) = 0.862$, and $\hat{q}(1) = 0.027$. This results in the following misclassification probabilities:

$$\widehat{FAP} = 0.138 \qquad \widehat{FRP} = 0.027.$$

Note that these estimates were obtained solely from the Y-values; the reference values were not used in the computations, as they are treated as a latent class. Because we are dealing here with a false dichotomy and a sample of parts of which representativeness is questionable, the results are not reliable.

Treating the Measurand as Continuous: Logistic Regression and Latent Trait Analysis

The artificial dichotomy defined by the $\widetilde{X}$-values is a false dichotomy and, consequently, repeated ratings of the same part are not i.i.d. conditional on $\widetilde{X}$. One way to solve the resulting problems is to ensure a random sample. The other way to go about it is not to dichotomize the measurand but to treat it as continuous. First, we apply logistic regression. The example is slightly more involved, in that we are dealing here with a lower *and* an upper boundary. In fact, the situation is basically not binary but ordinal with three classes, in which the two extreme classes ('below LSL', and 'above USL') are collapsed into

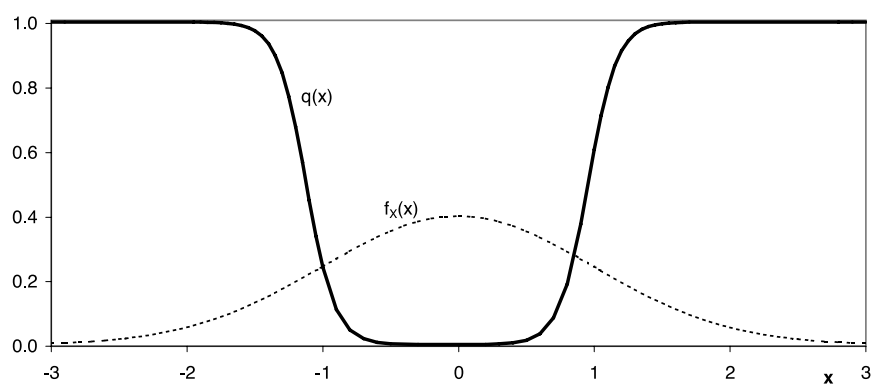

FIGURE 2. Characteristic Curve $q(x) = P(Y = 0 \mid X = x)$ and Density $f_X(x) = \phi(x)$.

one class ('reject'). We fit the curve

$$1 - q(x) = \frac{\exp((\delta_U - x)/\sigma)}{1 + \exp((\delta_U - x)/\sigma)} - \frac{\exp((\delta_L - x)/\sigma)}{1 + \exp((\delta_L - x)/\sigma)},$$

where δ_L and δ_U are the decision limits that are effectively used by the gauge (as opposed to *LSL* and *USL*, which are the nominal requirements). This characteristic curve is derived from a logistic regression model based on the logit link function for ordinal responses (McCullagh and Nelder (1989), p. 152). The maximum likelihood estimates are $\widehat{\delta_L} = 0.453$, $\widehat{\delta_U} = 0.547$, and $\hat{\sigma} = 0.00415$. Because the estimated δ_L and δ_U are close to the *LSL* and *USL*, we conclude that the gauge has negligible bias.

We estimate the parameters of the distribution of the measurand from the given X-values as $\hat{\mu}_X = 0.51$ and $\hat{\sigma}_X = 0.045$ (estimated from the sample average and standard deviation). The analytic method results in the following misclassification probabilities:

$$\widehat{FAP} = \frac{\int_{-\infty}^{LSL} (1 - \hat{q}(x))\hat{f}_X(x)dx}{\int_{-\infty}^{LSL} \hat{f}_X(x)dx} + \frac{\int_{USL}^{\infty} (1 - \hat{q}(x))\hat{f}_X(x)dx}{\int_{USL}^{\infty} \hat{f}_X(x)dx} = 0.093$$

$$\widehat{FRP} = \frac{\int_{LSL}^{USL} \hat{q}(x)\hat{f}_X(x)dx}{\int_{LSL}^{USL} \hat{f}_X(x)dx} = 0.051,$$

