Statistical evaluation of binary measurement systems

Erdmann, T.P.

Citation for published version (APA):
Erdmann, T. P. (2012). Statistical evaluation of binary measurement systems Amsterdam: Universiteit van Amsterdam

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.


2 Measurement system analysis for binary inspection: Continuous versus dichotomous measurands

2.1 Introduction

The literature describes a multitude of methods for studying the reliability of binary measurements (see Chapter 1). In this chapter, we aim to provide insight into the question 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 conditional independence 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 which 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.

2.2 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.

2.2.1 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.

2.2.2 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 light bulbs – the measurand \( X \) is whether the light bulb 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
2.2 General set-up

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,

$$(2.1) \quad \log \left( \frac{q(x)}{1 - q(x)} \right) = (x - \delta) / \sigma$$

(see Figure 2.1). Items with $X > USL$ are defective, while items with $X > \delta$ are likely to be rejected. Thus, the curve’s inflection point $\delta$ 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 $\delta$ strictly below $USL$. The value $\sigma$ 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.

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
Figure 2.1: Characteristic curve \( q(x) = P(Y=0|X=x) \) (solid curve) and density \( f_X(x) = \frac{dP(X \leq x)}{dx} \) (dashed curve).

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. 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. That is, they assume that \( Y_{ij} \) and \( Y_{kl} \) are independent conditional on the measurands \( X_i \) and \( X_k \) (conditional independence). Further, we need that \( P(Y_{ij} = 0 | X_i = x_i) = P(Y_{kl} = 0 | X_k = x_k) \) if \( x_i = x_k \). Note that this implies that (repeated) measurements on items with the same \( X \) are i.i.d. conditional on \( X \). In Bayesian terminology, we need the sequence to be exchangeable in \( Y \) conditional on \( X \) (Lindley and Novick, 1981). As we will see next, the assumption of conditional independence is easily violated in practice.

### 2.2.3 False dichotomies and conditional independence

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 \( \tilde{X} \), which is 1 if \( X \leq USL \) and 0 if \( X > USL \). Treating a
continuous measurand as dichotomous creates complications, as, in general, it creates an intrinsic reason for the conditional independence 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 \( \tilde{X}_i \):

\[
P(Y_{i1} = 0, Y_{i2} = 0 \mid \tilde{X}_i = 0) = \frac{\int_{-\infty}^{\infty} P(Y_{i1} = 0, Y_{i2} = 0, \tilde{X}_i = 0 \mid X_i = x) f_X(x) dx}{P(\tilde{X}_i = 0)}
\]

\[
= \frac{\int_{USL}^{\infty} P(Y_{i1} = 0, Y_{i2} = 0 \mid X_i = x) f_X(x) dx}{\int_{USL}^{\infty} f_X(x) dx}
\]

\[
= \frac{\int_{USL}^{\infty} q^2(x) f_X(x) dx}{\int_{USL}^{\infty} f_X(x) dx},
\]

which is, in general, not equal to

\[
P(Y_{i1} = 0 \mid \tilde{X}_i = 0)P(Y_{i2} = 0 \mid \tilde{X}_i = 0) = \left(\int_{USL}^{\infty} q(x) f_X(x) dx / \int_{USL}^{\infty} f_X(x) dx\right)^2,
\]

unless \( q \) is a step function with step at \( x = USL \). In words, \( Y \) depends not only on \( \tilde{X} \) (whether the item is good or defective), but in addition, it depends on \( X \) (the degree of goodness or defectiveness). We introduce the phrase false dichotomies for such measurands that are artificially dichotomized, and whose characteristic curve is not a step function.

