Assessment of a Binary Measurement System in Current Use

OANA DANILA, STEFAN H. STEINER, and R. JOCK MACKAY

University of Waterloo, Waterloo N2L 3G1, Canada

Binary measurement systems that classify parts as pass or fail are widely used in industry, especially for systematic inspection in high-volume processes. In this context, we are likely to have available a large number of previously measured passed and failed parts. To support production and quality improvement, it is important to assess the misclassification rates, e.g., the probability of failing a conforming part or passing a nonconforming part. We may also want to estimate the unknown conforming rate. Here we focus on the assessment of a binary measurement system when no gold-standard measurement system is available. The standard assessment plan is to repeatedly measure a sample of parts and use a latent class model. We demonstrate the substantial benefit of supplementing the standard plan with the available data from the previously measured parts. We propose new sampling plans and compare them with the standard plan with respect to the precision of the estimators of the misclassification rates. We also give recommendations for planning an assessment study when we can sample from a population of previously measured parts.

Key Words: Conditional Sampling; Gold Standard; Latent Class Analysis; Misclassification Rates; Pass–Fail Inspection.

In a manufacturing environment, critical decisions about process and product quality depend on the quality of the measurement systems. As a result, periodic assessment of measurement-system performance is required within many quality systems. When the characteristic of interest is continuous, Gauge R&R (AIAG (2002)) is the standard plan used to assess precision or measurement variation. This paper, on the other hand, focuses on the situation where the characteristic of interest is binary and parts are evaluated in a pass/fail inspection. We suppose that the binary measurement system (BMS) is automated so there are no operator effects and that there is high volume of parts so that a large number of measurements are made. In many cases, such a BMS is used for 100% inspection to help to protect customers from receiving nonconforming parts.

Examples of BMSs include on-line vision systems for optical inspection of integrated circuits (Boyles (2001)) and automated visual systems for checking manufactured parts from an injection molding process (van Wieringen and de Mast (2008)) or blank credit cards (Danila et al. (2008)). Note that these three examples describe measurement systems where the final binary classification is a summary based on one or more underlying continuous characteristics. The existence of an underlying continuous characteristic impacts the reasonableness of the part interchangeability assumption discussed later in this section. Examples that are arguably closer to purely binary measurement systems include testing functionality of an electronic device and, in a medical context, a pregnancy test.

A BMS classifies parts as either pass or fail. We denote the resulting classification by

$$y = \begin{cases} 1 & \text{if the part passes the inspection} \\ 0 & \text{if the part fails the inspection.} \end{cases}$$

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.

We denote the pass rate as $\pi_P = \Pr(\text{pass}) = \Pr(Y = 1)$ and assume that measurements on different parts are independent. We also assume that each part has a true quality state denoted by

$$x = \begin{cases} 1 & \text{if the part is conforming} \\ 0 & \text{if the part is nonconforming} \end{cases}$$

and the conforming rate is $\pi_C = \Pr(\text{conforming}) = \Pr(X = 1)$. Note that π_C is a function of the manufacturing process and not the measurement system.

In practice, a BMS is not error free and hence the need to assess its performance. A BMS can make two types of errors or misclassifications. It may pass a nonconforming part or fail a conforming part. The main goal of an assessment study is to estimate the probabilities associated with these two errors:

$$\alpha = \Pr(\text{pass} \mid \text{nonconforming}) = \Pr(Y = 1 \mid X = 0)$$
$$\beta = \Pr(\text{fail} \mid \text{conforming}) = \Pr(Y = 0 \mid X = 1).$$

This notation for α and β is consistent with Walter and Hui (1980), Walter and Irwin (1988), and Danila et al. (2008), though usually the definitions of α and β are reversed in the acceptance-sampling literature (Schilling (1982)). In most industrial applications, the consumer's risk, α, is of greater concern than is the producer's risk, β, because it quantifies the chance of a nonconforming item reaching the customer. In the definitions of α and β, we have implicitly assumed that, given the true state (conforming or nonconforming), all parts have the same probability of passing a single inspection. That is, the chance of passing any conforming (nonconforming) part is the same. This assumption that all conforming (and nonconforming) parts are interchangeable is questionable in cases where the measured characteristic y is based on an underlying continuous characteristic. In that case, parts close to the cut-off value on the continuous scale that distinguishes between conforming and nonconforming parts will be more difficult to correctly classify than other parts. Assessing the impact of a violation of the interchangeability assumption is beyond the scope of this paper.

Notice also that the pass rate π_P depends on both the performance of the BMS and the quality of the production process because

$$\begin{aligned} \pi_P &= \Pr(Y = 1 \mid X = 1)\Pr(X = 1) \\ &\quad + \Pr(Y = 1 \mid X = 0)\Pr(X = 0) \\ &= (1 - \beta)\pi_C + \alpha(1 - \pi_C). \end{aligned} \tag{1}$$

The statistical properties of a BMS can be assessed by measuring a sample of parts using the BMS and a gold-standard measurement system so that the true state of each part is determined. See Farnum (1994) and Danila et al. (2008), among others. However, a gold-standard system may not exist or be too time consuming or expensive. An alternative is to repeatedly measure a sample of parts using only the BMS and then use a latent class model in the analysis. Intuitively, this analysis can estimate the misclassification probabilities because, with a reasonable number of repeated measurements on each part, and critically assuming that α and β are small, we can classify each part as either conforming or nonconforming based on how often it passes inspection. That is, we conclude that parts that usually pass are conforming and parts that usually fail are nonconforming. Once we have a conforming/nonconforming classification for each part, it is straightforward to estimate α and β.

For latent class analysis, we assume the repeated measurements do not change the true state of the part, i.e., the measurement system is not destructive. Second, we assume that, conditional on the true quality state x, repeated measurements on each part are independent. Suppose we measure the ith part, $i = 1, 2, \ldots, n$, r times, and let y_{ij}, $j = 1, \ldots, r$ be the classification for the jth measurement. Then the conditional independence assumption can be expressed mathematically for each part i as

$$\Pr(Y_{i1}, Y_{i2}, \ldots, Y_{ir} \mid X = x) = \prod_{j=1}^{r} \Pr(Y_{ij} \mid X = x).$$

With these additional assumptions, we can use a latent class model to assess a BMS based on repeated measurements on a number of parts. Latent class models (Lazarsfeld and Henry (1968)) have been applied in many areas of research, such as psychology, sociology, and in assessment studies for medical tests (Hui and Walter (1980), Walter and Irwing (1988), Qu et al. (1996)). In the industrial context, Boyles (2001), Van Wieringen and Van den Heuvel (2005), and Van Wieringen and de Mast (2008) use latent class models in the assessment of a BMS.