with $\hat{q}$ and $\hat{f}_X$ the logit function and normal density based on $\hat{\delta}_L$, $\hat{\delta}_U$, $\hat{\sigma}$, $\hat{\mu}_X$, and $\hat{\sigma}_X$. Also here, the alleged nonrepresentativeness of the sample of parts creates some complications, but they are less serious. The estimation of the parameters of $q(x)$ is not impaired, but the estimated μ_X and σ_X may be biased. As a consequence, the estimated characteristic curve $q(x)$ represents reliably the behavior of the go/no-go inspections, but the translation into an *FAP* and an *FRP* is affected by the potential bias in μ_X and σ_X. Of course, one could collect a random sample of parts, apply the gold standard, and estimate μ_X and σ_X from the results. Substituting these estimates in the equations above would give estimates for *FAP* and *FRP*.

If the X values had not been available, one would have had to resort to latent trait modeling. The standard model in IRT for such a situation with both an *LSL* and an *USL* is Masters' partial credit model in the generalized form by Muraki (1992),

$$1-q(x) = \frac{\exp\left(\dfrac{x - \delta_L}{\sigma}\right)}{1 + \exp\left(\dfrac{x - \delta_L}{\sigma}\right) + \exp\left(\dfrac{2x - \delta_L - \delta_U}{\sigma}\right)}.$$

The normal distribution is assumed for X, with the origin and scale of the x-axis adjusted such that $\mu_X = 0$ and $\sigma_X = 1$. De Mast and Van Wieringen (2010) discuss how this model can be used for industrial applications and they propose a working algorithm for fitting the model and providing model diagnostics.

The fitted model parameters are $\hat{\sigma} = 0.106$, $\hat{\delta}_L = -1.12$, and $\hat{\delta}_U = 0.955$; note that the scale of the X-continuum is arbitrary and meaningless. The resulting characteristic curve is shown in Figure 2. The definitions for the probabilities of inconsistent ordering in Equation (7) become

$$\pi(1) = P(Y = 0 \mid \delta_L < X < \delta_U),$$
$$1 - \pi(0) = P(Y = 1 \mid X < \delta_L \text{ or } X > \delta_U).$$

The results are $\hat{\pi}(1) = 0.055$ and $1 - \hat{\pi}(0) = 0.099$. Also, in this case, the potential nonrandomness of the sample makes the results unreliable; the form of the characteristic curve in Figure 2 should well represent the behavior of the go/no-go gauge, but the distribution of X-values (indicated by their density) may not properly reflect the distribution in the items population.

Conclusions

The concept of a false dichotomy and its ramifications for the conditional i.i.d. property and estimation are this paper's most important novel contributions. The essential difference between binary inspections based on a truly dichotomous measurand versus a continuous measurand seems underappreciated in industry, as are the complications brought about by artificially dichotomizing a continuous measurand (although the problem was mentioned in Van Wieringen and De Mast (2008) and Danila et al. (2010)). We think that continuous measurands are far more common than truly dichotomous measurands and, therefore, complications for false dichotomies are a ubiquitous problem. A related issue is that many guidelines offered in industry are in conflict with our

conclusion that random sampling is in many cases crucial. An example of such misconceived advice is to sample items such that roughly one third is very bad, one third is very good, and one third is near the boundary (as quoted in, but not endorsed by, Mawby (2006), p. 122).

Our framework serves as a structure for a taxonomy of methods as shown in Table 1. Most of the mentioned methods are known in quality engineering, except for the latent trait modeling approach, which originates in the field of psychometrics. In the case of a false dichotomy, careful random sampling may allow safe use of nonparametric estimation, especially following Plan I. Random sampling may be difficult to achieve in Farnum's scheme and latent class modeling, and Plan II may result in a sample containing too few defective items. An alternative way to handle false dichotomies is to not dichotomize the continuous measurand at all but rather use logistic regression or latent trait analysis.