The complications brought about by false dichotomies stem from the fact that the \( Y_{ij} \) are no longer i.i.d. conditional on the event that \( \tilde{X}_i = 0 \) (or \( \tilde{X}_i = 1 \)). In fact, for each \( X = x \) value, \( Y \) has a different probability distribution, and the probabilities \( P(Y = 0 \mid \tilde{X} = 0) \) and \( P(Y = 0 \mid \tilde{X} = 1) \) are now weighted averages of \( q(x) \) values, weighted by the distribution \( F_X \) of the continuous measurand:

\[
\tilde{q}(0) = P(Y = 0 \mid \tilde{X} = 0) = \int_{USL}^{\infty} q(x) f_X(x) dx / \int_{USL}^{\infty} f_X(x) dx.
\]

\[
\tilde{q}(1) = P(Y = 0 \mid \tilde{X} = 1) = \int_{USL}^{\infty} q(x) f_X(x) dx / \int_{USL}^{\infty} f_X(x) dx.
\]

The practical ramification for sampling and estimation concerns exchangeability of the \( Y_{ij} \) values in the sample and the population. For a true dichotomy, exchangeability is given by the i.i.d. property (conditional on \( \tilde{X}_i = 0 \) or \( \tilde{X}_i = 1 \)). For a false dichotomy, exchangeability can be ensured by obtaining a sample of items in which the distribution of \( X \) values is representative of the distribution of \( X \) values in the whole population, for example, by random sampling. The complication, that we will discuss repeatedly in subsequent sections, is that in
many situations it is not clear how such random samples from the relevant subpopulation can be realized.

2.3 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, $p(1) = P(X = 0|Y = 1)$. The probabilities of interest are $q(1) = P(Y = 0|X = 1)$ and $q(0) = P(Y = 0|X = 0)$, 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 (AIAG, 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 ($X = 0$ or 1).

2.3.1 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 $m_{00}$, $m_{10}$, $m_{01}$, and $m_{11}$ (where, for example, $m_{10}$ is the number of items with $Y = 1$ and $X = 0$). Estimation is straightforward
from sample proportions: \( \hat{q}(0) = m_{00} / n_0 \) and \( \hat{q}(1) = m_{01} / n_0 \); the FAP and FRP are derived from these values.

We study what happens if the measurand is a false dichotomy, with \( \bar{X} = 1 \) 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

\[
E(\hat{q}(1)) = E(m_{01} / n_1) = \int_{-\infty}^{USL} q(x)f_x(x)dx / \int_{-\infty}^{USL} f_x(x)dx,
\]

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 (2.4), the distribution \( F_x^s \) of \( X \) in the sample equals the population distribution \( F_X \) and, therefore, \( E(\hat{q}(1)) = \bar{q}(1) = P(Y = 0 | \bar{X} = 1) \), per Equation (2.3).

**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 (2.2) and (2.4) show that, if \( q(x) \) is a step function, \( E(\hat{q}(1)) \) is independent of the sampling distribution \( F_x^s \), and there is no intrinsic reason for repeated classifications of an item not to be independent 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 \( F_X \). 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 (2.1) with \( \delta = USL \) and \( \sigma = 0.5 \), which gives \( FRP = 0.0284 \) (from Equation (2.3)). 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 $F_X$ 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 $E(FRP) = 0.0630$, 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 $F_X$ to be a normal distribution with mean 2.5 and standard deviation 0.5. The expected result is $E(FRP) = 0.325$, overestimating the FRP by nearly a factor 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 ($|E(FAP) - FAP| < 0.035$ and $|E(FRP) - FRP| < 0.030$) if the inflection point $\delta$ 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 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 \textit{naive sampling with swapping}. Also under this strategy, the estimated \textit{FAP} and \textit{FRP} are biased, with similar consequences as for naive sampling without swapping; see the Appendix.

\subsection*{2.3.2 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 the measurand for these items gives the numbers $m_{00}$, $m_{10}$, $m_{01}$, and $m_{11}$. From these, $\hat{p}(0) = m_{00} / m_0$ and $\hat{p}(1) = m_{10} / m_1$ and, for example,

\begin{align}
\hat{q}(1) &= \frac{\hat{q}(1 - \hat{p}(0))}{\hat{q}(1 - \hat{p}(0)) + (1 - \hat{q})(1 - \hat{p}(1))},
\end{align}

with $\hat{q}$ a 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 (2.5). 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 $m_{00} / m_0$ and $m_{10} / m_1$ 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.

\subsection*{2.3.3 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 $m_{00}$, $m_{10}$, $m_{01}$, and $m_{11}$. Estimation is done from sample proportions: $\hat{q}(0) = m_{00} / n_0$ and $\hat{q}(1) = m_{01} / n_1$. 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 (2.4) 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 $E(\hat{q}(1))$ does not depend on the sampling distribution $F_X^s$
and where conditional independence 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.