Boyles (2001) proposes selecting a random sample of n parts from the population of parts and measuring each part r times with the BMS. We call this the standard plan (SP). The results are summarized by the total number of passes $s_i = \sum_{j=1}^{r} y_{ij}$, $i = 1, \ldots, n$, for each part. The conditional distribution of S_i, given the part is conforming or noncon-

forming, is

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

or

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

The marginal distribution of S_i is a mixture of these two binomial distributions and the likelihood function for the standard plan is

$$\prod_{i=1}^{n}[(1-\beta)^{s_i}\beta^{r-s_i}\pi_C + \alpha^{s_i}(1-\alpha)^{r-s_i}(1-\pi_C)]. \quad (2)$$

For the parameters to be identifiable, we require $1-\beta > \alpha$ and at least three measurements per part, i.e., $r \geq 3$ (Boyles (2001), Van Weiringen and Van den Heuvel (2005), and Van Weiringen and de Mast (2008)). The assumption that $1-\beta > \alpha$ is reasonable because, for a useful BMS, the probability of passing a conforming part should be (much) larger than the probability of passing a nonconforming part. In fact, for most measurement systems currently in use for 100% inspection, we expect both α and β to be relatively small. Later in the paper, we compare various assessment plans in the region $0 < \alpha, \beta < 0.1$.

To find the maximum-likelihood estimates for α, β, and π_C, Boyles (2001) uses the EM algorithm (Dempster et al. (1977)). He recommends using the profile likelihood ratio, treating π_C as a nuisance parameter, to derive approximate confidence regions for (α, β). For sample-size calculation during the planning stage, Boyles uses the asymptotic variance–covariance matrix for the maximum-likelihood estimates assuming the complete data likelihood, i.e., the likelihood when the true state of the parts can be determined. Boyles (2001, p. 223) also notes that, if the sample is selected from previously inspected parts, then "the results of these inspections should be included in the study data". He does not pursue this point further.

Van Wieringen and de Mast (2008) use the latent class model in a similar context but suggest randomly selecting two samples of parts, one from the population of previously failed parts and one from the previously passed parts. They define the likelihood function in terms of the two misclassification probabilities, α and β, and the probability that a sampled part is conforming, i.e., $\pi_S = \Pr(X = 1 \text{ in the sample})$. They do not include the results of the initial inspections in their likelihood. With this approach, for any sampling plan, the likelihood is given by (2) with the population-based conforming rate π_C replaced by π_S. With the standard plan, $\pi_S = \pi_C$ and the two likelihoods are identical. If the two samples are selected at random from the populations of previously passed and failed parts, π_S is related to the population-based parameters by

$$\pi_S = \frac{(1-\beta)\pi_C}{(1-\beta)\pi_C + \alpha(1-\pi_C)} f + \frac{\beta\pi_C}{\beta\pi_C + (1-\alpha)(1-\pi_C)}(1-f), \quad (3)$$

where we define n_1 and n_0 as the number of passed and failed parts selected, respectively, and $f = n_1/(n_1 + n_0)$ is the proportion of previously passed parts in the sample. Van Weiringen and de Mast (2008) suggest and compare two estimation methods, maximum likelihood using the EM algorithm and the method of moments. For planning purposes, they recommend selecting parts so that there are roughly equal numbers of conforming and nonconforming parts in the sample. Although not considered by Van Weiringen and de Mast, it is possible to estimate the population conforming rate π_C by solving Equation (3) for $\hat{\pi}_C$, where α, β, and π_S are replaced by their corresponding estimates.

Both Boyles (2001) and Van Weiringen and de Mast (2008) recommend selecting parts for remeasurement to balance the number of conforming and nonconforming parts in the study. They present examples where the parts used for the assessment study were selected from populations of previously passed and rejected parts. In our experience with BMSs used for 100% inspection in high-volume processes, these two populations are usually available, especially the failed parts, which are not immediately shipped. Furthermore, the inspection system typically tracks the number of parts passed and failed; that is, we often have baseline data about the current pass rate π_P separate from the assessment study. Recall from Equation (1) that the pass rate is directly related to the parameters of interest and hence we can combine the baseline data with those collected in the assessment study to improve the overall estimation procedure. The goal of this article is to demonstrate the value of using these "free" data and to examine how its availability affects recommendations on the design of the assessment study.

Combining available baseline data with those from a standard assessment study was considered by Danila et al. (2008) for a BMS when a gold standard is available and by Browne et al. (2009) for a continuous measurement system.

The outline of this article is as follows. In the next section, we quantify the value of the baseline information for the following two classes of plans using the standard plan as a basis for comparison:

Random sampling (RS): n parts are selected at random from the population of previously inspected parts and each sampled part is remeasured r times.

Conditional Sampling (CS): n_0 parts are selected at random from the population of previously failed parts, n_1 parts are selected at random from the population of previously passed parts ($n_0+n_1 = n$) and each sampled part is remeasured r times.

Next, we compare similar CS and RS plans, showing that, when we have a large baseline sample, the CS plan is uniformly better than the RS plan. For CS plans, we then look at the effect of changing f, the fraction of previously passed parts in the sample. In the subsequent section, we discuss planning a BMS assessment with the preferred CS plan. The discussion addresses choosing the number of parts n and the number of repeated measurements r per part as well as the choice of the baseline sample. We conclude with some additional issues and a summary of our recommendations.

Effect of Using Baseline Data

Suppose we have a baseline population of parts each measured once for inspection purposes and we plan to repeatedly measure a sample of these parts r times each. That is, we have m parts measured once and n parts measured $r + 1$ times. For high-volume processes, m is typically large. We reasonably suppose that these baseline parts have been identified as pass or fail. For the m parts measured once only, the likelihood is

$$L_b(\pi_P) \propto \pi_P^z (1-\pi_P)^{m-z}, \tag{4}$$

where z is the number of passed parts in this group. Note that $L_b(\pi_P)$ can be rewritten in terms of α, β and π_C using the constraint (1), though it is not possible to separately estimate α, β, and π_C using this likelihood alone. There are two limiting cases. With m very large, we can assume π_P is known. With $m = 0$, we have the standard assessment plan with $r + 1$ measurements per part. In what follows, we consider the intermediate case with $m = 10{,}000$ and the limiting case when π_P is known.