An alternative class of methods used and prescribed commonly for MSA studies for binary inspection are methods based on agreement statistics and kappa-type indices. On the basis of a random sample of items that are judged repeatedly, one estimates the probability of agreement

$$\begin{aligned} P_a &= P(Y_1 = Y_2) \\ &= (1-p)\{q^2(1) + (1-q(1))^2\} \\ &\quad + p\{q^2(0) + (1-q(0))^2\}, \end{aligned}$$

where metrics such as FAP and FRP express agreement between observations (Y) and measurands (X), P_a expresses agreement among observations only (Y_1 to Y_2). The κ (kappa) statistic is the probability P_a of agreement rescaled such that $\kappa = 0$ corresponds to the probability of agreement achieved by noninformative chance ratings (De Mast and Van Wieringen (2007), De Mast (2007)). Our framework shows that agreement may not be the right measure to express the reliability of accept/reject inspections. Namely, in industry, p is typically very close to 0 and, in that case, $P_a \approx (1-p)(q^2(1)+(1-q(1))^2)$. In other words, P_a only reflects the false-rejection probability $q(1)$ and not the false-acceptance probability $1-q(0)$, and it is the latter that is typically more relevant (as it represents the consumer's risk).

Some binary inspections involve a hybrid between a continuous and a dichotomous measurand. For example, in visual inspection of items for scratches, 'no scratch' is a point ($x = 0$) but 'scratch' is a contin-

TABLE 1. Overview of Methods Discussed in this Paper. Gold standard is *Av* (available) or *UnAv* (unavailable). The table indicates whether methods are suited for *D* (dichotomous) or *C* (continuous) measurands.

Gold-stand.	Measurand	Method	Experimental design	Points of attention
A	D	Nonparametric: Farnum	Subsamples from the strata of (truly) good and defective items. Judge each item one or more times.	If the dichotomy is false, or if conditional i.i.d. is violated otherwise, the subsamples must be random, but this is unfeasible in practice.
A	D, C	Nonparametric: Plan I	Random subsamples from the strata of accepted and rejected items. Apply the gold-standard to each item.	Subsamples must be random. A historical estimate of q is needed to determine FAP and FRP. Also works if the measurand is continuous.
A	D, C	Nonparametric: Plan II	A sample from the population of items. Apply the gold-standard to each item, and judge each item one or more times	If the measurand is continuous, the sample must be random. The possibly small number of defectives in the sample may result in large standard errors.
U	D	Latent class modeling	A sample from the items population, preferably as balanced as possible. Judge each item multiple times.	May be difficult to obtain a sample with sufficient defectives. For false dichotomies, the subsamples must be random, but this is unfeasible in practice.
A	C	Logistic regression	Select items with equidistant X values. Judge each item several (typically 20) times.	The study allows the estimation of $q(x)$. For FAP and FRP, a separate sample is needed to determine the distribution F_X of the X values.
U	C	Latent trait modeling	A random sample from the items population. Judge each item multiple times.	If the sample is not random, the $q(x)$ can still be estimated, but the distribution F_X of the measurand (and FAP and FRP) cannot be determined.

uum ($x > 0$), ranging from small scratches that are hardly noticeable, to large, wide and deep scratches. A leak test is another example, where $x = 0$ corresponds to 'no leak' and positive values correspond to progressively larger leaks. Methods treating such measurands as dichotomous will encounter similar complications as in falsely dichotomous cases. But also the application of logistic regression or latent trait models is not straightforward, as the standard logit and probit characteristic curves are symmetric in their inflection point, whereas the true characteristic curve in such hybrid situations is likely to be strongly asymmetrical. Also, distributional assumptions for the X-values need to be critically revised in such situations where the X-continuum is bounded by zero. Further research will focus on the open question of how such hybrid situations are to be modeled.

In summary, we think that, given the currently available methods, the problematic situations are as follows:

- A gold standard is unavailable and the measurand is continuous. One has to turn to latent class modeling or latent trait analysis, but in the former, it is difficult to obtain random samples, and, in the latter, it is difficult to translate the fitted $q(x)$ into tangible results such as *FAP* and *FRP*.
- The measurand is a hybrid of a continuous and a discrete characteristic. Both logistic regression and latent trait modeling need nontrivial adjustments in that case.

These observations present an agenda for future research.

