Statistical Evaluation of **Binary Measurement Systems**

Tashi Erdmann

Statistical Evaluation of **Binary Measurement Systems**

Tashi Erdmann

Stellingen

Behorende bij het proefschrift

Statistical Evaluation of
Binary Measurement Systems

Tashi Erdmann

21 november 2012

1. De te meten grootheid van een binair meetsysteem is vrijwel altijd een gedichotomiseerd continuüm.

Hoofdstuk 2 van dit proefschrift

2. Foutkansen van een binair meetsysteem hebben geen betekenis zonder een specificatie van de populatie van te meten objecten.

Hoofdstuk 2 van dit proefschrift

3. Een meetsysteemanalyse is effectiever en de geschatte foutkansen zijn betrouwbaarder voor een meetsysteem dat goed is ontworpen, waarbij de te meten grootheid goed is gedefinieerd en zoveel mogelijk storende invloeden al door het meetprotocol zijn uitgesloten.

4. Meten is weten; classificeren is slechts samenvatten.

Hoofdstuk 4 van dit proefschrift

5. Een nadeel van het streven naar een kwaliteitsniveau van *zes sigma*, is dat binaire meetsysteemanalyse steeds lastiger wordt.

Hoofdstuk 5 van dit proefschrift

6. De herhaalbaarheid en reproduceerbaarheid van metingen aan veranderlijke objecten kan men achterhalen door elk object op een aantal tijdstippen simultaan te meten met meerdere meetsystemen.

M. Awad, T.P. Erdmann, Y. Shanshal en B. Bard (2009), "A Measurement System Analysis Approach for Hard-to-Repeat Events", Quality Engineering 21: 300-305

7. Temperatuurmetingen met een oorthermometer zijn preciezer als men aan beide oren meet en het maximum neemt.

T.P. Erdmann, R.J.M.M. Does en Søren Bisgaard (2010), "Quality Quandaries: A Gage R&R Study in a Hospital", Quality Engineering 22: 46-53

8. Ook bij zwart-witoordelen zoals in termen van goed en kwaad dient men zich af te vragen of de onderliggende grootheid dichotoom, continu of multidimensionaal is.

Hoofdstuk 2 van dit proefschrift

9. Ook de Boeddha ontmaskerde al valse dichotomieën:

"In the sky, there is no distinction of east and west; people create distinctions out of their own minds and then believe them to be true." Toegeschreven aan Siddhartha Gautama, grondlegger van het boeddhisme, 563-483 v.Chr.

10. Voor verbetering is het cruciaal dat men fouten niet onder het tapijt veegt, maar openlijk en dankbaar benut om van te leren.

"Herkennen, erkennen en verbeteren." In: I.J.H. van Vlodrop, "Nippon Karatedo Genwakai Doctrinair Handboek" (te verschijnen)

11. Een verbetervoorstel dat niet in praktijk wordt gebracht, verbetert niets.

"To know and to act are one and the same." Toegeschreven aan Wang Yangming, Chinees filosoof, 1472-1529

Statistical Evaluation of Binary Measurement Systems

Tashi Erdmann

IBIS UvA
Instituut voor Bedrijfs- en Industriële Statistiek

Dit proefschrift is mede mogelijk gemaakt door een financiële bijdrage van het Instituut voor Bedrijfs- en Industriële Statistiek van de Universiteit van Amsterdam

Omslagontwerp: Esther Ris (www.proefschriftomslag.nl)

ISBN: 978-94-6108-354-8

Statistical Evaluation of Binary Measurement Systems

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op woensdag 21 november 2012, te 14:00 uur

door

Tashi Paul Erdmann
geboren te Amsterdam

Promotiecommissie

Promotor:	Prof.dr. J. de Mast
Overige leden:	Prof.dr. H.P. Boswijk
	Prof.dr. R.J.M.M. Does
	Prof.dr. E.R. van den Heuvel
	Prof.dr. M.R.H. Mandjes
	Prof.dr. S.H. Steiner
	Dr. M.J.G. Bun
	Dr. W.N. van Wieringen

Faculteit Economie en Bedrijfskunde

Contents

1 Introduction

The statistical evaluation of binary measurement systems is a topic in *measurement system analysis*, as the field is called in industrial statistics, or *measurement theory*, as it is known in the behavioral sciences. Measurement system analysis attempts to evaluate the quality of measurements statistically, based on experiments.

1.1 Measurement

Measurement is the assignment of symbols to items in such a way that specified relations among the symbols represent empirical relations among the items with respect to a property under study (cf. classical definitions of measurement, Allen and Yen, 1979, p. 2; Lord and Novick, 1986, p. 17; Wallsten, 1988). For example, when a person measures his body weight on a scale, the scale assigns a value in kilograms to the person, reflecting the person's weight. The relations between values in kilograms on the scale represent empirical relations between the weight of persons: a higher value in kilograms represents a larger body weight, and if the value in kilograms is twice as large, this represents the body weight being twice as large. Thus, measurement is conceived as a mapping from items to measurement values (cf. Hand, 1996; De Mast and Trip, 2005). The underlying empirical property being measured is often referred to as the *true value* or *actual state*, but is called the *measurand* in measurement theory and metrology (see, for example, the International Standardization Organization's *Guide to the Expression of Uncertainty in Measurement* (ISO, 1995)). The symbol assigned to the items is called the *measurement value*. A *measurement system* is the collection of instruments, operating procedures, personnel, et cetera, used to do a measurement.

The extent to which relations between measurement values reflect empirical relations among the items is determined by the *level* or *scale* of measurement (Stevens, 1946; Allen and Yen, 1979): *nominal scale*, *ordinal scale*, *interval scale* or *ratio scale* measurement. Nominal scale measurement, such as a classification of people according to their race, only reflects an empirical equivalence relation among items: items assigned the same measurement value are equivalent with respect to the measurand (race), but the order between

measurement values has no empirical meaning. Ordinal scale measurement, such as a classification of people according to their level of education, in addition reflects an order relation between items: a larger measurement value implies more of the measurand (education). Interval scale measurement, such as the measurement of a person's body temperature in degrees Celsius, in addition reflects distances among items with respect to the measurand, and therefore addition and subtraction of measurement values have empirical meaning. Ratio scale measurement, such as the measurement of a person's body weight, is the richest in the information it conveys, as it also reflects proportions among the items with respect to the measurand (a measurement value of zero represents absence of the property being measured): all arithmetic operations have empirical meaning, including multiplication and division.

The extent to which the measurement value is not in accordance with the measurand, is called *measurement error*. There are various properties of a measurement system that influence the probability distribution of measurement error and, consequently, the quality of measurement. The properties that are relevant depend, among others, on the scale of measurement. In industrial statistics, the relevant properties of interval scale and ratio scale measurements (numerical measurements) are divided into *accuracy* and *precision* (cf. Eisenhart, 1968; AIAG, 2003). Accuracy is the degree to which the measurement system is subject to *systematic measurement error* or *bias*, that is, the degree to which there is a difference between the expected value of a measurement and the measurand. Accuracy can be assessed by a calibration study comparing measurements to an accepted *reference value*; a value determined by a measurement system of much higher accuracy, traceable to a metrological standard (cf. Kimothi, 2002). Precision is the degree to which the measurement system is subject to *random measurement error* or *measurement spread*, that is, to variability in measurement values for a constant measurand. The issue of random measurement error is known in the behavioral sciences as the reliability of measurements (Kerlinger and Lee, 2000). Several factors may contribute to random measurement error, such as raters, equipment and environmental conditions. Precision can be assessed by conducting an experiment in which items are repeatedly measured, and estimating the standard deviation of repeated measurements. Such an experiment is called a *gauge repeatability and reproducibility* (R&R) study (Montgomery and Runger, 1993a,b; Vardeman and Van Valkenburg, 1999; Burdick, Borror and Montgomery, 2003). *Repeatability* refers to the variation in repeated measurements conducted under identical circumstances, and

reproducibility refers to the additional variation in repeated measurements due to varying circumstances (factors).

For ordinal and nominal scale measurements (categorical measurements), relevant properties of a measurement system include the probability of consistent classification, the probability of consistent ordering, the probability of agreement, and the probability of misclassification (De Mast and van Wieringen, 2004 and 2010).

In industrial statistics, the statistical evaluation of measurement systems by means of experiments is called measurement system analysis (MSA). A leading standard for MSA in industry is the manual by the Automotive Industry Action Group (AIAG, 2003), which gives extensive guidelines for performing MSA experiments. Measurements reported by (suppliers of) companies in the American automotive industry (and far beyond) are required to conform to AIAG's standards. Another important standard is the International Standardization Organization's *Guide to the Expression of Uncertainty in Measurement* (ISO, 1995).

1.2 Binary measurement system analysis

This thesis is about a class of measurement systems whose measurement values are on a two-point scale, a so-called binary scale. In industry, binary measurements abound. Think of visual inspections of products where the outcome can be 'pass' or 'fail', functional tests where the outcome is 'ok' or 'nok', and automated tests where some parts are rejected and others are accepted. Also beyond industry binary classifications are omnipresent, as in diagnostic tests in medicine (think of a pregnancy test).

Binary measurement aims to classify items into two categories, $Y=0$ ('reject') or $Y=1$ ('accept'), reflecting a property X of the item, the measurand, that is not observed directly, such as 'good' versus 'defective'. Binary measurement may be regarded as nominal measurement, or, if one category is appreciated above the other, as a trivial form of ordinal measurement. A common industrial application is pass/fail inspection, where X is a quality characteristic such as the state of a light bulb (functional or defective), the amount of discoloration of a food product, or the displacement of a certain component. As will be discussed in Chapter 2, X can be binary itself, as in the case of light bulbs being functional or defective. But in many cases, X is a continuous property, as in the case of the discoloration of

a food product, and in some cases, binary inspections reflect a combination of properties, in which case the measurand is multi-dimensional.

Binary measurements are, as all measurements, subject to error, especially since they are often based on visual or other sensory assessments by humans. An MSA study, an assessment of the quality of a measurement procedure, is as important for binary inspections as it is for other types of measurements. The quality of binary measurements can be expressed as error rates, or, alternatively, as agreement. The probability that a defective item passes is the *false acceptance probability (FAP)*, and the probability that a good item fails is the *false rejection probability (FRP)*. Moreover, when the measurand X is continuous, it is often desirable to know the rejection probability for any value of the measurand $X = x$, which can be represented as a graph $q(x) = P(Y = 0 \mid X = x)$ called the *characteristic curve*. An alternative measure for the quality of binary measurements is agreement. The probability that two measurements of the same item are equal is the *probability of agreement* (P_A). Agreement is often expressed in the form of a so-called κ (*kappa*) index, often interpreted as the probability of agreement corrected for agreement 'by chance'.

Recently a flow of literature has appeared regarding MSA of binary measurements, describing methods to determine *FAP*, *FRP*, P_A and the κ index. Some methods assume that a higher order, authoritative measurement procedure is available to determine the measurand, while others do not. The result of such a higher order measurement procedure is called the *gold standard* or reference value. Also, some methods model the measurand as a binary variable, others as a continuum. Furthermore, some methods require a random sample of items, whereas other methods employ different sampling strategies. See for example Van Wieringen and Van den Heuvel (2005) for an overview, and Boyles (2001), De Mast (2007), De Mast and Van Wieringen (2007), Danila et al. (2008), Lyu and Chen (2008), Van Wieringen and De Mast (2008), Danila et al. (2010), Beavers et al. (2011) and Danila et al. (2012) for recent contributions in quality engineering. The topic is also studied extensively in the behavioral and medical sciences; see Pepe (2003) for a recent overview in the diagnostic sciences.

1.3 Motivation and objective of the thesis

Assessment and improvement of the error rates of binary measurement systems is of great importance for manufacturers. Inspection and testing procedures are omnipresent in industry, and industrial standards for quality control (ISO 9000:2005, ISO 9001:2008, ISO/TS 16949:2009) require studying and controlling the reliability of measurements and test results. Direct cost savings come from avoiding false rejects, as these may lead to unnecessary rework or scrap. Even more important is avoiding false accepts, or defective products reaching the customer. This motivates the need for accurate estimation of the error rates. While binary MSA is relevant in industry, it is in other fields as well. In medicine and psychology, the error rates and their complements (usually named sensitivity and specificity) are important properties of screening and diagnostic tests (Pepe, 2003). The practical ramifications of measurement errors in these fields range from impractical to dramatic.

Over the past years, the statistical evaluation of the reliability of binary measurement systems, inspections and test methods has received increasing attention in industrial statistics and quality engineering. The subject has also been researched extensively in the medical and psychometric world, and has proved to be challenging. It appears in practice that the type of random samples necessary for binary MSA studies is hard to realize, and premises concerning exchangeability and conditional independence are violated in complicated ways. Therefore, for many standard situations still no satisfactory method is available.

The objective of this thesis is to review existing methods and practices for binary MSA, and develop an understanding of their effectiveness and practical guidelines for their use. For situations where currently no satisfactory methods are available, we pursue the design of new methods.

1.4 Outline of the thesis

In Chapter 2, we review methods for assessing the quality of binary measurements. Our framework introduces two factors that are highly relevant in deciding which method to use: (1) whether a reference value (gold standard) can be obtained and (2) whether the underlying measurand is continuous or truly dichotomous. Artificially dichotomizing a continuous measurand, as is commonly done, creates complications that are underappreciated in the

literature and in practice. In particular, it introduces an intrinsic reason for the assumption of conditional independence of measurement values to be violated. For most methods, this is not crucial provided the samples are random (or at least representative). But, also for most methods, it is, in general, not clear how one can obtain a random sample from the relevant population. The taxonomy in Chapter 2 presents methods that are generally known in industry, such as nonparametric estimation of *FAP* and *FRP*, AIAG's analytic method (logistic regression), latent class modeling, and latent trait modeling. The methods discussed are applied to an example presented in AIAG's measurement system analysis manual.

Chapter 3 addresses issues that arise in measurement system analysis of a binary measurement system if the measurand is a hybrid between a dichotomy and a continuum. A case study is presented, which illustrates methods to assess the error rates of binary measurements with such a hybrid measurand. The case study concerns pass/fail inspection of laptop screens for scratches, where the measurand is the presence or absence of scratches. If a scratch is present, the measurand corresponds with a continuum of scratch sizes, but if no scratch is present, the measurand corresponds with a point. It is argued that if the measurand is hybrid, a standard logistic regression model is not suitable to estimate the characteristic curve relating the reject probability to the measurand. Several alternative specifications for the characteristic curve are introduced and compared. We conclude that many of the methods currently used for the assessment of a binary measurement system with a hybrid measurand are unsuited. This is a remarkable conclusion, given the frequent occurrence in industry of leak tests, inspections for defects, and other binary measurement systems with a hybrid measurand.

In Chapter 4, we signal and discuss common methodological errors in agreement studies and the use of κ indices, as found in publications in the medical and behavioral sciences. Our analysis is based on a proposed statistical model that is in line with the typical models employed in metrology and measurement theory. A first cluster of errors is related to nonrandom sampling, which results in a potentially substantial bias in the estimated agreement. Second, when class prevalences are strongly nonuniform, the use of the κ index becomes precarious, as its large partial derivatives result in typically large standard errors of the estimates. In addition, the index reflects rather one-sidedly in such cases the consistency of the most prevalent class, or the class prevalences themselves. A final cluster of errors concerns interpretation pitfalls, which may lead to incorrect conclusions based on agreement studies. These interpretation issues are clarified on the basis of the proposed statistical modeling. The signaled errors are illustrated from actual studies published in prestigious

journals. The analysis results in a number of guidelines and recommendations for agreement studies, including the recommendation to use alternatives to the κ index in certain situations. Reflecting the focus of this chapter on applications in the medical and behavioral sciences, the terminology used in this chapter deviates from the rest of the thesis in that the typical terminology of the diagnostic sciences is adopted (sensitivity, specificity, prevalence, and subjects) rather than quality engineering terminology (*FAP*, *FRP*, defect rate, and items).

Chapter 5 proposes a method for measurement system analysis of a binary measurement system if the measurand is an unobservable continuous variable. The measurand is modeled as a latent trait, and the quality of measurement is expressed in terms of probabilities of inconsistent ordering. Different sampling strategies for the items included in the MSA experiment are explored, aiming to obtain precise estimates of the model parameters. We argue that when the defect rate is low, it is optimal to sample (part of the) items selectively from the subpopulation of rejected items. Also, we present an estimation method for the latent trait model that takes into account this sampling procedure. Based on a simulation study, we investigate the bias and precision of the estimated probabilities of inconsistent ordering for different sampling strategies, the required sample sizes, and the robustness of the approach against model misspecification.

Chapter 6 briefly summarizes the main practical conclusions of the thesis, and formulates the author's vision of the current state of affairs and an outlook to the near future.

The material presented in this thesis has led to a number of papers published in peer-reviewed journals. The review of methods for binary MSA in Chapter 2 has been published in the *Journal of Quality Technology* (De Mast et al., 2011). Chapter 3, about methods for binary inspection with a hybrid measurand is based on a publication in *Quality and Reliability Engineering International* (Erdmann and De Mast, 2012). Chapter 4, about common errors of experimental design, interpretation and inference in agreement studies, will appear in *Statistical Methods in Medical Research* (Erdmann et al., in press). Chapter 5, about binary MSA with a latent continuous measurand, is based on a working paper by Erdmann et al. (2012). Additionally, our research on measurement system analysis has resulted in two publications in *Quality Engineering* (Awad et al., 2009; Erdmann et al., 2009) and a master thesis (Akkerhuis, 2012).

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

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),

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

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 $F_X(x) = P(X \le x)$, and $q(x)$ is typically an *S*-curve such as defined by the logit function,

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

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

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

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

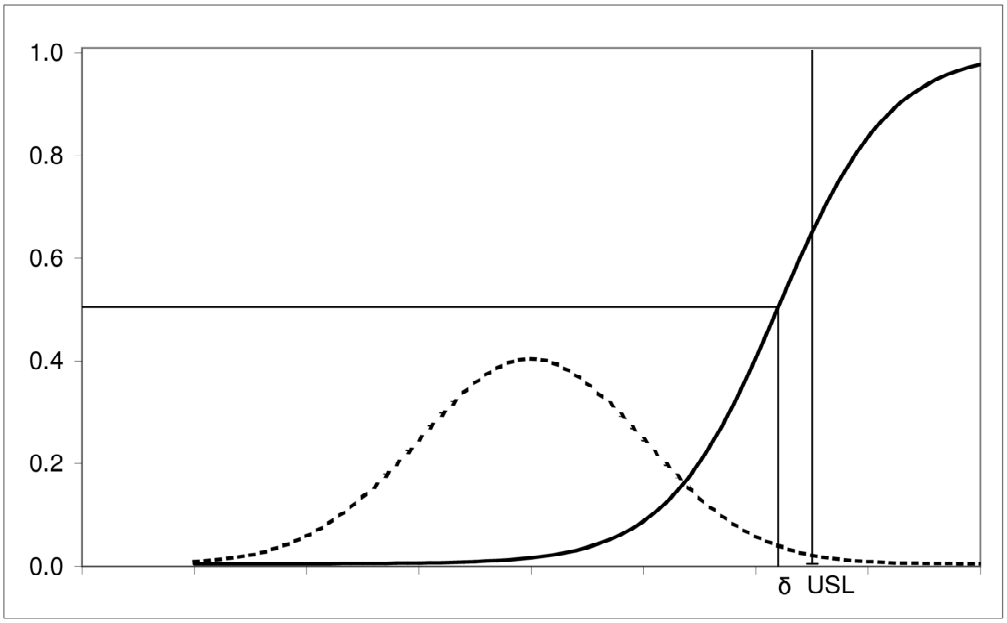

Figure 2.1: Characteristic curve $q(x)=P(Y=0|X=x)$ (solid curve) and density $f_X(x)=dP(X \le x)/dx$ (dashed curve).

observe a finite part of this sequence (the *sample*), namely, $\{Y_{ij}\}$ with $i = 1,...,n$ and $j = 1,...,m$, with n the number of items in the sample, and $m \ge 1$ 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 $P(Y_{ij} = 0 \mid X_i = x_i) = P(Y_{kl} = 0 \mid X_k = x_k)$ 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 $X \le USL$ 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 $\tilde{X}_i$:

$$
\begin{aligned}
P(Y_{i1}=0, Y_{i2}=0 \mid \tilde{X}_i = 0) &= \frac{\int_{-\infty}^{\infty} P(Y_{i1}=0, Y_{i2}=0, \tilde{X}_i = 0 \mid X_i = x) f_X(x)dx}{P(\tilde{X}_i = 0)} \\
&= \frac{\int_{USL}^{\infty} P(Y_{i1}=0, Y_{i2}=0 \mid X_i = x) f_X(x)dx}{\int_{USL}^{\infty} f_X(x)dx} \\
&= \int_{USL}^{\infty} q^2(x) f_X(x)dx \Big/ \int_{USL}^{\infty} f_X(x)dx,
\end{aligned}
\tag{2.2}
$$

which is, in general, not equal to

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

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:

$$
\begin{aligned}
\tilde{q}(0) &= P(Y=0 \mid \tilde{X}=0) = \int_{USL}^{\infty} q(x) f_X(x)dx \Big/ \int_{USL}^{\infty} f_X(x)dx. \\
\tilde{q}(1) &= P(Y=0 \mid \tilde{X}=1) = \int_{-\infty}^{USL} q(x) f_X(x)dx \Big/ \int_{-\infty}^{USL} f_X(x)dx.
\end{aligned}
\tag{2.3}
$$

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, $p(1) = P(X = 0 \mid Y = 1)$. The probabilities of interest are $q(1) = P(Y = 0 \mid X = 1)$ and $q(0) = P(Y = 0 \mid X = 0)$, and we have the rejection rate $q = P(Y = 0)$. The inspection procedure's error rates are given by the false acceptance probability, $FAP = 1 - q(0)$, and the false rejection probability, $FRP = q(1)$. Variants of nonparametric estimation of error rates are common in the diagnostic sciences; see, for instance, Pepe (2003, Chapter 2). Also the AIAG MSA Manual (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 ($\tilde{X} = 0$ 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 $m_{0|0}$, $m_{1|0}$, $m_{0|1}$, and $m_{1|1}$ (where, for example, $m_{1|0}$ is the number of items with $Y = 1$ and $X = 0$). Estimation is straightforward

from sample proportions: $\hat{q}(0) = m_{0|0} / n_0$ and $\hat{q}(1) = m_{0|1} / n_0$; the *FAP* and *FRP* are derived from these values.

We study what happens if the measurand is a false dichotomy, with $\tilde{X} = 1$ 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

$$(2.4) \quad E(\hat{q}(1)) = E(m_{0|1} / n_1) = \int_{-\infty}^{USL} q(x) f_X^s(x) dx \Big/ \int_{-\infty}^{USL} f_X^s(x) dx ,$$

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, $E(\hat{q}(1)) = \tilde{q}(1) = P(Y = 0 \mid \tilde{X} = 1)$, 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^s of X values in the sample might approach a uniform distribution on the interval $[-3,3]$. Such a sampling approach would give an expected result of $E(\widehat{FRP}) = 0.0630$, overestimating the *FRP* by more than a factor of two. Our second example considers a sample consisting mainly of difficult-to-judge parts, which we interpret by taking F_X^s to be a normal distribution with mean 2.5 and standard deviation 0.5. The expected result is $E(\widehat{FRP}) = 0.325$, 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 ($|E(\widehat{FAP}) - FAP| < 0.035$ and $|E(\widehat{FRP}) - FRP| < 0.030$) if the inflection point δ of the inspection procedure's characteristic curve is equal to *USL*. If $\delta \neq USL$, the bias can be arbitrarily large.

Naive sampling with swapping for false dichotomies

A variant of the previous is to take random samples from the streams of accepted and rejected items, but falsely accepted or rejected items are not removed, but added to the other sample.

We refer to this sampling plan as *naive sampling with swapping*. Also under this strategy, the estimated *FAP* and *FRP* are biased, with similar consequences as for naive sampling without swapping; see the Appendix.

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 $m_{0|0}$, $m_{1|0}$, $m_{0|1}$, and $m_{1|1}$. From these, $\hat{p}(0) = m_{0|0} / m_0$ and $\hat{p}(1) = m_{1|0} / m_1$ and, for example,

$$\hat{q}(1) = \frac{\hat{q}(1-\hat{p}(0))}{\hat{q}(1-\hat{p}(0)) + (1-\hat{q})(1-\hat{p}(1))}, \tag{2.5}$$

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 $m_{0|0}/m_0$ and $m_{1|0}/m_1$ are unbiased estimates of $p(0)$ and $p(1)$. Random sampling from the streams of accepted and rejected items will be straightforward in most cases, and we conclude that Plan I is feasible even in the case of a false dichotomy.