In an RS plan, suppose that, out of the n sampled parts, we have n_0 that failed the initial inspection and n_1 that passed. If we repeatedly measure these parts r times, then the overall likelihood is proportional to

$$\begin{aligned} L_b(\pi_P) \times \prod_{i=1}^{n_0} & [(1-\beta)^{s_i}\beta^{r+1-s_i}\pi_C \\ & + \alpha^{s_i}(1-\alpha)^{r+1-s_i}(1-\pi_C)] \\ \times \prod_{i=1}^{n_1} & [(1-\beta)^{s_i+1}\beta^{r-s_i}\pi_C \\ & + \alpha^{s_i+1}(1-\alpha)^{r-s_i}(1-\pi_C)], \end{aligned} \tag{5}$$

where s_i is the number of times part i passed inspection in the r repeated measurements. Note that the likelihood [Equation (5)] includes the information about all the baseline parts, and, relative to an SP, includes one extra measurement of the parts that are repeatedly measured because that additional result is available from the earlier inspection. In a CS plan, we randomly select a sample of n_0 parts from the population of previously failed parts, n_1 parts from the population of previously passed parts and remeasure each r times. In writing the likelihood for these data, we condition on the fact that the parts initially failed or passed. The overall likelihood is

$$\begin{aligned} L_b(\pi_P) \times \prod_{i=1}^{n_0} & [(1-\beta)^{s_i}\beta^{r+1-s_i}\pi_C \\ & + \alpha^{s_i}(1-\alpha)^{r+1-s_i}(1-\pi_C)]/(1-\pi_P) \\ \times \prod_{i=1}^{n_1} & [(1-\beta)^{s_i+1}\beta^{r-s_i}\pi_C \\ & + \alpha^{s_i+1}(1-\alpha)^{r-s_i}(1-\pi_C)]/\pi_P. \end{aligned} \tag{6}$$

Note that, for an RS plan, n_0 and n_1 are variable, while, for a CS plan, they are fixed.

To assess the value of the baseline information and to compare the sampling plans, we calculate the asymptotic standard deviation of the MLEs using the Fisher (expected) information matrix corresponding to the likelihoods of Equations (2), (5), and (6) with π_P replaced by the constraint (1). We omit the details of the calculations and the unappealing formulae but note the following:

- In an SP, using Equation (2), each part contributes equally to the information so that the overall information is $n\mathbf{I}(r)$, where $\mathbf{I}(r)$ is the information from repeatedly measuring a single randomly selected part r times.
- In an RS plan, using Equation (5), each of the once measured parts contributes equally to the information, as does each of the parts repeatedly measured, so the overall information has the form $m\mathbf{I}(1) + n\mathbf{I}(r + 1)$.

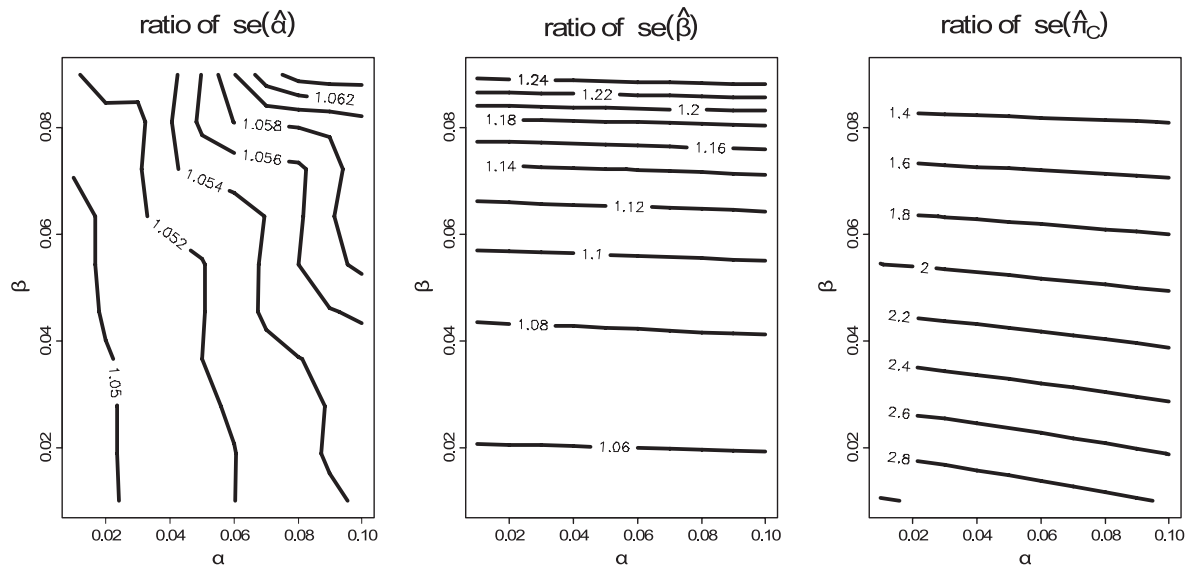

FIGURE 1. Contour Plots of sd(SP)/sd(RS, $m = 10{,}000$) for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_C$ when $\pi_P = 0.9$ and $r = 10$.

- In a CS plan, using Equation (6), the initially failed and passed parts contribute differently to the information, which has the form $m\mathbf{I}(1) + n_0\mathbf{I}_0(r) + n_1\mathbf{I}_1(r) = m\mathbf{I}(1) + n[(1 - f)\mathbf{I}_0(r) + f\mathbf{I}_1(r)]$. We calculate the contributions $\mathbf{I}_0(r)$ and $\mathbf{I}_1(r)$ using the conditional distribution of S_i given the initially measured value.
- In the limiting case with $m \to \infty$, we drop the first factor (4) of the likelihood (5) or (6) and treat π_P as known. For any of the plans we consider in this case, the information is a multiple of n.

We first consider RS plans. Figures 1 and 2 show contour plots of the ratios of the asymptotic standard deviations for the SP to the standard deviations for the RS plan with $m = 10{,}000$ and $m = \infty$ (i.e. π_P known). We give results for $\pi_P = 0.9$, $n = 1{,}000$, and $r = 10$, but the conclusions are valid for other values of π_P, n, and r. In particular, we checked with smaller values of r closer to the identifiability condition boundary $r = 3$ and larger values of π_P. The asymptotic approximations (and all estimation procedures) break down unless we select n large enough to ensure that there are some nonconforming parts in the sample. Note that the comparisons do not depend on the sample size n as $m \to \infty$.

From Figures 1 and 2, we see that, for the RS plan, using the baseline data improves the precision of all estimators, with substantial reductions in the standard deviations for estimating π_C and β, but little improvement for α. Also we see, as expected, that the gain in precision increases with m, the number of once-measured parts. 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.

Next, we conduct a similar comparison for CS plans where we sample equally ($f = 0.5$) from the two populations of previously inspected parts. Figures 3 and 4 compare CS plans with $m = 10{,}000$ and $m = \infty$ to the corresponding SP. The conclusions are quite different than for RS plans. Incorporating the baseline data and using conditional sampling greatly improves the estimation of π_C and α, and, to a lesser degree, improves the estimation of β. We expect this result because the conditional sampling increases the number of nonconforming parts in the sample and the baseline information helps to improve the estimation of all parameters. We see similar results for other values of π_P, n, and r.

