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.
5 Binary measurement system analysis with a latent continuous measurand

5.1 Introduction

This chapter studies measurement system analysis (MSA) for binary measurements \( Y \), that classify items as either ‘reject’ \( (Y = 0) \) or ‘accept’ \( (Y = 1) \), aiming to reflect a continuous measurand \( X \), an empirical property of the items that is not observed directly. Throughout this chapter we assume that a gold standard is unavailable, and therefore the measurand is treated as a latent continuous variable. An item is considered ‘defective’ if the measurand exceeds an upper specification limit \( (X > USL) \) and ‘good’ otherwise. The quality of the measurements can be expressed as error rates: the probability that a defective item is accepted is the false acceptance probability, \( FAP = P(Y = 1 \mid X > USL) \), and the probability that a good item is rejected is the false rejection probability, \( FRP = P(Y = 0 \mid X \leq USL) \).

The literature describes several methods for studying the reliability of binary measurements when a gold standard is unavailable. For an overview, see Chapter 2, or Van Wieringen and Van den Heuvel (2005). Many methods described for this situation are based on latent class models, which treat \( X \) as binary, such as the methods presented in Boyles (2001), Van Wieringen and De Mast (2008), Danila et al. (2010), and Beavers et al. (2011). However, if a method treats the measurand as binary when it is actually continuous (a false dichotomy), this brings about the complications analyzed in Chapter 2, suggesting that the choice of method should depend on the measurand being binary or continuous. In particular, Chapter 2 shows that such an artificial dichotomization of a continuous measurand creates an intrinsic reason for violation of the assumption that measurements are independent conditional on the measurand, and leads to biased estimates of the error rates. Recently, solutions to this problem have been proposed in the literature. One approach is based on a latent trait model, which treats the measurand as continuous; this approach was introduced for ordinal classifications in De Mast and Van Wieringen (2010). In the latent trait model, the reject probability is fitted as an increasing function of the continuous measurand \( q(x) = P(Y = 0 \mid X = x) \), called the characteristic curve. A second tentative solution is based
on a random effects model with varying error rates (Danila et al., 2012). The latter approach models the measurand as binary, but allows the error rates $FAP_i$ and $FRP_i$ to vary according to a beta distribution over the parts $i$. Note that the two approaches are not consistent with each other, as the latent trait model implies upper bounds for $FAP_i$ and $FRP_i$ at $1 - q(USL)$ and $q(USL)$ respectively, whereas a beta distribution has an upper bound of 1.

In the latent trait model as described by De Mast and Van Wieringen (2010) the sample of items is assumed to be randomly drawn from the population of all items. If it is not, the estimated values of the latent measurand and the corresponding reject probabilities represent the sampling distribution rather than the population distribution. In many manufacturing processes the defect rate is very low, and therefore, in a random sample, the number of defective items will be small or even zero. However, for precise estimation of the model parameters and $FAP$ in particular, it is desirable to have a sample with a substantial number of defective items. Danila et al. (2010, 2012) suggest taking two separate samples: one from the stream of accepted items and one from the stream of rejected items. In their model, they take into account the fact that the sampled items have been previously judged, by conditioning on the original classification outcome. However, they model the measurand as binary; their model with fixed error rates (Danila et al., 2010) leads to biased estimates if the measurand is actually continuous (cf. Chapter 2), and their model with random error rates (Danila et al., 2012) is not consistent with the common situation of a continuous measurand $X$ and an increasing characteristic curve $q(x)$. In the literature, currently no satisfactory method exists to assess a binary measurement system with a latent continuous measurand: the method by De Mast and Van Wieringen (2010) uses a random sample, but this gives imprecise estimates unless the sample size is enormous, and the methods by Danila et al. (2010, 2012) model the measurand as binary.

In this chapter, we try to find a suitable approach for binary MSA with a latent continuous measurand. We explore different sampling strategies for precise estimation of the parameters in a latent trait model. In particular, we aim to find an effective balance between a random sample and conditional samples taken from the streams of rejected and accepted items. Also, we present an estimation method for the latent trait model, which takes into account this sampling procedure.

The chapter is set up as follows. The next section introduces the experimental design and model and gives an interpretation of the model and its parameters. The estimation method is described in the third section. In the fourth section we explore the advantages and disadvantages of various sampling strategies, and in the fifth section we give a quantitative
comparison of these sampling strategies based on simulation. In the final section our conclusions are summarized.

5.2 Model and interpretation

We consider a binary measurement system with outcomes $Y=0$ (‘reject’) or $Y=1$ (‘accept’) and a latent measurand $X \in \mathbb{R}$. An item is ‘defective’ if $X$ exceeds an upper specification limit (USL) and otherwise it is ‘good’. To assess the reliability of the classification procedure one selects $I = I^{\text{ran}} + I^{\text{acc}} + I^{\text{rej}}$ items, of which $I^{\text{ran}}$ are a random sample from all items, $I^{\text{acc}}$ a random sample from the subpopulation of accepted items, and $I^{\text{rej}}$ a random sample from the subpopulation of rejected items. The sampled items are classified another $K$ times into the two categories applying similar procedures and under similar circumstances as the original measurement. The data are denoted $Y_{ik}$, with $i = 1, \ldots, I$ indexing items and $k = 0, \ldots, K$ indexing repeated measurements, where the original measurement, that determines whether an item is in the stream of accepted or rejected items, is indexed $k = 0$ and the measurements performed during the experiment are indexed $k = 1, \ldots, K$. Note that the $I^{\text{ran}}$ items sampled from the population of all items may not have been measured yet, and therefore may not have an original measurement $k = 0$. The repeated measurements may have been done by a number of appraisers, but, for simplicity, we treat measurements of different appraisers as replications of measurements by a single appraiser.

We model the $Y_{ik}$ using a latent variable model. The advantage of latent variable models is that the cause of the association among repeated measurements – the object’s measurand – is modeled explicitly. Consequently, the variation in the measurements is explicitly attributed to a systematic part (variation among items) and a random part (measurement variation), a practice which resembles the typical manner in which MSA studies for numerical measurements are modeled.