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 $m_{0|0}$, $m_{1|0}$, $m_{0|1}$, and $m_{1|1}$. Estimation is done from sample proportions: $\hat{q}(0) = m_{0|0} / n_0$ and $\hat{q}(1) = m_{0|1} / n_1$. If there is additional information about either p or q from historical data, then, in Plan II, the probabilities can be estimated more efficiently using an approach described in Danila et al. (2008).

Equations of the type of Equation (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) $$P(X=1|Y=0)=q(1)(1-p)/(q(1)(1-p)+q(0)p)$$

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 δ are underrepresented. As a consequence, latent class modeling comes across similar problems as the ones discussed for Farnum's set-up in the previous section; in fact, the situation is comparable to the naive sampling with swapping scenario.

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

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, …, X_n are more-or-less equidistant. Each item is to be classified a number m_i of times (with $m_{0|X_i}$ the resulting number of rejects). AIAG gives detailed guidelines for selecting these n items, including the requirements that items 1 and n should be selected extreme enough to ensure that $m_{0|X_1} = 0$ and $m_{0|X_n} = m_i$.

AIAG (2003) suggests 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 σ^{-1} and intercept $-\delta/\sigma$). For each (known) X_i, we have the corresponding observed proportion $\hat{q}(X_i) = m_{0|X_i}/m_i$ as an estimate of $q(X_i)$. From the observed proportions, δ and σ can be estimated. AIAG recommends plotting the $\hat{q}(X_i)$ against the X_i in a normal probability plot and fitting a straight line. The more conventional way to estimate δ and σ is by maximum likelihood, as is standard in logistic regression.

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

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

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

and

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

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

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 σ, δ, 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 δ, the inconsistent acceptance probability (*IAP*) and inconsistent rejection probability (*IRP*):

$$\begin{aligned} IAP &= P(Y=1 \mid X > \delta) = \int_{\delta}^{\infty} (1-q(x)) f_X(x)dx \Big/ \int_{\delta}^{\infty} f_X(x)dx, \\ IRP &= P(Y=0 \mid X \le \delta) = \int_{-\infty}^{\delta} q(x) f_X(x)dx \Big/ \int_{-\infty}^{\delta} f_X(x)dx. \end{aligned} \tag{2.7}$$

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 δ), 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 $0.450 \le X \le 0.545$ 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 $\tilde{X}$ 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 $\tilde{X}$ 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:

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

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

This analysis treats the measurand as dichotomous, but it is in fact continuous, and thus, we are dealing with a false dichotomy. As a consequence, q, $q(0)$ and $q(1)$ are estimated well only if the sample of parts is representative, and in fact, there is some evidence that refutes AIAG's claim that the sample is random. Namely, AIAG states that the process's performance is $P_p = P_{pk} = 0.50$ (where $P_p = (USL - LSL)/6\sigma_X$), suggesting that X has a mean of $\mu_X = 0.4975$ and a standard deviation of $\sigma_X = 0.032$ and that the defect rate is $p = 0.13$. However, the sample defect rate of $\hat{p} = 0.32$ is significantly different (p-value < 0.001) from 0.13, and also the sample's standard deviation $\hat{\sigma}_X = 0.045$ is significantly different from the process standard deviation $\sigma_X = 0.032$ (p-value < 0.001 based on a chi-square test). We

conclude that the given sample, being not representative, and given that the dichotomy is false, is not suited for this analysis. The miss rates (*FAP*) and false alarm rates (*FRP*) per appraiser as given by AIAG (2003, p. 132) may, therefore, be misestimated.

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

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

Note that these estimates were obtained solely from the *Y* values; the reference values were not used in the computations, as they are treated as a latent class. 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 $\tilde{X}$ values is a false dichotomy, and consequently, repeated ratings of the same part are not independent conditional on $\tilde{X}$. 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 δ_L and δ_U are the decision limits that are effectively used by the gauge (as opposed to *LSL* and *USL*, which are the nominal requirements). This characteristic curve is derived from a logistic regression model based on the logit link function for ordinal responses (McCullagh and Nelder, 1989, p.152). The maximum likelihood estimates are $\hat{\delta}_L = 0.453$, $\hat{\delta}_U = 0.547$,

and $\hat{\sigma} = 0.00415$. Because the estimated δ_L and δ_U are close to the *LSL* and *USL*, we conclude that the gauge has negligible bias.

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

$$\begin{aligned}\widehat{FAP} &= \int_{-\infty}^{LSL} (1-\hat{q}(x))\hat{f}_X(x)dx \Big/ \int_{-\infty}^{LSL} \hat{f}_X(x)dx \\ &\quad + \int_{USL}^{\infty} (1-\hat{q}(x))\hat{f}_X(x)dx \Big/ \int_{USL}^{\infty} \hat{f}_X(x)dx = 0.093, \\ \widehat{FRP} &= \int_{LSL}^{USL} \hat{q}(x)\hat{f}_X(x)dx \Big/ \int_{LSL}^{USL} \hat{f}_X(x)dx = 0.051.\end{aligned}$$

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 μ_X and σ_X may be biased. As a consequence, the estimated characteristic curve $q(x)$ represents reliably the behavior of the go/no-go inspections, but the translation into an *FAP* and an *FRP* is affected by the potential bias in $\hat{\mu}_X$ and $\hat{\sigma}_X$. Of course, one could collect a random sample of parts, apply the gold standard, and estimate μ_X and σ_X from the results. Substituting these estimates in the equations above would give estimates for *FAP* and *FRP*.

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

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

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

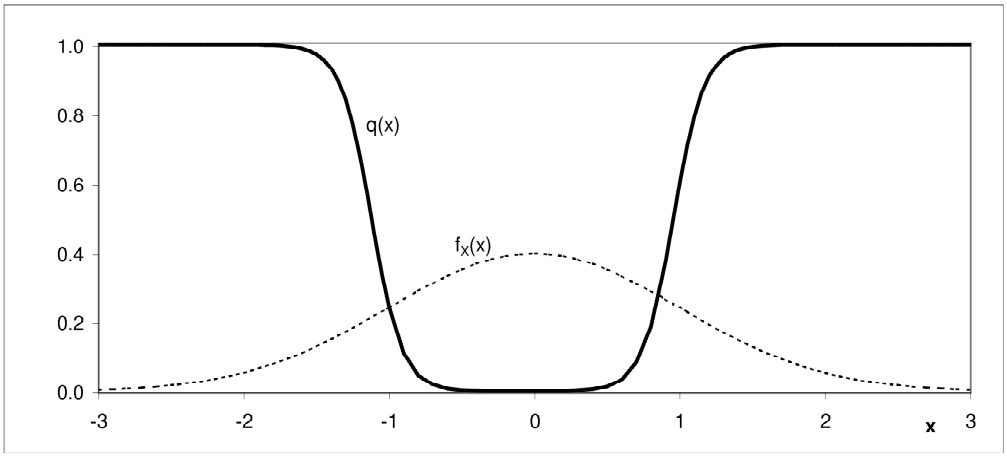

Figure 2.2: Characteristic curve $q(x)=P(Y=0|X=x)$ (solid curve) and density $f_X(x)=\varphi(x)$ (dashed curve).

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

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

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

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 to Y_2). The κ (kappa) statistic is the probability P_a of agreement rescaled such that $\kappa = 0$ corresponds to the probability of agreement achieved by noninformative chance ratings (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.

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

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:

$$\begin{aligned} F_X^0(x) &= P(X \le x \mid \tilde{X} = 0, Y = 0) \\ &= \int_{USL}^{x} \varphi(t) dt \,/\, \int_{USL}^{x} \varphi(t) dt \end{aligned}$$

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

$$\begin{aligned} F_X^{WoS,0}(x) &= P(X \le x \mid \tilde{X} = 0, Y = 0) \\ &= \int_{USL}^{x} q(t)\varphi(t) dt \,/\, \int_{USL}^{\infty} q(t)\varphi(t) dt \end{aligned}$$

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

$$\begin{aligned} E(\widehat{FAP}) &= E(m_{1|0} / n_0) \\ &= \int_{USL}^{\infty} (1 - q(x)) f_X^{WoS,0}(x) dx \ne FAP \end{aligned}$$

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

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

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

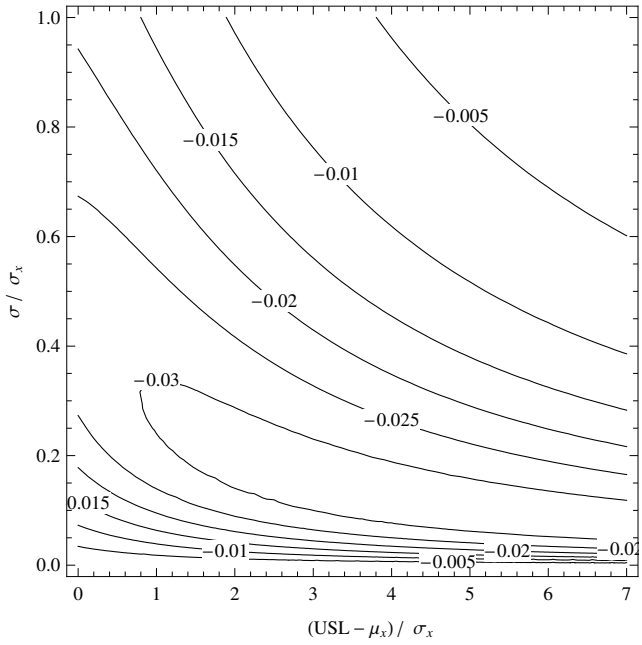

Figure 2.3: Contour plot of the bias $E(\widehat{FAP}) - FAP$ under the *naive sampling without swapping* scenario, and given that $\delta = USL$.

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

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

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

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

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

The distribution in the subsample of defective items is

$$
F_X^{WS,0}(x) = F_{X|\tilde{X}=0}^{WS}(x) = \frac{F_X^{WS}(x) - F_X^{WS}(USL)}{1 - F_X^{WS}(USL)}
$$

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.

3 Assessment of binary inspection with a hybrid measurand

3.1 Introduction

A common industrial application of binary measurement is pass / fail inspection. Items are classified as $Y = R$ ('reject') or $Y = A$ ('accept'), reflecting a measurand X, which is usually a quality characteristic.

As noted in the previous chapters, X can be binary itself, but in many cases, X is a continuous property. Chapter 2 gives an overview of methods to determine the error rates of binary measurements, the *false acceptance probability* (*FAP*) and the *false rejection probability* (*FRP*). It analyzes the complications brought about if a method treats X as binary when it is actually continuous, suggesting that the choice of method should depend on the measurand being binary or continuous. Moreover, when the measurand is continuous, it is often desirable to know the rejection probability for any value of the measurand $X=x$, as this allows one to make statements about measurement reliability without reference to a specific population of items.

In many cases, however, the measurand X is neither binary nor continuous, but a hybrid. A leak test is an example; the outcome $Y = A$ ('accept') reflects the point $X = 0$, whereas the outcome $Y = R$ ('reject') corresponds to a continuum $X > 0$, with X being the size of the leak. We call this a hybrid measurand. Note that, perhaps somewhat confusingly, $X = 0$ now corresponds to good items, in contrast to the previous chapter, where $X = 0$ denotes defective items.

The overview in Chapter 2 shows that there are currently no methods for situations with a hybrid measurand. Treating a hybrid measurand as binary creates the same complications as treating a continuous measurand as binary. In particular, it creates an intrinsic reason for violation of the assumption that (repeated) measurements on items with the same X be i.i.d. (independent and identically distributed) conditional on X. This results in a biased estimate of *FAP* unless defective items are randomly sampled, but it is difficult to obtain a random sample with sufficient defective items. Treating it as continuous, for

example by logistic regression, is more informative, but requires modifications of the used models.

The purpose of this chapter is, based on a somewhat simplified but yet realistic case study, to explore the problem of binary measurement system analysis (MSA) with a hybrid measurand and develop options towards a solution. The case study concerns visual inspection for scratches on a laptop screen. We propose and compare alternative models for the relationship between Y and X, and for the distribution of the measurand, allowing assessment of the quality of measurement. Throughout this chapter we will assume that a gold-standard measurement is available, allowing us to obtain the measurand (or reference value) of all items in the sample. It is often expensive to obtain such a gold-standard measurement, and therefore it is advantageous to use only a small number of items in the MSA experiment.

The next section introduces the case study. In the third section, the problems brought about with standard methods are discussed. Then in the fourth and fifth sections we propose alternative methods and apply them to the example of inspection for scratches. In the sixth section conclusions are drawn.

3.2 Case study: visual inspection for scratches

In order to illustrate the difficulties that arise when the measurand is a hybrid, we present an example of an MSA experiment involving a pass / fail inspection. We have in mind a visual inspection of laptop displays for scratches. A display is good if there is no scratch ($X = 0$), and defective if there are one or more scratches with positive scratch size ($X > 0$). Since we are not in a position to do such an experiment involving real scratches, we designed an experiment that mimics such scratch inspections. Instead of inspecting laptop screens for real scratches, respondents are asked to spot gray curves on otherwise white bitmap images, displayed on a computer screen. All such curves, which are intended to mimic scratches, have the same length and width, but varying grayness and varying location on the screen. We left the situation of multiple scratches out of consideration. The curves' grayness X, measured in percents, mimics the varying depths of real scratches, and we refer to it as scratch size.

The experiment was set up as follows. Twenty appraisers, assumed to be randomly selected from the population of appraisers, each inspected 100 different laptop screens, in random order. They sat in front of the screen with their eyes approximately 80 cm from the

screen, and judged within ten seconds whether a screen had a scratch on it. AIAG (2003) recommends, for its analytic method, to try to obtain a sample of items with more or less equidistant X values, and with that recommendation in mind, we created bitmaps with curves varying in grayness from 10% to 46% in steps of 4% (0% being white, and 100% corresponding to black). Further, these curves' positions vary over 10 locations on the screen / bitmap. The full factorial design in ten levels of grayness and ten locations (100 combinations in total) was divided in two blocks of 50 combinations each. Half of the appraisers inspected the combinations in one block, and the other half the combinations in the other block. In addition, the appraisers inspected 50 samples that had no scratch.

For each value of X, the numbers of rejected and accepted screens are displayed in Table 3.1. A peculiarity of the results is that at $X = 30$ fewer screens were rejected than might be expected. At that level of grayness, fourteen out of the 100 scratches were missed, whereas at $X = 34$ only two were missed. This is visible in Figure 3.1 as a peculiar kink in the curve that would result from connecting the points that give the empirical rejection fraction for each value of X. We have no explanation for this kink, but we know it is not an artifact of the location of the scratches on the screen, because the location on the screen was varied according to the experimental design described above.

Appraiser:			A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
X	***Y***	**Freq.**																				
0	***R***	**18**	1	0	2	0	3	0	0	2	0	1	0	1	2	0	0	4	0	0	0	2
10	***R***	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	***R***	**1**	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
18	***R***	**19**	0	1	1	3	1	1	1	3	1	1	0	0	0	2	1	1	0	0	1	1
22	***R***	**63**	2	2	2	5	1	2	4	3	3	4	1	3	2	5	5	3	5	2	4	5
26	***R***	**84**	3	5	3	5	4	5	5	5	5	5	3	4	2	5	3	4	4	4	5	5
30	***R***	**86**	4	3	1	5	5	4	5	4	4	5	3	5	4	5	5	5	4	5	5	5
34	***R***	**98**	5	5	5	5	5	4	5	5	5	5	4	5	5	5	5	5	5	5	5	5
38	***R***	**98**	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	4	5
42	***R***	**99**	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5
46	***R***	**100**	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
0	***A***	**982**	49	50	48	50	47	50	50	48	50	49	50	49	48	50	50	46	50	50	50	48
10	***A***	**100**	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
14	***A***	**99**	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5
18	***A***	**81**	5	4	4	2	4	4	4	2	4	4	5	5	5	3	4	4	5	5	4	4
22	***A***	**37**	3	3	3	0	4	3	1	2	2	1	4	2	3	0	0	2	0	3	1	0
26	***A***	**16**	2	0	2	0	1	0	0	0	0	0	2	1	3	0	2	1	1	1	0	0
30	***A***	**14**	1	2	4	0	0	1	0	1	1	0	2	0	1	0	0	0	1	0	0	0
34	***A***	**2**	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
38	***A***	**2**	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
42	***A***	**1**	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
46	***A***	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3.1: Data from the MSA experiment for inspection for scratches.

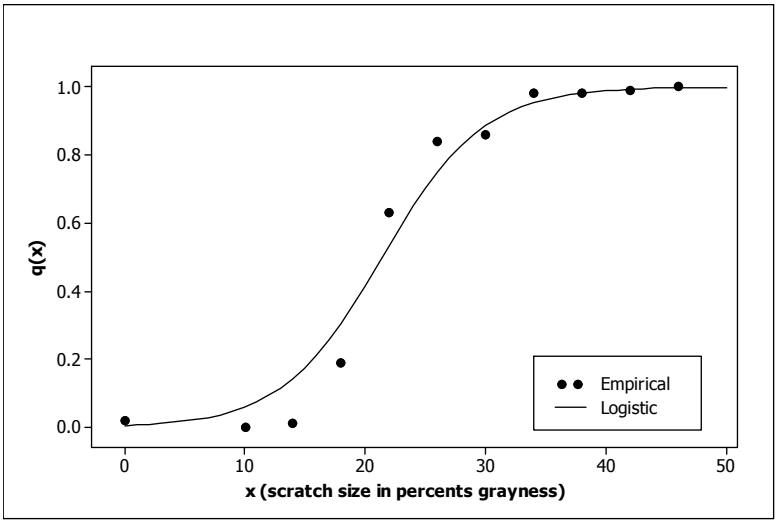

Figure 3.1: Empirical rejection fractions, and plot of the characteristic curve fitted by logistic regression.

3.3 Complications with current methods

In this section we explore complications when we apply standard MSA methods, as presented in Chapter 2, to a case with a hybrid measurand, such as the inspections for scratches. Following the set-up of Chapter 2, we note that the probability of rejecting a screen is a function, named *characteristic curve*, of the scratch size x:

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

Typically, $q(x)$ is an S-curve. Let $F_X^d(x) = P(X \le x \mid X > 0)$ be the distribution of scratch sizes in the subpopulation of defective items. The measurement system could be evaluated in terms of its error rates $FRP = q(0)$ and

$$FAP = \int_{x>0} q(x) f_X^d(x) dx, \tag{3.1}$$

that is, FAP is the average $q(x)$ over the interval $x > 0$, weighted by $f_X^d(x)$. Note that FAP depends on the population of items. In some applications, one needs to specify the measurement quality without reference to the population of items. For example, a manufacturer of leak testers must specify the performance of its products without knowing the population distribution of leaks. In such cases, instead of a characterization in FAP and FRP, the measurement system could better be characterized in terms of FRP and a limit value x° such that $q(x^\circ) = 0.90$ (or some other probability).

Two out of the methods described in Chapter 2 are candidates for analyzing the results in Table 3.1. The first is nonparametric estimation of error rates based on sample proportions, in which the measurand is treated as binary. The measurement system is evaluated in terms of $q(x)$, with $x = 0$ (no scratch) or $x = 1$ (scratch), instead of a continuum of positive x values representing scratch size. To get this approach off the ground, one needs to know the dichotomized measurand (scratch or no scratch) of each of the 100 screens in the sample, and the evaluations by the 20 appraisers. *FRP* and *FAP* are estimated from sample proportions: $\widehat{FRP} = 0.018$ and $\widehat{FAP} = 0.352$, because screens without scratches were incorrectly rejected 18 times (out of 1000 appraisals) and scratched screens were accepted 352 times (out of 1000 appraisals). Unfortunately, this analysis is misguided; since the defective items were not sampled randomly, but selectively at equidistant values of X at 4% increments, the estimate of *FAP* is inconsistent. In particular:

$$E(\widehat{FAP}) = \int_{x>0} q(x) f_X^s(x) dx,$$

where $F_X^s(x)$ is the sampling distribution of X in the subsample of defective items (i.e., the distribution determined by the sampling mechanism and conditional on $X > 0$). Since the defective items were not sampled randomly, but selectively at equidistant values of X, $E(\widehat{FAP}) \neq E(FAP)$ as defined in Equation (3.1), unless $q(x)$ is constant for positive x.

We consider next whether nonparametric estimation of error rates could work if we applied a different experimental set-up, in which items are not sampled at equidistant X values. The literature offers three standard set-ups (cf. Chapter 2; Danila et al., 2008), but these are either difficult to apply or require a large sample size. One set-up is to take random subsamples from the subpopulations of good and defective items separately. The proper way to do this is to use the gold-standard measurement to measure all items that are produced until a sufficiently large amount of defective items is obtained. This way, two subpopulations are created: one of good items and one of defective items. Then, from each of these subpopulations a sample is taken. However, if the defect rate is low, as is usually the case, it takes an unrealistically large effort to create such subpopulations. For example, if the defect rate is 1% and one wants to obtain subsamples of 30 items, thousands of items will need to be measured by the gold-standard measurement procedure. Plan I by Danila et al. (2008) solves this problem by sampling from the streams of rejected and accepted items instead. The measurand of the sampled items is determined by applying the gold-standard measurement. Then, one calculates the proportion of rejected items that were actually good, and the

proportion of accepted items that were actually defective. If one also knows the historical reject rate, one can calculate the error rates of the measurement system by applying Bayes' Law. However, Plan I only uses the original evaluation as accepted or rejected; it does not allow for repeated measurements of the same item and therefore it requires large sample sizes of, say, 300 items per subsample. Another approach (Plan II) is to sample directly from the population of all items, but, in case of a low defect rate, this requires an unrealistically large sample size in order to include sufficient defective items. Again, thousands of items need to be measured by the gold-standard measurement. Of these three standard set-ups, only Plan I is applicable if the defect rate is low.

Besides the sampling problems discussed above, a second disadvantage of the nonparametric approach based on sample proportions, is that it is less informative than logistic regression. It does not give the complete characteristic curve $q(x)$, but only the proportions *FRP* and *FAP*. The latter depends on the population of items per Equation (3.1). Therefore, in applications where the population of items (and the distribution of the measurand therein) is not fixed, the nonparametric approach is not applicable. In such applications one is typically interested in the limit value x°.

In summary: nonparametric estimation of error rates, treating the measurand as binary, is unsatisfactory for situations involving a hybrid measurand, as sampling schemes facilitating unbiased estimation are difficult to apply or require large sample sizes, and the reduction of the characteristic curve to an *FAP* and *FRP* value is an unsatisfactory summary of the stochastic behavior.

The other candidate method offered in Chapter 2 is logistic regression, which treats the measurand as continuous. One specific variant of this approach is AIAG's analytic method (AIAG, 2003). It estimates a parametric model $q^{\boldsymbol{\theta}}(x)$ for the characteristic curve, with parameter vector $\boldsymbol{\theta}$, from a sample obtained by selecting a number of items such that their measurands are more-or-less equidistant. Each item j is classified m_j times (AIAG, 2003, recommends $m_j = 20$ for all j). The number of rejects for covariate value x_j is r_j. The parameter vector $\boldsymbol{\theta}$ is then estimated by maximizing the log-likelihood function:

$$l(\boldsymbol{\theta}) = \sum_j \left(r_j \log q^{\boldsymbol{\theta}}(x_j) + (m - r_j) \log(1 - q^{\boldsymbol{\theta}}(x_j)) \right).$$

The regular logistic regression approach takes $x \in \mathbb{R}$, and this is also the treatment in AIAG's analytic method. We explore whether we can apply this approach to a case where $x \in \mathbb{R}_+$ (hybrid measurand) such as the MSA study for inspection for scratches, but we will

see that the approach requires more than straightforward modifications to be applicable in such situations.

In order to characterize the measurement system, *FRP* can be estimated as $q^{\hat{\boldsymbol{\theta}}}(0)$, but *FAP* is more difficult to estimate. Departing from Equation (3.1), for $q^{\boldsymbol{\theta}}$ one substitutes $q^{\hat{\boldsymbol{\theta}}}$, and for F_X^d one fits a parametric distribution function $\hat{F}_X^d$, whose parameters are estimated from a random sample of defective items. Natural choices for $\hat{F}_X^d$ are nonnegative distributions such as the exponential, lognormal, and Weibull distributions. Note that it is difficult to obtain a random sample of defective items. It cannot be taken randomly from the rejected items because if the defect rate is low, a substantial part of the rejected items will actually be good, and difficult-to-judge defective items will be underrepresented, as shown in Chapter 2. A possible solution is to select items randomly from the population of all items, apply a gold-standard measurement to determine whether each item is good or defective, and continue until a sufficient number of defective items are obtained. If the defect rate is low, an unrealistically large number of items need to be measured with the gold-standard measurement, for the same reasons as described above in the context of standard set-ups for nonparametric estimation of error rates based on sample proportions. As noted in the previous section, one can avoid estimating F_X^d altogether if one is not interested in *FAP* and instead evaluates the measurement system in terms of a limit value x°.