### 2.4 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 independence; see also Hui and Walter (1980), Boyles (2001), and Beavers et al. (2011). 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:
(2.6) \[ P(X = 1 | Y = 0) = \frac{q(1)(1 - p)}{(q(1)(1 - p) + q(0)p)} \]

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 independence 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 subsamples 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 \( \delta \) 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:

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 1) in scientific use, and of 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 independence will typically be violated per Equation (2.2)). However, if \( X \) is truly dichotomous, there is no intrinsic reason for conditional independence to be violated, and careful experimental design may enable the assumption to be fulfilled.
2.5 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 $m_{0X_i}$ 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 $m_{0X_1} = 0$ and $m_{0X_n} = m_i$.

AIAG (2003) suggests to assume 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 (2.1), which is a traditional logistic regression model (with slope $\sigma^{-1}$ and intercept $-\delta/\sigma$). For each (known) $X_i$, we have the corresponding observed proportion $\hat{q}(X_i) = m_{0X_i} / m_i$ as an estimate of $q(X_i)$. From the observed proportions, $\delta$ and $\sigma$ 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 $\delta$ and $\sigma$ 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 $\sigma_m$ the measurement spread of the numerical gauge.

Alternatively, one may compute similar metrics as in the previous section, such as
2.5 Gold standard available, continuous measurand

\[
FAP = \frac{\int_{-\infty}^{\infty} (1 - q(x))f_X(x)dx}{\int_{-\infty}^{\infty} f_X(x)dx}
\]

and

\[
FRP = \frac{\int_{-\infty}^{USL} q(x)f_X(x)dx}{\int_{-\infty}^{\infty} f_X(x)dx}
\]

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.

2.6 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 (2.1). The difference with logistic regression, is that the \(X\) values are unobservable, and they are treated as a latent variable. This type of models 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 \(\sigma\), \(\delta\), and the parameters of \(F_X\) are estimated simultaneously. An exposition of these algorithms is beyond the scope of this chapter, but the reader is referred to the IRT literature (with Embretson and Reise, 2000, a recent overview).

Since 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 $\delta$, the inconsistent acceptance probability (IAP) and inconsistent rejection probability (IRP):

\[
IAP = P(Y = 1 \mid X > \delta) = \int_{\delta}^{\infty} (1 - q(x)) f_X(x) dx / \int_{\delta}^{\infty} f_X(x) dx,
\]

\[
IRP = P(Y = 0 \mid X \leq \delta) = \int_{-\infty}^{\delta} q(x) f_X(x) dx / \int_{-\infty}^{\delta} f_X(x) dx.
\]

Whereas $FAP$ 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 $\delta$), these $IAP$ and $IRP$ express the random component only.

Like latent class modeling, also latent trait modeling has some unresolved difficulties. A random sample of items ensures consistent estimates for $IAP$ and $IRP$, 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. This matter is studied more thoroughly in Chapter 5, where we propose a solution. Also, it is difficult to interpret the fitted characteristic curve in tangible terms, as the $x$-axis is abstract and dimensionless.

### 2.7 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 $\tilde{X} = 1$ if $0.450 \leq X \leq 0.545$ and $\tilde{X} = 0$ otherwise. For normal inspection, the reference values $X$ are not available, and neither are the $\tilde{X}$. Instead, the appraisers use a go/no-go gauge which 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 $\bar{X}$ values, so, for the sake of the MSA experiment, a gold standard is available.

### 2.7.1 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 $\bar{X}$ 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:

$$\text{FAP} = 0.083,$$
$$\text{FRP} = 0.052.$$  

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 $\sigma_x = 0.045$ 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 $\bar{X}$ 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:

$$
\hat{FAP} = 0.138,
\hat{FRP} = 0.027.
$$

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. Since we are dealing here with a false dichotomy, and a sample of parts whose representativeness is questionable, the results are not reliable.