We follow the set-up that De Mast and Van Wieringen (2010) apply to ordinal classifications. In the general population of all items, the measurands are assumed i.i.d. (independent and identically distributed) and they have a distribution $F_X$. Note that in the subpopulations of accepted and rejected items, the measurands $X_i$ have a different distribution than in the general population. We assume that the repeated measurements $\{Y_{ik}\}_{k=1,\ldots,K}$ are
Binary measurement system analysis with a latent continuous measurand

independent conditional on the measurand $X_i$. This implies that besides $X_i$, there are no other properties of the items and no environmental factors that induce dependencies among the measurement results.

The rejection probability, conditional on the item’s measurand, is given by the characteristic curve:

\[
q(x) = P(Y_i = 0 | X_i = x).
\]

If $X$ is continuous, $q(x)$ is typically an S-curve. We will assume it is defined by the logit function,

\[
\log\left(\frac{q(x)}{1-q(x)}\right) = \alpha(x-\delta), \quad \alpha > 0.
\]

The curve’s inflection point $\delta$ can be interpreted as the threshold that appraisers appear to apply (with $q(\delta) = 0.5$). We will call this the decision threshold. Items with $x > \delta$ are more likely to be rejected than accepted. The value $\alpha > 0$ is a discrimination parameter, determining the steepness of the curve. Larger values of $\alpha$ correspond to better measurement reliability. Also, for the distribution $F_X$ we specify a parametric model. We will assume that in the general population of all items the measurands are normally distributed as $X_i \sim N(\mu_X, \sigma_X^2)$. Note that the origin and scale of the latent $X$-continuum are arbitrary and we will set them by fixing $\mu_X = 0$ and $\sigma_X = 1$, in which case, $F_X = \Phi$. Without these or similar restrictions, the model suffers from an identifiability problem. As an example, Figure 5.1 shows characteristic curves with all four combinations of the parameter values $\alpha = 5, 12$ and $\delta = 2, 3$ and the probability density function of $X$.
The purpose of the MSA study is to assess the quality of the measurement system. The quality of binary measurements can be expressed as the error rates \( FAP \) and \( FRP \) defined in Section 5.1. In the model described above, they are given by:

\[
FAP = \int_{USL}^{\infty} (1 - q(x)) \phi(x) dx / \int_{USL}^{\infty} \phi(x) dx
\]

\[
FRP = \int_{-\infty}^{USL} q(x) \phi(x) dx / \int_{-\infty}^{\infty} \phi(x) dx.
\]

Note that in the situation we consider in this chapter, where a gold standard is not available, the measurand’s continuum is often ill-defined, and only a conceptual entity, and it is treated here as a dimensionless scale. As a consequence, \( USL \) will in general not be known, and often it is not even possible to give it an operational definition. In turn, this makes it also impossible, in general, to give an operational definition of \( FAP \) and \( FRP \), as these are defined in terms of the \( USL \). 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 \( \delta \). These probabilities are the inconsistent acceptance probability (\( IAP \)) and inconsistent rejection probability (\( IRP \)):

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

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

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. This can be seen from the following decomposition of \( FRP \), where we assume that \( \delta \leq USL \) (and a similar decomposition can be given for \( FAP \) and for the case \( \delta > USL \)):

\[
FRP = P(Y = 0 \mid X \leq \delta) P(X \leq \delta \mid X \leq USL) + P(Y = 0 \mid \delta < X \leq USL) P(\delta < X \leq USL \mid X \leq USL).
\]

The last term is the contribution to \( FRP \) due to systematic measurement error, determined by the distance between \( \delta \) and \( USL \). The first term is \( IRP \) (both terms are multiplied by probability weights). Note that in the remainder, we will refer to \( USL \) and ‘defective items’, but the reader should bear in mind that although they play a role in our modeling as concepts, it may not be possible to give them an operational definition in practical situations.
In the model described above, \( IAP \) is (typically much) larger than \( IRP \), because for \( x > \delta \), \( \varphi(x) \) has the largest fraction of its mass in the range of hard-to-judge items (items with an \( X \) value close to \( \delta \)), and for \( x \leq \delta \) it does not. Figure 5.2 gives contour plots of \( IAP \) and \( IRP \) for varying parameter values \( \alpha \) and \( \delta \).

In an industrial process, binary inspections may be performed for different reasons. The Automotive Industry Action Group (AIAG, 2003) distinguishes process control and product control. If a measurement is used for process control it determines whether the process should be adjusted. If a measurement is used for product control, the measurement determines whether a part is sent to the customer or not. Depending on the purpose of the binary inspections, different aspects of the measurement quality may be relevant. In a process control situation one may be interested in the probability that the process is adjusted unnecessarily (Type I error) and the probability that the process is not adjusted when it should be (Type II error). Assuming that the decision rule to adjust the process depends on whether parts are rejected by the binary inspection, the Type I error probability will depend on \( FRP \) and the Type II error probability on \( FAP \). In a product control situation, on the other hand, it may be of interest what percentage of the parts that reach the customer are nonconforming, that is, \( P(X > USL | Y = 1) \), the probability that an item is defective, given that it has been accepted. Furthermore, one may want to know what percentage of the scrapped items are scrapped unnecessarily, that is, \( P(X \leq USL | Y = 0) \), the probability that an item is good, given that it has been rejected. If the reject rate \( P(Y = 0) \) is known, both

Figure 5.2: Contour plots of \( IAP \) and \( IRP \) as a function of \( \alpha \) and \( \delta \).
probabilities can be calculated from $FAP$ and $FRP$ by applying Bayes’ Law. They can also be calculated from the parameters of $q(x)$, if the USL is known:

$$P(X > USL \mid Y = 1) = \int_{-\infty}^{\infty} (1 - q(x))\phi(x)dx / \int_{-\infty}^{\infty} (1 - q(x))\phi(x)dx$$

$$P(X \leq USL \mid Y = 0) = \int_{-\infty}^{USL} q(x)\phi(x)dx / \int_{-\infty}^{\infty} q(x)\phi(x)dx.$$

Note that $P(X > USL \mid Y = 1)$ is small, even if the measurement system is completely uninformative (if $q(x)$ is constant for all $x$). In that case it equals the defect rate $P(X > USL)$. The percentage of good items in the stream of rejects $P(X \leq USL \mid Y = 0)$ is typically very large. If the measurement system is unbiased ($\delta = USL$ and thus $FAP > FRP$, see previous paragraph), it is larger than 50% whenever the defect rate $P(X > USL)$ is less than $FRP$, as shown in the Appendix. This has an interesting implication for MSA studies: It shows that a sample of $r^{(x)}$ items taken from the stream of rejects typically consists of a large proportion of good items.