The standard logistic regression model for $q(x)$ is the distribution function of a standard normal or a logistic distribution with a linear function of x as its argument:

(3.2) $q^{\text{Prob}}(x) = \Phi(\alpha + \beta x)$, or

(3.3) $q^{\text{Log}}(x) = \dfrac{1}{1 + e^{-(\alpha+\beta x)}}$,

both with $\beta > 0$.

These standard models are not appropriate if the measurand is a hybrid, as in the scratches inspection example. The particular forms of q^{Prob} and q^{Log} imply point symmetry in their inflection point, and $q^{\text{Prob}}(0)$ and $q^{\text{Log}}(0)$ are fixed given the location of and slope at the inflection point; these properties will be shown to be unsuited for situations with a hybrid measurand.

We illustrate these problems by applying standard logistic regression, based on model (3.3) and with the restriction $x \geq 0$, to the example. To simplify matters, we treat the appraisers as interchangeable, and treat the results of the 20 appraisers as replications. The

Parameter	Estimate	Std. error	P-value
α	-5.153	0.267	0.0000
β	0.2403	0.012	0.0000
Log-likelihood	-355.5		

Goodness-of-fit tests

Method	Chi-square	D.f.	P-value
Pearson	64.09	9	0.0000
Deviance	71.13	9	0.0000

Table 3.2: Parametric fit of characteristic curve by logistic regression, and goodness-of-fit tests.

maximum likelihood estimates are $\hat{\alpha} = 5.15$ and $\hat{\beta} = 0.24$. The estimation results are summarized in Table 3.2, and graphically displayed in Figure 3.1. For each value of x, the empirical rejection fraction is also displayed.

To assess the goodness of fit we compute the chi-square statistics based on Pearson and deviance residuals (Hosmer and Lemeshow, 1989). The Pearson residual u_j of the jth covariate value x_j equals

$$u_j = \frac{r_j - m_j \hat{q}(x_j)}{\sqrt{m_j \hat{q}(x_j)(1-\hat{q}(x_j))}},$$

where m_j is the total number of classifications with covariate value x_j of which r_j is the number of rejections, and $\hat{q}(x_j)$ is the estimated reject probability. The squared deviance residual d_j^2 equals

$$d_j^2 = 2\left(r_j \log\left(\frac{r_j}{m_j \hat{q}(x_j)} \right) + (m_j - r_j) \log\left(\frac{m_j - r_j}{m_j (1-\hat{q}(x_j))} \right) \right).$$

The chi-square statistics are obtained by summing the corresponding version of squared residuals over all covariate values. They asymptotically follow a chi-square distribution with degrees of freedom equal to $J - k$, where J is the number of different covariate patterns and k is the number of parameters in the model. Both goodness-of-fit tests indicate that the true reject probabilities are significantly different from the estimated ones (p-value = 0.000 for both tests). The model specification is clearly incorrect.

Further evidence of the inadequacy of the model is obtained by plotting *delta Pearson chi-square* against the reject probability $q(x)$ for each scratch size. Delta Pearson chi-square is a measure for the change in Pearson chi-square that results if a certain covariate pattern (scratch size) is left out of the data set. A plot of delta chi-square against the reject probability is shown in Figure 3.2.

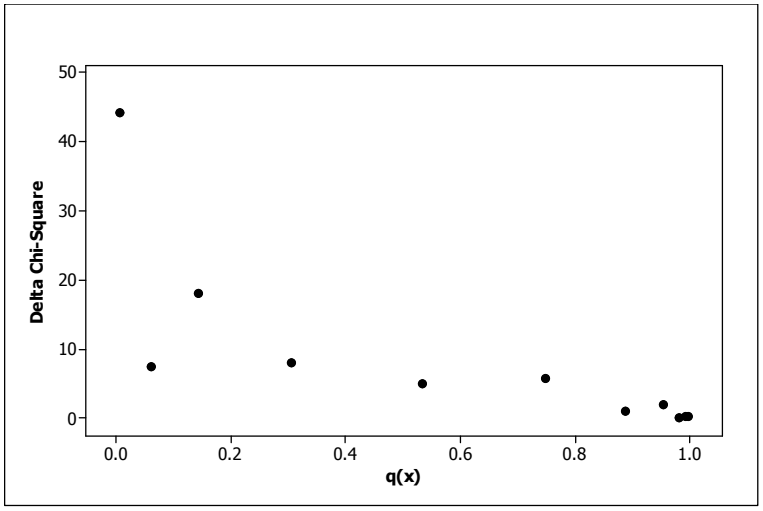

Figure 3.2: Plot of delta Pearson chi-square against the reject probability for each covariate pattern (scratch size).

A high delta chi-square indicates that for a certain scratch size the percentage of screens rejected deviates substantially from the estimated reject probability. Hosmer and Lemeshow (1989) state that if the model fits well, the upper 95% percentile of the delta chi-square is approximately 4. In our case most of the scratch sizes have delta chi-square values higher than 4, indicating a very poor fit. Especially for the covariate pattern $x = 0$ (screens with no scratch), delta chi-square is extremely high (44.3). The empirical rejection fraction of good screens was 0.018 in the experiment, whereas it is estimated as $\widehat{FRP} = 0.006$. The value of $\widehat{FRP}$ estimated by this model is too close to zero, because $\hat{q}(0)$ is fixed by the location and slope at the curve's inflection point. In particular, the distance of the inflection point from the *y*-axis, in combination with the steepness of the slope at the inflection point, force $\hat{q}(0)$ down toward the *x*-axis.

Clearly, both nonparametric estimation of error rates and the standard logistic regression model are unsuited for the analysis of our example of inspections for scratches. In the next section more adequate methods are proposed to model the characteristic curve $q(x)$.

3.4 Parametric solutions

We aim to modify the logistic regression model to overcome the problems described in the previous section, by exploring alternative functions for the characteristic curve $q(x)$. In order to identify and evaluate options for $q(x)$, we reflect on the physical mechanisms that

determine the stochastics of the inspections. Firstly, there is the possibility that an appraiser imagines seeing a scratch that is not actually there. Secondly, given that there is a scratch, the randomness in the inspection results is induced by such mechanisms as:

- Differences in eye sight between appraisers, and fluctuations in a single appraiser's eye sight over time, including fluctuations in concentration and motivation.
- Differences in circumstances, such as ambient light, angle of vision, and distance to the screen.
- The randomness of the search process, where the appraiser's eyes are scanning the screen, trying to locate a possible scratch in a limited amount of time. As a matter of fact, even a relatively deep scratch may be missed because the appraiser, in the restricted amount of time given, happens not to scan the area where it is located.

We believe that for values beyond the characteristic curve's inflection point, scratch size is progressively less influential in the randomness of the inspection results — the probability of detecting a scratch is almost completely determined by the random outcome of the search for its location, and increases in scratch size have virtually no further effect. This is one of the motivations to consider mathematical functions for $q(x)$ that are flexible in the degree of asymmetry that they can represent.

In view of these physical mechanisms, ideally, we have a mathematical function $q(x)$ with sufficient degrees of freedom to represent these four aspects:

1) The probability $q(0)$ that a good screen is rejected.
2) The location of the characteristic curve's inflection point $x^* = \{x : q''(x) = 0\}$.
3) The slope $q'(x^*)$ at the inflection point.
4) The height of the inflection point $q(x^*)$, determining the amount of symmetry of the curve.

A natural general form for $q(x)$ is:

$$q(x) = q(0) + (1 - q(0))G(x),$$

with G a nonnegative distribution function. We will refer to this form as a *zero-inflated* nonnegative distribution. That is, a nonnegative distribution rescaled such that x=0 with positive probability. For the logistic distribution in Equation (3.3), the location, height and slope at the inflection point are

$$x^* = -\frac{\alpha}{\beta},\ G(x^*) = \frac{1}{2},\ g(x^*) = \frac{\beta}{4};$$

where g is the derivative of G. The height of the inflection point is fixed at $G(x^*) = \frac{1}{2}$ and the curve is symmetrical, which makes this an unsuitable option for our purpose, as noted earlier. We discuss alternative options.

The Weibull distribution is a flexible nonnegative distribution. Its distribution function is given by

$$G(x) = 1 - e^{-(x/\beta)^{\alpha}}$$

with $\alpha, \beta > 0$. The parameter α is called the shape parameter and β the scale parameter. The distribution is positively skewed for $0 < \alpha < 3.6$. The location, height, and slope at the inflection point are:

$$x^* = \beta\left(\frac{\alpha-1}{\alpha}\right)^{\frac{1}{\alpha}},\ G(x^*) = 1 - e^{\frac{1-\alpha}{\alpha}},\ g(x^*) = \frac{\alpha-1}{\beta}\left(\frac{\alpha}{\alpha-1}\right)^{\frac{1}{\alpha}} e^{\frac{1-\alpha}{\alpha}}.$$

A disadvantage of the Weibull distribution for our purposes is that the coordinates of the inflection point fix the slope $g(x^*)$, through the relation:

$$g(x^*) = \frac{(G(x^*)-1)\log((1-G(x^*))}{x^*(1+\log(1-G(x^*)))}.$$

For positively skewed Weibull distributions ($\alpha < 3.6$), $g(x^*)$ is bounded above by $1.26/x^*$. The empirical rejection fractions of the scratch inspections, however, suggest a positively skewed distribution with x^* around 20 and $g(x^*)$ around 0.11, properties unobtainable by the Weibull function.

A function that is capable of fitting these results is the log-logistic distribution, with the following distribution function:

$$G(x) = \frac{1}{1+(x/\beta)^{-\alpha}}$$

with $\alpha, \beta > 0$. The log-logistic distribution is skewed to the right for all parameter values, although technically its third moment only exists for $\alpha > 3$. The location, height, and slope at the inflection point are:

$$x^* = \beta\left(\frac{\alpha-1}{\alpha+1}\right)^{\frac{1}{\alpha}},\ G(x^*) = \frac{\alpha-1}{2\alpha},\ g(x^*) = \frac{\alpha^2-1}{4\alpha\beta}\left(\frac{\alpha+1}{\alpha-1}\right)^{\frac{1}{\alpha}}.$$

Again, the slope is fixed by the coordinates of the inflection point:

$$g(x^*) = \frac{G(x^*)(1-G(x^*))}{x^*(1-2G(x^*))}.$$

The log-logistic distribution ascends more steeply at the inflection point than the Weibull for any given coordinate pair of the inflection point obtainable by both distributions (i.e. $0 < G(x^*) < 0.5$.

A distribution that has an additional parameter is the generalized logistic distribution (Zelterman, 1987). This distribution is less commonly used than the distributions discussed so far. It has the distribution function:

$$G(x) = \frac{1}{\left(1+e^{-(\alpha+\beta x)}\right)^{\gamma}}$$

with β, $\gamma > 0$. For $\gamma = 1$ this distribution is equal to the logistic distribution as in Equation (3.3). For $\gamma > 1$, the distribution is positively skewed. The location, height, and slope at the inflection point are:

$$x^* = \frac{\log(\gamma)-\alpha}{\beta},\ G(x^*) = \frac{1}{(1+\gamma^{-1})^{\gamma}},\ g(x^*) = \frac{\beta}{(1+\gamma^{-1})^{\gamma+1}}.$$

The three parameters allow the distribution to represent all three aspects separately. A disadvantage of the generalized logistic distribution is that the height of its inflection point $G(x^*)$ cannot be less than $e^{-1} \approx 0.37$. Furthermore, if γ is large such that $G(x^*) \approx e^{-1}$ and $g(x^*) \approx \beta e^{-1}$, the distribution has a nearly identical shape for any combination of α and γ that keeps the value of x^* constant. This creates problems regarding identification of γ and α if the height of the inflection point is close to e^{-1} or less.

A three-parameter distribution that is less restrictive for the height of the inflection point is the generalized extreme value distribution introduced by Jenkinson (1955). Its distribution function is:

$$G(x) = e^{-(1+\gamma(\alpha+\beta x))^{-1/\gamma}}$$

for $1 + \gamma(\alpha + \beta x) > 0$ with $\beta > 0$. Special cases of this distribution are the Gumbel (limit for $\gamma \to 0$), Frechet ($\gamma > 0$) and reversed Weibull ($\gamma < 0$) distributions. The location, height, and slope at the inflection point are:

$$x^* = \frac{(\gamma+1)^{-\gamma}-\alpha\gamma-1}{\beta\gamma},\ G(x^*) = e^{-\gamma-1},\ g(x^*) = \beta(\gamma+1)^{\gamma+1}e^{-\gamma-1}.$$

Model:	Logistic	Weibull	Log-Log.	Gen Log	G.E.V.	Tr. Weib.
Parameter						
$q(0)$	0.01465	0.01534	0.01531	0.01548	0.01579	0.01583
α	-7.278	3.997	7.744	11.129	-5.854	1.011
β	0.3285	24.661	21.600	0.2475	0.2973	5.4145
γ	-	-	-	9500000	0.1637	16.9086
Log-likelihood	-338.7	-348.4	-329.8	-327.4	-325.4	-324.7
Goodness-of-fit test based on Pearson residuals						
Chi-square	36.60	86.98	17.71	13.82	9.45	7.81
D.f.	8	8	8	7	7	7
P-value	0.0000	0.0000	0.0235	0.0544	0.2219	0.3498
Goodness-of-fit test based on deviances						
Chi-square	37.49	56.93	19.78	14.76	11.00	9.52
D.f.	8	8	8	7	7	7
P-value	0.0000	0.0000	0.0112	0.0391	0.1388	0.2175
Aspects of inflection point						
x^*	22.15	22.95	20.89	19.95	19.19	16.97
$q(x^*)$	0.5073	0.5348	0.4441	0.3777	0.3232	0.0263
$q'(x^*)$	0.08092	0.06076	0.08974	0.08963	0.10901	0.17319

Table 3.3: Parametric fit of $q(x)$, goodness-of-fit tests, and aspects of the inflection point for six specifications for $q(x)$: the zero-inflated logistic, Weibull, log-logistic, generalized logistic, generalized extreme value, and translated Weibull distributions.

By changing γ, the inflection point can obtain any height between 0 and 1. For $\gamma > 0$ it is lower than e^{-1}, the minimum height of the inflection point for the generalized logistic distribution.

The last option that we consider is the translated Weibull distribution with a third parameter γ that determines the horizontal translation. Both the translated Weibull and generalized extreme value distributions have a discontinuity in one of their (higher order) derivatives at the point where they start increasing, $x = \gamma$ for translated Weibull and $x = -\frac{\alpha\gamma + 1}{\beta\gamma}$ for the generalized extreme value distribution, although it may occur for negative x and thus be outside the domain of $q(x)$. In the scratch example there is no physical explanation for such a discontinuity.

Table 3.3 gives the maximum likelihood estimation results for the six different specifications of $q(x)$ proposed: the zero-inflated logistic, Weibull, log-logistic, generalized logistic, generalized extreme value, and translated Weibull distributions. The resulting fitted characteristic curves are shown in Figure 3.3. The figures also display the empirical rejection fractions for comparison.

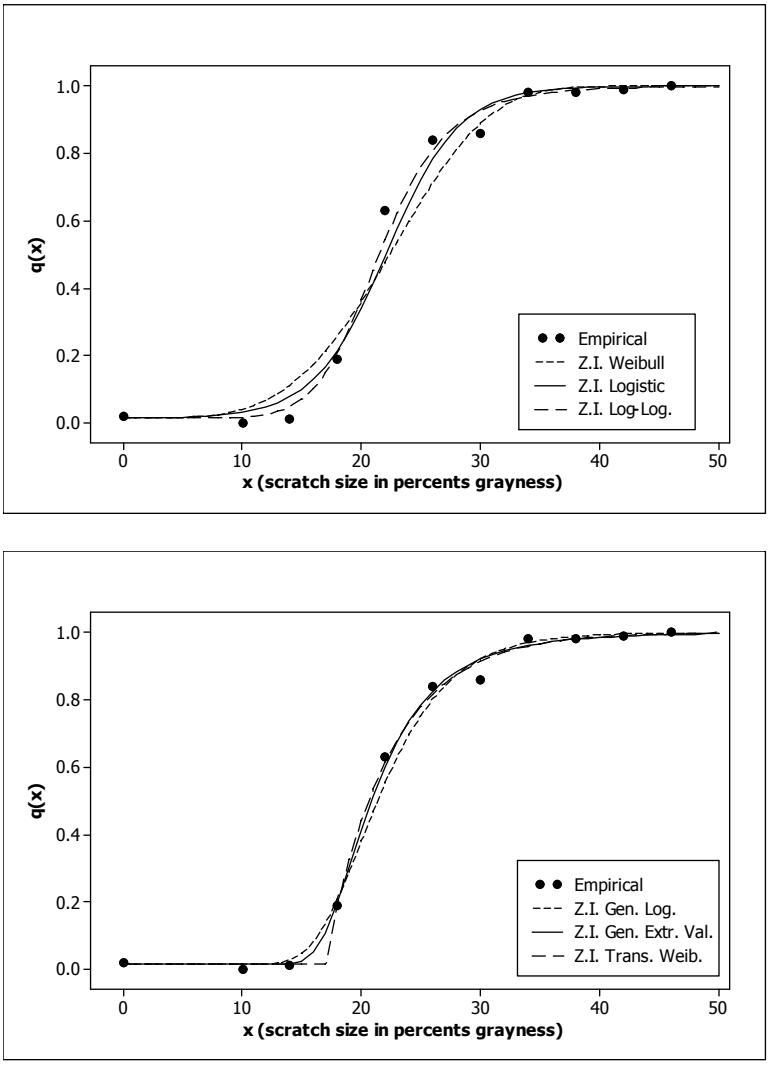

Figure 3.3: Plots of the characteristic curve for six specifications for $q(x)$: the zero-inflated logistic, Weibull, log-logistic, generalized logistic, generalized extreme value, and translated Weibull distributions.

The curves resulting from the zero-inflated logistic, Weibull and log-logistic models do not seem to adequately model the asymmetry in $q(x)$ and the steep slope at its inflection point. This is confirmed by the goodness-of-fit tests. Both the chi-square value based on Pearson residuals and the chi-square value based on deviances indicate significant lack of fit for all three distributions. The zero-inflated Weibull distribution results in an even worse fit than the zero-inflated logistic distribution.

The curve resulting from the zero-inflated generalized logistic model is much closer to the empirical rejection fractions. However, because the height of the inflection point is close to $q(0)+(1-q(0))e^{-1}$, the parameters of this model are hard to identify by maximum

likelihood due to the near-unidentifiability problem described in the previous section. After trying several different starting values for the parameter estimates, the highest log-likelihood that we were able to obtain results in the estimates in the table. Despite the estimation problem, the zero-inflated generalized logistic distribution model gives a good fit. It has a high log-likelihood and low goodness-of-fit chi-square statistics. The goodness-of-fit chi-square value based on Pearson residuals does not indicate significant lack of fit at the 5% significance level, but the one based on deviances does.

The generalized extreme value distribution and the translated Weibull distribution give the best fit. They have the highest log-likelihoods and the lowest goodness-of-fit chi-square statistics. Both tests do not reject the fit. The translated Weibull distribution fits slightly better according to these measures, but has some questionable properties. The characteristic curve fitted with the translated Weibull model has a kink at $x = 16.9$, where the second order derivative is discontinuous and jumps from zero to infinity. We cannot think of a physical explanation for such discontinuity, and therefore see it as an undesirable estimation artifact. Furthermore, the coordinates of the inflection point of the Weibull differ substantially from those of all other models: they are located near this discontinuity at $\hat{x}^* = 17.0$ and $\hat{q}(\hat{x}^*) = 0.026$. The fitted generalized extreme value distribution does not have a discontinuity in its derivatives for positive x, and therefore seems the better choice. It leads to a characteristic curve $\hat{q}(x)$ with $\hat{q}(0) = 0.0158$, $\hat{x}^* = 19.2$, $\hat{q}(\hat{x}^*) = 0.323$ and $\hat{q}'(\hat{x}^*) = 0.109$. *FRP* is estimated as $\hat{q}(0) = 0.0158$ and if the distribution of scratch sizes in the population of defective items $F_X^d(x)$ were known, *FAP* could be estimated by substituting the fitted $q(x)$ in Equation (3.1). Note that if the defect rate is low, it is practically very difficult – if not impossible – to estimate $\hat{F}_X^d(x)$ reliably, because of the sampling problems explained in the previous section. Note that in our (artificial) experiment *FAP* cannot be evaluated, because there is no well-defined population of items. The limit value x° such that $q(x^\circ) = 0.90$ is estimated as $\hat{x}^\circ = 28.8$, so scratches with scratch size larger than 28.8% (expressed as percent grayness) are detected with more than 90% probability.

3.5 Nonparametric solutions

If no parametric model for $q(x)$ fits well, the curve could be estimated by nonparametric logistic regression (Hastie and Tibshirani, 1986 and 1990). The logistic regression model is generalized to

$$q(x) = \frac{1}{1+e^{-s(x)}}, \tag{3.4}$$

where s is a smooth function, such as a cubic spline. A cubic spline is a twice differentiable continuous function consisting of piecewise cubic polynomials, whose pieces are separated by a sequence of breakpoints called *knots*. A useful function in the software pakage *S* to perform nonparametric logistic regression is the function *gam* (cf. Hastie and Tibshirani, 1990).

Hastie and Tibshirani (1986 and 1990) fit s using a procedure that they call the *local scoring* algorithm, because it is based on the Fisher scoring procedure used to find the maximum likelihood estimates in linear logistic regression. They start with initial values of $s(x_i)$ for all observations i, and then compute adjusted values z_i from the formula

$$z_i = s_{old}(x_i) + \frac{y_i - q_{old}(x_i)}{q_{old}(x_i)(1-q_{old}(x_i))}.$$

The new function s_{new} is obtained by fitting z to x using a scatterplot smoother and

$$q_{new}(x) = \frac{1}{1+e^{-s_{new}(x)}}.$$

Then in the next iteration s_{new} and q_{new} become s_{old} and q_{old}. This procedure is repeated until the change in deviance is less than a specified amount.

The scatterplot smoother that we will use is a cubic smoothing spline. The cubic smoothing spline that fits z to x is defined as the function s that minimizes the penalized residual sum of squares

$$RSS(s,\lambda) = \sum_i \left(z_i - s(x_i)\right)^2 - \frac{1}{2}\lambda \int (s''(t))^2 dt,$$

where λ is the smoothing parameter determining the degree of smoothing. It can be shown that the solution to this minimization problem is a natural cubic spline with knots at all distinct values of x_i. That is, it is a cubic polynomial in each interval (x_i, x_{i+1}), its first two derivatives are continuous, and it is linear below x_1 and above x_n. Numerical algorithms to find fitted values of s are given in De Boor (1978).

Effective d.f.:	**4**	**5**
x	***q*(*x*)**	***q*(*x*)**
0	0.0149	0.0166
10	0.0282	0.0147
14	0.0668	0.0407
18	0.2145	0.1929
22	0.5428	0.5832
26	0.8015	0.8196
30	0.9150	0.9082
34	0.9647	0.9602
38	0.9855	0.9835
42	0.9941	0.9934
46	0.9976	0.9975
Log-likelihood	-331.8	-326.0
Goodness-of-fit test based on Pearson residuals		
Chi-square	18.37	9.53
D.f.	7	6
P-value	0.0104	0.1461
Goodness-of-fit test based on deviances		
Chi-square	23.70	12.12
D.f.	7	6
P-value	0.0013	0.0594
Aspects of inflection point (approximations)		
*x**	20.96	20.45
q(x)*	0.451	0.425
q'(x)*	0.090	0.107

Table 3.4: Fit of $q(x)$ by nonparametric logistic regression using smoothing splines with four and five effective degrees of freedom, goodness-of-fit tests, and (numerical approximations of) aspects of the inflection point.

