Routine Assessment of a Binary Measurement System

Oana Danila, Stefan H. Steiner and R. Jock MacKay

Business and Industrial Statistics Research Group (BISRG)

Dept. of Statistics and Actuarial Sciences

University of Waterloo

Waterloo, N2L 3G1, Canada

Abstract

Binary measurement systems that classify parts as pass or fail are widely used in industry especially for systematic inspection purposes. To support production and quality improvement, it is critical that the misclassification rates of such binary measurement systems are assessed. In this paper, we focus on the situation where a routine assessment investigation of the binary system is conducted and therefore the pass rate is known prior to the study. Also, large populations of previously passed and failed parts are readily available for the assessment study. We consider the case where there is no available gold standard and parts are repeatedly classified using the binary measurement system. We assess the gain of knowing the pass rate in the precision of the estimators of the misclassification rates and the conforming rate. We propose new sampling plans and compare them with the existing ones with respect to the precision of the estimators, and also give recommendations for planning an assessment study for a binary measurement system.

Key words: Conditional Sampling; Gold standard; Latent Class Analysis; Misclassification Rates; Pass-Fail Inspection;

Business and Industrial Statistics Research Group Report RR-09-01

Ms. Danila is a Ph.D. student in the Dept. of Statistics and Actuarial Science at the University of Waterloo. Her email address is omdanila@math.uwaterloo.ca.

Dr. Steiner is an Associate Professor in the Dept. of Statistics and Actuarial Science at the University of Waterloo as well as director of the Business and Industrial Statistics Research Group. He is a senior member of ASQ. His email address is shsteiner@uwaterloo.ca.

Dr. MacKay is an Associate Professor in the Dept. of Statistics and Actuarial Science at the University of Waterloo. He is a member of ASQ. His email address is rjmackay@uwaterloo.ca.

Introduction, background and goals

In a manufacturing environment, critical decisions about process and product quality depend on the quality of the measurement systems. Therefore, a rigorous assessment study of the measurement system performance is required. When the quality characteristic of interest is continuous, Gauge R&R (AIAG, 2002) is the standard statistical plan used to assess precision or measurement variation. This paper focuses on the situation where the characteristic of interest is binary; for example, parts may be evaluated as conforming or nonconforming in a pass/fail inspection. These systems are called Binary Measurement Systems (BMS) and often involve automated inspection of manufactured parts. In many cases, a BMS is used for 100% inspection to protect customers against receiving nonconforming parts.

An example of a BMS is an automated measurement system used to inspect blank credit cards. The system takes a digital picture of the front of each card and calculates hundreds of summary measures based on comparing the picture to a template of the ideal card. The cards are checked for many defects such as missing parts, surface scratches, bleeding of colors, fuzzy letters and numbers. If any of the summary measures fall outside a pre-specified range, the card fails the inspection. In this case, although some of the characteristics are continuous, because the final decision is to pass or fail the card, we consider the measurement system a BMS.

The main goal of a BMS assessment study is to estimate the “customer’s risk”, the probability of passing a nonconforming part, a false positive, and the “producer’s risk”, the probability of rejecting a conforming part, a false negative. It may also be of interest to estimate the conforming rate, that is, the probability of producing a conforming part.

The statistical properties of a BMS can be assessed by measuring sampled parts using the BMS and a gold standard (a definitive measurement system, e.g. a human inspector in the credit card example), or by repeatedly measuring parts using only the BMS. BMS assessment based on the gold-standard approach was considered by Farnum (1994) and Danila, Steiner and MacKay (2008) among others. The essential requirement for this approach is that the true status of a part can be determined with no classification error. Farnum (1994) suggested a sampling plan where parts are independently selected from the population of conforming and nonconforming parts as determined by the gold standard, and then measured once by the BMS. This method is impractical for processes with a high conforming rate since it would require extensive use of the gold standard system to get the required number of nonconforming parts. Danila et al. (2008) proposed two alternative sampling schemes. The first involves independently sampling from the population of previously passed and failed parts and then measuring these parts with the gold standard system. In the second plan, parts are randomly selected and then measured using both the BMS and the gold standard system.

The assessment of a BMS with repeated measurements and without a gold standard uses a latent class (LC) model. Latent class analysis methods have been known for decades (Lazarsfeld and Henry, 1968) and they have been applied in many areas of research, such as psychology, sociology, and more recently in assessment studies for medical tests (Hui and Walter, 1980, Walter and Irwing, 1988, Qu et al., 1996). In the industrial context, Boyles (2001) and Van Wieringen and Van den Heuvel (2005) use a LC model in the assessment of a BMS.

This paper focuses on a particular context with two conditions often found in a high-volume production process. First, we assume the goal is the evaluation of the performance of the BMS after it has been in use for a period of time. Such assessments are often required as part of the Quality System.

Second, we suppose that the BMS is used for 100% inspection so that large populations of passed and failed parts are generated. As a consequence, we can assume that we have a precise estimate of the pass rate. In the credit card example, thousands of cards are inspected every day so it is reasonable to assume that the pass rate is known. It is the second condition that sets this work apart from the earlier literature.

With the increasing use of automated inspection systems, this context is now widely found in industry but there is little available research. Danila et al. (2008) propose assessment methods for this context when a gold standard is available and convenient to use. Browne et al. (2009) consider assessment of a continuous measurement system in this context. In this paper, we consider the case when the gold standard does not exist or is too expensive or time-consuming for an assessment study. We describe efficient sampling schemes when parts are repeatedly measured with the BMS using the available populations of passed and failed parts and a known pass rate. We use a LC approach to model the data and maximum likelihood for estimating the model parameters.

Throughout the paper, we make use of the following important assumptions and notation for the parameters and available data. First we assume the BMS under study is nondestructive so that repeated measurements can be made without changing the part characteristics and, during the time of the evaluation, the production process is under statistical control and the BMS itself is stable.