Comparing Random and Conditional Selection Plans

Here we compare RS and CS plans when we have baseline information. First consider the risk of having no nonconforming parts in those that are repeatedly measured. No analysis method will be able to esti-

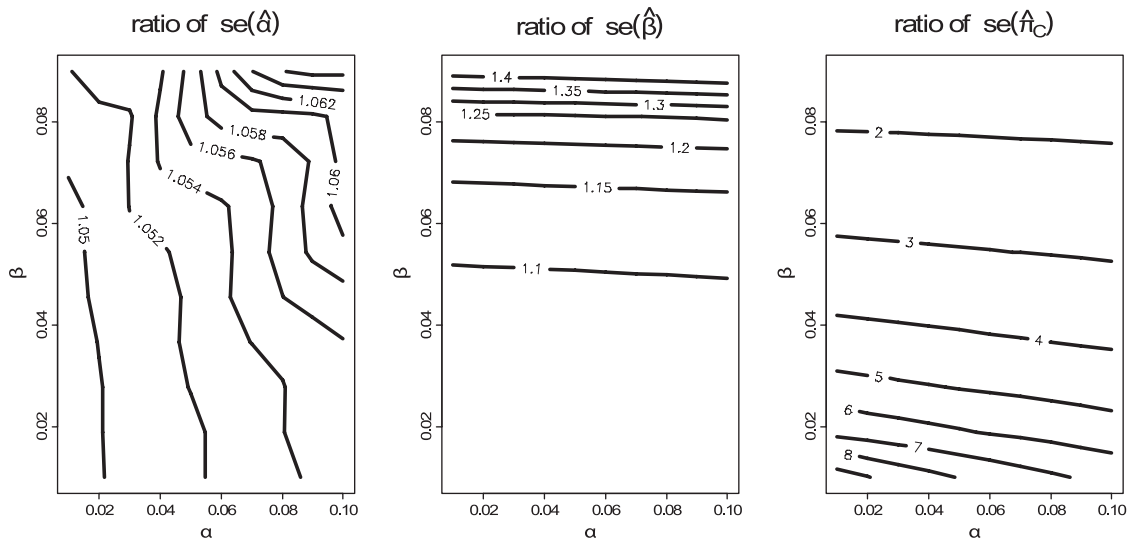

FIGURE 2. Contour Plots of sd(SP)/sd(RS, $m = \infty$) for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_C$ when $\pi_P = 0.9$ and $r = 10$.

mate both α and β unless the sample of repeatedly measured parts contains both conforming and nonconforming parts. In particular, when there are no nonconforming parts, α is not identifiable. When the conforming rate π_C is close to one, which is typical for existing high-performance processes, an RS plan with n small can produce samples with no or only a few nonconforming parts. The probabilities of no nonconforming parts in a RS and CS plan are

$$\Pr(\text{no nonconforming parts in RS}) = \pi_C^n$$

$$\Pr(\text{no nonconforming parts in CS}) = \left[\frac{(1-\beta)\pi_C}{\pi_P} f + \frac{\beta\pi_C}{1-\pi_P}(1-f)\right]^n.$$

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 the RS plan, 0.0001 for CS with $f = 0.5$ and 1.56e^{-09} for CS with $f = 0$. In general, for reasonable n, the probability of having no nonconforming parts in a CS plan is negligible.

We also compare the RS and CS plans in terms of precision of the estimators. When we omit the fac-

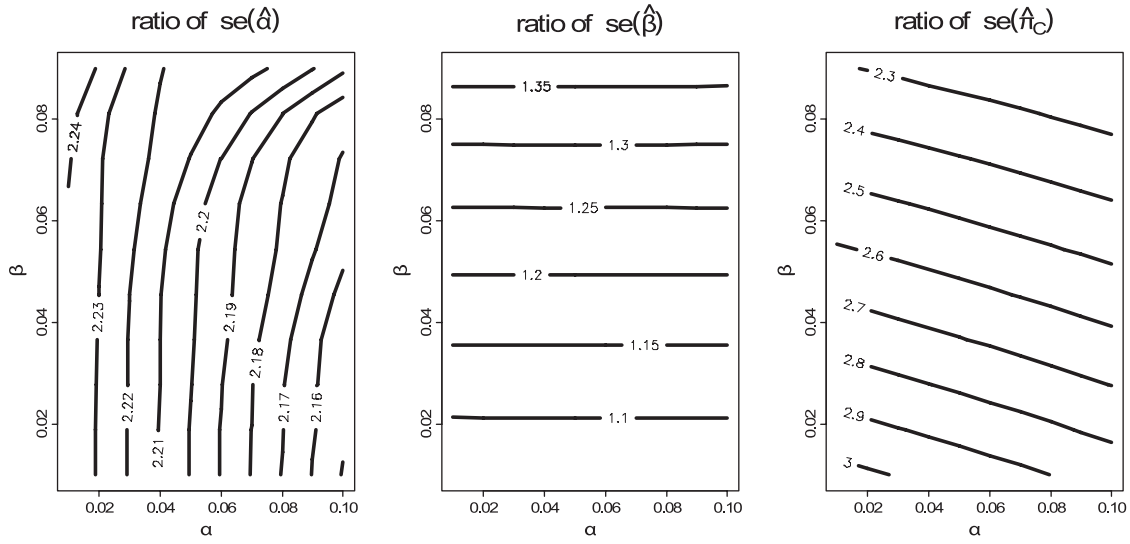

FIGURE 3. Contour Plots of sd(SP)/sd(CS, $m = 10{,}000$) for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_C$ when $\pi_P = 0.9$, $f = 0.5$, and $r = 10$.

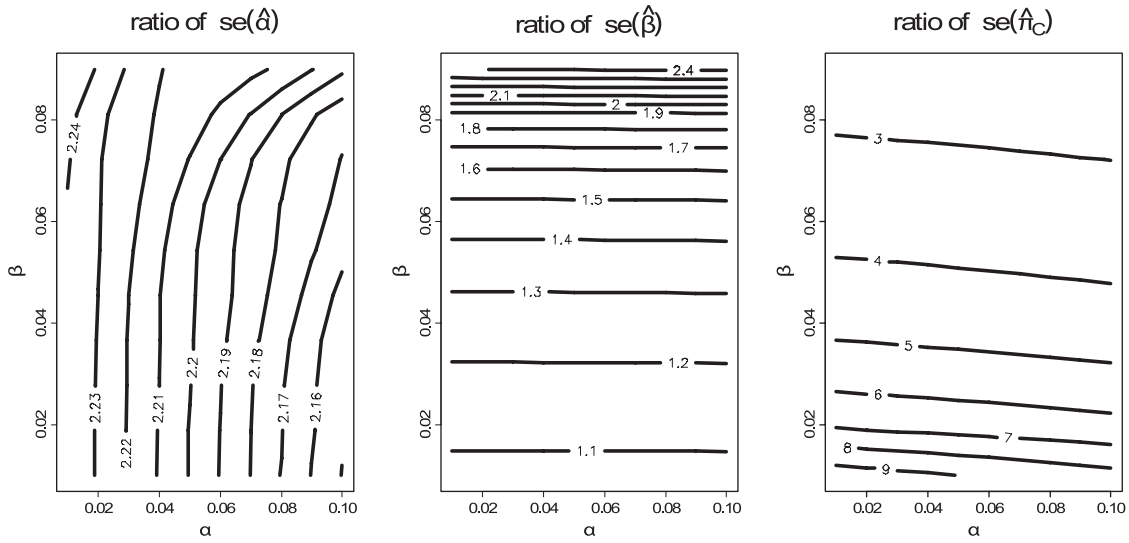

FIGURE 4. Contour Plots of sd(SP)/sd(CS, $m = \infty$) for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_C$ when $\pi_P = 0.9$, $f = 0.5$ and $r = 10$.