Besides in estimation of the error rates defined above, one may also be interested in comparing measurement systems, or different appraisers, in terms of the location of the decision threshold $\delta$ and the discrimination (or precision) as reflected by $\alpha$. The bias of a measurement system is given by $|\delta - USL|$, but as said, USL will generally not be known. Still, differences in $\delta$ between appraisers or measurement systems indicate systematic differences. The precision of a binary measurement system is reflected by $\alpha$, which determines the ‘gray area’, the range of items that are hard to judge. This can be operationalized as the range of those items for which the probability of misclassification is larger than $m$ (e.g., $m = 0.005$). Thus defined, the gray area is given by the interval $\left(\delta - \log(\frac{l-m}{m})/\alpha, \delta + \log(\frac{l-m}{m})/\alpha\right)$, and therefore its width for $m = 0.005$, called gauge repeatability and reproducibility (GRR) by AIAG (2003, pp. 125-140), is equal to $GRR = 2\log(\frac{0.995}{0.005})/\alpha$. Following AIAG, we define the GRR percentage ($%GRR$) as the GRR divided by the width of a 99% prediction interval for $X$, that is:

$$%GRR = \frac{q^{-1}(0.995) - q^{-1}(0.005)}{\Phi^{-1}(0.995) - \Phi^{-1}(0.005)}$$

In the model based on (5.1) and (5.2), it equals $%GRR = 2.055/\alpha$. According to AIAG’s acceptability criteria for precision, $%GRR < 0.10$ is generally considered to be acceptable and $%GRR > 0.30$ to be unacceptable. In our model, this corresponds with $\alpha > 20.5$ being considered acceptable and $\alpha < 6.8$ unacceptable.
5.3 Estimation

The parameters $\alpha$ and $\delta$ of model (5.2) are estimated from the experimental data by means of the maximum likelihood method. Conditional on $X$, the probability of a specific set of outcomes of $K$ repeated measurements on a single item, is

$$P(Y_{it} = y_{it}, \ldots, Y_{ik} = y_{ik} \mid X_i = x) = q(x)^r(1 - q(x))^{K - r},$$

where $r_i = \sum_{k=1}^{K}(1 - y_{ik})$ is the number of rejections of item $i$ in the MSA experiment. The measurand $X_i$ is unknown and therefore we treat it as a latent variable.

Following the terminology in Danila et al. (2010), we will refer to a sample that is representative for the general population of all items as a ‘random sample’. If a random sample is taken, the unconditional probability of the $K$ measurement outcomes of a single item is

$$P(Y_{i} = y_{i1}, \ldots, Y_{ik} = y_{ik}) = \int_{-\infty}^{\infty} \phi(x)P(Y_{it} = y_{it} = y_{ik} \mid X_i = x)dx$$

$$= \int_{-\infty}^{\infty} \phi(x)q(x)^r(1 - q(x))^{K - r} dx,$$

where, as mentioned in the previous section, we take the $X_i$ to be standard normally distributed. The log-likelihood of the experimental outcomes based on a random sample of $I_{ran}$ items is

$$\log L^{ran}(\theta) = \sum_{i=1}^{I_{ran}} \log \int_{-\infty}^{\infty} \phi(x)q^\theta(x)^r(1 - q^\theta(x))^{K - r} dx,$$

where $\theta$ is the parameter vector $\theta = (\alpha, \delta)$. We approximate the integrals in the likelihood by means of an adaptive quadrature as implemented in R in the function integrate (Piessens, 1983). Taking the derivative of the log-likelihood to each parameter $\theta_i$, one obtains the elements of the gradient:

$$\frac{\partial \log L^{ran}(\theta)}{\partial \theta_i} = \sum_{i=1}^{I_{ran}} \int_{-\infty}^{\infty} \phi(x)q^\theta(x)^r(1 - q^\theta(x))^{K - r - 1}(r_i - Kq^\theta(x)) \frac{\partial q^\theta(x)}{\partial \theta_i} \frac{dx}{dx}. $$

This expression is true for all choices of characteristic function $q^\theta(x)$. If $q^\theta(x)$ is defined by (2), then its partial derivatives with respect to the parameters $\alpha$ and $\delta$ are

$$\frac{\partial q^\theta(x)}{\partial \alpha} = \frac{(x - \delta)e^{\alpha(x - \delta)}}{1 + e^{\alpha(x - \delta)^2}} = (x - \delta)q^\theta(x)(1 - q^\theta(x))$$

$$\frac{\partial q^\theta(x)}{\partial \delta} = \frac{-\alpha e^{\alpha(x - \delta)}}{(1 + e^{\alpha(x - \delta)^2})^2} = -\alpha q^\theta(x)(1 - q^\theta(x)).$$
As announced earlier, another sampling strategy is to select \( I^{acc} \) items from the stream of accepted items, and \( I^{rej} \) from the stream of rejects. These two subsamples combined are called a ‘conditional sample’, again following Danila et al. (2010). For each of the \( I^{con} = I^{acc} + I^{rej} \) items in this conditional sample, it is known whether it has been rejected or accepted in the original measurement \( Y_{i0} \). Conditioning on \( Y_{i0} = y_{i0} \) and taking expectations over the latent variable \( X_i \), the joint probability of the \( K \) measurement outcomes of an item in the conditional sample is:

\[
P(Y_i = y_{i1}, \ldots, Y_i = y_{ik} | Y_{i0} = y_{i0}) = \frac{P(Y_{i0} = y_{i0}, Y_i = y_{i1}, \ldots, Y_i = y_{ik})}{P(Y_{i0} = y_{i0})}
\]

\[
= \int_{-\infty}^{\infty} \phi(x)P(Y_{i0} = y_{i0}, Y_i = y_{i1}, Y_i = y_{ik} | X_i = x) dx
\]

\[
= \int_{-\infty}^{\infty} \phi(x)q(x)^{y_{i1} - y_{i0}} (1 - q(x))^{k - y_{i0}} dx
\]

The log-likelihood of the experimental outcome based on a conditional sample of \( I^{con} = I^{acc} + I^{rej} \) items, becomes

\[
\log L^{con} = \sum_{i=1}^{I^{con}} \left( \log \int_{-\infty}^{\infty} \phi(x)q(x)^{y_{i1} - y_{i0}} (1 - q(x))^{k - y_{i0}} dx \right)
\]