In addition, we assume that each part has a true quality state, conforming or non-conforming, denoted by

$$X_i = \begin{cases} 1, & \text{if part i is conforming} \\ 0, & \text{if part i is nonconforming} \end{cases}$$

When there is no gold standard, the X_i's are unobserved. The data that arises from the repeated measurement of each part, denoted by y_{ij}, are the realizations of the random variables

$$Y_{ij} = \begin{cases} 1, \text{ if the i}^{th} \text{ part passes the j}^{th} \text{ inspection} \\ 0, \text{ otherwise} \end{cases}, \; i = 1,..,n \text{ and } j = 1,...,r$$

where n is the sample size and r is the number of time each part is measured. The data can be summarized as the total number of passes $s_i = \sum_{j=1}^{r} y_{ij}$, $i = 1,...,n$ for each of the repeatedly measured parts.

Using this notation, the performance of a BMS is characterized by the two misclassification probabilities

$$\alpha = \Pr(Y_{ij} = 1 \mid X_i = 0)$$

$$\beta = \Pr(Y_{ij} = 0 \mid X_i = 1)$$

where α is the probability of passing a nonconforming part and β is the probability of failing a conforming part. In most applications, α is of greater concern than β since it quantifies the risk of nonconforming items reaching the customer.

Also of interest are

$$\pi_P = \Pr(Y_{ij} = 1)$$

$$\pi_C = \Pr(X_i = 1)$$

where π_P is the probability that any randomly selected part passes an inspection and π_C is the probability that the manufacturing process produces a conforming part. Note that π_C is a function of the production process whereas π_P depends on both the performance of the BMS and the quality of the production process. Also, since

$$\Pr(Y_{ij}=1)=\Pr(Y_{ij}=1\,|\,X_i=1)\Pr(X_i=1)+\Pr(Y_{ij}=1\,|\,X_i=0)\Pr(X_i=0)$$

we have the constraint

$$\pi_P=(1-\beta)\pi_C+\alpha(1-\pi_C) \tag{1}$$

As the notation suggests, we assume that, given their true state, items have the same probability of passing. That is, the chance of passing any conforming (nonconforming) part is the same. We also assume events defined on different parts are independent and also that, conditional on the true quality state, repeated measurements on each part are independent. This conditional independence can be expressed mathematically as follows

$$P(Y_{i1},Y_{i2},...,Y_{ir}\,|\,X_i)=\prod_{j=1}^{r}P(Y_{ij}\,|\,X_i)\text{, for each } i=1,...,n.$$

Using these assumptions and notation, the conditional distribution of the total number of times part i passes the BMS inspection, given the part is conforming, is

$$S_i\,|\,\{X_i=1\}\sim Binomial(r,1-\beta),$$

and given it is nonconforming is

$$S_i\,|\,\{X_i=0\}\sim Binomial(r,\alpha).$$

Boyles (2001) uses the LC approach to model the data from a BMS assessment study when there is no gold standard and parts are repeatedly measured with the BMS. Boyles proposes selecting a random sample of n parts from a population of parts produced. Then, each part is measured r times with the BMS and the number of passes is recorded.

In this paper, we often refer to Boyles' paper, as he discusses not only parameter estimation using the LC model, but also other issues such as the selection of the sampling plan and sample size determination.

For Boyles' random selection (RS) plan, using the notation and assumptions given above and noting that the distribution of the number of passes for part i is a mixture of two Binomial distributions, the likelihood function is:

$$L(\alpha,\beta,\pi_C)\propto\prod_{i=1}^{n}[(1-\beta)^{s_i}\beta^{r-s_i}\pi_C+\alpha^{s_i}(1-\alpha)^{r-s_i}(1-\pi_C)] \qquad (2)$$

To make the parameters identifiable, we have to assume that $1-\beta>\alpha$ and that there are at least three measurements per part, i.e. $r\geq 3$ (Boyles, 2001, van Weiringen and van den Heuvel, 2005). The assumption $1-\beta>\alpha$ is reasonable, since we expect the probability of passing a conforming part to be larger than the probability of passing a nonconforming part. In fact, for most measurement systems in use, we expect both α and β to be close to zero.

To find the maximum likelihood estimates for α, β and π_C, Boyles (2001) uses the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). The algorithm uses the likelihood function (2), which is also called the observed or incomplete-data likelihood, as it is based on the observed measurements of the BMS, and the complete-data likelihood function which includes the number of passes, s_i, and the (unobserved) true state of the part, x_i, given by

$$L_C(\alpha,\beta,\pi_C\mid(s_i,x_i))\propto\prod_{i=1}^{n}[(1-\beta)^{s_i}\beta^{r-s_i}\pi_C]^{x_i}[\alpha^{s_i}(1-\alpha)^{r-s_i}(1-\pi_C)]^{(1-x_i)}$$

Boyles derives confidence regions and confidence intervals based on the asymptotic properties of the likelihood. For constructing the confidence intervals, the observed-data information matrix is obtained using the missing information principle (Meng and Rubin, 1991, McLachlan and Krishnan, 1997). The inverse of this matrix gives the asymptotic variance-covariance matrix of the Maximum Likelihood (ML) estimators.

For sample size determination when planning an assessment study, Boyles uses the inverse of the Fisher information matrix for the complete-data likelihood, that is

$$J = E(I_c) = diag(\frac{nr(1-\pi_C)}{\alpha(1-\alpha)}, \frac{nr\pi_C}{\beta(1-\beta)}, \frac{n}{\pi_C(1-\pi_C)}) \quad (3)$$

as the asymptotic variance-covariance matrix of the estimators. Therefore, for pre-specified parameters values and some desired precision of the estimators, we can obtain the minimum number of parts n and the number of measurements per part, r.