### 2.7.2 Treating the measurand as continuous: Logistic regression and latent trait analysis

The artificial dichotomy defined by the $\bar{X}$ values is a false dichotomy, and consequently, repeated ratings of the same part are not independent conditional on $\bar{X}$. 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 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 $\delta_l$ and $\delta_u$ 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 $\hat{\delta}_l = 0.453$, $\hat{\delta}_u = 0.547$. 
and $\sigma = 0.00415$. Because the estimated $\delta_L$ and $\delta_U$ 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:

$$\hat{FAP} = \int_{-\infty}^{L_{SL}} (1 - \hat{q}(x)) \hat{f}_X(x) dx / \int_{-\infty}^{L_{SL}} \hat{f}_X(x) dx$$

$$+ \int_{L_{SL}}^{\infty} (1 - \hat{q}(x)) \hat{f}_X(x) dx / \int_{L_{SL}}^{\infty} \hat{f}_X(x) dx = 0.093,$$

$$\hat{FRP} = \int_{L_{SL}}^{\infty} \hat{q}(x) \hat{f}_X(x) dx / \int_{L_{SL}}^{\infty} \hat{f}_X(x) dx = 0.051.$$

with $\hat{q}$ and $\hat{f}_X$ the logit function and normal density based on $\hat{\delta}_L$, $\hat{\delta}_U$, $\hat{\sigma}_L$, $\hat{\mu}_X$ 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 $\mu_x$ and $\sigma_x$ 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 $\hat{\mu}_X$ and $\hat{\sigma}_X$. Of course, one could collect a random sample of parts, apply the gold standard, and estimate $\mu_x$ and $\sigma_x$ 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),

$$1 - q(x) = \frac{\exp((x - \hat{\delta}_L) / \hat{\sigma})}{1 + \exp((x - \hat{\delta}_L) / \hat{\sigma}) + \exp((2x - \hat{\delta}_L - \hat{\delta}_U) / \hat{\sigma})}.$$  

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 2.2. The definitions for the probabilities of inconsistent ordering in Equation (2.7) become

\[
\begin{align*}
IAP &= P(Y = 1 \mid X < \delta_L \text{ or } X > \delta_U), \\
IRP &= P(Y = 0 \mid \delta_L \leq X \leq \delta_U).
\end{align*}
\]

The results are \( IAP = 0.099 \) and \( IRP = 0.055 \). Also in this case, the potential nonrandomness of the sample makes the results unreliable; the form of the characteristic curve in Figure 2.2 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.

### 2.8 Conclusions

The concept of a false dichotomy, and its ramifications for conditional independence and estimation are this chapter’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
2.9 Appendix

many cases crucial. An example of such a 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 2.1. Most of the mentioned methods have been 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 a 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, which are judged repeatedly, one estimates the probability of agreement

\[ P_a = P(Y_1 = Y_2) = (1 - p)(q^2(1) + (1 - q(1))^2) + p(q^2(0) + (1 - q(0))^2), \]

where metrics such as \( FAP \) and \( FRP \) express agreement between observations \( Y \) and measurands \( X \), \( P_a \) expresses agreement among observations only \( (Y_1 \text{ to } Y_2) \). The \( \kappa \) (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 (cf. Chapter 4, 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 which is typically more relevant (as it represents the consumer’s risk). This matter is studied more thoroughly in Chapter 4.

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 continuum \( (x > 0) \), 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. Chapter 3 focuses on the question how such hybrid situations are to be modeled.

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

- 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. We propose solutions in Chapters 3 and 5.