(5.6)

\[
= \sum_{i=1}^{I^{con}} \left( \log \int_{-\infty}^{\infty} \phi(x)q(x)^{y_{i1} - y_{i0}} (1 - q(x))^{k - y_{i0}} dx \right)
\]

and the elements of the gradient become

\[
\frac{\partial \log L^{con}}{\partial \theta_i} = \sum_{i=1}^{I^{con}} \left( \frac{\partial}{\partial \theta_i} \int_{-\infty}^{\infty} \phi(x)q(x)^{y_{i1} - y_{i0}} (1 - q(x))^{k - y_{i0}} dx \right)
\]

(5.7) \( \log L^{his} = r^{his} \log \int_{-\infty}^{\infty} \phi(x)q(x) dx \). Danila et al. (2010, 2012) call such a historical dataset ‘baseline data’. They show that incorporating these data in the estimation substantially increases the precision of the estimators in the models for binary MSA they discuss. The log-likelihood of the number of rejects in the historical dataset is
where \( r_{\text{his}} = \sum_{i=1}^{n_{\text{his}}} (1 - y_{i0}) \). The gradient is

\[
\frac{\partial \log L_{\text{his}}(\theta)}{\partial \theta_i} = \frac{r_{\text{his}} - I_{\text{his}} \int_{-\infty}^{\infty} \varphi(x) q^\theta(x) \, dx}{\left(1 - \int_{-\infty}^{\infty} \varphi(x) q^\theta(x) \, dx\right) \int_{-\infty}^{\infty} \varphi(x) q^\theta(x) \, dx} \int_{-\infty}^{\infty} \varphi(x) \frac{\delta \varphi(x)}{\delta \theta_i} \, dx.
\]

If a combination of random and conditional samples is used, and a historical dataset is available, the log-likelihood of all data is the sum of the different log-likelihood functions defined in Equations (5.5), (5.6) and (5.7):

\[
(5.8) \quad \log L(\theta) = \log L_{\text{ran}}(\theta) + \log L_{\text{com}}(\theta) + \log L_{\text{his}}(\theta).
\]

We maximize the log-likelihood function \( \log L(\alpha, \delta) \) with gradient \( \partial \log L(\alpha, \delta) / \partial \alpha \), \( \partial \log L(\alpha, \delta) / \partial \delta \) under the constraint \( \alpha > 0 \), using the BFGS algorithm by Broyden, Fletcher, Goldfarb and Shanno as implemented in R in the function \text{maxbfgs} (see, for example, Fletcher, 1970) with the starting values of both parameters set to 1. After having estimated the parameters \( \alpha \) and \( \delta \), they can be plugged into \( q(x) \) in Equation (5.4) in order to obtain the estimates \( \widehat{IAP} \) and \( \widehat{IRP} \).

### 5.4 Various sampling strategies and intuitive motivation

Before a more rigorous and quantitative evaluation of various sampling strategies, we first explore the advantages and disadvantages of the sampling procedures introduced in the previous sections, aiming to build an intuitive understanding.

In their paper about the latent trait model for ordinal classifications, De Mast and Van Wieringen (2010) assume that the sample of items is a random sample from the general population of all items. In typical situations the defect rate is low, and such a random sample will contain only a small number of defective items. Figure 5.3a shows empirical 95% confidence bounds for the values of the characteristic curve with parameters \( \alpha = 10 \) and \( \delta = 3 \), if the curve is estimated using a random sample of \( I_{\text{ran}} = 200 \) items and \( K = 9 \) repeated measurements per item (note that the 95% lower confidence bound is flat on the x-axis and therefore difficult to see). The percentiles are based on 1000 simulated datasets of MSA experiments. The box plot below the graph shows the quartiles of the probability distribution of \( X \) for items in the random sample, where the box represents the 25%, 50% and 75% quartiles, and the end-points of the whiskers delineate a 99% interval. It can be seen that
only a small part or even none of the items in a sample are defective (assuming \(\text{USL} = \delta = 3\)), or even within the gray area, and therefore a random sample provides little or no information about the shape of the curve \(q(x)\). As a consequence, it does not allow for precise estimation of the model parameters; they may not even be identified. In the situation depicted in Figure 5.3a, in 62% of the cases \(\delta = 3\) is estimated as \(\hat{\delta} > 7\) indicating that no or very few items have been rejected in the experiment, and this causes the 95% lower bound of the characteristic curve to lie almost flat on the \(x\)-axis.

Aiming to obtain a larger number of defective items, an approach proposed by Danila et al. (2010) in the context of a latent class model, is to take a \textit{conditional sample}: one subsample from the stream of accepted items and one from the stream of rejected items. A common procedure intended to obtain a sample with equal numbers of good and defective items, is to sample equally from the streams of accepted and rejected items. However, if the defect rate is low, this may result in a sample with only very few defective items, as even the stream of rejected items often contains many good items that were incorrectly rejected (cf. Section 5.2). Danila et al. (2010) suggest sampling exclusively from the stream of rejected items. This approach typically leads to a more evenly balanced sample. The box plot below the graph in Figure 5.3b shows the distribution of \(X\) in a conditional sample taken from the subpopulation of rejected items. In this example, the median \(X\) in a sample of rejected items
is 3.14. The sample has substantial numbers of defective items and of items within the gray area. However, because all sampled items are in the tail of the distribution of $X$, it is hard to estimate at which percentile of $X$ the decision threshold $\delta$ is located, as can be seen from the wide empirical confidence bounds in Figure 5.3b. Furthermore, the simulations in the next section show that a conditional sample from the rejected items allows for precise estimation of $IAP$, but $IRP$ is estimated more precisely when a random sample is used. A practical risk of a conditional sample is that the experimental data become useless if the circumstances during the experiment are different from the circumstances during the original classifications that sent the items to the streams of rejected or accepted items, whereas a random sample still allows conclusions regarding the measurement system under the circumstances used during the experiment.