In the next section, we comment on the advantages and shortcomings of this model and investigation plan. Then, we consider the situation when the pass rate π_P is assumed known as it is commonly the case for routine assessment studies of a BMS. In Section 3, we assess the value of incorporating the known pass rate into the analysis when using Boyle's sampling plan (RS). In Section 4, we propose new sampling plans that use the available populations of previously passed and failed parts and show that they improve the precision of the estimators compared to the RS plan when the pass rate is known. In Section 5, we propose an algorithm to determine sample sizes for the new sampling plans and in Section 6 we discuss some issues related to the model assumptions used in the LC analysis and provide a summary of our results.

Assessment of Boyles' Latent Class Model

Using Boyles' latent class approach, all parameters of interest, that is α, β and π_C, can be estimated from the available data. Therefore, provided that the initial assumptions hold, both the performance of the BMS and the quality of the production process can be obtained without knowing the true state of the selected parts. This result, which is somewhat surprising at a first look, is based on the fact that for a sufficiently large number of repeated measurements and when the assumption $1-\beta > \alpha$ holds, the BMS is able to classify parts almost surely as conforming or nonconforming based on the results of the repeated measurements. Furthermore, when α and β are small (e.g. less than 0.1), the BMS can distinguish between conforming and nonconforming parts for a relatively small number of repeated measurements (e.g. $r = 5$).

Boyles and other authors in the medical field (Dawid and Skene, 1979) use the EM algorithm to estimate the parameters as the score equations of the observed-data log-likelihood function do not have a simple form. This is not the only option for maximizing the likelihood function as other optimization algorithms can also be used, such as Broyden-Fletcher-Goldfarb-Shanno (1970) or Nelder-Mead (1965). These methods are not more computationally intensive than the EM algorithm and can also automatically provide the observed information matrix. In general, the EM algorithm is preferred when the number of parameters in the model is large, which is not the case here as there are only three parameters to estimate. Therefore, in our ML estimation, we use direct optimization methods to maximize the likelihood.

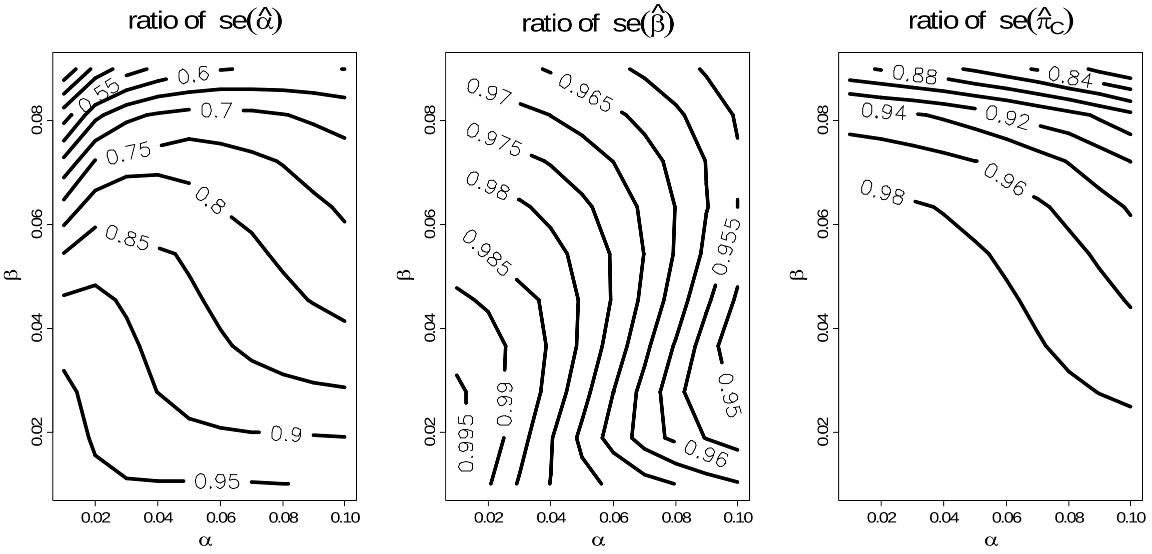
Boyles proposes using a random sample of parts from the process. When the conforming rate π_C is close to one, which is usually the case with high-performance processes, this scheme can lead to samples with no or only a few nonconforming parts. When there are no nonconforming parts in the sample, the parameter α is not identifiable. In Section 4, we propose Conditional Selection (CS) to avoid this situation by sampling parts based on the result of a previous measurement. This sampling plan virtually eliminates the chance of having no nonconforming parts in the sample and, compared to the random selection, better balances the expected number of conforming and nonconforming parts in the sample.

As noted above, Boyles also proposed the use of the inverse of the Fisher information matrix from the complete-data likelihood for sample size determination. Using the complete-data information for sample size determination is convenient because we can obtain the minimum sample size n and/or the number of the repeated measurements r by solving simple equations. Boyles is assuming that the Fisher information from the complete data is a good approximation to the one from the incomplete (observed) data. The missing information principle gives the following relationship between the information matrices from complete and incomplete-data likelihood (MacLachlan and Krishnan, 1997):

$$J(\alpha,\beta,\pi_C) = J_{comp}(\alpha,\beta,\pi_C) - E[J_{miss}(\alpha,\beta,\pi_C;s_i)],$$

where J is the information from the incomplete or observed-data likelihood, J_{comp} is the information from the complete-data likelihood and $E[J_{miss}(\alpha,\beta,\pi_C;s_i)]$ is the missing information. J_{comp} is a good approximation to J only if the missing information is negligible. In the BMS assessment context, this is not the case for all values of the parameters or any number of repeated measurements.