tor $L_b(\pi_P)$ from the likelihoods in Equations (5) and (6), i.e., set $m = 0$, the ratio of the standard deviations for the RS and CS plans is given in Figure 5. These ratios do not depend on the sample size n. In this comparison, we are ignoring any information provided by the once-measured parts. The results for the precision of α and β are as expected; we can estimate β better with the RS plan and α better with the CS plan. This makes sense because the expected proportion of nonconforming parts $N_{\bar{C}}/n$ given by RS and CS plans, respectively, are

$$E_{\mathrm{RS}}(N_{\bar{C}}/n) = 1 - \pi_C$$
$$E_{\mathrm{CS}}(N_{\bar{C}}/n) = \frac{\alpha(1-\pi_C)}{(1-\beta)\pi_C + \alpha(1-\pi_C)} f + \frac{(1-\alpha)(1-\pi_C)}{\beta\pi_C + (1-\alpha)(1-\pi_C)}(1-f). \quad (7)$$

For example, with $\alpha = \beta = 0.05$, $\pi_P = 0.9$, and $f = 0.5$, $E_{\mathrm{RS}}(N_{\bar{C}}/n) = 0.055$ while $E_{\mathrm{CS}}(N_{\bar{C}}/n) = 0.265$. Because α is the probability of passing a nonconforming part, it will be better estimated in plans that, on average, contain more nonconforming parts and vice

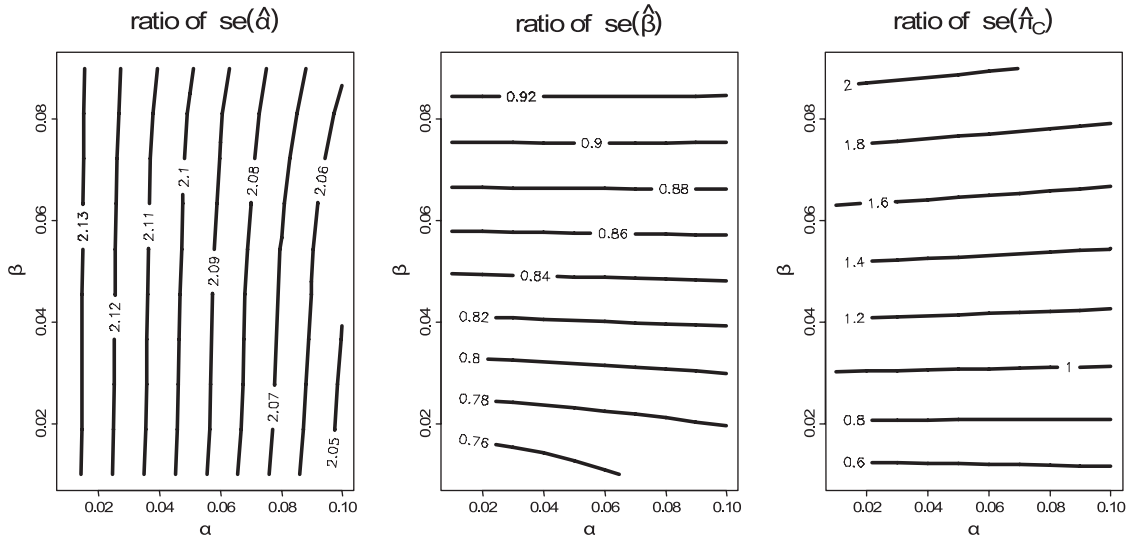

FIGURE 5. Contour Plots of sd(RS, $m = 0$)/sd(CS, $m = 0$) for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_C$ when $\pi_P = 0.9$, $f = 0.5$, and $r = 10$.

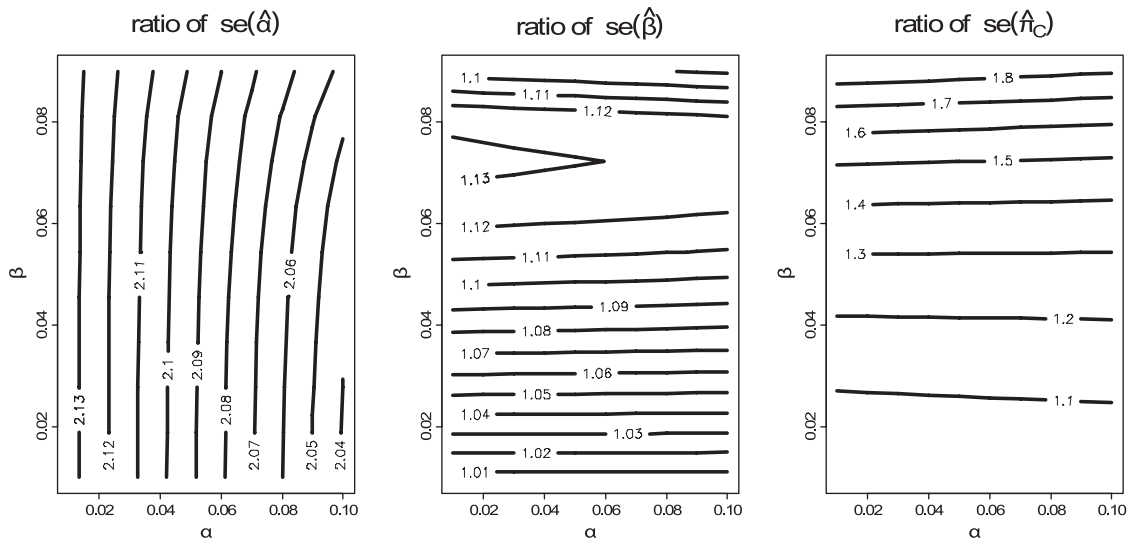

FIGURE 6. Contour Plots of sd(RS, $m = 10{,}000$)/sd(CS, $m = 10{,}000$) for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_C$ when $\pi_P = 0.9$, $f = 0.5$, and $r = 10$.

versa for β. For fixed n, there is a tradeoff because more nonconforming parts means fewer conforming parts. The CS plan in this comparison formalizes the recommendations of Boyles (2001) and Van Wierigen and de Mast (2008) to increase the number of nonconforming parts in the sample.