Hastie and Tibshirani (1990) use the notion of *effective degrees of freedom* of a smoother in order to be able to compare them with parametric models. The fitted values of a smoothing spline at the observed points x_i can be written as linear combinations of the values of the dependent variable z_i:

$$\hat{\mathbf{s}} = \mathbf{S}\mathbf{z},$$

where the matrix $\mathbf{S}$ depends only on the x_i and on λ, and is called the *smoother matrix*. The number of effective degrees of freedom is calculated as the trace of $\mathbf{S}$. This is because of the analogy of $\mathbf{S}$ to the projection matrix in linear regression analysis that transforms the observed values of the dependent variable into the fitted values, often called the *hat matrix*. The effective degrees of freedom are determined by the smoothing parameter λ. If $\lambda = 0$ the smoothing spline simply interpolates the data and the number of degrees of freedom of the

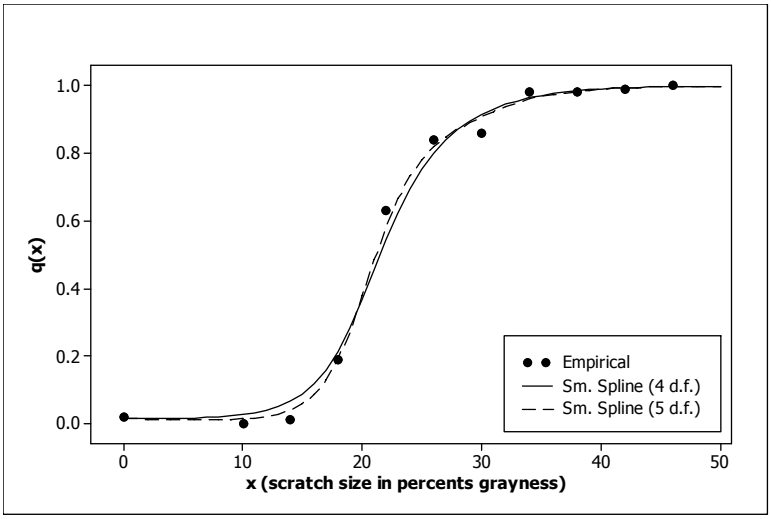

Figure 3.4: Plots of the characteristic curve fitted by nonparametric logistic regression using smoothing splines with four and five effective degrees of freedom.

curve equals the number of distinct values of x. If λ approaches infinity, the smoothing spline approaches the ordinary least squares line, and the effective degrees of freedom approach two. The smoothing parameter λ can be chosen to fix the degrees of freedom to a certain number. A common criterion is to choose it such that the cross-validation sum of squares is minimized (cf. Hastie and Tibshirani, 1990).

An advantage of nonparametric logistic regression is that it can be applied generally, in contrast to the parametric models discussed in the previous section, which were chosen specifically for the dataset in this example. A disadvantage is that the resulting curve is defined implicitly as the solution of a minimization problem and no explicit function of x is obtained (cf. Silverman, 1985).

Nonparametric logistic regression using a smoothing spline leads to the fitted values of $q(x)$ given in Table 3.4. Two splines are fitted with four and five effective degrees of freedom, respectively. The characteristic curves are plotted in Figure 3.4. The spline with four degrees of freedom does not fit well, as indicated by both chi-square statistics. The spline with five degrees of freedom gives a good fit. Its log-likelihood and chi-square statistics are comparable to the generalized extreme value distribution, although it uses one (effective) degree of freedom more. Both chi-square statistics do not reject a good fit at the 5% level. The approximate characteristics of the inflection point are $\hat{x}^* = 20.5$, $\hat{q}(\hat{x}^*) = 0.425$, and $\hat{q}'(\hat{x}^*) = 0.107$, indicating that the inflection point is located slightly higher and more to the right than for the parametric distributions with the best fit. The

estimated *FRP* is $\hat{q}(0) = 0.0166$. Again, *FAP* cannot be evaluated for our experiment because we do not have a relevant population of defective items. The fitted $\hat{q}(\hat{x}^\circ) = 0.90$ for $\hat{x}^\circ = 29.5$, reasonably close to the value of x° based on the zero-inflated generalized extreme value distribution.

3.6 Summary and conclusions

The case study discussed in this chapter shows that assessing a binary measurement system with a hybrid measurand is a difficult task. Even common binary inspections such as a leak test or an inspection for scratches are hard to evaluate. If *FRP* and *FAP* are determined using sample proportions of incorrectly rejected and incorrectly accepted items, the estimate of *FAP* critically depends on the randomness of the sample, but a sufficiently large random sample of defective items is practically difficult to obtain if the defect rate is low. The approach elaborated in this chapter is based on estimation of the characteristic curve $q(x)$, representing the rejection probability as a function of the measurand x, for $x \geq 0$. Because standard logistic regression models are not appropriate for a nonnegative measurand such as scratch size, an appropriate asymmetric function for $q(x)$ is fitted to the data. In the case study discussed in this chapter three specifications for $q(x)$ are found to model the specific shape of $q(x)$ reasonably well. They all have four parameters, and consequently fitting them requires a large sample size. Another possible approach, which is more generally applicable, is nonparametric logistic regression, but the disadvantage is that it does not give an explicitly defined function for $q(x)$. Once the function $q(x)$ has been estimated, *FRP* and a value x° such that $q(x^\circ) = 0.90$ could be used to evaluate the measurement system. If one is interested in *FAP*, the distribution $F_X^d(x)$ of the measurand in the population of defective items needs to be estimated as well, again requiring a random sample of defective items, which is difficult to obtain.

Estimating $q(x)$ and $F_X^d(x)$ becomes even more difficult if the measurand is not a one-dimensional property such as grayness, but a multidimensional set of properties, such as for example: grayness, length of the scratch, number of scratches, etc. Reliably estimating *FAP*, *FRP*, or x° in such cases is a formidable challenge.

We conclude that the methods currently used for assessment of a binary measurement system with a hybrid measurand are often unsuited. This is a remarkable conclusion, given the frequent occurrence in industry of leak tests, inspections for defects, and other binary measurement systems with a hybrid measurand. Methods aimed at expressing *FAP* and *FRP* will necessarily be bound to a certain study population of items, and often require unrealistic sampling plans. In order to correctly assess the quality of measurements, one either needs to:

1) apply nonparametric estimation using sample proportions based on subpopulations of accepted/rejected rather than good/defective items, as is done in Plan I by Danila et al. (2008), or
2) estimate the curve $q(x)$ rather than *FAP*. The curve $q(x)$ could be estimated based on repeated measurements at a number of values of x, using an appropriate zero-inflated distribution function as a model, or using nonparametric logistic regression.

4 Some common errors of experimental design, interpretation and inference in agreement studies

4.1 Introduction

Diagnostic tests, clinical diagnoses, and ratings can be perceived as measurements on a nominal scale. They classify subjects into a set of unordered categories, aiming to reflect an empirical property of the subjects, the measurand, that is not observed directly.

The quality or reliability of nominal measurements is often expressed in terms of agreement, typically in the form of a κ (kappa) index. Introduced by Cohen (1960), it is a measure of agreement between repeated classifications that corrects for agreement 'by chance', that is, for agreement achieved by blind classifications (Fleiss, 1971; Conger, 1980; Davies and Fleiss, 1982). The κ index is surrounded by quite some controversy, and a number of papers have identified paradoxical behavior (Feinstein and Cicchetti 1990; Byrt et al., 1993; Thompson and Walter, 1988; Uebersax, 1987; Grove et al., 1981; De Mast, 2007; Warrens, 2010). Still, it has a prominent place in literature, education and practice in the social and medical sciences, and has appeared in the practice and scientific literature of quality engineering as well.

This chapter aims to signal and comment on a number of common methodological errors made in agreement studies and in applications of the κ index found in scientific publications in prestigious journals in medical science especially. Some of these errors concern the design of agreement studies, and their ramifications include a potentially serious bias in the estimated κ index. Other errors are related to the standard error of the estimated agreement; the parameters of many agreement studies are such that this standard error, and consequently the confidence margins on the estimated κ index, is so large as to make the studies' results fairly useless. Finally, there are a number of interpretation pitfalls, stemming from the ambiguity of the concept of chance agreement, and from the strong under-weighting of the agreement on less prevalent classes in the case of strongly nonuniform class

prevalences. As a consequence of these interpretation pitfalls, reported κ values may not reflect the authors' intention.

The next section gives a statistical model for nominal measurements, and defines the κ index in terms of this statistical model. Three clusters of problematic issues concerning the κ index are introduced, which are explained in the three subsequent sections and illustrated with examples of agreement studies in the literature of the social and medical sciences. The final section presents our recommendations for agreement studies. Throughout this chapter issues are illustrated with actual papers from the literature. Please note that these examples are not in any way intended to discredit the work of their authors, but instead, are presented to draw attention to methodological shortcomings in practices in the medical and behavioral sciences.

The terminology used in this chapter deviates from the rest of the thesis in that the typical terminology of the diagnostic sciences is adopted (sensitivity, specificity, prevalence, and subjects) rather than quality engineering terminology (*FAP*, *FRP*, defect rate, and items).

4.2 Experimental design and statistical model

As we see it, many of the errors and much of the confusion to be discussed are rooted in the weak and superficial theoretical foundation of many expositions and discussions of agreement. Many discussions are framed in terms of sample statistics, without reference to a population model, and the few population models that are introduced do not capture what we see as essential characteristics of measurement and measurement reliability. We see the conceptualization and modeling that we present in this section as an important contribution to the theory of agreement studies.

To assess the quality of a nominal measurement procedure**,** one could collect data in the following manner. A sample of n subjects, randomly selected from the population of subjects, are independently classified once by each of m appraisers on an unordered scale $\{0,1,\ldots,a-1\}$. The results are denoted Y_{ij}, with $i = 1,\ldots,n$ indexing subjects, and $j = 1,\ldots,m$ indexing appraisers. Measurements are intended to reflect an underlying empirical property of the subjects, named the measurand, and denoted X_i, which assumes the same values $\{0,1,\ldots,a-1\}$. In the population of subjects the value of the measurand are assumed stochastically independent with a discrete distribution given by

(4.1) $p(k) = P(X_i = k)$, $k = 0,...,a-1$, with $\sum_{k=0}^{a-1} p(k) = 1$

(the *class prevalences*). As for the distribution of the Y_{ij}, we assume that given a subject's true state X_i the m measurements Y_{i1}, Y_{i2}, …, Y_{im} are stochastically independent (the assumption of *conditional independence*). Moreover, the distribution of the Y_{i1}, Y_{i2}, ..., Y_{im} depends on X_i, and we define

(4.2) $q(k \mid l) = P(Y_{ij} = k \mid X_i = l)$,

thus specifying the distribution of the measurement errors. Note that, in case of a dichotomous test resulting in $Y = 0$ (negative) or $Y = 1$ (positive), $q(0|0)$ is the test's specificity (the probability of a correct negative), and $q(1|1)$ is the test's sensitivity (the probability of a correct positive), while $p(1)$ is the prevalence of the disorder. The model parameters $p(k)$, $k = 0, 1, \ldots, a–1$, and $q(k \mid l)$, $k, l = 0, 1, \ldots, a–1$, determine the distribution of the Y_{ij} and we have

(4.3) $P(Y_{ij} = k) = \sum_{l=0}^{a-1} p(l)q(k \mid l) = q(k)$ (marginal distribution).

Situations may deviate from the assumptions above in numerous ways, and one's objectives may motivate alternative study designs. For example, the abovementioned assumption of conditional independence is often violated in practice due to nuisance factors affecting the results. For the purposes of our exposition, we think it is productive to keep the basic model relatively simple, and comment, where suitable, on possible extensions.

We now turn to the evaluation of nominal measurements in terms of a probability of agreement. Two measurements of a subject agree if they are identical (the subject is classified in the same category both times). $P_{Agreement}$ (or short: P_A) is the probability that two arbitrary measurements of an arbitrary subject agree. Under the model specified by Equations (4.1) and (4.2), we have for a subject with actual state $X_i = l$:

$$P_A(l) = P(Y_{ij_1} = Y_{ij_2} \mid X_i = l) = \sum_{k=0}^{a-1} q^2(k \mid l),$$

and for an arbitrary subject:

(4.4) $$P_A = P(Y_{ij_1} = Y_{ij_2}) = \sum_{l=0}^{a-1}\sum_{k=0}^{a-1} p(l)q^2(k \mid l).$$

Fleiss (1971) introduced the sample statistic

$$\hat{P}_A = \frac{1}{nm(m-1)} \sum_{i=1}^{n}\sum_{k=0}^{a-1} N_{ik}(N_{ik} - 1),$$

where $N_{ik} = \{\#j\colon Y_{ij} = k\}$. De Mast and Van Wieringen (2003) show that $\hat{P}_A$ is an unbiased estimator of P_A (that is, $\mathrm{E}(\hat{P}_A) = P_A$).

The probability of agreement is positive, even if measurements are completely unrelated to the measurand they intend to reflect. To correct for this phenomenon, Cohen (1960), Fleiss (1971), Conger (1980) and numerous others have introduced κ indices as a recentered and rescaled version of P_A. They are defined as

$$\kappa = \frac{P_{\mathrm{A}} - P_{\mathrm{A|C}}}{1 - P_{\mathrm{A|C}}}, \tag{4.5}$$

where $P_{A|C}$ is the probability of agreement of random measurements 'by chance'. The value $\kappa = 1$ corresponds to the agreement that a perfect measurement system would attain, and 0 corresponds to the agreement that 'chance measurements' would attain. The most common conception of chance measurements is the one by Fleiss (1971), where chance measurements are defined as independent of the measurand and with a probability distribution equal to the marginal distribution of the measurement system under study when applied to the subjects population under study (as given in (3)). Denoting chance measurements by Z_{ij}, this amounts to

Z_{ij} are i.i.d. and $P(Z_{ij} = k) = q(k)$ for all i, j, and k.

Under these premises, the probability of agreement of chance measurements equals $P_{\mathrm{A|C}} = P(Z_{ij_1} = Z_{ij_2}) = \sum_{k=0}^{a-1} q^2(k)$. Fleiss's (1971) sample statistic

$$\hat{P}_{\mathrm{A|C}} = \sum_{k=0}^{a-1} \frac{N_k^2}{(mn)^2},$$

(with $N_k = \{\#(i, j)\colon Y_{ij} = k\}$) estimates $P_{\mathrm{A|C}}$ with a minor bias (De Mast and Van Wieringen, 2007). The sample $\hat{\kappa}$ is defined as in (4.5), but with the sample statistics $\hat{P}_A$ and $\hat{P}_{\mathrm{A|C}}$ instead of the corresponding population parameters.

Agreement studies are commonly done as part of scientific endeavor in the social and medical sciences (and beyond), and the results are frequently reported in the form of κ indices. Upon reviewing a large number of publications reporting on the results of agreement studies, we identified a number of commonly made methodological errors. We present and discuss these errors in the next sections, and we give examples of such errors in the existing literature. First we speculate on the grounds for these errors.

The presentation and modeling above deviate from many expositions in the literature in two important aspects. First, they define an experimental model with population parameters, and define P_A and κ in terms of these population parameters. Sample statistics $\hat{P}_A$ and $\hat{\kappa}$ are presented as estimators for P_A and κ. In the literature, expositions are often framed in terms of sample statistics only, without referring to a population model (notable exceptions include Landis and Koch (1977), Kraemer (1979), Tanner and Young (1985), De Mast (2007)). This makes it difficult to assess the properties of κ statistics as estimators of a population parameter. For example, inferences based on sample statistics should include an assessment of the estimate's standard error or confidence margins.

Another noteworthy characteristic of the given exposition, setting it apart from the population models mentioned above, is that it attributes total dispersion in the measurements to dispersion in the measurand X, and dispersion in the measurements Y conditional on X. This is in line with the typical models employed in metrology and measurement theory (ISO, 1995; Pepe, 2003). Much of the literature of agreement studies, however, introduces κ statistics in the context of classifications and cross tabulations, without reference to a measurand. By doing so, a mapping that is essentially a measurement is treated as merely a classification. The distinguishing characteristic of a measurement, is that it is a classification *aimed to reflect an empirical property of the subjects being measured* (cf. classical definitions of measurement in Lord and Novick, 1968, p.17; Allen and Yen, 1979, p.2; Wallsten, 1988). Including this empirical property as an element of the statistical model, as is done in the model presented above, is not only more natural, it also allows one to separate the behavior of the measurand (which is a characteristic of the subjects population) from the behavior of the measurement errors (a characteristic of the classification procedure), and to state assumptions about both explicitly and separately. By defining κ indices in the context of classification and cross tabulation, as is typically done in the literature, the assumptions about the measurand's behavior are obscured. We speculate that the conception of agreement in the context of classification rather than measurement is one of the causes of many of the interpretation problems discussed in a later section.

Kraemer's (1979) population model (her 'Case 1'), which is restricted to dichotomous classifications, allows a comparable distinction between "characteristics of the population" and "decision-making errors". Contrary to our conceptualization, however, Kraemer sees the marginals $q(0)$ and $q(1)$ as population characteristics, rather than our $p(0)$ and $p(1)$, and this line of reasoning permeates, implicitly or explicitly, much of the literature on agreement.

However, the $q(k)$ reflect both the distribution of true values in the subject population *and* the distribution of classification errors (per Equation (4.3)). We think it is better to regard the class prevalences $p(k)$ and the conditional probabilities $q(k|l)$ as the intrinsic characteristics of the subject population and measurement errors respectively, and the marginal distribution given by the $q(k)$ as their combined consequence.

We discuss three problematic issues concerning agreement studies:

1) Study design: nonrandom sampling
2) Problems and errors related to nonuniform class prevalences
3) Interpretation pitfalls

4.3 Study design: Nonrandom sampling

The main commonly made error in the design of an agreement study, is that the sample is unrepresentative for the subjects population. It is crucial to work with a representative sample. If one selects a sample in which the numbers of subjects in the different classes are not representative for the study population, $\hat{P}_A$ will be biased, because P_A depends on the class prevalences $p(l)$ (per Equation (4.4)), unless $P_A(l)$ is equal for all l. This bias will mostly be modest, depending on the differences between the $P_A(l)$ for different classes l. However, if one expresses the result in the form of κ, this bias is leveraged substantially by the rescaling based on $P_{A|C}$, and $\hat{\kappa}$ will be strongly biased even if the $P_A(l)$ are equal for all l.

We illustrate the large bias that can result from unrepresentative sampling by a study conducted by Spiteri et al. (1988) that appeared in *The Lancet*, which researched the agreement in detecting the presence or absence of certain respiratory signs. The authors do not state clearly how and from what population the patients were sampled. Most of the patients in the study had respiratory disorders and all patients had "stable well-defined features and a definitive diagnosis confirmed by investigations", suggesting that the patients were not a random sample from the general population. If that is so, the estimated κ values are biased, or only representative for that specific population from which the sample was taken. We illustrate the possible magnitude of this bias from a numerical example. In the abovementioned study a certain chest sign, an 'increased percussion note', has a $\hat{\kappa}$ value of 0.50. The 24 patients were each inspected by four physicians. The number of physicians that

indicated an increased percussion note was 0 for 16 patients, 1 for five patients, 2 for one patient, 3 for none of the patients, and 4 for two patients. This could correspond to the following statistical properties of the test procedure (values chosen for illustration, but not based on the original study): a prevalence of $p(1) = 0.10$, a specificity of $q(0|0) = 0.93$ and a sensitivity of $q(1|1) = 0.92$, which gives the reported value $\kappa = 0.50$. Now suppose that the sample is unrepresentative for the population that the researchers have in mind, because in fact the prevalence is not 0.10 but 0.01. Then, the population value of κ is only 0.10, and the study is likely to overestimate agreement by about a factor 5. Clearly, κ depends heavily on prevalence and on the sampling method, and it is crucial that the sample is representative. Therefore, when conducting an agreement study as in the paper on detecting respiratory signs, the population of subjects for which the measurement procedure is intended, must be clearly defined, and when using κ, the sample of subjects must be a random sample from this subjects population.

In the statistical model given above, the assumption was made that the m measurements $Y_{i1}, Y_{i2}, \ldots, Y_{im}$ are stochastically independent given the subject's true value X_i and, therefore, that X_i is the only factor affecting the stochastic properties of the measurements (the assumption of conditional independence). If the sample is representative, this assumption is not crucial. However, if the sample is unrepresentative, a violation of the assumption of conditional independence creates an even further bias (cf. Chapter 2). In the study about detecting respiratory signs by Spiteri et al., suppose that an increased percussion note is easier to detect for some patients than for others. This would be a violation of the conditional independence assumption. If the sample is unrepresentative in the sense that patients are overrepresented whose increased percussion note is relatively easy to detect, the expected value of $\hat{\kappa}$ will increase even beyond 0.50, leading to an even more serious overestimation.

The importance of a random sample is underappreciated in literature. The Food and Drug Administration (2007) mentions that in studies evaluating diagnostic tests the subjects should include the complete spectrum of patient characteristics, but does not mention that the sample of subjects should otherwise be representative for the study population. Several papers about agreement studies based on κ recommend a balanced sample, in which sample prevalences are uniform (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990; Byrt et al., 1993; Hripcsak and Heitjan, 2002), clearly in conflict with our observation that a representative sample is essential. Other misconceived sampling strategies include sampling subjects from the so-called *gray area*, i.e. subjects that are hard to judge, or sampling subjects

such that roughly one third is clearly positive, one third is clearly negative, and one third is hard to judge.

4.4 Problems and errors related to nonuniform class prevalences

Kraemer, Periyakoil and Noda (2002) point out that in the case of scales with 3 or more classes, κ indices may obscure a poor consistency on two classes because it is averaged out with a possibly good consistency on the remaining classes. We think the situation becomes even more tricky when class prevalences are nonuniform (that is, $p(k) >> p(l)$ for some $k \neq l$). Especially for tests for disorders this is typically the situation, because typically $p(0) >> p(1)$ (the fraction of subjects unaffected by the disorder is much larger than the fraction affected).

If class prevalences are nonuniform, κ and P_A by approximation only reflect the consistency in the most prevalent class. This is because the $P_A(l)$ are weighted by the prevalence of each of the classes as in Equation (4.4). This is not an error in itself, but it should be born in mind in interpreting the κ index. For example, if a certain disease has a small prevalence ($p(1) \approx 0.0$), the P_A estimated from a random sample of diagnoses almost exclusively reflects the specificity $q(0|0)$:

$$\begin{aligned} P_A &= p(0)\left(q^2(0|0)+q^2(1|0)\right)+p(1)\left(q^2(0|1)+q^2(1|1)\right) \\ &\approx p(0)\left(q^2(0|0)+(1-q(0|0))^2\right) \end{aligned}$$

and similarly for κ. Reporting the quality of the diagnostic procedure solely as a κ index, would fail to reflect the procedure's sensitivity $q(1|1)$, which is an equally important aspect of diagnostic quality. De Mast, Erdmann and Van Wieringen (2011) give a similar warning for pass/fail inspections in industry, where κ indices evaluate an inspection exclusively in terms of the producer's risk (the probability of a false rejection), ignoring the consumer's risk (a false acceptance).

Interpretation of κ becomes even more precarious if class prevalences are extremely nonuniform (one $p(l) > 0.95$). In such cases, the partial derivatives of κ with respect to the parameters $q(k|\ l)$ approach 0, while the partial derivative with respect to the $p(l)$ becomes very large (Figure 4.1). This implies that, in the case of extremely nonuniform class prevalences, κ responds strongly and rather one-sidedly to changes in class prevalences.

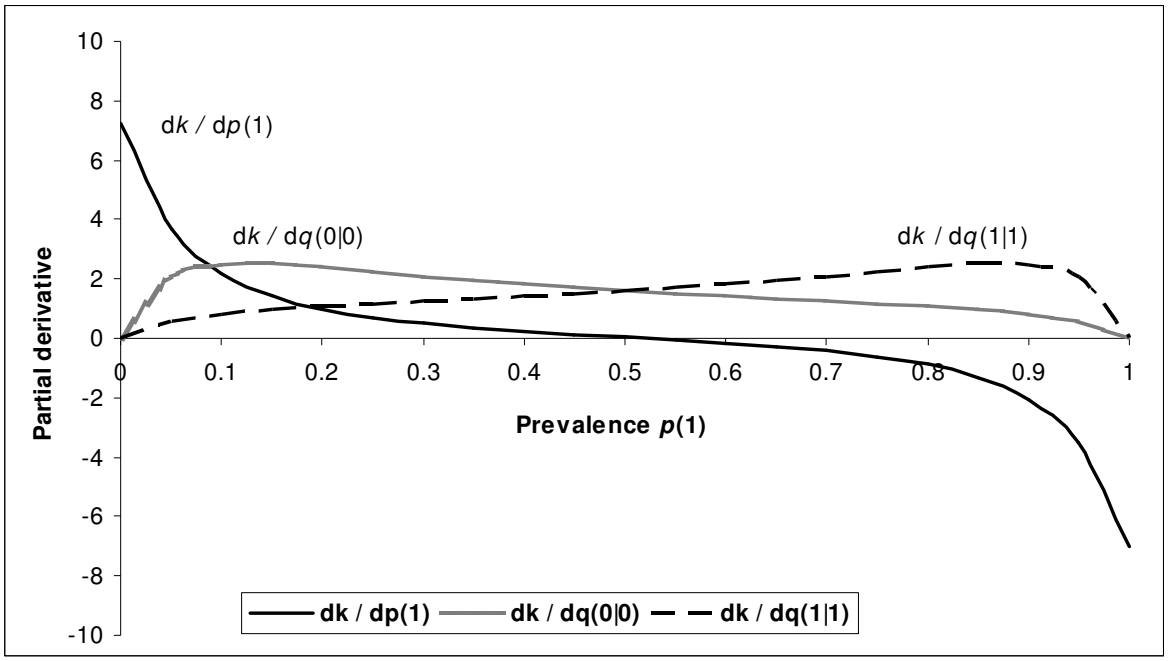

Figure 4.1. Partial derivatives of κ for various values of the prevalence $p(1)$, in the situation of a dichotomous scale, and assuming that $q(0|0) = q(1|1) = 0.95$.