We investigated this issue by comparing the asymptotic standard deviations of the estimators for the complete and incomplete case for a reasonable range of parameter values. Figure 1 shows the ratios of asymptotic standard deviations of the estimators derived from the complete and incomplete-data likelihood for the case $r = 4$, α and β in the interval $[0.01, 0.1]$ and $\pi_P = 0.9$. We note that the ratio of the standard deviations given by the two methods for each estimator do not depend on the number of parts, n, but do depend on the parameter values and the number of repeated measurements r.


ratio of $se(\hat{\alpha})$

ratio of $se(\hat{\beta})$

ratio of $se(\hat{\pi}_C)$

Figure 1: Contour plots of $\frac{sd(complete)}{sd(incomplete)}$ *of the estimators of* α , β *and* π_C ,

Random Selection plan $\pi_P = 0.9$ *and* $r = 4$

For example, from Figure 1, we see that when $\alpha = 0.02$, $\beta = 0.08$, $\pi_P = 0.9$ and $r = 4$ the standard deviation of the estimator for α given by the complete-data information is half that of the incomplete-data. Clearly in this case, the missing information is not negligible. To determine when the missing information is negligible, we computed the asymptotic standard deviations for the estimators of α, β and π_C using both the complete and incomplete-data information matrices for larger values of r and other values of π_P. In our investigation, we noted that for each choice of the parameter values, as the number of repeated measurements increases, the ratio of the two standard errors approaches one. Therefore, for r large the complete-data information provides a good approximation for the standard deviation of the estimators. Figure 2 shows the minimum value of r so that the ratio of the two standard deviations is larger than 0.999 for a range of α, β and $\pi_P = 0.9$.

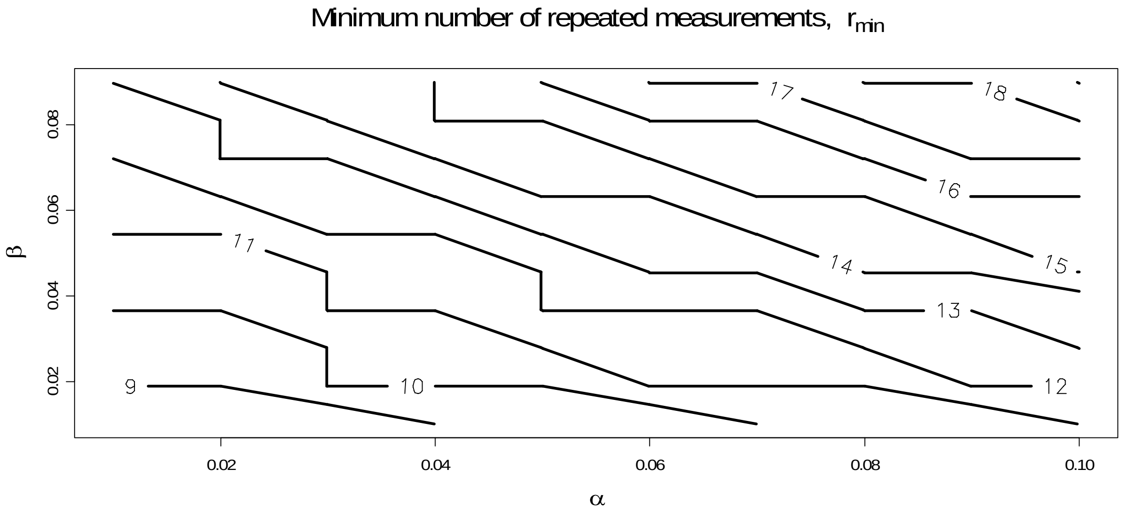


Figure 2: Minimum number of repeated measurements that guarantees $sd(\hat{\alpha})$, $sd(\hat{\beta})$ and $sd(\hat{\pi}_C)$ given by the observed-data information are close to the ones given by the complete-data information (ratio ≥ 0.999), $\pi_P = 0.9$

From Figure 2 and other such plots, we conclude that for reasonable parameters values and a large number of repeated measurements, e.g. $r > 15$, the complete-data information provides a good approximation for the asymptotic standard deviations of the estimators. We can explain this conclusion intuitively because, for large numbers of repeated measurements and small misclassification probabilities, the BMS can identify the true state of the parts.

In this paper, we discuss and compare cases where the number of repeated measurements is relatively small, i.e. $r < 10$, the pass rate is high ($\pi_P \geq 0.9$) and the misclassification rates are small (α and β between 0.01 and 0.1). The asymptotic standard deviations of all estimators are based on the observed-data information. For the case where r is large ($r > 15$), we recommend using Boyle's approximation (3) based on the complete-data information.

Latent Class Model with Random Selection and Known Pass Rate

When the BMS is used for 100% inspection, it is reasonable to assume the pass rate π_P known. In this section, we assess the value of including the a priori information about π_P in the analysis when parts are selected using a RS plan. The ML estimates are derived by substituting $\pi_C = \dfrac{\pi_P - \alpha}{1 - \beta - \alpha}$ in the likelihood (2). To quantify the gain in precision when π_P is known, we compare the asymptotic standard derivations of the estimators when π_P is unknown and known, using the Fisher information.

In Figure 3, we show the ratios of the standard deviations of the estimators (π_P unknown versus π_P known) when the number of repeated measurements is $r = 10$.