When we incorporate the data from the once-measured parts, the comparison of the RS and CS plans yields a different conclusion. Figures 6 and 7 compare the asymptotic standard derivations of the RS plan and the CS plan with $f = 0.5$ when the baseline has $m = 10{,}000$ parts and when π_P is known ($m = \infty$).

Figures 6 and 7 show that, when there is baseline information, the CS plan with $f = 0.5$ provides substantially better estimators for all parameters than does the RS plan. Also (not shown here), we see that, as r increases, the relative advantage of CS over RS plans increases for the estimator of α and decreases for the estimators of β and π_C. We also compared

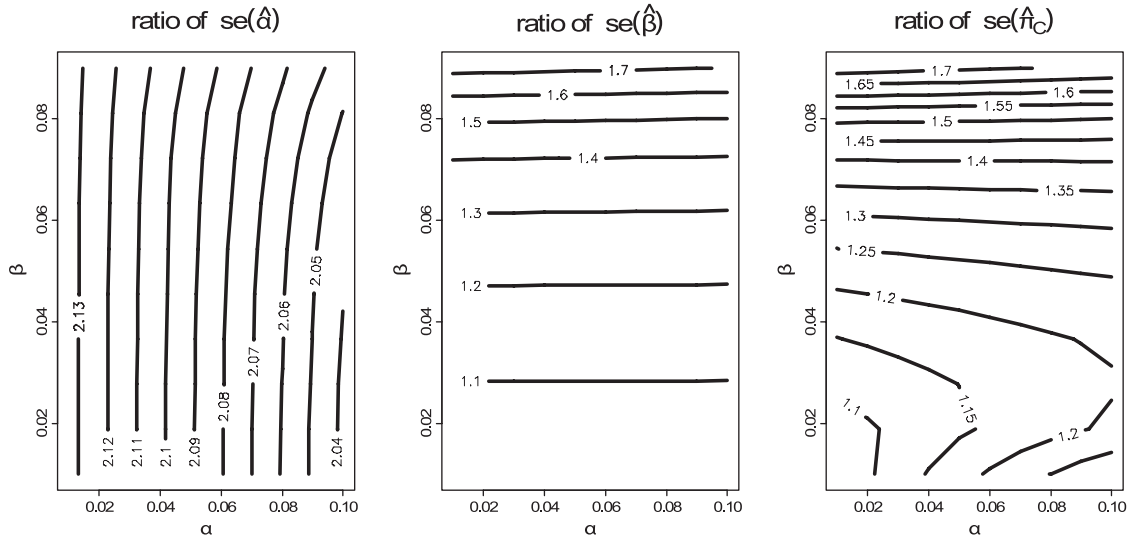

FIGURE 7. Contour Plots of sd(RS, $m = \infty$)/sd(CS, $m = \infty$) for $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\pi}_C$ when $\pi_P = 0.9$, $f = 0.5$, and $r = 10$.

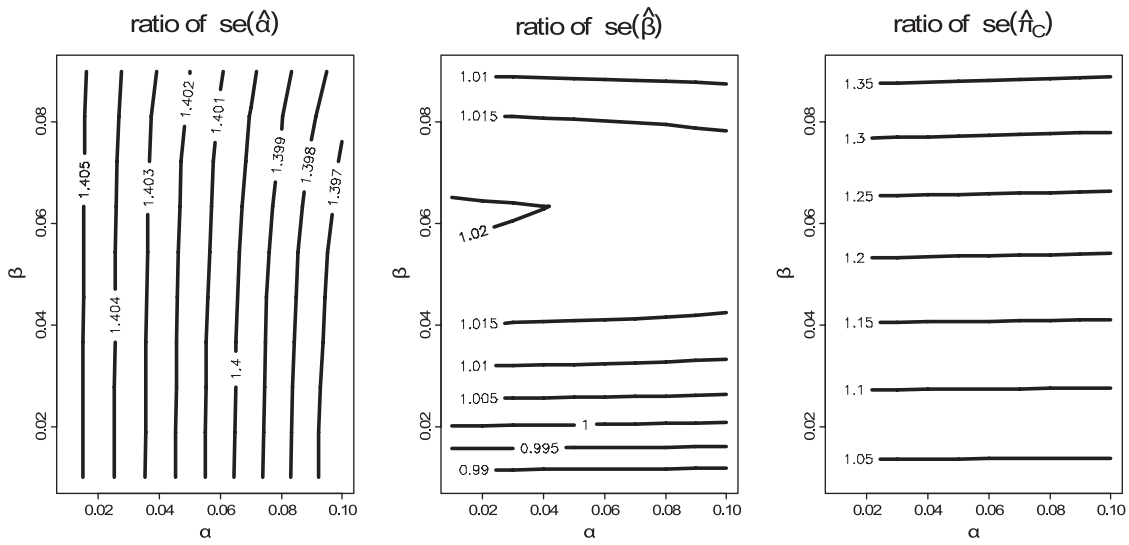

FIGURE 8. Contour Plots of sd(CS, $m = 10{,}000$, $f = 0.5$)/sd(CS, $m = 10{,}000$, $f = 0$) for $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}_C$ when $\pi_P = 0.9$ and $r = 10$.

the precision of the estimators for higher values of the pass rate (e.g., $\pi_P = 0.95$) and concluded that the CS plan becomes even more efficient than the corresponding RS plan when the pass rate increases with r fixed.

Comparing Conditional Sampling Plans as f Changes

In a CS plan, we can choose the proportion of initially passed parts f and hence change the expected number of nonconforming parts in the sample of parts that are to be remeasured. For an SP, Van Wieringen and de Mast (2008) recommend selecting the sample so that $\pi_S = 0.5$, i.e., the sample is expected to have equal numbers of conforming and nonconforming parts. Choosing $f = 0.5$ does not in general achieve this goal. For example, from Equation (7), when $\alpha = 0.05$, $\beta = 0.1$, $\pi_C = 0.95$, and $f = 0.5$, the expected proportion of nonconforming parts in the sample is 0.17. With $f = 0.5$, the expected proportion in Equation (7) is always less than 0.5 for the assumed range of values for α, β, and π_C. Achieving roughly equal numbers of conforming and nonconforming parts requires f values less than 0.5 and is not possible for some values of α, β, and π_C.

When we incorporate the baseline information, the intuition that suggests α (β) will be better estimated by a plan with more nonconforming (conforming) parts no longer holds. To address the question of which conditional sampling plan is the best, in Figures 8 and 9, we compare the precision of each estimator for CS plans with $f = 0.5$ and $f = 0$.