Appendix

Naive Sampling Without Swapping

The continuous measurand is dichotomized by defining $\widetilde{X} = 1$ if $X < USL$ and $\widetilde{X} = 0$ otherwise. Let $F_X = \Phi$, the normal distribution with $\mu_X = 0$ and $\sigma_X = 1$, and q defined as in (1). The *FAP* is the proportion of accepted items in the subpopulation of defective items:

$$FAP = \int_{x=USL}^{\infty} (1 - q(x)) f_X^0(x) dx,$$

with F_X^0 the distribution of X in the subpopulation of defective items:

$$F_X^0(x) = P(X \le x \mid \widetilde{X} = 0) = \int_{t=USL}^{x} \phi(t)dt \Big/ \int_{t=USL}^{\infty} \phi(t)dt$$

(for $x \ge USL$). Under naive sampling without swapping, one obtains a sample of items from the stream of rejects and next removes the wrongly rejected items. The distribution of X in the resulting subsample of n_0 defective items is

$$F_X^{WoS:0}(x) = P(X \le x \mid \widetilde{X} = 0, Y = 0) = \int_{t=USL}^{x} q(t)\phi(t)dt \Big/ \int_{t=USL}^{\infty} q(t)\phi(t)dt$$

(for $x \ge USL$). Obviously, $F_X^{WoS:0} \ne F_X^0$ and, consequently,

$$\mathrm{E}(\widehat{FAP}) = \mathrm{E}\left(\frac{m_{1|0}}{n_0}\right) = \int_{-\infty}^{\infty} (1 - q(x)) f_X^{WoS:0}(x) dx \ne FAP$$

(a similar derivation can be given for *FRP*). The bias $\mathrm{E}(\widehat{FAP}) - FAP$ depends on $(USL - \mu_X)/\sigma_X$, $(\delta - USL)/\sigma_X$, and σ/σ_X. In cases where $F_X = \Phi$, q is the logit link function and $\delta = USL$. Plots of this bias (Figure 3) show that

- *FAP* is always underestimated in expectation; this is caused by the fact that items with X values close to δ are underrepresented.
- The bias is generally modest and never exceeds -0.035.

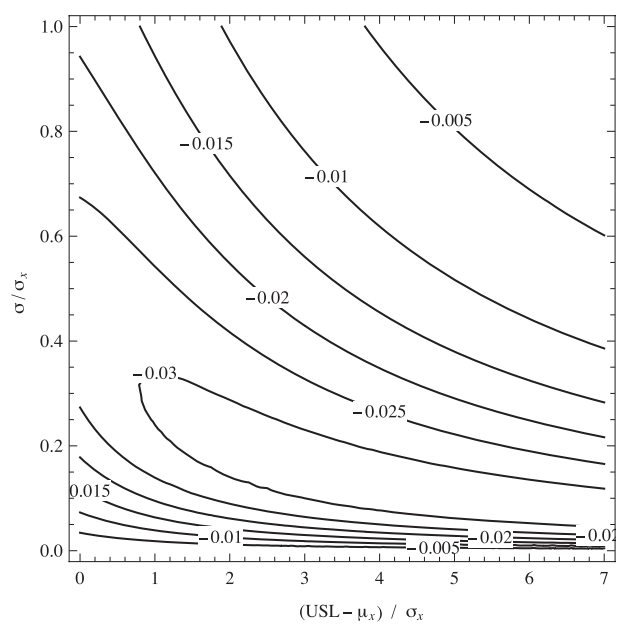

FIGURE 3. Contour Plot of the Bias $\mathrm{E}(\widehat{FAP}) - FAP$ Under the Naive Sampling Without Swapping Scenario and Given That $\delta = USL$.

If $\delta < USL$, the bias is even smaller; but if $\delta > USL$, the bias can become quite large (as large as 0.5).

A similar expression can be derived for the bias $\mathrm{E}(\widehat{FRP}) - FRP$. Plots of this bias show that

- FRP is always underestimated in expectation; this is caused by the fact that items with X values close to δ are underrepresented.
- The bias is generally quite small and never exceeds -0.030. This maximum is attained when USL is close to μ_X and σ is close to σ_X.
- The bias becomes negligibly small (below 0.005) when $(USL-\mu_X)/\sigma_X > 3$; when $\sigma/\sigma_X < 0.5$ the bias becomes negligibly small when $(USL-\mu_X)/\sigma_X > 2$.

If $\delta > USL$, the bias is even smaller; but if $\delta < USL$, the bias can become quite large (as large as 0.5). Note that, as a consequence, the bias in FRP and FAP is modest if $\delta = USL$; but if $\delta \neq USL$, either the bias in FRP is substantial or the bias in FAP is substantial.