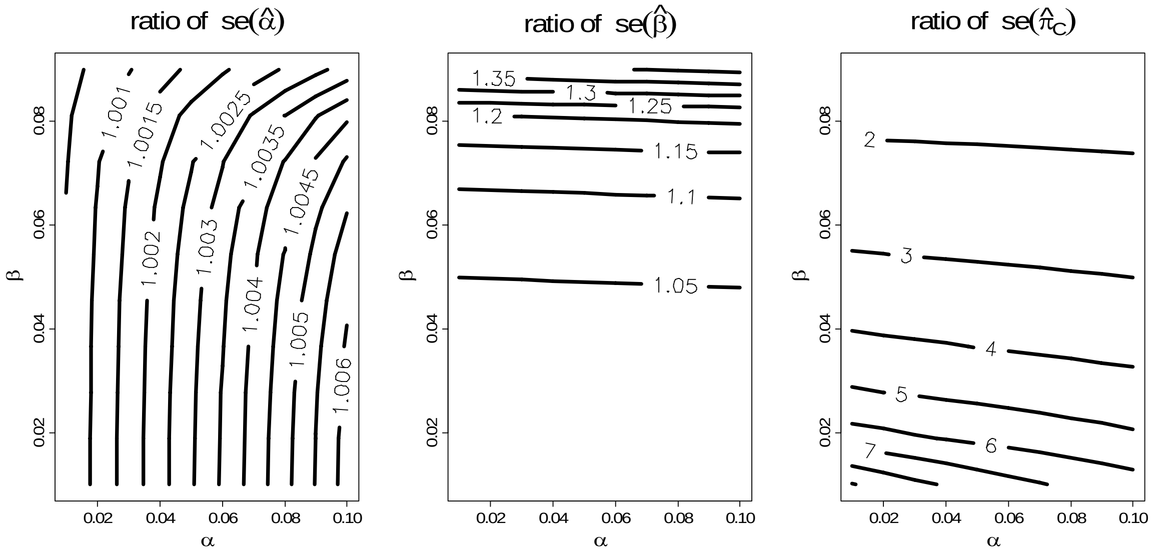


Figure 3: Contour plots of $\frac{sd(RS\ \pi_P\ unknown)}{sd(RS\ \pi_P\ known)}$ *of the estimators of* α *,* β *and* π_C*, when* $\pi_P = 0.9$ *and* $r = 10$

In Figure 3, we see that knowing π_P improves the precision of all estimators, with a substantial gain for the estimator of π_C. We also notice that the ratio of the standard deviations for the estimators of β gets larger when β increases, whereas for the estimator of π_C the ratio decreases with larger values of β. These results are also valid for smaller values of r closer to the identifiability condition boundary, e.g. $r = 5$. Looking at a wide range of such comparisons, we conclude that including the available information about the pass rate improves the precision of the estimators of α, β and especially π_C when parts are sampled using the RS plan.

Latent Class Model with Conditional Selection and Known Pass Rate

In the context we are considering, there are large populations of passed and failed parts readily available for the assessment study. We propose a new sampling plan, called Conditional Selection (CS), where we randomly sample from these two populations of previously classified parts. As we see later in this section, the Conditional Selection is a solution to the "no nonconforming parts" situation, as we have better control over the expected number of nonconforming parts in the sample. We also show that CS provides uniformly better estimators than RS.

The CS plan involves independently selecting a pre-determined number of parts from the populations of previously passed and failed parts. The likelihood function for the CS plan is:

$$\begin{aligned} L_C(\alpha,\beta,\pi_C \mid s_i) = \prod_{i=1}^{n_1} & \frac{(1-\beta)^{s_i+1}\beta^{r-s_i}\pi_C + \alpha^{s_i+1}(1-\alpha)^{r-s_i}(1-\pi_C)}{(1-\beta)\pi_C + \alpha(1-\pi_C)} \times \\ & \prod_{i=1}^{n_0} \frac{(1-\beta)^{s_i}\beta^{r-s_i+1}\pi_C + \alpha^{s_i}(1-\alpha)^{r-s_i+1}(1-\pi_C)}{\beta\pi_C + (1-\alpha)(1-\pi_C)} \end{aligned} \qquad (4)$$

where n_1 is the number of previously passed parts and n_0 the number of previously failed parts in the sample. For the CS plan, we can choose the proportion of passed parts, $f = \frac{n_1}{n_1 + n_0}$, where $n_1 + n_0 = n$ is the total sample size.

To explore the CS plan, we first investigate how the probability of having no nonconforming parts in the sample compare for RS and CS for the same sample size and parameter values. For RS, Pr(no nonconforming)=π_C^n, whereas for CS we have

$$\text{Pr(no nonconforming)} = [\frac{(1-\beta)\pi_C}{\pi_P} f + \frac{\beta\pi_C}{1-\pi_P}(1-f)]^n \qquad (5)$$

For example, when $\beta = 0.1$, $\alpha = 0.05$, $\pi_C = 0.95$ and $n = 50$, the probability of having no nonconforming parts is 0.08 for RS plan, 0.0001 for CS with $f = 0.5$ and $1.56e^{-09}$ for CS with $f = 0$. In general, for reasonable CS plans, the probability of having no nonconforming parts is very small.

Next, we compare the precision of the estimators for the two sampling plans, RS with $r = 5$ and CS with $r = 4$, when the pass rate is known and $f = 0.5$. We use one fewer repeated measurement for the CS plan to take into account the initial BMS measurement made before conducting the assessment study. This makes the comparison fair in some sense but we could consider the initial measurements in the CS plan free. The differences between the two plans are even greater if we use the same value of r in each case. Figure 5 shows the ratio of the asymptotic standard derivations for the case where the total number of measurements with the BMS is 5 and the proportion of previously passed parts, f, is 0.5.