In summary, the κ index confounds various properties of the measurements and the class prevalences into a single number, and especially for nonuniform class prevalences is driven rather one-sidedly by either the $q(k|l)$ of the most prevalent class l, or (for extremely nonuniform prevalences) responds almost exclusively to the class prevalences themselves. As a consequence, it may be an oversimplification to reject or accept a measurement procedure on the basis of κ, following criteria for values of κ as in Landis and Koch (1977), Cicchetti and Sparrow (1981), Fleiss et al. (2003). For example, Fleiss et al. (2003) judge $\kappa < 0.40$ as poor, κ between 0.40 and 0.74 as fair to good and κ between 0.75 and 1.00 as excellent. To illustrate how uncritical application of such criteria leads to practically questionable decisions, consider a dichotomous test with specificity and sensitivity $q(0|0) = q(1|1) = 0.95$. For many practical applications, these error probabilities may be acceptable. However, if the prevalence $p(1) = 0.01$ (extremely nonuniform, but a common order of magnitude), then $\kappa = 0.14$, 'poor' according to the abovementioned criteria. Note that as a first order (Taylor) approximation around these values,

$$\kappa \approx 0.14 + 12.3\,(p(1) - 0.01) + 2.6\,(q(0|0) - 0.95) + 0.3\,(q(1|1) - 0.95),$$

illustrating the fact that for such nonuniform prevalences κ is largely driven by the prevalence (slope of 12.3), and that it is nearly insensitive to $q(1|1)$ (slope of 0.3).

Instead of the sensitivity and specificity, in different contexts it might be more important to consider the positive and negative predictive values of a diagnosis (Pepe, 2003). Suppose a diagnosis has a positive predictive value $P(X = 1 | Y = 1) = 0.8$ and a negative predictive value $P(X = 0 | Y = 0) = 0.95$. Assuming a prevalence of $p(1) = 0.10$, the $\kappa = 0.39$ is poor according to Fleiss et al., but the predictive values may be acceptable for the application under study. A situation with worse positive predictive value (0.7 instead of 0.8) but higher prevalence (0.25) gives $\kappa = 0.48$, 'fair to good' according to the criteria.

Besides an interpretation problem, nonuniform class prevalences also result in a strong sensitivity of the estimator $\hat{\kappa}$ to sampling variations, and thus a large standard error of the estimator $\hat{\kappa}$. De Mast (2007) gives an example for a two-point scale, involving 100 subjects rated by two appraisers, where changing a single data point reduces the estimated agreement $\hat{\kappa}$ from 1.0 to 0.66.

We conducted a simulation study to determine the standard error of $\hat{\kappa}$ statistics for dichotomous tests where subjects are diagnosed as either negative (0) or positive (1). The standard error depends on the prevalence $p(1)$, specificity $q(0|0)$, sensitivity $q(1|1)$ and the sample size (the numbers of subjects n and appraisers m). For each combination of sample size and model parameters, 10,000 samples were created and for each sample the $\hat{\kappa}$ statistic was computed, taking the sample standard deviation of these 10,000 realizations as the estimated standard error. With 99% confidence, the results in Tables 4.1 and 4.2 have a relative error of at most 2% (leading to an absolute error of 0.007 in extreme cases). Alternatively, the standard error could be approximated based on a multinomial distribution for the cell counts in a cross tabulation (Fleiss et al., 1969, Bloch and Kraemer, 1989, Kraemer et al., 2002).

Tables 4.1 and 4.2 show that the standard error may be unacceptably large, especially if prevalence is below 0.10 and specificity above 0.90, a very common situation. In this situation, if the number of subjects is $n = 50$ and the number of appraisers is $m = 4$, the standard error is above 0.08 (and can get as large as 0.35). The same holds for $n = 100$ subjects and $m = 2$ appraisers, even for lower values of specificity. Bootstrapping shows that a 95% confidence interval on κ has a width larger than 0.35 in almost all of these cases, which makes the estimate practically useless.

A potentially unacceptably large standard error of $\hat{\kappa}$ is a common problem in applications of κ in literature. A study by Meining et al. (2003), that appeared in *Endoscopy*,

	spec.:	0.50	0.75	0.95	0.99	0.50	0.75	0.95	0.99
prev.	**sens.**		***n* = 30, *m* = 2**				***n* = 100, *m* = 2**		
0.01	**0.50**	0.18	0.18	0.18	0.22	0.10	0.10	0.12	0.25
	0.75	0.18	0.18	0.22	0.31	0.10	0.10	0.14	0.31
	0.95	0.18	0.18	0.25	0.37	0.10	0.10	0.16	0.34
	0.99	0.18	0.18	0.26	0.38	0.10	0.10	0.16	0.35
0.10	**0.50**	0.18	0.18	0.26	0.35	0.10	0.10	0.15	0.20
	0.75	0.18	0.19	0.28	0.33	0.10	0.10	0.14	0.16
	0.95	0.18	0.19	0.25	0.26	0.10	0.10	0.12	0.10
	0.99	0.18	0.19	0.24	0.23	0.10	0.10	0.11	0.08
0.25	**0.50**	0.18	0.19	0.23	0.26	0.10	0.10	0.12	0.14
	0.75	0.18	0.19	0.20	0.20	0.10	0.10	0.10	0.10
	0.95	0.19	0.17	0.14	0.10	0.10	0.09	0.07	0.05
	0.99	0.19	0.17	0.13	0.07	0.10	0.09	0.07	0.04
0.50	**0.50**	0.18	0.19	0.20	0.20	0.10	0.10	0.11	0.11
	0.75	0.19	0.18	0.16	0.16	0.10	0.10	0.09	0.08
	0.95	0.20	0.16	0.11	0.09	0.11	0.09	0.06	0.05
	0.99	0.20	0.16	0.09	0.05	0.11	0.08	0.05	0.03
prev.	**sens.**		***n* = 50, *m* = 2**				***n* = 200, *m* = 2**		
0.01	**0.50**	0.14	0.14	0.16	0.25	0.07	0.07	0.09	0.20
	0.75	0.14	0.14	0.19	0.33	0.07	0.07	0.10	0.24
	0.95	0.14	0.14	0.22	0.39	0.07	0.07	0.11	0.25
	0.99	0.14	0.14	0.22	0.39	0.07	0.07	0.12	0.25
0.10	**0.50**	0.14	0.14	0.21	0.29	0.07	0.07	0.10	0.14
	0.75	0.14	0.15	0.21	0.24	0.07	0.07	0.10	0.11
	0.95	0.14	0.15	0.18	0.16	0.07	0.07	0.08	0.06
	0.99	0.14	0.15	0.17	0.14	0.07	0.07	0.08	0.05
0.25	**0.50**	0.14	0.14	0.18	0.20	0.07	0.07	0.09	0.09
	0.75	0.14	0.14	0.15	0.15	0.07	0.07	0.07	0.07
	0.95	0.14	0.13	0.11	0.08	0.07	0.07	0.05	0.04
	0.99	0.14	0.13	0.09	0.05	0.07	0.07	0.05	0.03
0.50	**0.50**	0.14	0.14	0.15	0.15	0.07	0.07	0.07	0.07
	0.75	0.14	0.14	0.13	0.12	0.07	0.07	0.06	0.06
	0.95	0.15	0.13	0.09	0.07	0.07	0.06	0.04	0.03
	0.99	0.15	0.12	0.07	0.04	0.08	0.06	0.03	0.02

Table 4.1: Standard errors of $\hat{\kappa}$ for m = 2 for various values of prevalence, sensitivity and specificity, based on simulations with R = 10,000 (empirical 99% confidence interval bounds for the standard error do not exceed +/- 0.007).

is an example of this problem. It assesses inter-observer agreement of the findings from a certain type of endoscopy. Four endoscopists evaluated video sequences recorded during endoscopies of 51 patients with reflux symptoms. This is very close to the sample size discussed above. The prevalence is not reported in the paper, but if it is low, the standard errors of the $\hat{\kappa}$ statistics may be extremely large. For example, the estimated $\hat{\kappa} = 0.36$ for the detection of 'Methylene blue positivity' might have a standard error of 0.18 if prevalence were 0.05, and the true agreement would be 'somewhere in between 0.09 and 0.79' (based on

	spec.:	0.50	0.75	0.95	0.99	0.50	0.75	0.95	0.99
prev.	**sens.**		***n* = 30, *m* = 4**				***n* = 100, *m* = 4**		
0.01	**0.50**	0.07	0.08	0.10	0.17	0.04	0.04	0.07	0.17
	0.75	0.08	0.08	0.15	0.27	0.04	0.04	0.10	0.25
	0.95	0.08	0.09	0.20	0.35	0.04	0.05	0.13	0.30
	0.99	0.08	0.09	0.21	0.36	0.04	0.05	0.14	0.32
0.10	**0.50**	0.07	0.08	0.15	0.21	0.04	0.04	0.09	0.11
	0.75	0.08	0.10	0.19	0.22	0.04	0.05	0.10	0.10
	0.95	0.08	0.12	0.21	0.21	0.05	0.06	0.10	0.07
	0.99	0.09	0.12	0.20	0.21	0.05	0.07	0.09	0.06
0.25	**0.50**	0.07	0.08	0.13	0.13	0.04	0.05	0.07	0.07
	0.75	0.08	0.10	0.12	0.12	0.04	0.06	0.07	0.06
	0.95	0.09	0.11	0.10	0.07	0.05	0.06	0.05	0.04
	0.99	0.09	0.11	0.10	0.06	0.05	0.06	0.05	0.03
0.50	**0.50**	0.08	0.08	0.10	0.10	0.04	0.05	0.05	0.05
	0.75	0.08	0.10	0.10	0.10	0.05	0.05	0.05	0.05
	0.95	0.10	0.10	0.08	0.06	0.05	0.05	0.04	0.03
	0.99	0.10	0.10	0.06	0.04	0.05	0.05	0.03	0.02
prev.	**sens.**		***n* = 50, *m* = 4**				***n* = 200, *m* = 4**		
0.01	**0.50**	0.06	0.06	0.09	0.19	0.03	0.03	0.05	0.13
	0.75	0.06	0.06	0.13	0.28	0.03	0.03	0.07	0.18
	0.95	0.06	0.07	0.17	0.35	0.03	0.03	0.09	0.22
	0.99	0.06	0.07	0.18	0.37	0.03	0.03	0.10	0.23
0.10	**0.50**	0.06	0.06	0.12	0.16	0.03	0.03	0.06	0.07
	0.75	0.06	0.08	0.14	0.15	0.03	0.04	0.07	0.07
	0.95	0.07	0.09	0.15	0.12	0.03	0.05	0.07	0.05
	0.99	0.07	0.09	0.14	0.11	0.03	0.05	0.06	0.04
0.25	**0.50**	0.06	0.07	0.09	0.10	0.03	0.03	0.05	0.05
	0.75	0.06	0.08	0.09	0.09	0.03	0.04	0.05	0.04
	0.95	0.07	0.09	0.08	0.05	0.03	0.04	0.04	0.03
	0.99	0.07	0.09	0.07	0.04	0.04	0.04	0.03	0.02
0.50	**0.50**	0.06	0.06	0.08	0.08	0.03	0.03	0.04	0.04
	0.75	0.06	0.08	0.08	0.07	0.03	0.04	0.04	0.04
	0.95	0.08	0.08	0.06	0.05	0.04	0.04	0.03	0.02
	0.99	0.08	0.07	0.05	0.03	0.04	0.04	0.02	0.01

Table 4.2: Standard errors of $\hat{\kappa}$ for m = 4 for various values of prevalence, sensitivity and specificity, based on simulations with R = 10,000 (empirical 99% confidence interval bounds for the standard error do not exceed +/- 0.007).

a bootstrapped 95% confidence interval, and taking for illustration that $q(0|0) = q(1|1) = 0.94$).

Another illustration of excessive standard errors of $\hat{\kappa}$ is a paper by Weisz et al. (2005) that appeared in the *Annual Review of Psychology*, which analyzed 236 studies on youth psychotherapy. In order to assess the coding procedures used to categorize the studies, a 'master coder' and two students coded 30 randomly selected studies. Then $\hat{\kappa}$ values were computed between the 'master coder' and each of the students and the mean of the two resulting $\hat{\kappa}$ statistics was taken. For this small sample size (30 subjects and 2 coders), if there

are two categories, the standard error of $\hat{\kappa}$ is larger than 0.10 for almost all parameter values, which is unacceptably large. If prevalence (which is not reported in the paper) is 0.05, the standard error can get as large as 0.31, again making the value of $\hat{\kappa}$ almost entirely uninformative. (The standard error of the mean of the two $\hat{\kappa}$ statistics is smaller, but still unacceptable.)

4.5 Interpretation pitfalls

The problematic interpretation of κ and its sometimes paradoxical behavior have been much discussed in the literature (Feinstein and Cicchetti, 1990; Byrt et al., 1993; Thompson and Walter, 1988; Uebersax, 1987; Grove et al., 1981; De Mast, 2007; Warrens, 2010). The usual interpretation given to κ, is as the probability of agreement corrected for agreement by chance, in the sense that the value of κ is zero if P_A is equal to the probability of agreement of chance measurements. However, the concept of 'chance measurements' is too ambiguous to provide a well-defined zero point. Chance measurements are a nonexistent hypothetical concept, and therefore, anything said about their distribution is bound to be hopelessly arbitrary, and it does not make sense to argue about how appraisers would conduct 'chance measurements' in practice.

In fact, an analysis of the chance correction $P_{A|C}$ shows that its premises have implausible or even irreconcilable implications (De Mast, 2007). Chance measurements are assumed to have a distribution equal to the marginal distribution of the measurements under study (that is, $P(Z_{ij} = k) = q(k)$). It is hard to imagine why or by what sort of mechanism blind measurements would happen to have the same distribution as the marginal distribution of the measurement procedure under study. But besides being an implausible choice, a problematic consequence is that κ, thus interpreted, is unsuited to compare two different measurement procedures. For instance, Naranjo et al. (1981) compare the agreement of a new procedure for classifying adverse drug reactions to current practice. However, the chance correction applied to the agreement of the new procedure is based on the marginal distribution of the new procedure, whereas the chance correction applied to the agreement in current practice is based on the marginals in current practice. Thus, the two κ indices employ different chance corrections and therefore have values on scales with different zero points. The same comment holds for a study by O'Keefe et al. (1994) reported in *the Lancet*, that compares the

agreement of two procedures for assessing the presence or absence of the ankle jerk in elderly people.

Another problematic consequence is that on the one hand chance measurements are conceived as blind (that is, uninformative about the measurand), but on the other hand their distribution given by the $q(k)$ is related to the class prevalences $p(k)$ of the measurand (since the marginals $q(k)$ are related to the $p(k)$ via (4.3)); to the authors these seem two irreconcilable implications (see De Mast, 2007).

One possible solution is to use an unambiguous zero-point for correcting the raw P_A. De Mast and Van Wieringen (2007) propose to define $P_{A|C}$ as the agreement of a maximally non-informative measurement system. Defining chance measurements as having a uniform distribution, that is, $P(Z_{ij} = k) = 1/a$ for all k, the resulting chance correction is $P_{A|C}^{\text{Unif}} = 1/a$, and

$$\kappa^{\text{Unif}} = \frac{P_A - 1/a}{1 - 1/a} \ .$$

This metric was proposed earlier by Bennett et al. (1954) and advocated by Brennan and Prediger (1981), and others. It has at least two clear and unambiguous interpretations. First, $P_{A|C} = 1/a$ is the lower bound for the probability of agreement attainable by measurement systems on a scale with a classes (De Mast and Van Wieringen, 2007). Second, chance measurements thus defined represent maximally non-informative measurements, in the information theoretic sense where information is defined as the negation of entropy, and the uniform distribution has maximal entropy (Berger, 1988). Thus, κ^{Unif} is the probability of agreement in excess of minimal agreement on the given scale, or in excess of the agreement of maximally non-informative measurements.

An alternative solution is to interpret κ not as a measure of agreement corrected for agreement by chance, but as a measure of intraclass association. The problematic term $P_{A|C}$ is not interpreted in itself. Instead, the κ index can be shown to have the form of a measure of predictive association by rearranging its term (De Mast, 2007). Let $\Delta_Z^G = 1 - \sum_{k=0}^{a-1} p_k^2$ be the Gini dispersion of a categorical variable Z with a probability distribution $(p_0, p_1, \ldots, p_{a-1})$ (Gilula and Haberman, 1995). Then

$$\kappa = 1 - \frac{1 - \sum_{l=0}^{a-1} \left(p(l) \sum_{k=0}^{a-1} q^2(k \mid l) \right)}{1 - \sum_{k=0}^{a-1} q^2(k)} = 1 - \frac{\Delta_{Y|X}^G}{\Delta_Y^G} \ .$$

The form $1-\Delta_{Y|X}/\Delta_Y$ on the right is the general expression of a coefficient of predictive association, where Δ can be any measure of dispersion (Hershberger and Fisher, 2005), and κ thus turns out to be a measure of association based on Gini's dispersion measure. Taking for Δ the entropy $\Delta_Z^E = -\sum_{k=0}^{a-1} p_k \log p_k$ instead of the Gini dispersion, one finds Theil's uncertainty coefficient, which is thus a direct cousin of κ (De Mast, 2007). It is also similar in form to the intraclass correlation coefficient (*ICC*) used to express the reliability of interval or ratio scale measurements:

$$ICC = Cor(Y_{i,1}, Y_{i,2}) = 1 - \frac{\sigma^2_{Y_{ij}|X_i}}{\sigma^2_{Y_{ij}}} = 1 - \frac{\Delta^V_{Y|X}}{\Delta^V_Y},$$

with Δ^V now the variance instead of the Gini dispersion. Interpreted in this way, κ represents the association between two measurements $Y_{i,1}$ and $Y_{i,2}$ of the same subject i. Another analogue is the coefficient of determination R^2 in regression analysis, where $Y_{ij} = X_i + \varepsilon_{ij}$:

$$R^2 = Cor^2(Y_{ij}, X_i) = \frac{\sigma^2_{X_i}}{\sigma^2_{Y_{ij}}} = 1 - \frac{\sigma^2_{Y_{ij}|X_i}}{\sigma^2_{Y_{ij}}} = 1 - \frac{\Delta^V_{Y|X}}{\Delta^V_Y}.$$

This gives the interpretation that κ represents the fraction of the total dispersion in the measurements Y_{ij} that can be attributed to dispersion in the measurands X_i, that is, as a measure of reliability. Interpreting κ as a measure of predictive association, much of its paradoxical behavior makes sense.

Kraemer et al. (2002) dismiss the chance corrected agreement interpretation as a historical curiosum, but focus on an interpretation as an *ICC*. Since their elaboration only holds for $a = 2$ classes, they recommend against the use of κ if $a \geq 3$. We think that our elaboration, based on the Gini dispersion, shows that κ *can* be interpreted as a reliability measure when $a \geq 3$.

Working with the interpretation of κ as a measure of association, it is important to be aware that such measures express measurement dispersion *in relation to a population of subjects* (and the class prevalences or distribution of the measurand in that population). If the same diagnostic test is applied in another population of subjects, with different class prevalences, then κ will be different. Consider, as an example, a dichotomous diagnostic test, with specificity and sensitivity $q(0|0) = q(1|1) = 0.95$. Depending on the prevalence $p(1)$ of the disorder, κ ranges from 0.00 to 0.81 in this case. In view of this fact that agreement is expressed in relation to prevalences in the subjects population, when expressing the results of

an agreement study in terms of κ, it is crucial to define and delineate the study population of subjects to which it applies, and failure to do so makes the reported κ meaningless.

In the third section, we have given an example of a paper about respiratory disorders (Spiteri et al., 1988) that does not clearly specify the relevant subjects population. Another interesting example is Naranjo et al. (1981), who assess the reliability of classifying alleged adverse drug reactions (ADRs) by the probability that they were caused by drug therapy: definite, probable, possible, or doubtful. The aim of the authors is to "develop a simple method to assess the causality of ADRs in a variety of clinical situations". They take a sample of 63 randomly selected cases published in a number of prestigious journals. The results therefore only apply to those ADRs that are published in prestigious journals, which cannot be assumed to be representative for the ADRs in clinical situations. The values of κ that they report are therefore meaningless for clinical situations.

4.6 Conclusion and recommendations

We conclude this chapter by listing our recommendations for agreement studies. Throughout the chapter we have emphasized the role of the measurand, and our first recommendation is that a prerequisite for an agreement study is that the measurand is well (that is, clinically) defined. Without a clinical definition of the measurand, the whole concept of measurement error becomes meaningless. Second, it is important to clearly define and delineate the population of subjects in which the measurement procedure should be discriminating. Especially if the κ index is used, the results are meaningless if the subjects population is not well defined. Third, when using κ, the sample of subjects must be a random sample from the defined subjects population, however impractical that may be, since the estimation bias in $\hat{\kappa}$ may be substantial otherwise.

To evaluate the quality of the measurements, one may use the κ index if one wants a measure of reliability. We recommend against the interpretation as agreement corrected for agreement by chance, as the notion of chance agreement is too problematic and ambiguous. Instead, it can be interpreted as a reliability measure, much alike to the intraclass correlation coefficient used for interval and ratio scale measurements. However, if a gold standard is available to determine the measurand of each subject in the study, we recommend against the use of the κ index, and instead, propose to estimate the classification probabilities $q(k \mid l)$

conditional on the measurand; in the case of a dichotomous test, this amounts to establishing the sensitivity and specificity. Such studies are described in Pepe (2003) and in Chapter 2, and since they establish the intrinsic parameters of the measurement errors, we consider them more informative than agreement studies. If no gold standard is available, latent class or latent trait methods may be used to estimate the same parameters (Van Wieringen and De Mast, 2008), although such methods critically depend on the viability of the conditional independence assumption.

Also in applications with strongly nonuniform class prevalences, such as diagnoses of disorders with low prevalences, we recommend against using the κ index, because of three problematic properties in such cases. Firstly, the standard errors of the estimates are typically unacceptably large. Secondly, the κ index is very sensitive to the class prevalences, and reflects prevalences more than it reflects the quality of the measurement procedure. Thirdly, κ reflects rather one-sidedly the consistency on the most prevalent class. In such cases numerical criteria for κ intended to evaluate a measurement procedure may be an oversimplification and lead to practically questionable decisions. One alternative is a sensitivity/specificity study mentioned above. Another option is to estimate κ^{Unif} as defined in an earlier section. It has a clear interpretation, and it does not suffer from the first two problems related to nonuniform class prevalences.

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 $I = I^{ran} + I^{acc} + I^{rej}$ items, of which I^{ran} are a random sample from all items, I^{acc} a random sample from the subpopulation of accepted items, and I^{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^{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

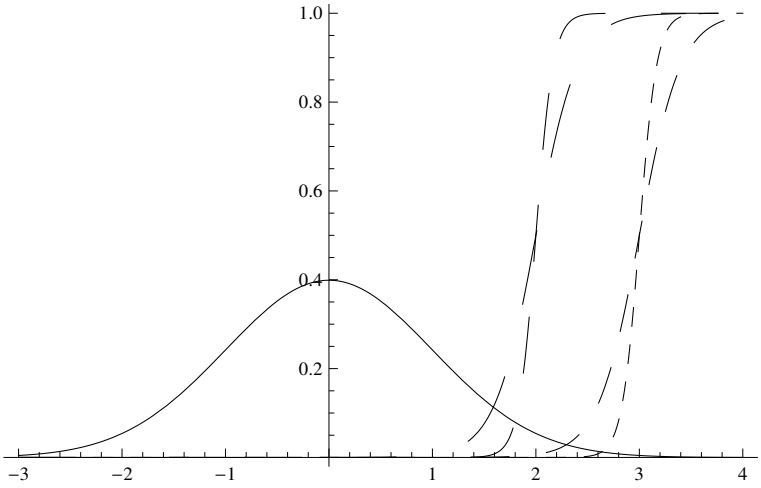

Figure 5.1: Probability density $f_X(x)$ (solid curve) and characteristic curves $q(x)$ (dashed) with α=5, 12; and δ = 2, 3.

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:

(5.1) $\quad q(x) = P(Y_{ik} = 0 \mid X_i = x)$.

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

(5.2) $\quad \log\left(\dfrac{q(x)}{1-q(x)}\right) = \alpha(x-\delta)$, $\alpha > 0$.