Figures 8 and 9 suggest that, when π_P is known, $f = 0$ is uniformly more efficient than a CS plan with $f = 0.5$, especially when β is large. We also compared the precision of the estimators for smaller values of r (e.g., $r = 5$) and concluded 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 with $f = 0.5$. Also, for larger values of the pass rate ($\pi_P \geq 0.9$), the ratios of the standard deviations are larger for all parameters so that a CS plan with $f = 0$ is even more efficient in that case.

In conclusion, in cases when there is baseline information, α and β are small and π_C (and thus π_P) is close to one, we recommend a Conditional Selection plan with $f = 0$, i.e., all parts are sampled from the population of previously failed parts. For plausible sample sizes, with this plan, there is a negligible chance of having no nonconforming parts in the sample. The plan is substantially more efficient in estimating the parameters α, β, and π_C compared with the other plans we have investigated. In Figure 10, we demonstrate the substantial gain provided by the recommended CS plan compared with the SP that uses random selection and ignores any available baseline information.

We see similar large gains for other values of π_P and r. The choice $f = 0$ may not be optimal for some

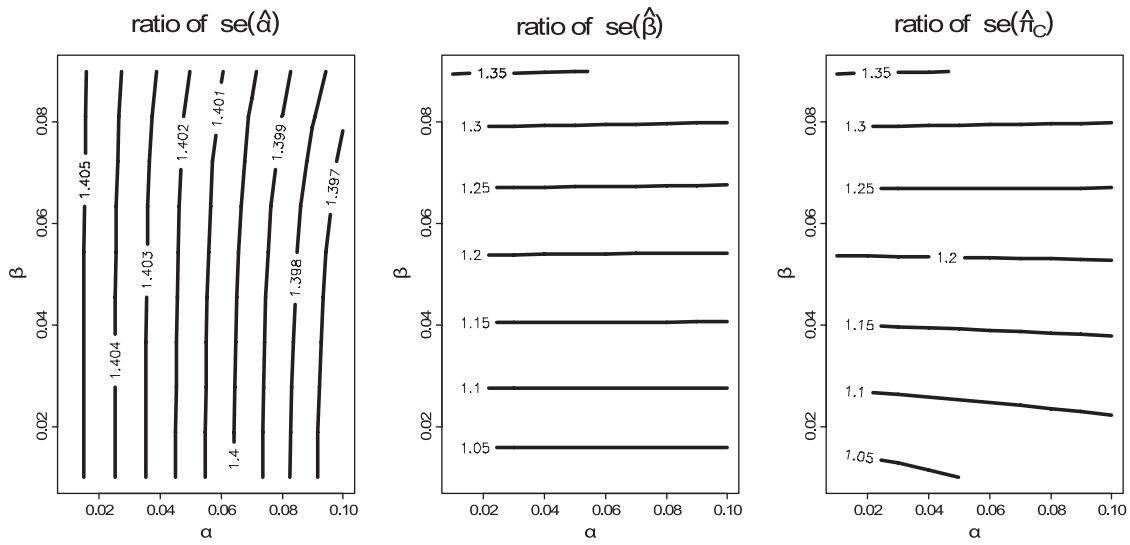

FIGURE 9. Contour Plots of sd(CS, $m = \infty$, $f = 0.5$)/sd(CS, $m = \infty$, $f = 0$) for $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}_C$ when $\pi_P = 0.9$ and $r = 10$.

values of α, β, and π_P, but it is always a good choice that is easy to implement.

Planning the BMS Assessment Study Using Conditional Selection with $f = 0$

Next, we address the design of the recommended CS plan. Because we are assuming that the baseline data are freely available, we suggest that the number of parts in the baseline be as large as possible. One caveat is that we have assumed that α, β, and π_C are constant over the sampling period, i.e., the process is stable. Note that we need only to know the total number of parts inspected and the proportion passing. As well, because the recommended plan has $f = 0$, we need to save a sample of the parts that fail the initial inspection. Failed parts are typically set aside in any case to be repaired or scrapped.

We choose n and r using an algorithm coded in R (2005) that provides feasible combinations that

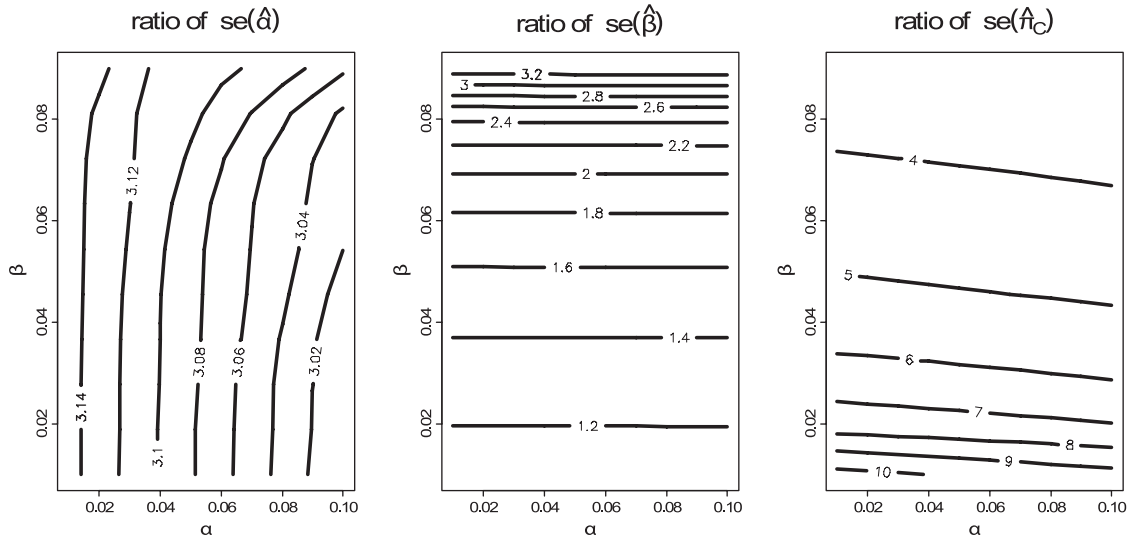

FIGURE 10. Contour Plots of sd(SP)/sd(CS, $m = \infty$, $f = 0$) for $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}_C$ when $\pi_P = 0.9$ and $r = 10$.