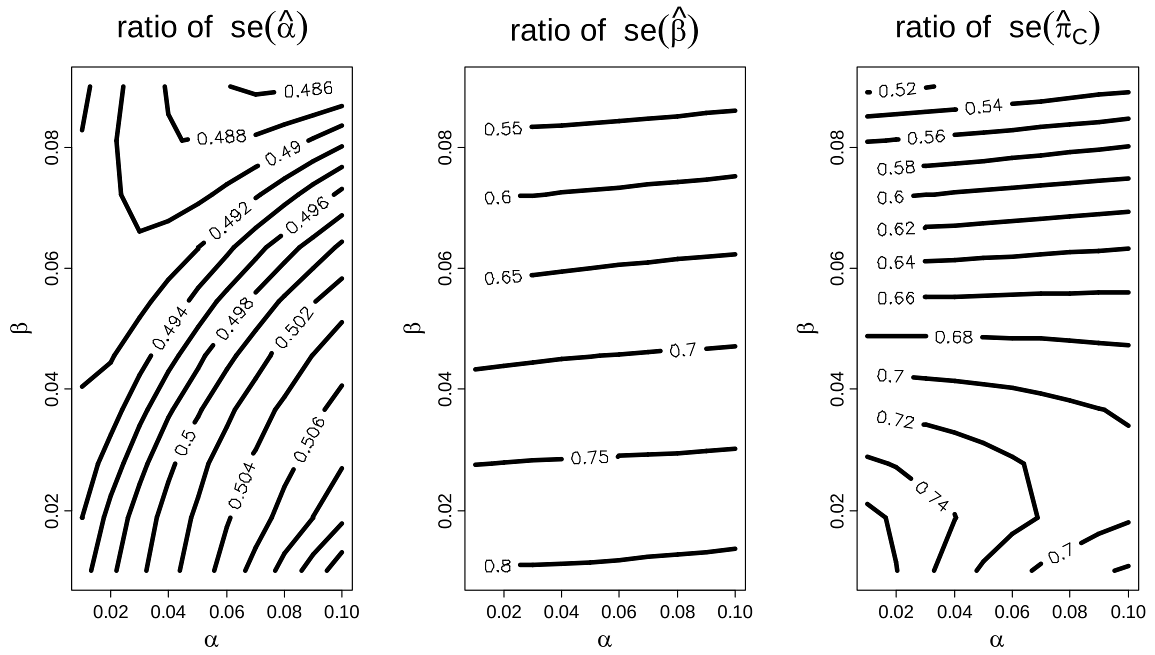


Figure 5: Contour plots of $\frac{sd(CS\ \ \pi_P\ \ known)}{sd(RS\ \ \pi_P\ \ known)}$ *of the estimators of* α *,* β *and* π_C *,*

$f = 0.5$, $\pi_P = 0.9$ *and five total measurements by the BMS*

Figure 5 suggests that the CS plan with $f = 0.5$ gives substantially better estimators for all parameters.

Next, we compare the precision of the estimators for ten repeated measurements and $\pi_P = 0.9$.

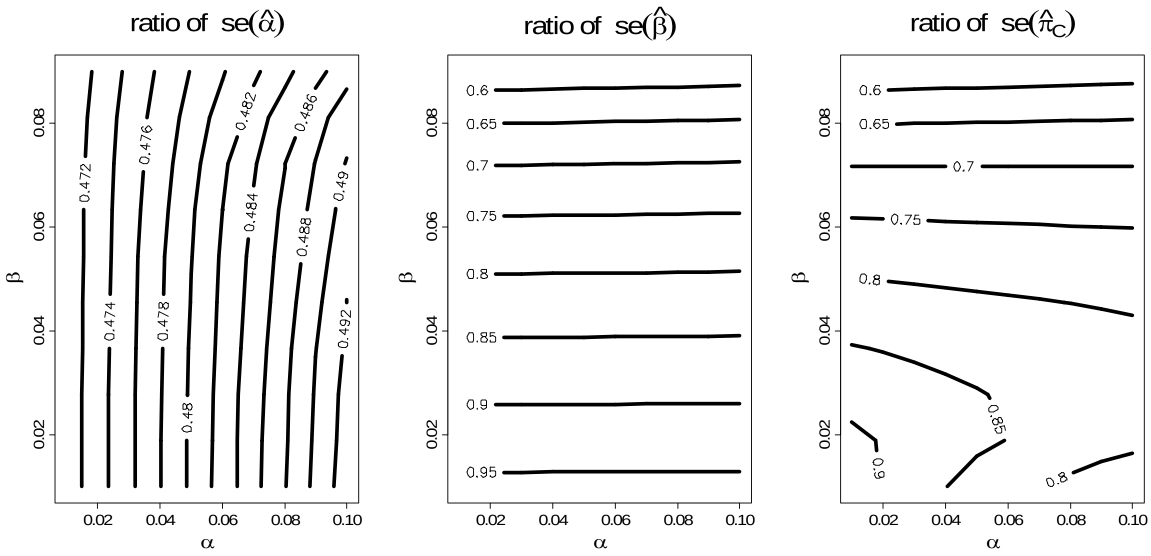


Figure 6: Contour plots of $\dfrac{sd(CS\ \ \pi_P\ \ known)}{sd(RS\ \ \pi_P\ \ known)}$ *of the estimators of* α *,* β *and* π_C *,*

$f = 0.5$*,* $\pi_P = 0.9$ *and ten measurements by the BMS*

From Figures 6 and 5, we conclude that as the number of repeated measurements increases, the relative advantage of CS over RS increases for the estimator of α and decreases for the estimators of β and π_C. We also compared the precision of the estimators for higher values of the pass rate (e.g., $\pi_P = 0.95$) and conclude that CS becomes even more efficient than RS when the pass rate increases with r fixed.

For reasonable α and β values, a CS with $f = 0.5$ will better balance the number of conforming and nonconforming parts in the sample than the RS plan. For small values of α and β, even better balance is given by selecting only from previously failed parts, i.e. using $f = 0$. In Figure 7, we compare the precision of each estimator given by the CS with $f = 0.5$ and $f = 0$, when the pass rate π_P is known.