Naive Sampling with Swapping

Here one starts with subsamples of sizes m_1 and m_0 from the streams of accepted and rejected items ($m = m_0 + m_1$), but now, erroneously classified items are not removed but added to the other subsample. The distribution of X values in the total sample of m items is

$$\begin{aligned} F_X^{WS}(x) &= \frac{m_0}{m} P(X \le x \mid Y = 0) \\ &\quad + \frac{m_1}{m} P(X \le x \mid Y = 1) \\ &= \frac{m_0 \int_{-\infty}^{x} q(t)\phi(t)dt}{m \int_{-\infty}^{\infty} q(t)\phi(t)dt} \\ &\quad + \frac{m_1 \int_{-\infty}^{x} (1-q(t))\phi(t)dt}{m \int_{-\infty}^{\infty} (1-q(t))\phi(t)dt}. \end{aligned}$$

The distribution in the subsample of defective items is

$$\begin{aligned} F_X^{WS:0}(x) &= F_{X|\widetilde{X}=0}^{WS}(x) \\ &= \frac{F_X^{WS}(x) - F_X^{WS}(USL)}{1 - F_X^{WS}(USL)}. \end{aligned}$$

Again, $F_X^{WS:0} \neq F_X^0$ and the estimates are biased. In cases where $F_X = \Phi$, q is the logit link function, and $\delta = USL$, plots of the bias show that

- FAP is always underestimated in expectation, but not by more than -0.035.
- FRP is always overestimated in expectation; because of the low defect rate, the stream of rejected items will consist in large proportion of falsely rejected items (from Equation (6)), which are then swapped to the subsample of good items, thus creating an overrepresentation of hard-to-judge items in the subsample of good items.
- The positive bias in FRP can be as large as 0.069.

If $\delta \neq USL$, the bias in either FAP or FRP can become substantial.

References

Automotive Industry Action Group (2003). *Measurement Systems Analysis: Reference Manual*, 3rd ed. Detroit, MI: Author.

Beavers, D. P.; Stamey, J. D.; and Bekele, B. N. (in press). "A Bayesian Model to Assess a Binary Measurement System When No Gold Standard System Is Available". *Journal of Quality Technology*, to appear.

Boyles, R. A. (200 . "Gauge Capability for Pass–Fail Inspection". *Technometrics* 43, pp. 223–229.

Danila, O.; Steiner, S. H.; and MacKay, R. J. (2008). "Assessing a Binary Measurement System". *Journal of Quality Technology* 40, pp. 310–318.

Danila, O.; Steiner, S. H.; and MacKay, R. J. (2010). "Assessment of a Binary Measurement System in Current Use". *Journal of Quality Technology* 42, pp. 152–164.

De Mast, J. (2007). "Agreement and Kappa Type Indices". *The American Statistician* 61, pp. 148–153.

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, pp. 191–202.

De Mast, J. and Van Wieringen, W. N. (2010). "Modeling and Evaluating Repeatability and Reproducibility of Ordinal Classifications". *Technometrics* 52, pp. 94–106.

Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists*. London: Law Erlbaum Associates.

Farnum, N. R. (1994). *Modern Statistical Quality Control and Improvement*. Belmont, CA: Duxbury Press.

Hui, S. L. and Walter, S. D. (1980). "Estimating the Error Rates of Diagnostic Tests". *Biometrics* 36, pp. 167–171.

International Organization for Standardization (1995). *Guide to the Expression of Uncertainty in Measurement*, 1st ed. Geneva, Switzerland: Author.

Lindley, D. V. and Novick, M. R. (1981). "The Role of Exchangeability in Inference". *The Annals of Statistics* 9, pp. 45–58.

Mawby, W. D. (2006). *Make Your Destructive, Dynamic, and Attribute Measurement System Work for You*. Milwaukee, WI: ASQ Quality Press.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press.

Van Wieringen, W. N. (2005). "On Identifiability of Certain Latent Class Models". *Statistics and Probability Letters* 75, pp. 211–218.

Van Wieringen, W. N. and De Mast, J. (2008). "Measurement System Analysis for Binary Data". *Technometrics* 50, pp. 468–478.

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, pp. 495–507.