In summary, on the one hand, a random sample of all items has too few items in the gray area to determine the shape of $q(x)$ and does not estimate $IAP$ precisely, and on the other hand, a conditional sample of only rejected items makes it difficult to determine $\delta$ (the location of $q(x)$ with respect to $f_X(x)$) and does not allow precise estimation of $IRP$. A combination of a random sample and a conditional sample from the rejected items combines the advantages of both sampling strategies: both the location and the shape of $q(x)$, and both $IAP$ and $IRP$ can be estimated precisely. Figure 5.3c shows the empirical 95% confidence
5.4 Various sampling strategies and intuitive motivation

bounds of the values of the characteristic curve based on estimation with such a combined sample, with $I_{ran}=100$ items sampled randomly and $I_{rej}=100$ of the items taken from the rejected items. The confidence bounds have become much narrower, and $IAP$ and $IRP$ are both estimated with reasonable precision, with empirical 95% confidence intervals of $0.1387 < IAP < 0.1900$ and $0.0000 < IRP < 0.0027$ (true values are $IAP = 0.1655$, $IRP = 0.0005$).

An easy way to improve the precision of the estimates, is by incorporating a historical dataset of inspection results in the estimation, as explained at the end of the previous section. In particular, such a historical dataset helps to estimate the decision threshold $\delta$. As the simulations in the next section show, if a sample of rejected items is supplemented with a historical dataset, it is no longer necessary to include items sampled randomly from all items. Of course, it is essential that during the period over which this historical dataset is obtained, the measurement system and the circumstances are identical as during the MSA experiment. A historical dataset is typically easy to obtain. Even if it is not readily available, it can be obtained during the collection of rejected items for the MSA study (which typically involves several thousand inspections).
5.5 Quantitative evaluation by means of simulation

By means of Monte Carlo simulation, we investigate what combination of the sampling methods introduced in the previous section performs best in terms of bias and precision of the estimators \( \hat{IAP} \) and \( \hat{IRP} \). Secondly, we determine the precision for different sample sizes. Finally, we assess the robustness against model misspecification. The bias of \( \hat{IAP} \) and \( \hat{IRP} \) will be measured as the absolute deviation of the Monte Carlo mean from the true value. The precision will be quantified as the width of empirical 95% confidence intervals for \( IAP \) and \( IRP \) (based on the empirical 2.5% and 97.5% percentiles of \( \hat{IAP} \) and \( \hat{IRP} \)).

The simulations are performed by simulating \( R \) datasets as follows: we draw a random sample of \( I^{\text{run}} \) realizations of \( X \) from the distribution \( F_X \) (standard normal, unless specified otherwise). In order to create the conditional samples, we continue drawing \( X \) values, and for each of these, a measurement \( Y \) is drawn from a Bernoulli distribution with 
\[
P(Y = 0 \mid X = x) = q(x),
\]
as specified in Equations (5.1) and (5.2). We retain the first \( I^{\text{rej}} \) realizations of \( X \) for which \( Y = 0 \) (‘reject’), and the first \( I^{\text{acc}} \) realizations for which \( Y = 1 \) (‘accept’), and remove the rest. Thus, we obtain a sample of \( I = I^{\text{run}} + I^{\text{rej}} + I^{\text{acc}} \) items \( i = 1, \ldots, I \). For each \( X_i = x \), the \( K \) measurements \( Y_{i1}, \ldots, Y_{iK} \) are drawn based on (5.1) and (5.2). In addition, we draw the number of rejects in the historical dataset from a binomial distribution with \( I^{\text{his}} \) trials and rejection probability 
\[
P(Y = 0) = \int_{-\infty}^{\infty} f_X(x)q(x)dx.
\]
For each simulated dataset, the parameters \( \alpha \) and \( \delta \) of the characteristic curve \( q(x) \) are estimated by maximum likelihood, with the log-likelihood defined by Equations (5.5) through (5.8), after which \( \hat{IAP} \) and \( \hat{IRP} \) are calculated by plugging these estimates into Equation (5.4).

In order to explore which sampling strategy works best, we first investigate a number of different sampling strategies for the case \( \alpha = 5, \delta = 2, I = 200, K = 9 \). For these parameter values, the probabilities of inconsistent ordering are \( IAP = 0.2154 \) and \( IRP = 0.0125 \), based on Equation (5.4). We consider all combinations of random and conditional samples where \( I^{\text{run}}, I^{\text{rej}} \) and \( I^{\text{acc}} \) are multiples of 50 and add up to \( I = 200 \). For each sampling strategy, we consider both the situation where no historical dataset is available and the situation where \( I^{\text{his}} = 100,000 \). Table 5.1 gives Monte Carlo mean estimates and confidence interval widths for \( IAP \) and \( IRP \) for the various proportions between the three sampling methods, based on
5.5 Quantitative evaluation by means of simulation

Quantitative evaluation by means of simulation. It is also indicated whether the Monte Carlo means are significantly different from the true values $I_{\text{AP}} = 0.2154$ and $I_{\text{RP}} = 0.0125$ based on a $t$-test.

The simulation results in Table 5.1 show that differences between sampling strategies are in the precision of the estimates rather than the bias. In about a third of the cases there is evidence of a slight bias, but this is to be expected in a finite sample and its magnitude is generally not concerning. As for the precision, all samples with $I_{\text{rej}} = 0$ give an unacceptably large confidence interval for $I_{\text{AP}}$, as is to be expected because of the lack of defective items in those samples. This motivates the importance of sampling at least part of the items selectively from the subpopulation of rejects. The results also show that incorporating a