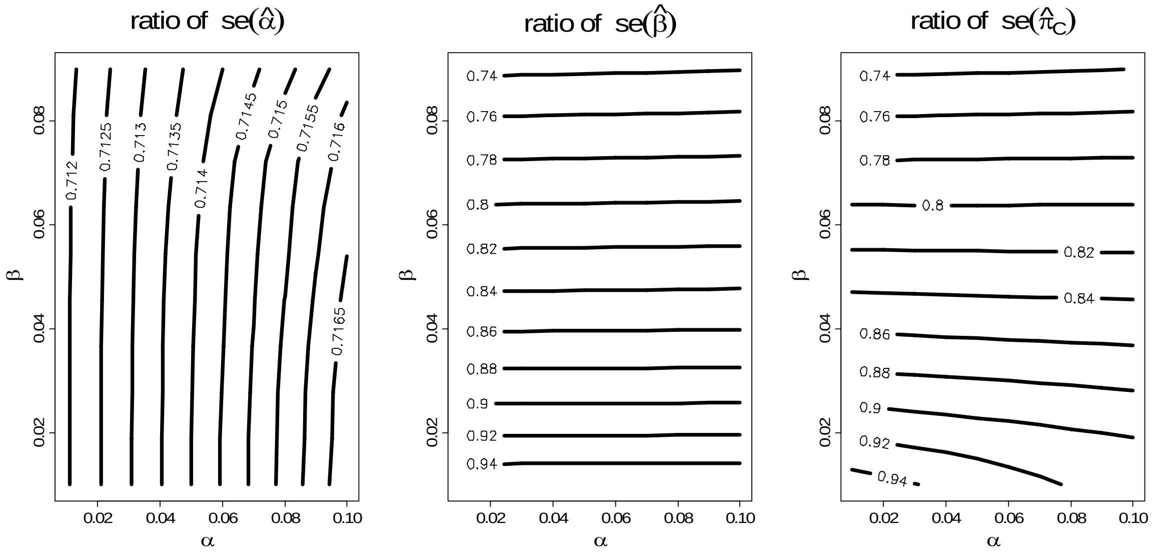


Figure 7: Contour plots of $\dfrac{sd(CS\ \ \pi_P\ \ known\ f = 0)}{sd(CS\ \ \pi_P\ \ known\ f = 0.5)}$ *of the estimators of* α *,* β *and* π_C *,* $\pi_P = 0.9$ *and ten total measurements by the BMS*

Figure 7 shows that, for the range of α and β given, the CS plan with π_P known and $f = 0$ is uniformly more efficient than CS with $f = 0.5$. We also compare the precision of the estimators for smaller values of r (e.g. $r = 5$) and conclude that the gain in precision when $f = 0$ is higher for the estimator of α and lower for the estimators of β and π_C, when compared to $f = 0.5$. Also, for larger values of the pass rate ($\pi_P \geq 0.9$) , the ratios of the standard deviations are smaller for all parameters, therefore CS with $f = 0$ is even more efficient in that case.

In conclusion, in cases where π_P is known, the BMS has good performance and the manufacturing process is high quality, i.e. α and β are small and π_C (and π_P) is close to one, we recommend the

Conditional Selection plan with $f = 0$, i.e. all parts are sampled from the population of previously failed parts. This plan has a negligible chance of having no nonconforming units in the sample and is substantially more efficient in estimating the parameters, when compared to the other plans we have investigated.

Planning the Assessment Study Using Conditional Selection

In this section, we discuss how to choose the sample size and the total number of repeated measurements in order to achieve a pre-specified precision for the estimators of α and β when using a CS plan. As mentioned before, a reasonable BMS is characterized by small misclassification probabilities, e.g. α and β less than 0.1. For such small parameter values, small changes can have large impact on both the producer and consumer and hence we need high-precision estimators. In that case, the required number of parts and the total number of measurements will be large, as we see in the following example.

We now describe an algorithm to determine n and r for the CS plan where we assume the pass rate π_P is known. As in any sample size calculation, we start with some conjectured values for the unknown parameters α and β. The pass rate is a priori known and therefore the value of π_C can be determined using equation (1). We select a value of f the proportion of previously passed parts ($f = 0$ is recommended) and then enter the required precision for the estimators of α and β. Here, we focus on the precision of α and β, as they are the main parameters of interest. The algorithm provides several combinations of the total number of parts n and the number of repeated measurements r. Also, the asymptotic standard deviations for the estimators of α, β and π_C are provided, along with the probability of having no nonconforming parts in the sample. R code for the algorithm can be found at www.bisrg.uwaterloo.ca/.

The algorithm computes asymptotic standard deviations for the parameters based on the observed (incomplete data) information. This is different from Boyles' approach where he uses an approximation to these standard deviations based on the complete-data information. As discussed earlier, we found this approximation not appropriate for small numbers of repeated measurements. After entering the parameter values and the required precision, the algorithm starts with the minimum number of repeated measurements, i.e. $r = 3$, and the minimum of parts $n = 2$, and then increases n until the required precision for the estimators of α and β are achieved. Next, r is increased by one-unit increments up to $r = 10$ and then by increments of five. For each r value the corresponding n is determined.

The following example illustrates the output of the algorithm for the case $f = 0$. Suppose we know that the pass rate $\pi_P = 0.9$. We also assumed that the parameter values are $\alpha = 0.01$ and $\beta = 0.02$. Using identity (1), solving for π_C gives 0.92. Suppose the desired precisions for the estimators of α and β are $std_0(\hat{\alpha}) = 0.005$ and $std_0(\hat{\beta}) = 0.005$. The corresponding sample size n, minimum number of repeated measurements r, total number of measurements $n \times r$ and actual standard deviations are given in Table 1. Note, in general, $n_1 = fn,\ n_0 = (1 - f)n$.