The curve's inflection point δ 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 α 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:

$$\begin{aligned} FAP &= \int_{USL}^{\infty} (1-q(x))\varphi(x)dx \Big/ \int_{USL}^{\infty} \varphi(x)dx \\ FRP &= \int_{-\infty}^{USL} q(x)\varphi(x)dx \Big/ \int_{-\infty}^{USL} \varphi(x)dx. \end{aligned} \tag{5.3}$$

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 δ. These probabilities are the inconsistent acceptance probability (*IAP*) and inconsistent rejection probability (*IRP*):

$$\begin{aligned} IAP &= P(Y=1 \mid X > \delta) = \int_{\delta}^{\infty} (1-q(x))\varphi(x)dx \Big/ \int_{\delta}^{\infty} \varphi(x)dx, \\ IRP &= P(Y=0 \mid X \le \delta) = \int_{-\infty}^{\delta} q(x)\varphi(x)dx \Big/ \int_{-\infty}^{\delta} \varphi(x)dx. \end{aligned} \tag{5.4}$$

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 δ), these *IAP* and *IRP* express the random component only. This can be seen from the following decomposition of *FRP*, where we assume that $\delta \le USL$ (and a similar decomposition can be given for *FAP* and for the case $\delta > USL$):

$$\begin{aligned} FRP = {} & P(Y=0 \mid X \le \delta) P(X \le \delta \mid X \le USL) \\ & + P(Y=0 \mid \delta < X \le USL) P(\delta < X \le USL \mid X \le USL). \end{aligned}$$

The last term is the contribution to *FRP* due to systematic measurement error, determined by the distance between δ 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.

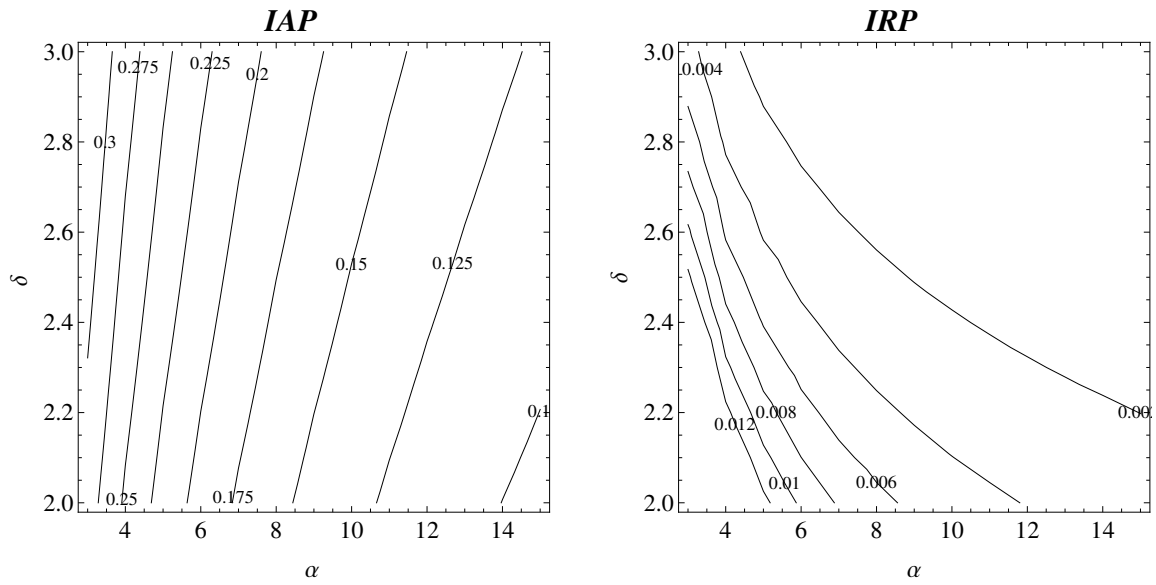

Figure 5.2: Contour plots of *IAP* and *IRP* as a function of α and δ.

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 δ), and for $x \leq \delta$ it does not. Figure 5.2 gives contour plots of *IAP* and *IRP* for varying parameter values α and δ.

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, $P(X > USL \mid Y = 1)$, 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, $P(X \leq USL \mid Y = 0)$, the probability that an item is good, given that it has been rejected. If the reject rate $P(Y = 0)$ is known, both

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:

$$P(X > USL \mid Y = 1) = \int_{USL}^{\infty} (1 - q(x))\varphi(x)dx \Big/ \int_{-\infty}^{\infty} (1 - q(x))\varphi(x)dx$$
$$P(X \leq USL \mid Y = 0) = \int_{-\infty}^{USL} q(x)\varphi(x)dx \Big/ \int_{-\infty}^{\infty} q(x)\varphi(x)dx.$$

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 I^{rej} 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 δ and the discrimination (or precision) as reflected by α. The *bias* of a measurement system is given by $|\delta - USL|$, but as said, *USL* will generally not be known. Still, differences in δ between appraisers or measurement systems indicate systematic differences. The *precision* of a binary measurement system is reflected by α, 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 $\left(\delta - \log(\frac{1-m}{m})/\alpha,\ \ \delta + \log(\frac{1-m}{m})/\alpha\right)$, 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 $\%GRR \approx 2.055/\alpha$. 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 α and δ of model (5.2) are estimated from the experimental data by means of the maximum likelihood method. Conditional on X_i, the probability of a specific set of outcomes of K repeated measurements on a single item, is

$$P(Y_{i1} = y_{i1},...,Y_{iK} = y_{iK} \mid X_i = x) = q(x)^{r_i}(1-q(x))^{K-r_i},$$

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

$$\begin{aligned} P(Y_{i1} = y_{i1},...,Y_{iK} = y_{iK}) &= \int_{-\infty}^{\infty} \varphi(x) P(Y_{i1} = y_{i1},...,Y_{iK} = y_{iK} \mid X_i = x)dx \\ &= \int_{-\infty}^{\infty} \varphi(x) q(x)^{r_i}(1-q(x))^{K-r_i} dx, \end{aligned}$$

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

$$\log L^{ran}(\boldsymbol{\theta}) = \sum_{i=1}^{I^{ran}} \log \int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{r_i}(1-q^{\boldsymbol{\theta}}(x))^{K-r_i} dx, \tag{5.5}$$

where $\boldsymbol{\theta}$ is the parameter vector $\boldsymbol{\theta} = (\alpha, \delta)$. 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 θ_t, one obtains the elements of the gradient:

$$\frac{\partial \log L^{ran}(\boldsymbol{\theta})}{\partial \theta_t} = \sum_{i=1}^{I^{ran}} \frac{\int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{r_i-1}(1-q^{\boldsymbol{\theta}}(x))^{K-r_i-1}(r_i - Kq^{\boldsymbol{\theta}}(x)) \frac{\partial q^{\boldsymbol{\theta}}(x)}{\partial \theta_t} dx}{\int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{r_i}(1-q^{\boldsymbol{\theta}}(x))^{K-r_i} dx}.$$

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

$$\begin{aligned} \frac{\partial q^{\boldsymbol{\theta}}(x)}{\partial \alpha} &= \frac{(x-\delta)e^{\alpha(x-\delta)}}{(1+e^{\alpha(x-\delta)})^2} = (x-\delta)q^{\boldsymbol{\theta}}(x)(1-q^{\boldsymbol{\theta}}(x)) \\ \frac{\partial q^{\boldsymbol{\theta}}(x)}{\partial \delta} &= \frac{-\alpha e^{\alpha(x-\delta)}}{(1+e^{\alpha(x-\delta)})^2} = -\alpha q^{\boldsymbol{\theta}}(x)(1-q^{\boldsymbol{\theta}}(x)) \end{aligned}.$$

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:

$$\begin{aligned}
&P(Y_{i1} = y_{i1},...,Y_{iK} = y_{iK} \mid Y_{i0} = y_{i0}) = P(Y_{i0} = y_{i0}, Y_{i1} = y_{i1},...,Y_{iK} = y_{iK}) \big/ P(Y_{i0} = y_{i0}) \\
&= \int_{-\infty}^{\infty} \varphi(x) P(Y_{i0} = y_{i0}, Y_{i1} = y_{i1},...,Y_{iK} = y_{iK} \mid X_i = x) dx \Big/ \int_{-\infty}^{\infty} \varphi(x) P(Y_{i0} = y_{i0} \mid X_i = x) dx \\
&= \int_{-\infty}^{\infty} \varphi(x) q(x)^{r_i + 1 - y_{i0}} (1 - q(x))^{K - r_i + y_{i0}} dx \Big/ \int_{-\infty}^{\infty} \varphi(x) q(x)^{1 - y_{i0}} (1 - q(x))^{y_{i0}} dx.
\end{aligned}$$

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

$$\begin{aligned}
\log L^{con}(\boldsymbol{\theta}) = \sum_{i=1}^{I^{con}} &(\log \int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{r_i + 1 - y_{i0}} (1 - q^{\boldsymbol{\theta}}(x))^{K - r_i + y_{i0}} dx \\
&- \log \int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{1 - y_{i0}} (1 - q^{\boldsymbol{\theta}}(x))^{y_{i0}} dx,
\end{aligned} \tag{5.6}$$

and the elements of the gradient become

$$\begin{aligned}
&\frac{\partial \log L^{con}(\boldsymbol{\theta})}{\partial \theta_t} \\
&= \sum_{i=1}^{I^{con}} \left(\frac{\int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{r_i - y_{i0}} (1 - q^{\boldsymbol{\theta}}(x))^{K - r_i + y_{i0} - 1} (r_i + 1 - y_{i0} - (K+1) q^{\boldsymbol{\theta}}(x)) \frac{\partial q^{\boldsymbol{\theta}}(x)}{\partial \theta_t} dx}{\int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{r_i + 1 - y_{i0}} (1 - q^{\boldsymbol{\theta}}(x))^{K - r_i + y_{i0}} dx} \right. \\
&\qquad \left. - \frac{(-1)^{y_{i0}} \int_{-\infty}^{\infty} \varphi(x) \frac{\partial q^{\boldsymbol{\theta}}(x)}{\partial \theta_t} dx}{\int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x)^{1 - y_{i0}} (1 - q^{\boldsymbol{\theta}}(x))^{y_{i0}} dx} \right).
\end{aligned}$$

In addition, often there is a *historical dataset* of the inspection results ('reject' or 'accept') of a large number of items $i = 1,..., I^{his}$. Such a historical dataset can be used to estimate the reject rate $P(Y_{ik} = 0) = \int_{-\infty}^{\infty} \varphi(x) q(x) dx$. 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

$$\log L^{his}(\theta) = r^{his} \log \int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x) dx + (I^{his} - r^{his}) \log \int_{-\infty}^{\infty} \varphi(x)(1 - q^{\boldsymbol{\theta}}(x)) dx, \tag{5.7}$$

where $r^{his} = \sum_{i=1}^{I^{his}} (1 - y_{i0})$. The gradient is

$$\frac{\partial \log L^{his}(\boldsymbol{\theta})}{\partial \theta_t} = \frac{r^{his} - I^{his} \int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x) dx}{\left(1 - \int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x) dx\right) \int_{-\infty}^{\infty} \varphi(x) q^{\boldsymbol{\theta}}(x) dx} \int_{-\infty}^{\infty} \varphi(x) \frac{\partial q^{\boldsymbol{\theta}}(x)}{\partial \theta_t} dx .$$

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) $\log L(\boldsymbol{\theta}) = \log L^{ran}(\boldsymbol{\theta}) + \log L^{con}(\boldsymbol{\theta}) + \log L^{his}(\boldsymbol{\theta})$.

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 *maxbfgs* (see, for example, Fletcher, 1970) with the starting values of both parameters set to 1. After having estimated the parameters α and δ, 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 $I^{ran} = 200$ 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

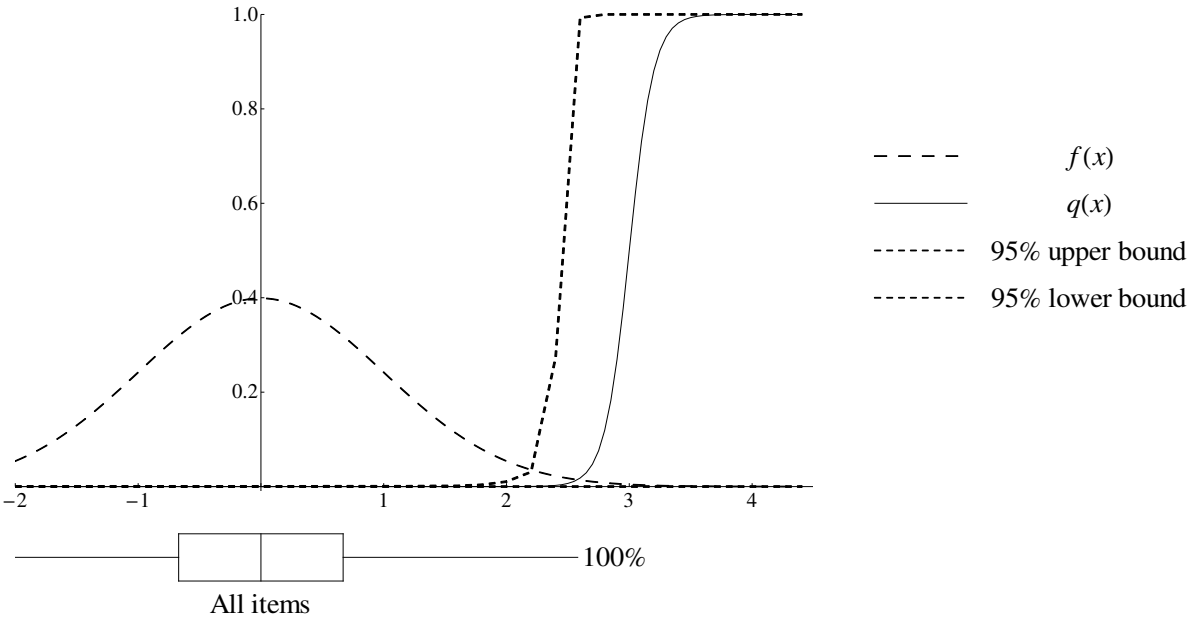

Figure 5.3a: Probability density $\varphi(x)$, characteristic curve $q(x)$ with α=10, δ=3, empirical 95% confidence bounds for estimation with I^{ran}=200, K=9, and box plot representing X in a random sample.

only a small part or even none of the items in a sample are defective (assuming $USL \approx \delta = 3$), 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 *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

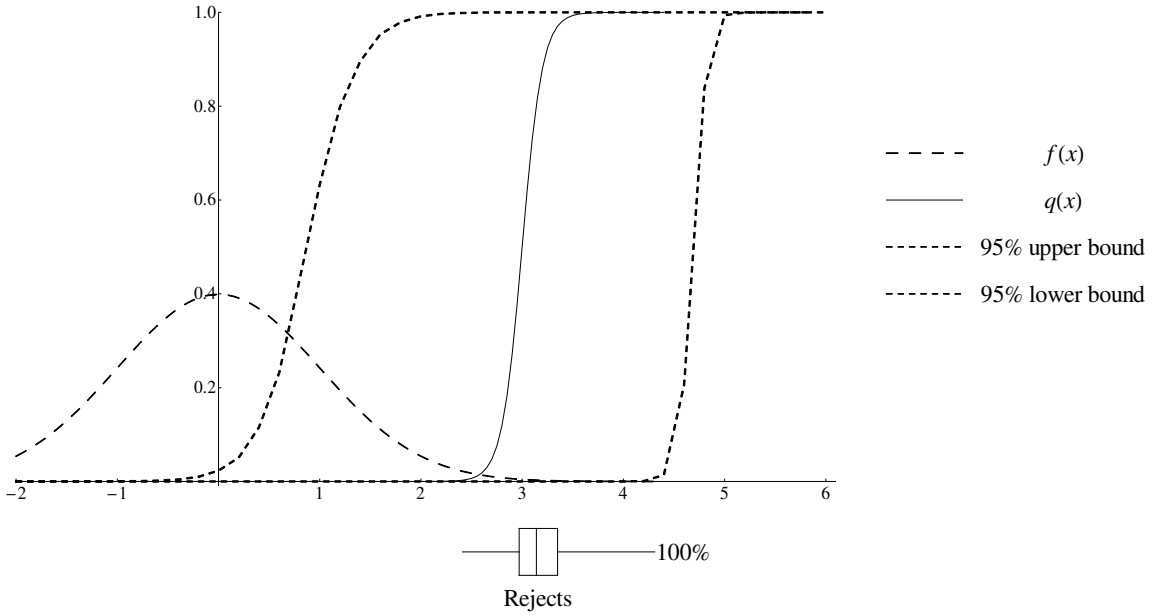

Figure 5.3b: Probability density $\varphi(x)$, characteristic curve $q(x)$ with α=10, δ=3, empirical 95% confidence bounds for estimation with I^{rej}=200, K=9, and box plot representing X in a sample of rejected items.

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 δ 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 δ (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

bounds of the values of the characteristic curve based on estimation with such a combined

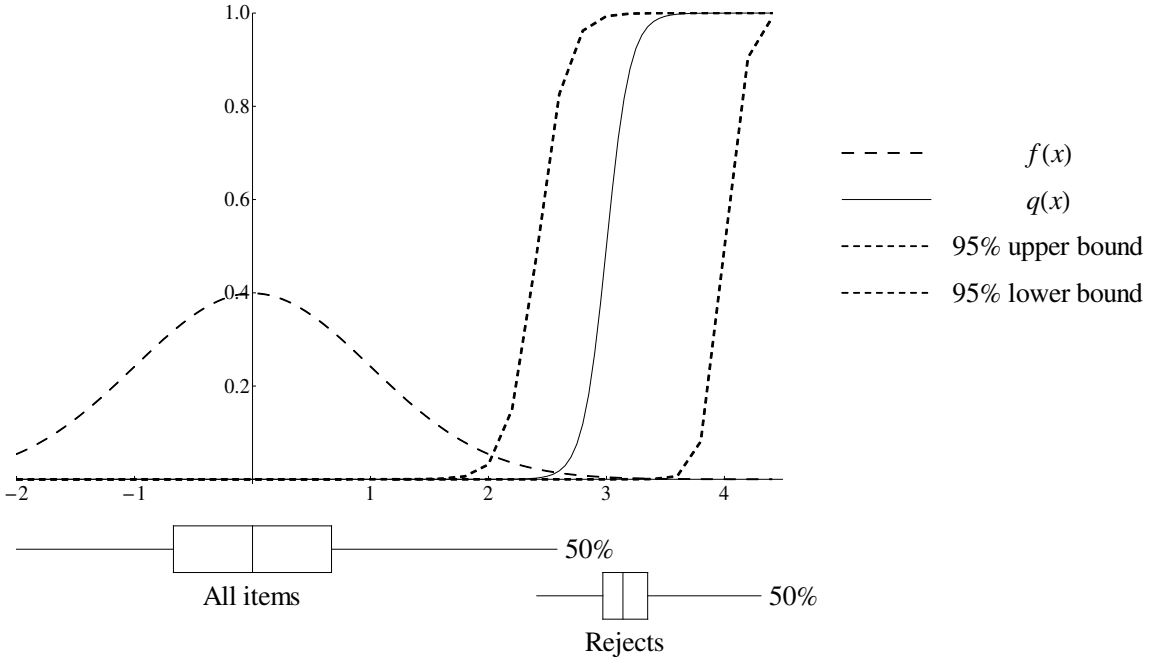

Figure 5.3c: Probability density $\varphi(x)$, characteristic curve $q(x)$ with α=10, δ=3, empirical 95% confidence bounds for estimation with I^{ran}=100, I^{rej}=100, K=9, and box plots representing X in a random sample and in a sample of rejected items, respectively.

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 δ. 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 $\widehat{IAP}$ and $\widehat{IRP}$. Secondly, we determine the precision for different sample sizes. Finally, we assess the robustness against model misspecification. The bias of $\widehat{IAP}$ and $\widehat{IRP}$ 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 $\widehat{IAP}$ and $\widehat{IRP}$).

The simulations are performed by simulating *R* datasets as follows: we draw a *random* sample of I^{ran} 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^{rej} realizations of *X* for which $Y = 0$ ('reject'), and the first I^{acc} realizations for which $Y = 1$ ('accept'), and remove the rest. Thus, we obtain a sample of $I = I^{ran} + I^{rej} + I^{acc}$ items $i = 1,...,I$. 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^{his} trials and rejection probability $P(Y=0) = \int_{-\infty}^{\infty} f_X(x)q(x)dx$. For each simulated dataset, the parameters α and δ 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 $\widehat{IAP}$ and $\widehat{IRP}$ 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^{ran}, I^{rej} and I^{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 $I^{his} = 100,000$. 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

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

Table 5.1: Monte Carlo mean estimates and 95% confidence interval widths for *IAP* and *IRP* for α=5, δ=2, l=200, K=9, R=2500.

*Monte Carlo mean estimate significantly different from true value (*IAP*=0.2154, *IRP*=0.0125)

$R = 2500$ simulations. It is also indicated whether the Monte Carlo means are significantly different from the true values $IAP = 0.2154$ and $IRP = 0.0125$ 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 $I^{rej} = 0$ give an unacceptably large confidence interval for *IAP*, 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

				I^{his}=100,000	
	I^{ran}	I^{rej}		*IAP*	*IRP*
(α,δ)=(5,2)	**150**	**50**	**Width C.I.**	0.0590	0.0055
***IAP*=0.2154**	**100**	**100**	**Width C.I.**	0.0486	0.0435
***IRP*=0.0125**	**50**	**150**	**Width C.I.**	0.0410	0.0038
	0	**200**	**Width C.I.**	0.0371	0.0035
(α,δ)=(5,3)	**150**	**50**	**Width C.I.**	0.0559	0.0006
***IAP*=0.2562**	**100**	**100**	**Width C.I.**	0.0435	0.0005
***IRP*=0.0015**	**50**	**150**	**Width C.I.**	0.0354	0.0005
	0	**200**	**Width C.I.**	0.0282	0.0005
(α,δ)=(12,2)	**150**	**50**	**Width C.I.**	0.0679*	0.0031*
***IAP*=0.1134**	**100**	**100**	**Width C.I.**	0.0466	0.0022
***IRP*=0.0039**	**50**	**150**	**Width C.I.**	0.0414	0.0019
	0	**200**	**Width C.I.**	0.0351	0.0017
(α,δ)=(12,3)	**150**	**50**	**Width C.I.**	0.0719	0.0003
***IAP*=0.1447**	**100**	**100**	**Width C.I.**	0.0497	0.0002
***IRP*=0.0004**	**50**	**150**	**Width C.I.**	0.0391	0.0002
	0	**200**	**Width C.I.**	0.0351	0.0002

Table 5.2: 95% confidence interval widths for *IAP* and *IRP* for α=5, 12; δ=2, 3; I=200; K=9; I^{his}=100.000; R=1000.

*Monte Carlo mean estimate significantly different from true value

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 $I^{rej} > 0$. With a historical dataset, a sample consisting of rejected items exclusively $(I^{rej} = 200)$ is uniformly the best strategy. If a historical dataset is unavailabe, one should use $0 < I^{ran} < 200$, $0 < I^{rej} < 200$, and $I^{acc} = 0$. In that case, the best strategy depends on whether one is more interested in *IAP* or *IRP*, but arguably a sample with $I^{ran} = 150$, $I^{rej} = 50$ and $I^{acc} = 0$ 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^{ran} items and a selective sample of I^{rej} rejected items, again adding up to I = 200. A historical dataset based on $I^{his} = 100,000$ items is assumed to be available. The resulting empirical confidence interval widths for *IAP* and *IRP* are presented in

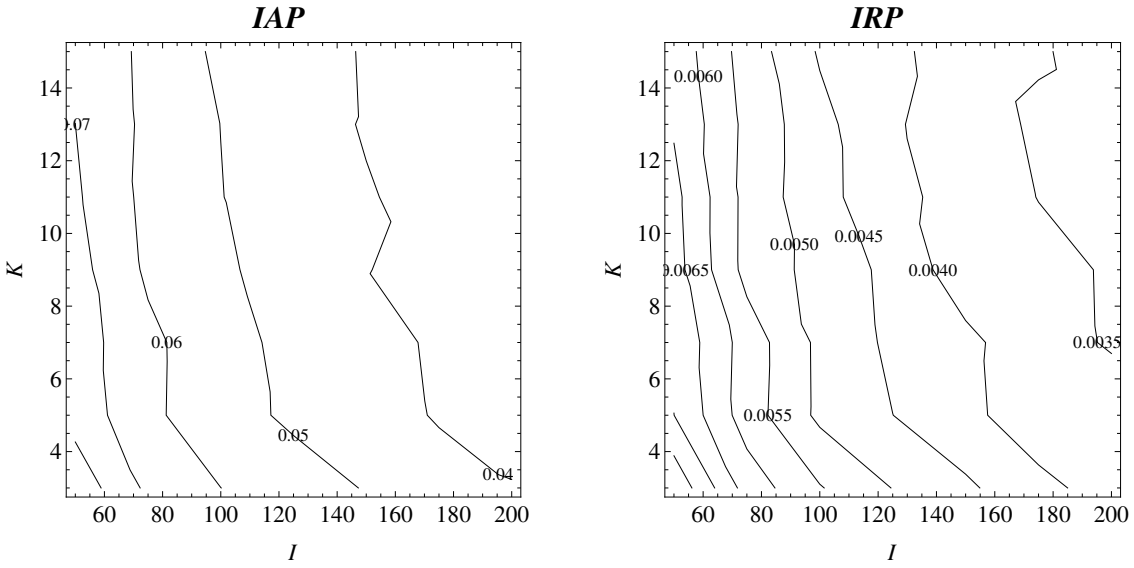

Figure 5.4a: Contour plots of the 95% confidence interval width for *IAP* and *IRP* as a function of *I* and *K*, for α=5, δ=2, *IAP*=0.2154, *IRP*=0.0125, *R*=2500.