TABLE 1. Sample-Size Determination Using the Fisher Information when $\alpha = 0.01$, $\beta = 0.02$, $\pi_P = 0.9$, $f = 0$, $\text{sd}(\hat{\alpha}) = 0.005$, and $\text{sd}(\hat{\beta}) = 0.005$

Sample size n	Repeated measurements r	Total measurements $n \times r$	$\text{sd}(\hat{\alpha})$	$\text{sd}(\hat{\beta})$	$\text{sd}(\hat{\pi}_C)$
170	3	510	0.0050	0.0031	0.0031
125	4	500	0.0050	0.0036	0.0035
98	5	490	0.0050	0.0040	0.0039
81	**6**	**486**	**0.0050**	**0.0044**	**0.0042**
70	7	490	0.0050	0.0047	0.0045
61	8	488	0.0050	0.0050	0.0048
60	9	540	0.0047	0.0050	0.0048
59	10	590	0.0045	0.0050	0.0048

achieve prespecified precision for the estimators of α and β. See www.bisrg.uwaterloo.ca/ for the code. We focus on the precision of the estimates of α and β, rather than π_C, because the misclassification rates are the main parameters of interest. We determine sample-size requirements based on the asymptotic standard deviations for α and β based on the expected information using the likelihood of Equation (6). Boyles (2001) uses an approximation to these standard deviations based on the information if we knew the true status of each part. This leads to simple formulae for choosing n and r. We found the Boyles' approximation inappropriate for our recommended plan. Note that, for reasonable precision requirements, the suggested number of parts and repeated measurements should be large enough for the asymptotic results to be reasonable.

As in most sample-size calculations, we must provide some conjectured values for the unknown parameters α and β as well as the required precision (asymptotic standard deviations) for the estimators of α and β. We also specify the available number of baseline measurements and the proportion of previously passed parts f ($f = 0$ is recommended) in the sample. The output of the algorithm provides a table of combinations of the total number of parts n and the number of repeated measurements r. The output also includes the asymptotic standard deviations for the estimators of α, β, and π_C, along with the probability of having no nonconforming parts in the sample.

To find feasible values for n and r, the algorithm uses a simple search strategy. It starts with the minimum number of repeated measurements $r = 3$ and the minimum of parts $n = 2$ and then increments n until the required precision for the estimators of α and β is achieved. Next, r is increased in one-unit increments and for each r value the corresponding minimum n is determined. The following example illustrates the use of the algorithm. Suppose we know the pass rate $\pi_P = 0.9$ (i.e., we assume $m = \infty$) and select $f = 0$. We also assume that the true (unknown) parameter values are $\alpha = 0.01$ and $\beta = 0.02$. Using Equation (1), solving for π_C gives 0.92. Suppose also the desired precision for the estimators of α and β are $\text{sd}(\hat{\alpha}) = 0.005$ and $\text{sd}(\hat{\beta}) = 0.005$. The corresponding sample size n, the number of repeated measurements r, the total number of measurements $n \times r$, and the resulting asymptotic standard deviations as provided by the algorithm are given in Table 1.

To choose the best combination of n and r, we can select the combination that results in the fewest total number of measurements, in this case $n = 81$ and $r = 6$, or some other combination that takes into account the relative costs of measuring and sampling a part. Note that, in Table 1, the plans with r between 5 and 8 all have roughly the same total number of measurements.

Discussion and Conclusions

We have not discussed the analysis of the data from the recommended CS plan. Following Boyles (2001), we recommend using maximum-likelihood estimation and profile likelihoods to generate confidence intervals for the parameters of interest. With α and β small, confidence intervals using the estimated asymptotic standard deviations are likely

to be inaccurate. We found maximizing the likelihood of Equation (6) was straightforward using the EM algorithm (Van Weiringen and de Mast (2008)) or even easier direct optimization algorithms, such as Broyden–Fletcher–Goldfarb–Shanno (Broyden (1970)) or Nelder–Mead (Nelder and Mead (1965)).

Many authors in both industrial (van Wierigen and de Mast (2008)) and medical (Pepe (2003), Qu et al. (1996), Vacek (1983), Torrance-Rynard and Walter (1997)) contexts have questioned the conditional independence and interchangeability assumptions, key requirements in the use of the LC model. In our context with an automated gauge and no operator effects, the independence assumption is not at issue if the BMS has no memory. The second assumption of interchangeability, i.e., all (non)conforming parts have the same chance of failing an inspection, is hard to justify when the binary measurement is based on one or many underlying characteristics. For our results to be reasonable, we should ensure that there are many conforming and nonconforming parts repeatedly measured so that the estimated α and β are at least estimates of the average misclassification rates across the range of conforming and nonconforming parts. Our proposed sampling scheme is designed to increase the number of presumably rare nonconforming parts.

Van Wierigen and de Mast (2008) recommend using a goodness-of-fit test after the parameter estimation. The proposed method can be adapted to a CS plan, but a simpler less formal approach is to classify each of the remeasured parts as conforming or nonconforming. Assuming the classifications are correct, the number of times each conforming (or nonconforming) part passes the inspection is binomial. It is then easy to build a likelihood-ratio test to look at the hypothesis that there is a common success rate among the conforming and nonconforming parts.

It is common practice in many 100% inspection schemes to remeasure failed parts a second time and ship these parts if they pass the second inspection. Only double failures are set aside to be reworked or scrapped. If we sample parts to be remeasured from only the population of twice-failed parts, we can easily modify the likelihood of Equation (6). We have not investigated the advantages and disadvantages of this plan.

We based all of our comparisons on ratios of asymptotic standard deviations so the actual numbers should not be taken too seriously. We expect that the asymptotic results will break down for the small values of α and β and large values of π_C we considered, when n is small. However, the gains in information using the recommended CS plan incorporating the baseline data rather than the SP are huge and we expect any estimation procedure based on the likelihood of Equation (6) to do much better than that from the corresponding SP. The recommended CS plan can be executed with little or no extra work or cost.

In summary, we have provided a new method for assessing the statistical properties of a BMS when there is no gold-standard method available. The method makes effective use of data that are routinely available from 100% inspection operations. We recommend selecting parts to be remeasured from the parts that initially failed the inspection. Combining the conditional sampling scheme with the available data leads to large improvements in the estimation of the unknown parameters relative to the standard plan for a BMS assessment. We also provide an algorithm for planning a BMS assessment with the recommended CS plan that produces several combinations of the sample size n and number of repeated measurements r for a specified precision of the estimates of the unknown parameters.

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 for Processes with 100% Inspection". *Journal of Quality Technology* 41, pp. 376–388.

Broyden, G. C. (1970). "The Convergence of a Class of Double Rank Minimization Algorithms. Parts I and II". *Journal of Institute of Mathematics and Its Applications* 6, pp. 76–90 and pp. 222–236.

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

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm". *Journal of the Royal Statistical Society, Series B* 39, pp. 1–38.

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

Hui, S. L. and Walter, S. D. (1980). "Estimating the Error Rates of Diagnostic Test". *Biometrics* 36, pp. 167–171

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Boston, MA: Houghton Mifflin Co.

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

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

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.

R Development Core Team. (2005). "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, http://www.R-project.org.

Schilling, E. G. (1982). *Acceptance Sampling in Quality Control*. New York, NY: Marcel Dekker.

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 De Mast, J. (2008). "Measurement System Analysis for Binary Data". *Technometrics* 50, pp. 468–478.

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.