Sample size n	**Number of repeated measurements** r	**Total number of measurements** $n \times r$	$std(\hat{\alpha})$	$std(\hat{\beta})$	$std(\hat{\pi}_C)$
170	3	510	0.0031	0.0050	0.0030
125	4	500	0.0036	0.0050	0.0034
98	5	490	0.0040	0.0050	0.0038
81	**6**	**486**	**0.0044**	**0.0050**	**0.0042**
70	7	490	0.0047	0.0050	0.0045
61	8	488	0.0050	0.0050	0.0047
60	9	540	0.0050	0.0047	0.0047
59	10	590	0.0050	0.0045	0.0047
56	15	840	0.0050	0.0038	0.0047

Table 1: Sample size determination using the observe- data Fisher Information, when $\alpha = 0.01$, $\beta = 0.02$, $\pi_P = 0.9$, $f = 0$, $std_0(\hat{\alpha}) = 0.005$ and $std_0(\hat{\beta}) = 0.005$

When choosing a combination of the sample size n and the number of repeated measurements r, each planner has to make a compromise on how to allocate resources, based on the relative cost of measuring and sampling a part. In the example, the plans with $r = 5, 6, 7, 8$ all have roughly the same total number of measurements.

Discussion and Conclusions

BMSs are extensively used in medicine where they are known as diagnostic or screening tests. The literature on assessing medical tests is much broader (Pepe, 2003, Walter and Irwig, 1988 and Rutjes et al., 2007 and many others) than in an industrial context. Many authors (Pepe, 2003, Qu et al., 1996, Vacek, 1983, Torrance-Rynard and Walter, 1997) have questioned the conditional independence assumption, a key requirement in the use of the LC model. As a result, methods for accounting for conditional dependence have been proposed (Joseph et al., 1995, Fujisawa and Izumi, 2000, Qu et al., 1996, Kosinsky and Flanders, 1999). We do not pursue this matter further here.

In summary, we discuss assessment methods for a binary measurement system when there is no available gold standard. The goal of the study is to estimate the misclassification probabilities and the proportion of conforming parts coming from the process. In the study, parts are repeatedly classified by the BMS. We focus on a particular industrial context where the overall pass rate is a priori known and populations of previously passed and failed parts are available for use in the study. We conclude that using the known the pass rate improves the precision of all estimators, especially for the estimators of β and π_C. We also propose a new sampling plan, Conditional Selection, which involves random selections from the populations of the previously passed and failed parts. We conclude for a range of reasonable parameter values that it is best to sample only previously failed parts ($f = 0$). We give an algorithm for planning an assessment with Conditional Selection that provides several combinations of the sample size n and number of repeated measurements r for a specified precision of α and β.

References

Automotive Industry Action Group (AIAG) (2002). *Measurement Systems Analysis*, 3rd edition. Southfield MI.

Boyles, R.A. (2001). “Gage Capability for Pass-Fail Inspection”. *Technometrics*, 43, pp. 223-229.

Browne, R., MacKay, R. J. and Steiner, S. H. (2009). “Improved Measurement System Assessment”, submitted to *Journal of Quality Technology*

Danila, O., Steiner, S. H., and MacKay, R. J. (2008). “Assessing a Binary Measurement System”. *Journal of Quality Technology*, 40, 3, pp. 310-318

Dawid, A.P. and Skene, A.M. (1979). “Maximum likelihood estimation of observer error-rates using the EM algorithm”. *Appl. Statist.*, 28, pp. 20-28

Dempster, A. P., Laird, N. M. Rubin, D. B. (1977). “Maximum likelihood from incomplete data via the EM algorithm”. *Journal of the Royal Statistical Society*, B, 39, pp. 1-38.

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

Fujisawa, H. and Izumi, S. (2000). “Inference about the misclassification probabilities from repeated binary responses”. *Biometrics*, 56, pp. 706-711

Joseph, L., Gyorkos, T.W. and Coupal, L. (1995).”Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of gold standard”. *Am. Journal of Epidemiology.*, 141, pp. 263-272

Kosinki, A.S. and Flanders, W.D. (1999). “Evaluating the exposure and disease relationship with adjustment for different types of exposure misclassification: a regression approach”. *Statistics in Medicine*, 18, pp. 2975-808

McLachlan, G.J. and Krishnan, T. (1997). *The EM algorithm and Extensions*, Wiley, New York

Meng, X.L. and Rubin, D.B. (1991). “Using EM to obtain asymptotic variance-covariance matrices : the SEM algorithm". *Journal of the American Statistical Association*, 86, 416, pp. 899-909

Nelder, J.A. and Mead, R. "A Simplex Method for Function Minimization". *Computer Journal*, 1965, 7, pp 308-313

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. 1st Edition, Oxford University Press Inc., New York.

Qu, Y., Tan, M. and Kutner, M.H. (1996). "Random effects models in latent class analysis for evaluating accuracy of diagnostic tests". *Biometrics*, 52, pp. 797-810

Rutjes, A.W.S., REitsma, J.B., Coomarasamy, K.S, Khan, K.S. and Bossuyt, P.M.M. (2007). "Evaluation of diagnostic tests when there is no gold standard. A review of methods". *Health Technology Assessment*, II, 50

Torrance-Rynard, V.L. and Walter, S.D. (1997). "Effects of dependent errors in the assessment of diagnostic test performance". *Statistics in Medicine*, 16, pp. 2157-2175

Vacek, P. (1983). "The effect of conditional dependence on the evaluation of diagnostic tests". *Biometrics*, 41, pp. 959-968

Van Wieringen, W. N. and van der Heuvel, E. R. (2005). "A Comparison of Methods for the Evaluation of Binary Measurement Systems". *Quality Engineering*, 17, pp. 495-507.

Walter, S. D. and Irwig, L. M. (1988). "Estimation of Test Error Rates, Disease Prevalence and Relative Risk for Misclassified Data: a Review". *Journal of Clinical Epidemiology*, 41, pp. 923-937.