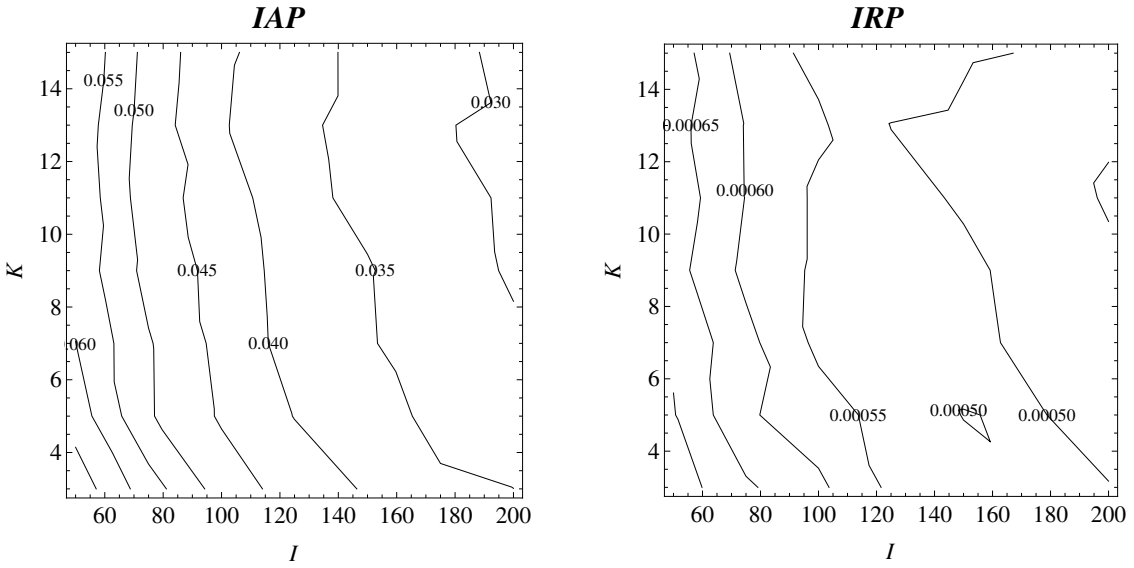

Figure 5.4b: Contour plots of the 95% confidence interval width for *IAP* and *IRP* as a function of *I* and *K*, for α=5, δ=3, *IAP*=0.2562, *IRP*=0.0015, *R*=2500.

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 $\widehat{IAP}$ and $\widehat{IRP}$. Figures 5.4a through 5.4d contain contour plots of the empirical

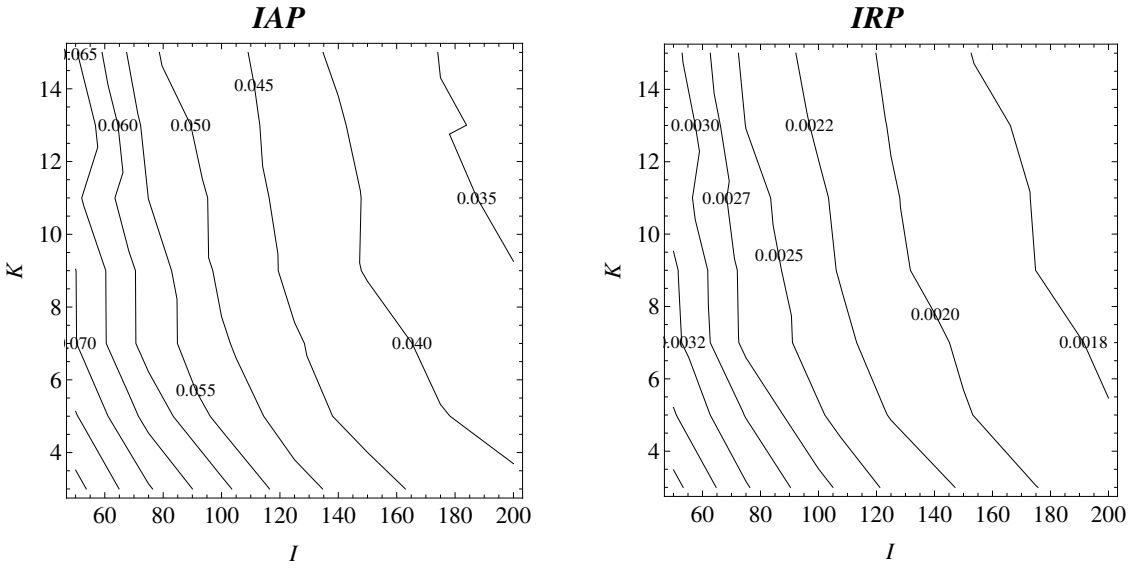

Figure 5.4c: Contour plots of the 95% confidence interval width for *IAP* and *IRP* as a function of *I* and *K*, for α=12, δ=2, *IAP*=0.1134, *IRP*=0.0039, *R*=2500.

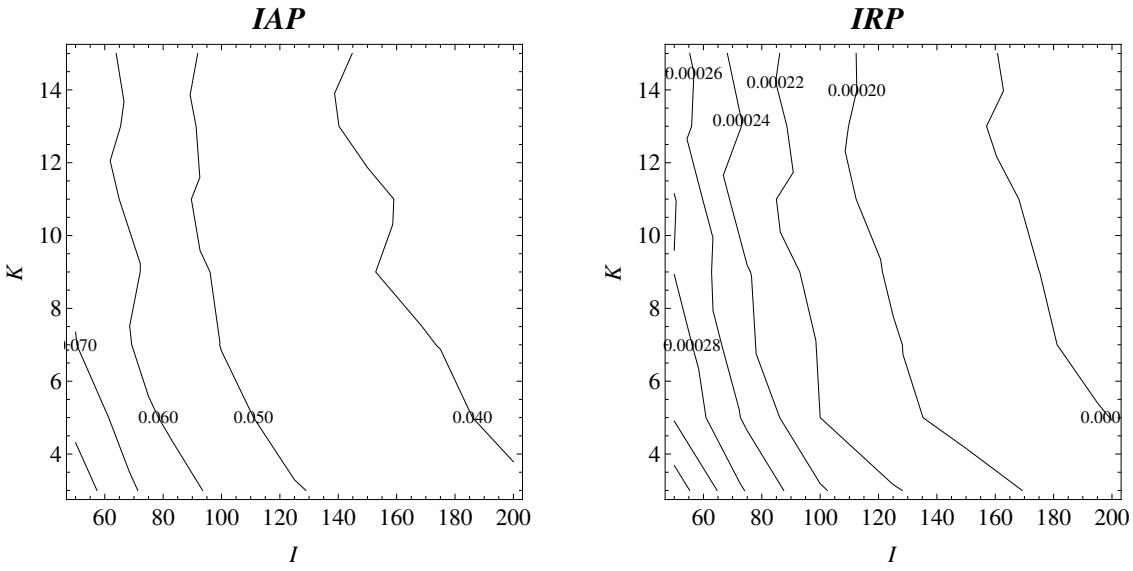

Figure 5.4d: Contour plots of the 95% confidence interval width for *IAP* and *IRP* as a function of *I* and *K*, for α=12, δ=3, *IAP*=0.1447, *IRP*=0.0004, *R*=2500.

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 ($I = I^{rej}$) and a historical dataset of $I^{his} = 100{,}000$ items. The contour plots are drawn using linear interpolation based on all 7×7 combinations of $K = 3,\ 5,\ ...,\ 15$ and $I = 50,\ 75,\ ...,\ 200$. The figures show that, for all

four combinations of parameter values, the marginal effects of both I and K on the precision of $\widehat{IAP}$ and $\widehat{IRP}$ 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 $\widehat{IAP}$ and $\widehat{IRP}$ 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 χ^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, α and δ 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 α and δ 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 $F_X = \Phi$

I^{rej}=200,	$X \sim N(0,1)$		$X \sim t(3)$		$X \sim t(7)$		$X \sim logN(0,1)$		$X \sim \chi^2(1)$	
I^{his}=100,000	*IAP*	*IRP*	*IAP*	*IRP*	*IAP*	*IRP*	*IAP*	*IRP*	*IAP*	*IRP*
%*GRR*=0.15, $P(Y=0)$=0.005	**α=13.7, δ=2.60**		**α=6.04, δ=5.87**		**α=10.1, δ=3.52**		**α=5.40, δ=13.2**		**α=8.96, δ=7.89**	
True value	0.1194	0.0009	0.0489	0.0003	0.0789	0.0005	0.0266	0.0002	0.0398	0.0002
Bias	0.0003	0.0000	0.0013*	0.0000*	0.0014*	0.0000*	0.0080*	0.0001*	0.0010*	0.0000
Width C.I.	0.0336	0.0004	0.0272	0.0002	0.0352	0.0003	0.0254	0.0001	0.0203	0.0001
%*GRR*=0.15, $P(Y=0)$=0.01	**α=13.7, δ=2.35**		**α=6.04, δ=5.48**		**α=10.1, δ=3.02**		**α=5.40, δ=10.3**		**α=8.96, δ=6.65**	
True value	0.1120	0.0016	0.0589	0.0008	0.0824	0.0012	0.0312	0.0004	0.0404	0.0005
Bias	0.0006	0.0000	0.0010*	0.0001*	0.0009*	0.0001*	0.0011*	0.0000	0.0004	0.0000
Width C.I.	0.0342	0.0007	0.0293	0.0004	0.0318	0.0006	0.0239	0.0003	0.0239	0.0003
%*GRR*=0.30, $P(Y=0)$=0.005	**α=6.85, δ=2.67**		**α=3.02, δ=5.96**		**α=5.04, δ=3.60**		**α=2.70, δ=13.2**		**α=4.48, δ=7.93**	
True value	0.2012	0.0019	0.0875	0.0007	0.1366	0.0012	0.0504	0.0003	0.0744	0.0005
Bias	0.0001	0.0000	0.0041*	0.0001*	0.0062*	0.0001*	0.0006*	0.0000	0.0006*	0.0000*
Width C.I.	0.0340	0.0006	0.0346	0.0003	0.0377	0.0005	0.0455	0.0003	0.0298	0.0003
%*GRR*=0.30, $P(Y=0)$=0.01	**α=6.85, δ=2.41**		**α=3.02, δ=4.69**		**α=5.04, δ=3.10**		**α=2.70, δ=10.3**		**α=4.48, δ=6.68**	
True value	0.1912	0.0035	0.1030	0.0018	0.1422	0.0026	0.0583	0.0008	0.0753	0.0010
Bias	0.0001	0.0000	0.0070*	0.0002*	0.0065*	0.0002*	0.0007*	0.0001	0.0014*	0.0000*
Width C.I.	0.0343	0.0012	0.0378	0.0008	0.0362	0.0009	0.0304	0.0005	0.0335	0.0005

Table 5.1: Monte Carlo mean estimates and 95% confidence interval widths for *IAP* and *IRP* for α=5, δ=2, I=200, K=9, R=2500.

*Monte Carlo mean estimate significantly different from true value (*IAP*=0.2154, *IRP*=0.0125)

in the likelihood function. For each distribution F_X and for each combination of $P(Y = 0)$ and *%GRR*, Table 5.3 gives the absolute deviation of the Monte Carlo means of $\widehat{IAP}$ and $\widehat{IRP}$ from their population values and the widths of empirical 95% confidence intervals. The population values *IAP* and *IRP* are calculated as

$$IAP = \int_{\delta}^{\infty} (1 - q(x)) f_X(x) dx \Big/ \int_{\delta}^{\infty} f_X(x) dx,$$
$$IRP = \int_{-\infty}^{\delta} q(x) f_X(x) dx \Big/ \int_{-\infty}^{\delta} f_X(x) dx.$$

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.

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., $\mathbf{X} \in \mathbb{R}^M$) with joint probability density function $f_{\mathbf{X}}(\mathbf{x})$. 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:

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

where we assume that each of the $q_m(x_m)$ is defined by the logit function (2) with parameters α_m and δ_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

$$IAP = P(Y = 1 \mid \overline{\mathbf{X} \le \boldsymbol{\delta}}) = \frac{P(Y = 1) - P(Y = 1, \mathbf{X} \le \boldsymbol{\delta})}{1 - P(\mathbf{X} \le \boldsymbol{\delta})}$$

$$= \frac{\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} (1 - q(\mathbf{x})) f_{\mathbf{X}}(\mathbf{x}) dx_1 \ldots dx_m - \int_{-\infty}^{\delta_1} \ldots \int_{-\infty}^{\delta_m} (1 - q(\mathbf{x})) f_{\mathbf{X}}(\mathbf{x}) dx_1 \ldots dx_m}{1 - \int_{-\infty}^{\delta_1} \ldots \int_{-\infty}^{\delta_m} f_{\mathbf{X}}(\mathbf{x}) dx_1 \ldots dx_m},$$

$$IRP = P(Y = 0 \mid \mathbf{X} \le \boldsymbol{\delta}) = \frac{\int_{-\infty}^{\delta_1} \ldots \int_{-\infty}^{\delta_m} (1 - q(\mathbf{x})) f_{\mathbf{X}}(\mathbf{x}) dx_1 \ldots dx_m}{\int_{-\infty}^{\delta_1} \ldots \int_{-\infty}^{\delta_m} f_{\mathbf{X}}(\mathbf{x}) dx_1 \ldots dx_m},$$

where $\boldsymbol{\delta}$ is the vector of decision thresholds δ_m, and $\overline{\mathbf{X} \le \boldsymbol{\delta}}$ denotes the complement of the event $\mathbf{X} \le \boldsymbol{\delta}$. 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:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{+} = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 1 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 1 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}^{-} = \begin{pmatrix} 1 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & 1 & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & 1 & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 1 \end{pmatrix}.$$

For the four curves $q_m(x_m)$, we consider the same parameters as chosen in Table 5.3 for the normally distributed measurand: (α_1, δ_1) = (13.7, 2.60), (α_2, δ_2) = (13.7, 2.35), (α_3, δ_3) = (6.85,

I^{rej}=200,	$\mathbf{X} \sim N(0,\Sigma)$		$\mathbf{X} \sim N(0,\Sigma^{+})$		$\mathbf{X} \sim N(0,\Sigma^{-})$	
I^{his}=100,000	*IAP*	*IRP*	*IAP*	*IRP*	*IAP*	*IRP*
True value	0.1489	0.0074	0.1438	0.0075	0.1506	0.0081
Bias	0.0049*	0.0005*	0.0072*	0.0006*	0.0049*	0.0005*
Width C.I.	0.0378	0.0027	0.0365	0.0025	0.0378	0.0027

Table 5.4: Absolute deviations of Monte Carlo mean estimates from true value and 95% confidence interval widths for *IAP* and *IRP* 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

2.67) and $(\alpha_4,\ \delta_4) = (6.85,\ 2.41)$. Table 5.4 shows bias and precision for this form of misspecification. Both bias and precision are acceptable for all three cases considered: the absolute value of the bias is less than 0.008 for *IAP* and less than 0.0007 for *IRP*, and the width of the confidence interval is less than 0.038 for *IAP* and less than 0.0028 for *IRP*. This indicates that the approach for binary MSA described in this chapter may still perform reasonably well if items are inspected on several properties simultaneously. Alternatively, separate MSA studies could be performed for the inspections on each of the distinct properties.

5.6 Summary and conclusions

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 *IAP* and *IRP*.

Simulations show that the estimators $\widehat{IAP}$ and $\widehat{IRP}$ 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 $P(X \le USL \mid Y = 0)$ 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 $x \to \frac{x}{1-x}$ 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 $P(X \le USL \mid Y = 0) = \frac{FRP(1-P(X > USL))}{FRP(1-P(X > USL)) + (1-FAP)P(X > USL)} > 0.5$, as claimed.

6 Current state of affairs and outlook to the future

In this final chapter of the thesis, I summarize its main practical conclusions, and formulate a vision on the current state of affairs as well as an outlook to the future.

For a long time most of the attention in the literature on MSA in industry has been given to MSA for interval and ratio scale measurements. The complexities of MSA for ordinal and nominal scale measurements, and binary measurements in particular, have long been overlooked. Over the past decade, a number of publications in the literature on binary MSA, from the statistics groups at Waterloo University and the University of Amsterdam in particular, have uncovered that the assessment of binary measurement systems is not as straightforward as it may seem. A number of these complicating issues, specifically for binary MSA, have been studied in detail in the previous chapters.

Unfortunately, both industrial standards and practitioners largely ignore these issues in their recommendations and practices. The most popular methods prescribed and applied for binary MSA seem to be the calculation of the κ index based on a contingency table of repeated measurements, and of *FAP* and *FRP* based on sample proportions. Based on the studies in the previous chapters, we formulate a number of practical recommendations and issues.

1) In the application of MSA methods, it is important that the population of items is defined. The κ index and also, in case of a continuous measurand, *FAP* and *FRP* are dependent on the distribution of the measurand in the items population, and therefore, inferences about κ or *FAP* and *FRP* are meaningless without a well-defined reference population of items (Chapters 2 and 4). Also, in the scientific study of methods for binary MSA, it is important that a population model is formulated as a frame of reference for inference, and that indices are defined not only as sample statistics, but also as population parameters (Chapter 4).

2) Estimation of *FAP* and *FRP* from sample proportions requires that the empirical property under study (the measurand) is well understood and operationally defined, and that a reliable gold standard assessment is available for the items in the sample (Chapter 2). In many practical situations, the measurand is poorly understood, and a reliable gold standard is substituted by the assessments of an "experienced" or "senior" operator. In our view, in situations where the measurand cannot reliably be represented by a gold standard, the measurand should be treated as a latent variable. Industrial standards do not (yet) mention latent variable models for MSA studies, and therefore such models are only rarely used in industry.

3) The κ index is not recommendable for pass/fail inspections in industry, as it mainly reflects the producer's risk and not the consumer's risk. Other problems are the typically very large standard errors, and the difficulty of interpretation of its value (Chapter 4), although the latter two problems can be resolved by using κ^{Unif} or P_A.

4) Most methods require a random sample of items. But on the other hand, a sample containing a fair number of defective items is needed to avoid excessive standard errors of the estimates. Given the very low defect rates common in industrial processes, the combination of these two requirements results in impractically large sample sizes. Therefore, in many practical instances, MSA studies are performed with nonrandom samples, leading to biased estimates (Chapters 2 and 4).

5) For most binary measurements, the measurand is not dichotomous but continuous. If such a continuous measurand is treated as dichotomous, nonrandom samples generally lead to (possibly seriously) biased estimates of error rates (Chapter 2).

6) If the measurand is treated as continuous, it is usually modeled as a normally distributed variable, whereas in reality, the measurand often has an asymmetric distribution, and also the characteristic curve is typically not symmetrical (Chapter 3).

The difficulties described above illustrate that the design and analysis of binary MSA experiments is more complex than one might expect. Below, we characterize the current state of affairs, aiming to specify for which situations effective techniques are available, and which situations need further research.

In the situation that a gold standard is available, by now effective techniques are available. In particular, estimation of the characteristic curve by logistic regression is a recommended option for continuous measurands (Chapter 2) or hybrid measurands (Chapter 3). Nonparametric estimation of *FAP* and *FRP* on the basis of a study design referred to as Plan I (Chapter 2) seems a reliable technique whether the measurand is binary or continuous (Chapter 2), although recommendations 1), 2) and 5) above should be kept in mind.

In the situation that a gold standard is unavailable, the current state of knowledge is far less satisfactory. The general idea is to use latent variable models, such as a latent class model or a latent trait model. In the past decade, there has been increasing attention to latent variable models for binary MSA in the literature (Boyles, 2001; Van Wieringen and De Mast 2008; Danila et al., 2010; De Mast and Van Wieringen, 2010; Danila et al., 2012; and Chapters 2, 4 and 5 of this thesis). However, estimation techniques based on latent variable models suffer from stubborn problems:

7) Generally, a random sample of items is required, but this typically leads to a sample with a very small number of defective items. If nonrandom samples are taken, this needs to be modeled appropriately (Danila et al., 2012; Chapter 5). This is a topic of ongoing research.

8) For latent variable models, the assumption of conditional independence is crucial. Unfortunately, this assumption is easily violated in practice. For example, if a continuous measurand is treated as binary, the assumption of conditional independence is generally violated (Chapter 2).

9) Although our first robustness studies in Chapter 5 suggest that inferences are reasonably robust against some forms of model misspecification, more extensive studies are needed, in view of the complexity of the models.

10) Aggravating the problems of violation of conditional independence and possible non-robustness mentioned under 8) and 9), it is not clear whether effective techniques for diagnosing model misspecifications can be developed.

11) Currently proposed techniques for binary MSA based on latent variable modeling have not been applied much in practice yet, and consequently, there is virtually no evidence base for their usefulness in practice.

Traditional methods for the gold standard unavailable situation are based on a latent class model (e.g., Boyles, 2001; Van Wieringen and De Mast, 2008), which makes the rather precarious assumptions that the measurand is dichotomous and that conditional independence holds. Current research explores two directions for relaxing these assumptions. One direction is exemplified by the work in Chapter 5 of this thesis, where the stochastic properties of the measurements are explicitly attributed to the stochastic properties of a continuous measurand by means of a characteristic curve. The other direction, as in Danila et al. (2012), explores the use of a random effects model, which aggregates the effects of the underlying properties and other (nuisance) variables into a flexible probability distribution for the error rates, typically a beta distribution. Both approaches still require quite some research in order to investigate in which situation they are suitable and they need to be tried and tested in practice.

Observations 7) through 11) motivate what we consider is a crucial principle in the design and analysis of MSA studies for binary measurements. This principle is:

12) The design of MSA studies for binary measurements is more straightforward, and the analysis and interpretation of their results more reliable, to the extent that the measurand is better understood and defined in operational terms.

If the measurand is poorly understood, it is unclear what empirical properties the measurements reflect, and consequently, measurement degenerates into mere classification, and the concept of measurement error becomes meaningless. Failure to properly accommodate the measurand in the models used to understand agreement and the κ index have resulted, in our view, in much of the confusion over their interpretation (Chapter 4). Failure to define the measurand in operational terms appears to us as an important cause for the unavailability of a gold standard, and consequently, the resort to latent variable modeling and its associated problems. Before practitioners resort too quickly to methods based on latent variable modeling, they are well advised, in our opinion, to go quite some way in trying to improve understanding of the measurand, and developing an operational definition and thus a gold-standard measurement.

Finally, we enumerate the steps that we think are needed in the immediate future to advance the field of binary MSA.

- Practical guidelines and advice, as framed in our principle 12) and recommendations 1) through 6), should be incorporated in industrial MSA standards.
- Recent advances and new techniques, especially for the gold standard available situation, should be incorporated in industrial MSA standards and implemented in statistical software packages.
- Techniques for the gold standard unavailable situation should be further studied, using simulation studies and field trials, in order to gain better understanding of their usefulness.

References

AIAG (2003). *Measurement systems analysis: Reference manual* (3rd ed.). Detroit, MI: Automotive Industry Action Group.

Akkerhuis, T.S. (2012). *Bias and precision of error rate estimates of binary measurement systems*. MSc Thesis, University of Amsterdam.

Allen, M.J., and Yen, W.M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Awad, M., Erdmann, T.P., Shanshal, Y., and Barth, B. (2009). "A measurement system analysis approach for hard-to-repeat events". *Quality engineering* 21, pp. 300-305.

Beavers, D.P., Stamey, J.D., and Bekele, B.N. (2011). "A Bayesian model to assess a binary measurement system when no gold standard system is available". *Journal of Quality Technology* 43, pp. 16-27.

Bennett, E.M., Alpert, R., and Goldstein, A.C. (1954). "Communications through limited response questioning". *Public Opinion Quarterly* 18, pp. 303-308.

Berger, T. (1988). "Information theory and coding theory". In: Kotz, S., and Johnson, N. (eds.), *Encyclopedia of Statistical Sciences* (8th ed., Vol. 5). New York, NY: Wiley.

