The statistical evaluation of binary tests without gold standard: Robustness of latent variable approaches

Thomas Akkerhuis a,*, Jeroen de Mast a, Tashi Erdmann b

a Department of Operations Management, Amsterdam Business School, University of Amsterdam, P.O. Box 15953, 1001 NL Amsterdam, The Netherlands
b Shell Technology Centre Amsterdam, P.O. Box 38000, 1030 BN Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 20 June 2016
Received in revised form 12 October 2016
Accepted 15 October 2016
Available online 17 October 2016

Keywords:
Binary measurement
Binary data
Precision
Reproducibility and repeatability
Pass/fail inspection

ABSTRACT

Binary tests classify items into two categories such as reject/accept or positive/negative. Such tests are usually evaluated in terms of their misclassification probabilities FAP (false acceptance probability) and FRP (false rejection probability). A common complication arises when there is no gold standard or reference standard. Various methods based on latent variable modelling have been proposed for this situation. We present the results of a simulation study in which these methods are tried in a range of scenarios, to study how robust they are to departures from the assumptions on which they are based.

The study convincingly shows that in general, the ambition of estimating FAP and FRP without gold standard is unattainable, since all methods easily derail when assumptions are not precisely met. The study also shows that the random components of the FAP and FRP can be reliably estimated by a straightforward modification of one of the tested methods.

1. Introduction

Binary tests are common in industrial processes, and classify items in two categories such as ‘reject’ (Y = 0) or ‘accept’ (Y = 1). Examples include visual quality inspections and automated tests where some parts fail and others pass. Diagnostic and screening tests in medicine, yielding the results ‘positive’ or ‘negative’, are closely related. We conceive of such tests as a form of measurement (‘binary measurement’, see Suzuki et al. [1]), and thus, these classifications aim to reflect an underlying true state X of the items called the measurand [2,3], which can be ‘truly defective’ (X = 0) or ‘conforming to specifications’ (X = 1).

A measurement system analysis (MSA) experiment is an experiment to evaluate how reliably the test results Y reflect the measurand X. Measurement error is generally defined as the discrepancy between measured and true value. Binary scales are equipped with only the simplest of algebraic structures and in particular, subtraction and addition are usually not meaningfully defined [4]. Consequently, it is problematic to define discrepancy in terms of a difference Y – X (or derived statistics such as standard deviations). For binary measurements, therefore, the measurement error is usually expressed as a misclassification (Y ≠ X), and a statistical evaluation is in terms of the misclassification probabilities:

\[ FAP = P(Y = 1 | X = 0) \] (False Acceptance Probability),

\[ FRP = P(Y = 0 | X = 1) \] (False Rejection Probability).

In medicine, one sometimes works with the complements of these probabilities: the sensitivity and specificity. Traditional methods for estimating the FAP and FRP (as described in AIAG [5], Pepe [6], Danila et al. [7,8] and many other articles), require a so-called gold standard or reference standard: a higher-order, authoritative test that is accepted to constitute a faithful representation of the measurand [9]. A common problem occurs when such gold standard is not available, in which case the true state of the items is practically unobservable. Reasons for this include the absence of a sufficiently capable authoritative test, ambiguity of the specifications (such as when human perception is involved), prohibitive cost, or damage resulting from applying the gold standard to the tested items.

When a gold standard is unavailable, the traditional solution is to fit a latent class model, in which FAP and FRP are assumed constant in the subpopulations of defective and conforming items (see Hui and Walter [10], Boyles [11], Van Wieringen and De Mast [12], Danila et al. [13] and many others). However, this assumption has been discredited as generally not realistic [14–17]. For example, if there are various degrees of defectiveness such that some parts are harder to judge than others, this assumption is violated and estimators may have a substantial bias.
To allow for variability in the misclassification probabilities in each subpopulation, more complex latent variable models have been proposed. In industrial statistics, Danila et al. [18] treat FAP and FRP as random effects with beta distributions, and Erdmann et al. [19] explicitly attribute variability in the misclassification probability to an underlying continuous property of the items. Also in medical statistics various more complex approaches have been suggested [20,21]. However, Albert and Dodd [22] warn that these models are not robust against model misspecification, which is typically difficult to detect from the binary observations. Mathematical analysis by Akkerhuis [23] reveals that the parameters FAP and FRP are in general unidentifiable from the binary observations when a gold standard is unavailable. This suggests that the estimation of FAP and FRP is inherently problematic without a gold standard.

This paper compares the main approaches proposed in industrial statistics for the evaluation of binary tests without a gold standard, to learn how robust they are to model misspecification. The study is based on a crossed design where methods M1, M2 and M3 are applied in a range of scenarios S1, S2, . . . . The estimation bias of the methods in the various scenarios is analyzed with three goals:

- Establish which method is most reliable across a range of realistic scenarios.
- Empirically establish how problematic estimation of FAP and FRP is without a gold standard.
- If the problems turn out to be severe: identify alternative ways to evaluate binary tests.