<table>
<thead>
<tr>
<th>$\mu_{\text{ran}}$</th>
<th>$\mu_{\text{acc}}$</th>
<th>$\mu_{\text{rej}}$</th>
<th>$I_{\text{his}}=0$</th>
<th>$I_{\text{his}}=100,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$I_{\text{AP}}$</td>
<td>Width C.I.</td>
<td>$I_{\text{RP}}$</td>
<td>Width C.I.</td>
</tr>
<tr>
<td>200 0 0</td>
<td>0.2213*</td>
<td>0.0125</td>
<td>0.2127*</td>
<td>0.0124</td>
</tr>
<tr>
<td></td>
<td>0.1564</td>
<td>0.0131</td>
<td>0.1229</td>
<td>0.0109</td>
</tr>
<tr>
<td>150 50 0</td>
<td>0.2254*</td>
<td>0.0124</td>
<td>0.2131*</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td>0.1706</td>
<td>0.0138</td>
<td>0.1254</td>
<td>0.0112</td>
</tr>
<tr>
<td>150 0 50</td>
<td>0.2162</td>
<td>0.0126</td>
<td>0.2155</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0656</td>
<td>0.0137</td>
<td>0.0620</td>
<td>0.0057</td>
</tr>
<tr>
<td>100 100 0</td>
<td>0.2251*</td>
<td>0.0123*</td>
<td>0.2119*</td>
<td>0.0124*</td>
</tr>
<tr>
<td></td>
<td>0.1976</td>
<td>0.0135</td>
<td>0.1331</td>
<td>0.0117</td>
</tr>
<tr>
<td>100 50 50</td>
<td>0.2162</td>
<td>0.0125</td>
<td>0.2156</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0674</td>
<td>0.0146</td>
<td>0.0637</td>
<td>0.0059</td>
</tr>
<tr>
<td>100 0 100</td>
<td>0.2159</td>
<td>0.0127</td>
<td>0.2155</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td>0.0486</td>
<td>0.0177</td>
<td>0.0483</td>
<td>0.0045</td>
</tr>
<tr>
<td>50 150 0</td>
<td>0.2275*</td>
<td>0.0121*</td>
<td>0.2109*</td>
<td>0.0123*</td>
</tr>
<tr>
<td></td>
<td>0.2361</td>
<td>0.0140</td>
<td>0.1344</td>
<td>0.0116</td>
</tr>
<tr>
<td>50 100 50</td>
<td>0.2161</td>
<td>0.0125</td>
<td>0.2155</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0702</td>
<td>0.0158</td>
<td>0.0641</td>
<td>0.0059</td>
</tr>
<tr>
<td>50 50 100</td>
<td>0.2158</td>
<td>0.0126</td>
<td>0.2155</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td>0.0490</td>
<td>0.0189</td>
<td>0.0479</td>
<td>0.0044</td>
</tr>
<tr>
<td>50 0 150</td>
<td>0.2157</td>
<td>0.0130*</td>
<td>0.2158</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0427</td>
<td>0.0248</td>
<td>0.0402</td>
<td>0.0037</td>
</tr>
<tr>
<td>0 200 0</td>
<td>0.2176</td>
<td>0.0121*</td>
<td>0.2120*</td>
<td>0.0124</td>
</tr>
<tr>
<td></td>
<td>0.3625</td>
<td>0.0150</td>
<td>0.1350</td>
<td>0.0117</td>
</tr>
<tr>
<td>0 150 50</td>
<td>0.2161</td>
<td>0.0124</td>
<td>0.2154</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0708</td>
<td>0.0170</td>
<td>0.0643</td>
<td>0.0059</td>
</tr>
<tr>
<td>0 100 100</td>
<td>0.2157</td>
<td>0.0125</td>
<td>0.2156</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0493</td>
<td>0.0204</td>
<td>0.0488</td>
<td>0.0045</td>
</tr>
<tr>
<td>0 50 150</td>
<td>0.2155</td>
<td>0.0129*</td>
<td>0.2158</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>0.0432</td>
<td>0.0291</td>
<td>0.0408</td>
<td>0.0037</td>
</tr>
<tr>
<td>0 0 200</td>
<td>0.2163*</td>
<td>0.0200*</td>
<td>0.2155</td>
<td>0.0125</td>
</tr>
<tr>
<td></td>
<td>0.0475</td>
<td>0.0743</td>
<td>0.0360</td>
<td>0.0035</td>
</tr>
</tbody>
</table>

Table 5.1: Monte Carlo mean estimates and 95% confidence interval widths for $I_{\text{AP}}$ and $I_{\text{RP}}$ for $\alpha=5$, $\delta=2$, $l=200$, $K=9$, $R=2500$.

*Monte Carlo mean estimate significantly different from true value ($I_{\text{AP}}=0.2154$, $I_{\text{RP}}=0.0125$)
Binary measurement system analysis with a latent continuous measurand

Historical dataset in the likelihood results in a big improvement of the precision. Especially the confidence interval for \( IRP \) becomes much smaller: its width is at least cut in half for all samples with \( \text{IRP} > 0 \). With a historical dataset, a sample consisting of rejected items exclusively \( (I^{\text{rej}} = 200) \) is uniformly the best strategy. If a historical dataset is unavailable, one should use \( 0 < I^{\text{ran}} < 200, \quad 0 < I^{\text{rej}} < 200, \quad \text{and} \quad I^{\text{acc}} = 0 \). In that case, the best strategy depends on whether one is more interested in \( IAP \) or \( IRP \), but arguably a sample with \( I^{\text{ran}} = 150, \quad I^{\text{rej}} = 50 \) and \( I^{\text{acc}} = 0 \) gives the best overall precision. Note that generally a historical dataset will be available, as argued in the previous section, and thus a sample with only rejected items is the best choice.

Next, we investigate whether these results also hold for different parameter values. We perform simulations with all four combinations of \( \alpha = 5 \) (poorer reliability), \( \alpha = 12 \) (better reliability), and \( \delta = 2 \) (decision threshold at the 97.7% percentile of \( X \) values) and \( \delta = 3 \) (in the remote tail). Since the results in Table 5.1 suggest that sampling from the subpopulation of accepted items is never optimal, we only consider sampling strategies that combine a random sample of \( I^{\text{ran}} \) items and a selective sample of \( I^{\text{rej}} \) rejected items, again adding up to \( I = 200 \). A historical dataset based on \( I^{\text{his}} = 100,000 \) items is assumed to be available. The resulting empirical confidence interval widths for \( IAP \) and \( IRP \) are presented in