Bloch, D.A., and Kraemer, H.C. (1989). "2 x 2 kappa coefficients: Measures of agreement or association". *Biometrics* 45, pp. 269-287.

Boyles, R.A. (2001). "Gauge capability for pass-fail inspection". *Technometrics* 43, pp. 223-229.

Brennan, R.L., and Prediger, D.J. (1981). "Coefficient kappa: Some uses, misuses, and alternatives". *Educational and Psychological Measurement* 41, pp. 687-699.

Burdick, R.K., Borror, C.M., Montgomery, D.C. (2003). "A review of methods for measurement systems capability analysis". *Journal of Quality Technology* 35, pp. 342-354.

Byrt, T., Bishop. J, and Carlin, J.B. (1993). “Bias, prevalence and kappa”. *Journal of Clinical Epidemiology* 46, pp. 423-429.

Cicchetti, D.V., and Feinstein. A.R. (1990). “High agreement but low kappa: II. Resolving the paradoxes”. *Journal of Clinical Epidemiology* 43, pp. 551-558.

Cicchetti, D.V., and Sparrow, S.S. (1981). “Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior”. *American Journal of Mental Deficiency* 86, pp. 127-137.

Cohen, J. (1960). “A coefficient of agreement for nominal scales”. *Educational and Psychological Measurement* 20, pp. 37-46.

Conger, A.J. (1980). “Integration and generalization of kappas for multiple raters”. *Psychological Bulletin* 88, pp. 322-328.

Danila, O., Steiner, S.H., and MacKay, R.J. (2008). “Assessing a binary measurement system”. *Journal of Quality Technology* 40, pp. 310-318.

Danila, O., Steiner, S.H., and MacKay, R.J. (2010). “Assessment of a binary measurement system in current use”. *Journal of Quality Technology* 42, pp. 152-164.

Danila, O., Steiner, S.H., and MacKay, R.J. (2012). “Assessing a binary measurement system with varying misclassification rates”. *Journal of Quality Technology* 44, pp. 179-191.

Davies, M., and Fleiss, J.L. (1982). “Measuring agreement for multinomial data”. *Biometrics* 38, pp. 1047-1051.

De Boor, C. (1978). *A Practical Guide to Splines.* New York, NY: Springer Verlag.

De Mast, J. (2007). “Agreement and kappa type indices”. *The American Statistician* 61, pp. 148-153.

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

De Mast, J., and Trip, A. (2005). “Gauge R&R studies for destructive measurements”. *Journal of Quality Technology* 37, pp. 40-49.

De Mast, J., and Van Wieringen, W.N. (2004). “Measurement system analysis for bounded ordinal data”. *Quality and Reliability Engineering International* 20, pp. 383-395.

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

De Mast, J., and Van Wieringen, W.N. (2010). "Modeling and evaluating repeatability and reproducibility of ordinal classifications". *Technometrics* 52, pp. 94-106.

Eisenhart, C. (1968). "Expression of the uncertainties of final results". *Science* 160, pp. 1201-1204.

Embretson, S.E., and Reise, S.P. (2000). *Item response theory for psychologists*. London, UK: Law Erlbaum Associates.

Erdmann, T.P., Akkerhuis, T.S., De Mast, J., and Steiner, S.H. (2012). "Binary measurement system analysis with a latent continuous measurand". *Manuscript in preparation.*

Erdmann, T.P., and De Mast, J. (2012). "Assessment of binary inspection with a hybrid measurand". *Quality and Reliability Engineering International* 28, pp. 47-57.

Erdmann, T.P., De Mast, J., and Warrens, M.J. (in press). "Some common errors of experimental design, interpretation and inference in agreement studies". *Statistical Methods in Medical Research*.

Erdmann, T.P., Does, R.J.M.M., Bisgaard, S. (2009). "Quality quandaries: A gage R&R study in a hospital". *Quality Engineering* 22, pp. 46-53.

Farnum, N.R. (1994). *Modern statistical quality control and improvement*. Belmont, CA: Duxbury Press.

Feinstein, A.R., and Cicchetti, D.V. (1990). "High agreement but low kappa: I. The problems of two paradoxes". *Journal of Clinical Epidemiology* 43, pp. 543-549.

Fleiss, J.L. (1971). "Measuring nominal scale agreement among many raters". *Pychological Bulletin* 76, pp. 378-382.

Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969). "Large sample standard errors of kappa and weighted kappa". *Psychological Bulletin* 72, pp. 323-337.

Fleiss, J.L., Levin, B., and Paik, M.C. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York, NY: Wiley.

Fletcher, R. (1970). "A new approach to variable metric algorithms". *The Computer Journal* 13, pp. 317-322.

Food and Drug Administration (2007). *Guidance for industry and FDA staff – Statistical guidance on reporting results from studies evaluating diagnostic tests.* Rockville, MD: U.S. Department of Health and Human Services.

Gilula, Z., and Haberman, S.J. (1995). "Dispersion of categorical variables and penalty functions: Derivation, estimation, and comparability". *Journal of the American Statistical Association* 90, pp. 1447-1452.

Grove, W.M., Andreasen, N.C., McDonald-Scott, P., Keller, M.B., and Shapiro, R.W. (1981). "Reliability studies of psychiatric diagnosis: Theory and practice". *Archives of General Psychiatry* 38, pp. 408-413.

Hand, D.J. (1996). "Statistics and the theory of measurement". *Journal of the Royal Statistical Society, Series A* 159, pp. 445-492.

Hastie, T., and Tibshirani R. (1986). "Generalized additive models". *Statistical Science* 3, pp. 297-310.

Hastie, T., and Tibshirani R. (1990). *Generalized additive models*. London, UK: Chapman & Hall.

Hershberger, S.L., and Fisher, D.G. (2005). "Measures of association". In: Everitt, B., and Howell, D. (eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1183-1192). Chichester, UK: Wiley.

Hosmer, D.W., and Lemeshow, S. (1989). *Applied logistic regression* (2nd ed.). New York, NY: Wiley.

Hripcsak, G., and Heitjan, D.F. (2002). "Measuring agreement in medical informatics reliability studies". *Journal of Biomedical Informatics* 35, pp. 99-110.

Hui, S.L., and Walter, S.D. (1980). "Estimating the error rates of diagnostic tests". *Biometrics* 36, pp. 167-171.

ISO (1995). *Guide to the expression of uncertainty in measurement* (1st ed.). Geneva, Switzerland: International Organization for Standardization.

ISO 9000:2005 (2005). *Quality management systems – Fundamentals and vocabulary*. Geneva, Switzerland: International Organization for Standardization.

ISO 9001:2008 (2008). *Quality management systems – Requirements*. Geneva, Switzerland: International Organization for Standardization.

ISO/TS 16949:2009 (2009). *Quality management systems – Particular requirements for the application of ISO 9001:2008 for automotive production and relevant service part organizations.* Geneva, Switzerland: International Organization for Standardization.

Jenkinson, A.F. (1955). "The frequency distribution of the annual maximum (or minimum) of meteorological elements". *Quarterly Journal of the Royal Meteorological Society* 81, pp. 158-171.

Kerlinger, F.N., and Lee, H.B. (2000). *Foundations of behavioral research* (4th ed.). New York, NY: Harcourt College Publisher.

Kimothi, S.K. (2002). *The uncertainty of measurements: Physical and chemical metrology: Impact and analysis*. Milwaukee, WI: ASQ Quality Press.

Kraemer, H.C. (1979). "Ramifications of a population model for κ as a coefficient of reliability". *Psychometrika* 44, pp. 461-472.

Kraemer, H.C., Periyakoil, V.S., and Noda, A. (2002). "Kappa coefficients in medical research". *Statistics in Medicine* 21, pp. 2109–2129.

Landis, J.R., and Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics* 33, pp. 159-174.

Lindley, D.V., and Novick, M.R. (1981). "The role of exchangeability in inference". *The Annals of Statistics* 9, pp. 45-58.

Lord, F.M., and Novick, M.R. (1968). *Statistical theories of mental test scores.* Oxford, UK: Addison-Wesley.

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

Mawby, W.D. (2006). *Make your destructive, dynamic, and attribute measurement system work for you*. Milwaukee, WI: ASQ Quality Press.

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

Meining, A., Rösch, T., Kiesslich, R., Muders, M., Sax, F., and Heldwein, W. (2004). "Inter- and intra-observer variability of magnification chromoendoscopy for detecting specialized intestinal metaplasia at the gastroesophageal junction". *Endoscopy* 36, pp. 160-164.

Montgomery, D.C., and Runger, G.C. (1993a). "Gauge capability and designed experiments. Part I: Basic methods". *Quality Engineering* 6, pp. 115-135.

Montgomery, D.C., and Runger, G.C. (1993b). "Gauge capability and designed experiments. Part II: Experimental design methods and variance component estimation". *Quality Engineering* 6, pp. 289-305.

Naranjo, C.A., Busto, U., Sellers, E.M., Sandor, P., Ruiz, I., Roberts, E.A., et al. (1981). "A method for estimating the probability of adverse drug reactions". *Clinical Pharmacology and Therapeutics* 30, pp. 239-245.

O'Keefe, S.T., Smith, T., Valacio, R., Jack, C.I.A., Playfer, J.R., and Lye, M. (1994). "A comparison of two techniques for ankle jerk assessment in elderly subjects." *Lancet* 344, pp. 1619-1620.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Oxford, UK: Oxford University Press.

Piessens, R. (1983). *Quadpack: A subroutine package for automatic integration.* Berlin, Germany: Springer-Verlag.

Silverman, B.W. (1985). "Some aspects of the spline smoothing approach to non-parametric regression curve fitting". *Journal of the Royal Statistical Society. Series B* 47, pp. 1-52.

Spiteri, M.A., Cook, D.G., and Clarke, S.W. (1988). "Reliability of eliciting physical signs in examination of the chest". *Lancet* 331, pp. 873-875.

Stevens, S.S. (1946). "On the theory of scales of measurement". *Science* 103, pp. 677-680.

Tanner, M.A., and Young, M.A. (1985). "Modeling agreement among raters". *Journal of the American Statistical Association* 80, pp. 175-180.

Thompson, W.D., and Walter, S.D. (1988). "A reappraisal of the kappa coefficient". *Journal of Clinical Epidemiology* 41, pp. 949-958.

Uebersax, J.S. "Diversity of decision-making models and the measurement of interrater agreement". *Psychological Bulletin* 101, pp. 140-146.

Van Wieringen, W.N. (2005). "On identifiability of certain latent class models". *Statistics and Probability Letters* 75, pp. 211-218.

Van Wieringen, W.N., and De Mast, J. (2008). "Measurement system analysis for binary data". *Technometrics* 50, pp. 468-478.

Van Wieringen, W.N., and Van den Heuvel, E.R. (2005). "A comparison of methods for the evaluation of binary measurement systems". *Quality Engineering* 17, pp. 495-507.

Vardeman, S.B., and Van Valkenburg, E.S. (1999). "Two-way random effects analyses and gauge R&R studies". *Technometrics* 41, pp. 202-211.

Wallsten, T.S. (1988). "Measurement theory". In: Kotz, S., and Johnson, N. (eds.), *Encyclopedia of Statistical Sciences* (8th ed., Vol. 5). New York, NY: Wiley.

Warrens, M.J. (2010). "A formal proof of a paradox associated with Cohen's kappa". *Journal of Classification* 27, pp. 322-332.

Weisz, J.R., Jensen Doss, A.J., and Hawley, K.M. (2005). "Youth psychotherapy outcome research: A review and critique of the evidence base". *Annual Review of Psychology* 56, pp. 337-363.

Zelterman, D. (1987). "Parameter estimation in the generalized logistic distribution". *Computational Statistics and Data Analysis* 5, 177-184.

Samenvatting

Een binaire meting classificeert objecten in twee categorieën, bijvoorbeeld goedkeur of afkeur, met het doel daarmee een empirische eigenschap van de objecten te weerspiegelen, bijvoorbeeld of zij goed of defect zijn. Deze onderliggende empirische eigenschap, de *werkelijke waarde*, wordt in de metrologie de *te meten grootheid* (Engels: *measurand*) genoemd. De te meten grootheid kan zelf binair zijn, zoals de staat van een gloeilamp: werkend of defect. In veel gevallen is zij echter een continue eigenschap, bijvoorbeeld de mate van verkleuring van een voedselproduct: een continuüm dat varieert van nauwelijks verkleuring tot zeer sterke verkleuring.

Binaire metingen zijn, net als alle metingen, onderhevig aan *meetfout*: de mate waarin de meting erin slaagt of faalt de te meten grootheid weer te geven. Het experimenteel onderzoeken en evalueren van meetsystemen noemt men in de industriële statistiek meetsysteemanalyse (MSA). Meetsysteemanalyse voor binaire metingen is van hetzelfde belang als voor andere typen metingen. Industriestandaards voor kwaliteitsbeheersing eisen dat de betrouwbaarheid van meet- en testresultaten wordt onderzocht en beheerst. Daarnaast leidt het voorkomen van onterechte afkeur tot kostenbesparingen, en het voorkomen van onterechte goedkeur tot betrouwbaardere kwaliteit. Ook buiten de industrie is binaire meetsysteemanalyse belangrijk, zoals in vakgebieden als geneeskunde en psychologie, waar het dan bijvoorbeeld gaat om de betrouwbaarheid van diagnostische tests.

De kwaliteit van binaire metingen kan op verschillende manieren worden uitgedrukt. De kans dat een defect object wordt goedgekeurd is de *kans op onterechte goedkeur*, en de kans dat een goed object wordt afgekeurd is de *kans op onterechte afkeur*. Als de te meten grootheid continu is, is het vaak wenselijk om de kans op afkeur te kennen voor elke waarde van de te meten grootheid. De kans op afkeur kan dan worden weergegeven in een grafiek, die de *karakteristieke curve* wordt genoemd. Een alternatieve maat voor de kwaliteit van binaire metingen is de *kappa-index*, gebaseerd op de *kans op overeenstemming* (Engels: *agreement*), de kans dat twee metingen aan hetzelfde object overeenkomen.

Recentelijk is een aantal artikelen verschenen over MSA voor binaire metingen, waarin methoden worden beschreven voor het bepalen van de kans op onterechte afkeur en goedkeur, de kappa-index, en de kans op overeenstemming. Sommige methoden

veronderstellen dat een zogenaamde *referentiewaarde* of *gouden standaard* (Engels: *gold standard*) beschikbaar is, een meetprocedure van hogere orde die de waarde van de te meten grootheid (vrijwel) exact kan bepalen, en sommige niet. Daarnaast modelleren sommige methoden de te meten grootheid als een binaire variabele en sommige als een continuüm. Verder vereisen sommige methoden een aselecte steekproef van objecten, terwijl andere methoden afwijkende steekproefstrategieën hanteren. Toch bestaat er voor verschillende veelvoorkomende situaties nog geen geschikte methode.

In dit proefschrift worden bestaande methoden voor binaire MSA vergeleken en onderzocht, om inzicht te krijgen in hun effectiviteit. Daarbij worden praktische richtlijnen gegeven voor het gebruik van de verschillende methoden. Voor een aantal situaties waarin momenteel geen geschikte methode bestaat, worden nieuwe methoden geïntroduceerd.

In hoofdstuk 2 onderzoeken en vergelijken wij bestaande methoden voor het beoordelen van de kwaliteit van binaire metingen. Ons raamwerk introduceert twee factoren die zeer relevant zijn om te beslissen welke methoden te gebruiken: (1) of een referentiewaarde beschikbaar is, en (2) of de onderliggende te meten grootheid continu is of werkelijk binair. Het kunstmatig dichotomiseren van een continue grootheid – wat vaak gedaan wordt – creëert complicaties die onderschat worden in zowel de wetenschappelijke literatuur als de praktijk. Meer in het bijzonder introduceert dit een intrinsieke oorzaak voor het schenden van de veelgemaakte aanname dat herhaalde metingen stochastisch onafhankelijk zijn geconditioneerd op de te meten grootheid. Voor de meeste methoden is dit niet cruciaal, mits steekproeven aselect zijn (of op zijn minst representatief). Echter, voor de meeste methoden is het in het algemeen onduidelijk hoe men een aselecte steekproef kan verkrijgen uit de relevante populatie. De taxonomie in hoofdstuk 2 presenteert methoden die algemeen bekend zijn in de industrie, zoals niet-parametrische schatting van de kansen op onterechte goedkeur en afkeur, de zogenaamde *analytic method* van de *Automotive Industry Action Group (AIAG)* en latente-variabele-modellen. De onderzochte methoden worden toegepast op een voorbeeld uit de handleiding voor meetsysteemanalyse van de AIAG, een belangrijke industriestandaard.

Hoofdstuk 3 behandelt complicaties die ontstaan bij meetsysteemanalyse voor binaire metingen waarvan de te meten grootheid hybride is: een grootheid die zowel kenmerken van een dichotomie als van een continuüm heeft. Aan de hand van een casus worden methoden verkend voor deze veelvoorkomende situatie. De casus betreft visuele inspectie van laptopschermen op krassen, waarbij de te meten grootheid de aan- of afwezigheid van krassen is. De meetuitkomst “goedkeur” moet nu corresponderen met een punt (geen kras), maar

"afkeur" correspondeert met een heel continuüm van kleine, ondiepe krasjes tot grote, diepe krassen. Wij laten zien, dat als de te meten grootheid hybride is, een eenvoudig logistisch regressiemodel ongeschikt is voor het schatten van de karakteristieke curve die het verband weergeeft tussen de te meten grootheid en de kans op afkeur. Verschillende alternatieve modelspecificaties voor de karakteristieke curve worden geïntroduceerd en vergeleken. Wij concluderen dat veel van de methoden die in deze situatie doorgaans worden toegepast, ongeschikt zijn. Dit is een opvallende conclusie, gezien de alomtegenwoordigheid in de industrie van lektesten, inspecties op defecten, en andere binaire meetsystemen waarvan de te meten grootheid hybride is. Wij verkennen daarnaast opties die wel toepasbaar zijn.

In hoofdstuk 4 signaleren en bespreken wij veelvoorkomende methodologische fouten bij het bepalen en interpreteren van de kans op overeenstemming en de kappa-index, aan de hand van studies die zijn gepubliceerd in de medische en sociale wetenschappen. Onze analyse is gebaseerd op een voorgesteld statistisch model dat in lijn is met typische modellen uit de metrologie. Een eerste type fouten is gerelateerd aan het gebruik van een onrepresentatieve steekproef, hetgeen kan leiden tot een aanzienlijke onzuiverheid in de geschatte kans op overeenstemming en kappa-index. Ten tweede is het gebruik van de kappa-index precair wanneer de prevalenties van de categorieën sterk van elkaar verschillen, aangezien de kappa-index dan zeer gevoelig is voor steekproeffluctuaties, resulterend in een grote standaardfout van de schatting. Bovendien weerspiegelt de index in dergelijke gevallen bijna uitsluitend de consistentie van de meest-voorkomende categorie. Een laatste type fouten betreft interpretatieproblemen, die kunnen leiden tot verkeerde conclusies op basis van de kappa-index. Deze interpretatieproblemen worden verhelderd aan de hand van het voorgestelde statistische model. De gesignaleerde fouten worden geïllustreerd met studies die zijn gepubliceerd in vooraanstaande wetenschappelijke tijdschriften. De analyse leidt tot een aantal richtlijnen en aanbevelingen, waaronder de aanbeveling om in bepaalde situaties alternatieven voor de kappa-index te gebruiken.

Hoofdstuk 5 introduceert een methode voor meetsysteemanalyse voor binaire metingen waarvan de te meten grootheid een niet-observeerbare continue eigenschap is. De te meten grootheid wordt gemodelleerd als een latente variabele, en de kwaliteit van de meting wordt uitgedrukt in termen van *kansen van inconsistente classificatie*. Verschillende steekproefstrategieën voor de objecten in het MSA-experiment worden verkend, met als doel de modelparameters zo precies mogelijk te schatten. Wij laten zien dat, als het defectpercentage laag is, het optimaal is om (een deel van de) objecten te selecteren uit de subpopulatie van afgekeurde objecten. Verder presenteren wij een schattingsmethode voor

het latente-variabele-model die rekening houdt met deze manier van steekproeftrekken. Door middel van een simulatiestudie onderzoeken wij de zuiverheid en precisie van de geschatte kansen van inconsistente classificatie voor verschillende steekproefstrategieën, de benodigde steekproefgrootten, en de robuustheid van de methode tegen modelmisspecificatie.

De bevindingen in de hoofdstukken 2 tot en met 5 van dit proefschrift motiveren een aantal conclusies en aanbevelingen, die zijn samengevat in hoofdstuk 6. De belangrijkste is het principe dat het ontwerpen van MSA-studies voor binaire metingen eenvoudiger wordt, en de analyse en interpretatie van de resultaten betrouwbaarder, naargelang er een beter begrip is van de te meten grootheid en deze beter operationeel gedefinieerd is.

Curriculum vitae

Tashi Erdmann werd geboren in 1983 te Amsterdam. Hij doorliep zijn middelbare-schoolopleiding aan het Vossius Gymnasium te Amsterdam. In 2001 zette hij zijn opleiding voort aan de Universiteit van Amsterdam, waar hij econometrie studeerde. Hij beëindigde zowel de bachelor- als de masteropleiding “cum laude”. Zijn afstudeerscriptie over financieel risicomanagement schreef hij tijdens een stage bij Watson Wyatt Insurance Consulting in 2006. Na zijn studie was hij twee jaar werkzaam als docent wiskunde op Keera-Pat International School in Bangkok, Thailand.

In 2008 trad Tashi in dienst van het Instituut voor Bedrijfs- en Industriële Statistiek van de Universiteit van Amsterdam (IBIS UvA) om daar training en advisering te combineren met wetenschappelijk onderzoek op het gebied van industriële statistiek. Bij IBIS UvA verrichtte hij onderzoek naar methoden voor meetsysteemanalyse, waarvan dit proefschrift het resultaat is. Als statistisch adviseur trainde hij daarnaast projectleiders in het bedrijfsleven in onderzoeksvaardigheden en leidde hen op in het gebruik van statistische methoden. Tevens begeleidde hij procesverbeterprojecten. Deze activiteiten vonden met name plaats in het kader van de invoering van een Lean Six Sigma-programma. Hij is als adviseur werkzaam geweest bij onder meer Ziggo, het Medisch Spectrum Twente, de Hogeschool van Arnhem en Nijmegen, TenneT en Eurocross Assistance. Aan de Universiteit van Amsterdam doceerde Tashi het mastervak Industriële Statistiek, onderdeel van de wiskundeopleiding, en was hij assistent-docent bij het vak Operations and Process Management in de bedrijfskundeopleiding.

In zijn vrije tijd doet Tashi intensief aan karate bij Nippon Karatedo Genwakai, waar wordt getraind op traditioneel Japanse wijze. Verder is hij geïnteresseerd in cinema, reist graag en veel, en spreekt zes talen, waaronder Thais.

Acknowledgements

This thesis is based on the research I have conducted with my supervisor Jeroen de Mast since 2008. I owe a lot to the people who contributed to the research and the thesis.

First of all, I would like to thank Jeroen, who guided me through the four-year journey that resulted in this thesis, allowing me to "sit in the back seat" at the beginning, and then supporting me to become increasingly independent over the years. Besides being the instigator of the research and fellow author of all publications on which this thesis is based, Jeroen contributed to its content through numerous discussions, suggestions and comments. The research profited tremendously from his creativity, conceptual thinking and writing skills, and from his focus, confidence and enthusiasm.

Next, I thank Ronald Does, the director of the Institute for Business and Industrial Statistics, for creating the opportunity to combine this research with consultancy in a dynamic and productive working environment, and for motivating me through his energetic leadership.

I would also like to thank Thomas Akkerhuis, who made a major contribution to Chapter 5, and the other co-authors of the papers this thesis is based on, Wessel van Wieringen, Matthijs Warrens and Stefan Steiner. Furthermore, I thank Rob Cuperus, who made it possible to apply the method of Chapter 5 in practice, and all participants of the experiment in Chapter 3 about inspection for scratches on a laptop screen. I thank Esther Ris for designing the cover. Also, I am indebted to my current and former colleagues at IBIS UvA, Atie Buisman, Marit Schoonhoven, Benjamin Kemper, Inez Zwetsloot, Alex Kuiper Henk de Koning, Sander Gerritsen, Joran Lokkerbol, Tjarko de Vree and Wouter Vink, for the pleasant working atmosphere.

I am grateful to my friends and family, and to the members of the Phoenix Juku clan of Nippon Karatedo Genwakai, especially sensei Igor van Vlodrop. Finally, my partner Bob, my parents Ferry and Tia, Tycho, Toom, Bernadette, Jorge, thank you for your love and support.

Tashi Erdmann

Amsterdam, September 2012