The next section explains the set-up of the study in detail. Section 3 presents the findings of the study, which results in specific conclusions discussed in the final section.

2. Theory and methods

The current literature describes three classes of approaches for estimating misclassification probabilities of binary tests under absence of a gold standard: traditional latent class methods (M1), latent class random effects approaches (M2), and approaches based on characteristic curves (M3). In the comparison study we try these methods M1, M2 and M3 in a number of scenarios (S1, S2, . . .) to learn how sensitive they are to violations of their assumptions.

In this section, we present a novel statistical modelling framework that is general enough to describe all methods M1, M2 and M3 under study as special cases. Also the test scenarios will be defined in terms of this modelling framework. The descriptions of the methods are cursory, only highlighting the main idea, but full descriptions can be found in the references.

2.1. Statistical modelling framework

When a test is applied in regular production, it produces results $Y_i = 0$ (rejection) or $Y_i = 1$ (acceptance) for tested items $i = 1, 2, \ldots$. The unobservable true states of the items are $X_i = 0$ (truly defective) or $X_i = 1$ (conforming to specifications), and the (unknown) defect rate is $p = P[X_i = 0]$. The probability that an item $i$ is rejected is $R_i = P[Y_i = 0]$, which depends on $X_i$ and possibly on other properties affecting the measurement.

To determine the error probabilities of the test, one executes an MSA experiment, in which a sample of $I$ items is tested $J$ times, producing the results $Y_{ij} \in \{0, 1\}$. When a gold standard is available, the corresponding true values $X_1, \ldots, X_I$ are established, and a comparison of $Y_{ij}$ to $X_i$ allows the calculation of FAP and FRP. In the problem under consideration, however, a gold standard is unavailable and the true states $X_i$ of the items in the experiment are unobservable. In fact, one can only observe the per-item rejection counts $U_i = \sum_{j=1}^{J} (1 - Y_{ij}) \in \{0, 1, \ldots, J\}$, from which the rejection probabilities $R_i$ can be estimated.

We will specify statistical models in terms of the $R_i$, avoiding the unobservable $X_i$. The rejection probabilities of items vary from 0 to 1 and have a statistical distribution $F_R(r) = P[R_i \leq r], r \in [0, 1]$. We assume that given an item’s rejection probability, repeated tests are independent: conditional on the event $(R_i = r)$, the $Y_{ij}, \ldots, Y_{ij}$ are i.i.d. Bernoulli (with parameter $1 - r$) and the rejection counts $U_i$ have a binomial $(J, r)$ distribution. A special case is the traditional latent class model, where the rejection probabilities assume only two values: $R_i = FRP$ for all conforming items and $R_i = 1 - FAP$ for all defective items (with FRP and FAP two constants).

The distribution of rejection probabilities can be interpreted as a mixture of two component distributions: $F_R(r) = pF^*_R(r) + (1 - p)F^*_R(r)$, with

$$F^*_R(r) = P[R_i \leq r|X_i = 0]$$

(distribution of rejection probabilities of defective items),

$$F^*_R(r) = P[R_i \leq r|X_i = 1]$$

(distribution of rejection probabilities of conforming items).

Fig. 1 gives an example. The solid curve is the density $f_R$ of rejection probabilities in the population of items. The gray and white areas show the components $f^*_R$ and $f^*_R$ associated with defective and conforming items. In this example (produced by scenario S2a explained later), the rejection probabilities of conforming and defective items are clearly separated and $f^*_R(r) \approx 0$ around $r = 0.6$.

The FAP and FRP are the mean probabilities of misclassification of conforming or defective items:

$$FAP = 1 - E[R_i|X_i = 0],$$

$$FRP = E[R_i|X_i = 1].$$

In MSA studies for quantitative measurements, the measurement error is often decomposed into systematic and random measurement error (or similar concepts such as trueness/precision). Recent literature has proposed a similar decomposition for binary MSA [19,23]. Let $X_i = 0$ if $R_i > 0.5$ and $X_i = 1$ if $R_i \leq 0.5$ be the modal (most likely) outcome for item $i$. Then, $X_i \neq X_i$ implies that test results for item $i$ are systematically off, and $Y_{ij} \neq X_i$ is a random deviation of a test result from the modal outcome. Akkerhuis [23] shows how the misclassification probabilities can be decomposed; the terms corresponding to random measurement error are

![Fig. 1. Density $f_R$ of rejection probabilities in scenario S2a (solid curve). Estimated pdf $f_R$ fitted by method M3 (dashed curve).](image-url)