<table>
<thead>
<tr>
<th>Gold-stand. Measurand</th>
<th>Method</th>
<th>Experimental design</th>
<th>Points of attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>A D</td>
<td>Nonparametric Farnum</td>
<td>Subsamples from the strata of (truly) good and defective items. Judge each item one or more times.</td>
<td>If the dichotomy is false, or if conditional independence is violated otherwise, the subsamples must be random, but this is unfeasible in practice.</td>
</tr>
<tr>
<td>A D, C</td>
<td>Nonparametric Plan I</td>
<td>Random subsamples from the strata of accepted and rejected items. Apply the gold-standard to each item.</td>
<td>Subsamples must be random. A historical estimate of ( q ) is needed to determine ( FAP ) and ( FRP ). Also works if the measurand is continuous.</td>
</tr>
<tr>
<td>A D, C</td>
<td>Nonparametric Plan II</td>
<td>A sample from the population of items. Apply the gold-standard to each item, and judge each item one or more times.</td>
<td>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.</td>
</tr>
<tr>
<td>U D</td>
<td>Latent class modeling</td>
<td>A sample from the items population, preferably as balanced as possible. Judge each item multiple times.</td>
<td>May be difficult to obtain a sample with sufficient defectives. For false dichotomies, the subsamples must be random, but this is unfeasible in practice.</td>
</tr>
<tr>
<td>A C</td>
<td>Logistic regression</td>
<td>Select items with equidistant ( X ) values. Judge each item several (typically 20) times.</td>
<td>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.</td>
</tr>
<tr>
<td>U C</td>
<td>Latent trait modeling</td>
<td>A random sample from the items population. Judge each item multiple times.</td>
<td>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.</td>
</tr>
</tbody>
</table>

Table 2.1: Overview of methods discussed in this paper. Gold standard is \( A \) (available) or \( U \) (unavailable). The table indicates whether methods are suited for \( D \) (dichotomous) or \( C \) (continuous) measurands.
2.9 Appendix

2.9.1 Naive sampling without swapping

The continuous measurand is dichotomized by defining $\tilde{X}$ if $X < USL$ and $X = 0$ otherwise. Let $F_X = \Phi$, the normal distribution with $\mu_X = 0$ and $\sigma_X = 1$, and $q$ defined as in Equation (2.1). The FAP is the proportion of accepted items in the subpopulation of defective items:

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

with $F_X^0(x)$ the distribution of $X$ in the subpopulation of defective items:

$$F_X^0(x) = P(X \leq x \mid \tilde{X} = 0, Y = 0)$$

$$= \int_{USL}^{x} \varphi(t) dt \int_{USL}^{\infty} \varphi(t) dt$$

(for $x \geq USL$). 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

$$F_X^{WoS,0}(x) = P(X \leq x \mid \tilde{X} = 0, Y = 0)$$

$$= \int_{USL}^{\infty} q(t) \varphi(t) dt \int_{USL}^{\infty} q(t) \varphi(t) dt$$

(for $x \geq USL$). Obviously, $F_X^{WoS,0} \neq F_X^0$ and, consequently,

$$E(\hat{FAP}) = E\left(\frac{m_{n_0}}{n_0}\right)$$

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

(a similar derivation can be given for FRP). The bias $E(\hat{FAP}) - FAP$ depends on $(USL - (USL - \mu_X) / \sigma_X$, $(\delta - USL) / \sigma_X$ and $\sigma / \sigma_X$. In cases where $F_X = \Phi$, $q$ is the logit link function and $\delta = USL$, plots of this bias (Figure 2.3) show that

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

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 $E(\hat{FAP}) - FAP$. 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 $\delta$ are underrepresented.

- The bias is generally quite small and never exceeds $-0.030$. This maximum is attained when $USL$ is close to $\mu_X$ and $\sigma$ is close to $\sigma_X$.

- 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.

### 2.9.2 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
\[ F_{X}^{WS}(x) = \frac{m_0}{m} P(X \leq x | Y = 0) + \frac{m_1}{m} P(X \leq x | Y = 1) \]
\[ = \frac{m_0}{m} \int_{-\infty}^{x} q(t) \varphi(t) dt + \frac{m_1}{m} \int_{-\infty}^{x} (1-q(t)) \varphi(t) dt \]
\[ = \frac{m_0}{m} \int_{-\infty}^{x} q(t) \varphi(t) dt + \frac{m_1}{m} \int_{-\infty}^{x} (1-q(t)) \varphi(t) dt . \]

The distribution in the subsample of defective items is
\[ F_{X}^{WS,0}(x) = F_{X|X=0}^{WS}(x) = \frac{F_{X}^{WS}(x) - F_{X}^{WS}(USL)}{1 - F_{X}^{WS}(USL)} \]

Again, \( F_{X}^{WS,0}(x) \neq F_{X}^{0}(x) \), 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 (2.6)), 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.