<table>
<thead>
<tr>
<th>((\alpha, \delta))</th>
<th>( I^{\text{ran}} )</th>
<th>( I^{\text{rej}} )</th>
<th>( I^{\text{his}} = 100,000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (5, 2) )</td>
<td>150</td>
<td>50</td>
<td>Width C.I. 0.0590</td>
</tr>
<tr>
<td>( \text{IAP} = 0.2154 )</td>
<td>100</td>
<td>100</td>
<td>Width C.I. 0.0486</td>
</tr>
<tr>
<td>( \text{IRP} = 0.0125 )</td>
<td>50</td>
<td>150</td>
<td>Width C.I. 0.0410</td>
</tr>
<tr>
<td>0</td>
<td>200</td>
<td>Width C.I. 0.0371</td>
<td>Width C.I. 0.0035</td>
</tr>
<tr>
<td>( (5, 3) )</td>
<td>150</td>
<td>50</td>
<td>Width C.I. 0.0559</td>
</tr>
<tr>
<td>( \text{IAP} = 0.2562 )</td>
<td>100</td>
<td>100</td>
<td>Width C.I. 0.0435</td>
</tr>
<tr>
<td>( \text{IRP} = 0.0015 )</td>
<td>50</td>
<td>150</td>
<td>Width C.I. 0.0354</td>
</tr>
<tr>
<td>0</td>
<td>200</td>
<td>Width C.I. 0.0282</td>
<td>Width C.I. 0.0005</td>
</tr>
<tr>
<td>( (12, 2) )</td>
<td>150</td>
<td>50</td>
<td>Width C.I. 0.0679*</td>
</tr>
<tr>
<td>( \text{IAP} = 0.1134 )</td>
<td>100</td>
<td>100</td>
<td>Width C.I. 0.0466</td>
</tr>
<tr>
<td>( \text{IRP} = 0.0039 )</td>
<td>50</td>
<td>150</td>
<td>Width C.I. 0.0414</td>
</tr>
<tr>
<td>0</td>
<td>200</td>
<td>Width C.I. 0.0351</td>
<td>Width C.I. 0.0017</td>
</tr>
<tr>
<td>( (12, 3) )</td>
<td>150</td>
<td>50</td>
<td>Width C.I. 0.0719</td>
</tr>
<tr>
<td>( \text{IAP} = 0.1447 )</td>
<td>100</td>
<td>100</td>
<td>Width C.I. 0.0497</td>
</tr>
<tr>
<td>( \text{IRP} = 0.0004 )</td>
<td>50</td>
<td>150</td>
<td>Width C.I. 0.0391</td>
</tr>
<tr>
<td>0</td>
<td>200</td>
<td>Width C.I. 0.0351</td>
<td>Width C.I. 0.0002</td>
</tr>
</tbody>
</table>

Table 5.2: 95% confidence interval widths for \( IAP \) and \( IRP \) for \( \alpha = 5, 12; \quad \delta = 2, 3; \quad I = 200; \quad K = 9; \quad I^{\text{his}} = 100,000; \quad R = 1000 \).

*Monte Carlo mean estimate significantly different from true value
5.5 Quantitative evaluation by means of simulation

Table 5.2, based on \( R = 1000 \) simulations. Again, the most precise results are obtained when only rejected items are included in the sample.

Now that we have established that, at least in the cases we considered, it is a good choice to use a sample exclusively of rejected items supplemented with a historical dataset, we investigate how the sample size \( I \) and the number of repeated measurements \( K \) affect the precision of \( \hat{IAP} \) and \( \hat{IRP} \). Figures 5.4a through 5.4d contain contour plots of the empirical

Figure 5.4a: Contour plots of the 95% confidence interval width for \( IAP \) and \( IRP \) as a function of \( I \) and \( K \), for \( \alpha = 5, \delta = 2, IAP = 0.2154, IRP = 0.0125, R = 2500. \)

Figure 5.4b: Contour plots of the 95% confidence interval width for \( IAP \) and \( IRP \) as a function of \( I \) and \( K \), for \( \alpha = 5, \delta = 3, IAP = 0.2562, IRP = 0.0015, R = 2500. \)
95% confidence interval widths as a function of $I$ and $K$ for all four combinations of $\alpha=5, 12$ and $\delta=2, 3$. The confidence interval widths are based on $R = 2500$ simulated datasets using samples of only rejected items ($I = I^{\text{rej}}$) and a historical dataset of $I^{\text{his}} = 100,000$ items. The contour plots are drawn using linear interpolation based on all $7 \times 7$ combinations of $K = 3, 5, \ldots, 15$ and $I = 50, 75, \ldots, 200$. The figures show that, for all
four combinations of parameter values, the marginal effects of both $I$ and $K$ on the precision of $\hat{IAp}$ and $\hat{IRP}$ diminish as $I$ and $K$ increase. For $K < 7$, an additional classification generally leads to a substantial improvement of precision, but using more than 7 repeated classifications hardly has an effect on precision. Therefore, we recommend using $K = 7$. The effect of the number of items $I$ on precision also diminishes for larger $I$, but less strongly so. $I = 150$ generally seems to give acceptably narrow confidence intervals (at least for the parameter values investigated). Of course, the required sample size depends on the preferred precision, and Figures 5.4a through 5.4d can be used as a reference when choosing $I$ and $K$ for an MSA experiment.

To assess the robustness of the proposed approach, we evaluate the bias and precision of $\hat{IAp}$ and $\hat{IRP}$ for several forms of misspecification. First, we vary the distribution $F_X$ of the measurand, and second, we investigate the situation where the measurand has several dimensions.

The distributions $F_X$ that we consider are $t$-distributions with 3 and 7 degrees of freedom (which are symmetric distributions with excess kurtosis), and a standard lognormal distribution and a $\chi^2$-distribution with 1 degree of freedom (which are asymmetric nonnegative distributions). To enable comparison of the bias and precision of the estimates over the different distributions, we use parameter values of $q(x)$ that represent measurement systems with comparable properties. Unfortunately, $\alpha$ and $\delta$ depend on the scale of $X$, which is different for each of the distributions. For example, an (unbiased) measurement system with $\delta = 2$ is associated with a defect rate of 0.023 in the standard normal case, but of 0.244 in the standard lognormal case. For comparison, we choose parameter values corresponding to constant reject rates $P(Y = 0) = \int_{-\infty}^{\infty} f_X(x)q(x)dx$ and constant $%GRR$, which we define, as earlier, as the range of those items for which the probability of misclassification is larger than 0.005 divided by the width of a 99% prediction interval for $X$ (cf. Section 5.2; AIAG, 2003), that is:

$$%GRR = \frac{q^{-1}(0.995) - q^{-1}(0.005)}{F_X^{-1}(0.995) - F_X^{-1}(0.005)}.$$  

For each distribution $F_X$, we consider all four combinations of $P(Y = 0) = 0.005$, 0.01 and $%GRR = 0.15$, 0.30. The corresponding $\alpha$ and $\delta$ are found by numerically solving a system of two equations (e.g. $P(Y = 0) = 0.01$, $%GRR = 0.15$) using the Newton-Raphson algorithm. The simulations are then performed drawing items from $F_X$, but incorrectly specifying $F_X = \Phi$.
in the likelihood function. For each distribution \( F_X \) and for each combination of \( P(\mathcal{Y} = 0) \) and %GRR, Table 5.3 gives the absolute deviation of the Monte Carlo means of \( \hat{IAP} \) and \( \hat{IRP} \) from their population values and the widths of empirical 95% confidence intervals. The population values \( IAP \) and \( IRP \) are calculated as

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

\[
IRP = \int_{-\infty}^{\delta} q(x) f_X(x) \, dx / \int_{-\infty}^{\delta} f_X(x) \, dx.
\]

The results indicate that for all distributions \( F_X \) we consider, the bias is very modest (less than 0.009 for \( IAP \) and less than 0.003 for \( IRP \)). As for the precision, some nonnormal distributions lead to slightly wider confidence intervals (relative to the values of \( IAP \) and \( IRP \)), but the confidence interval widths are generally acceptable (less than 0.046 for \( IAP \) and less than 0.0013 for \( IRP \)). This is reassuring, because nonnormal measurands, such as measurands with a nonnegative distribution, are common for binary inspections in industry (cf. Chapter 3); take as an example the size of a scratch, or the crookedness of a wrapping.

A second type of misspecification occurs when the measurand is multi-dimensional, that is, the inspection takes into account \( M \) properties instead of one property of the items.
5.5 Quantitative evaluation by means of simulation

For example, in injection molding, parts are inspected for splay marks (property 1), scratches (property 2), short shots (property 3), and more properties. Thus, the measurand is not a single continuum; in fact, it is an $M$-tuple of variables (i.e., $X \in \mathbb{R}^M$) with joint probability density function $f_X(x)$. An item is considered good if none of the elements $X_m$ exceeds its upper specification limit $USL_m$ and otherwise it is defective. The probability of rejecting an item, conditional on $X = x$, is the complement of the probability that an item is accepted on all properties:

$$q(x) = 1 - \prod_{m=1}^{M} (1 - q_m(x_m))$$

where we assume that each of the $q_m(x_m)$ is defined by the logit function (2) with parameters $\alpha_m$ and $\delta_m$. (For simplicity, we assume that the $M$ simultaneous inspections per property are independent conditional on $X$ and that each inspection depends only on the property it measures). The probabilities of inconsistent classification are

$$IAP = P(Y = 1 | X \leq \delta) = \frac{P(Y = 1) - P(Y = 1, X \leq \delta)}{1 - P(X \leq \delta)}$$

$$IRP = P(Y = 0 | X \leq \delta) = \frac{\int_{-\infty}^{\delta} \int_{-\infty}^{\delta} \cdots \int_{-\infty}^{\delta} (1 - q(x)) f_X(x) dx_1 \cdots dx_m}{\int_{-\infty}^{\delta} \int_{-\infty}^{\delta} \cdots \int_{-\infty}^{\delta} f_X(x) dx_1 \cdots dx_m},$$

where $\delta$ is the vector of decision thresholds $\delta_m$, and $X \leq \delta$ denotes the complement of the event $X \leq \delta$. We investigate the effects on bias and precision if these $m$ properties are incorrectly treated as a single continuous measurand (i.e. they are estimated by maximizing the log-likelihood defined by Equations (5.5) through (5.8), after which $\widehat{IAP}$ and $\widehat{IRP}$ are calculated by plugging these estimates into Equation (5.4)). In the simulation, we consider a four-dimensional measurand with a multivariate normal distribution with mean zero and one of the following three covariance matrices:

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Sigma^+ = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}, \quad \Sigma^- = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}.$$
Table 5.4: Absolute deviations of Monte Carlo mean estimates from true value and 95% confidence interval widths for \( I_{AP} \) and \( I_{RP} \) under misspecification, for a four-dimensional measurand, with \((\alpha_1, \delta_1)=(13.7, 2.60), \ (\alpha_2, \delta_2)=(13.7, 2.35), \ (\alpha_3, \delta_3)=(6.85, 2.67), \ (\alpha_4, \delta_4)=(6.85, 2.41), \ I=200, \ K=9, \ I_{his}=100,000, \ R=1000. \)

*Estimated bias significantly different from zero

In this chapter, we study a latent trait model for binary MSA with a latent continuous measurand. To obtain a sufficient number of defective items in the sample, we propose taking a conditional sample from the subpopulation of rejected items, and taking this sampling procedure into account in the estimation procedure. The precision of a measurement system is expressed in terms of probabilities of inconsistent classification \( I_{AP} \) and \( I_{RP} \).

Simulations show that the estimators \( \widehat{I_{AP}} \) and \( \widehat{I_{RP}} \) have the highest precision if only rejected items are included in the MSA experiment and the data are supplemented with a historical dataset. Furthermore, simulations show that this procedure is robust to certain forms of misspecification: Even if the distribution of the measurand has fat tails or is asymmetric, or if the measurand is multi-dimensional, the estimators perform reasonably well in terms of bias and precision.
The approach still lacks effective model diagnostics to assess the fit and to detect unusual observations, and this is a topic that further research should focus on. Another interesting question is how this method compares to other models such as the latent class model (Van Wieringen and De Mast (2008), Danila et al. (2010)) and the random effects model proposed by Danila et al. (2012). Finally, and perhaps most importantly, the approach needs to be tried and tested in practice, in order to provide evidence for its applicability.

Appendix

We show that the percentage of good items in the stream of rejects $P(X \leq USL \mid Y = 0)$ is larger than 50% whenever the defect rate $P(X > USL)$ is less than $FRP$, assuming $\delta = USL$ and thus $FAP > FRP$.

Let $FAP > FRP$ and $FRP > P(X > USL)$. Because $x \rightarrow \frac{x}{1-x}$ is an increasing function for $0 < x < 1$, it follows that $\frac{FRP}{1-FRP} > \frac{P(X > USL)}{1-P(X > USL)}$, and therefore $FRP(1-P(X > USL)) > (1-FRP)P(X > USL)$. Then, using $FAP > FRP$, it follows that $FRP(1-P(X > USL)) > (1-FAP)P(X > USL)$. This implies that $P(X \leq USL \mid Y = 0) = \frac{FRP(1-P(X > USL))}{FRP(1-P(X > USL)) + (1-FAP)P(X > USL)} > 0.5$, as claimed.