Assessing a Binary Measurement System

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

Business and Industrial Statistics Research Group (BISRG), Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo N2L 3G1, Canada

Assessing measurement systems is a necessary task in all industrial contexts. While a great deal has been written about assessing measurement systems that yield continuous outputs, little work addresses binary measurement systems despite their widespread use. This article proposes two new plans for assessing a binary measurement system that are applicable when we can assume the pass rate of the system is known. This assumption is often reasonable when we need to assess a system used for 100% inspection in a production process. The plans provide estimates of the misclassification rates as well as the proportion of conforming items produced. The two methods are compared to existing plans.

Key Words: Gauge R&R; Go/No-Go Testing; Inspection Errors; Misclassification Rates; Pass–Fail Inspection.

Binary measurement systems (BMSs) that classify items as pass or fail are widely used in industry, especially for 100% inspection purposes. Good measurement systems are essential both for problem solving (i.e., reducing the rate of defectives) and to protect customers from receiving defective products. As a result, it is desirable to assess the performance of the BMS as well as to separate the effects of the measurement system and the production process on the observed classifications.

The automated visual inspection of credit cards provides our motivating example. The measurement system is used to check blank credit cards before they are personalized. In the measurement process, cards are checked for many defects, such as missing parts, surface scratches, bleeding of colors, fuzzy letters and numbers, etc. 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. If any of the summary measures falls outside a prespecified range, the card is rejected.

We define $\pi_c = \Pr(c)$ as the probability of a conforming item, that is, an acceptable credit card; $\pi_p = \Pr(p)$ as the probability the measurement system passes an item, $\alpha = \Pr(p/\bar{c})$ as the probability of passing a nonconforming item; and $\beta = \Pr(\bar{p}/c)$ as the probability of rejecting a conforming item. The performance of the measurement system is summarized by the misclassification rates α and β, while the quality of the production process is determined by π_c. Finally, π_p depends on the properties of both the measurement system and the production process. Note that, in most applications, α is of greater concern than β because α quantifies the risk of nonconforming items reaching the customer.

The misclassification rates α and β, also called "miss rate" (consumer's risk) and "false alarm" (producer's risk), respectively, are commonly used to quantify the performance of a binary measurement system in industry. See, for example, AIAG (2002), Johnson et al. (1991), and Boyles (2001). In medical literature, the performance of diagnostic tests is assessed using the misclassification rates α and β or their complements, called specificity and sensitivity. See, for example, Fleiss (1981), Walter and Irwing (1988), and Pepe (2003). In this context, similar to the industrial setting, estimating the misclassification rates is important because they quantify the

---

Ms. Danila is a PhD Student in the Department of Statistics and Actuarial Science. Her email address is omdanila@math.uwaterloo.ca.

Dr. Steiner is an Associate Professor in the Department of Statistics and Actuarial Science and 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 Department of Statistics and Actuarial Science. He is a member of ASQ. His email address is rjmackay@uwaterloo.ca.

costs and risks due to misclassification of patients. Also, it has been shown (Barron (1977), Quade et al. (1980)) that, when the diagnostic tests are not error free, ignoring the existence of such errors can have serious consequences on estimating indices of association, such as relative risks, odds ratios, etc. Therefore, the measurement error has to be estimated and incorporated in further analyses.

While there is an extensive literature on the assessment of continuous measurement systems (see AIAG (2002), Wheeler and Lyday (1989), Burdick et al. (2005)), much less research has addressed assessing a BMS. AIAG (2002) provides a method that assumes there is an underlying continuous measure that has been discretized. This approach is not feasible in some cases, such as the motivating example, where the classification is based on a large number of tests for conformance.

Farnum (1994) suggests an assessment method for a BMS where two equally sized independent samples of conforming and nonconforming items are selected and then evaluated by the BMS. This plan allows direct estimation of α and β, the misclassification probabilities. However, Farnum's plan is often not practical because it requires sampling from large populations of conforming and nonconforming items, the status of which can only be determined by using a gold-standard measurement device. The cost of creating these two populations can be prohibitive for a process with a small nonconforming rate.

Boyles (2001) presents a latent-class approach that does not require knowledge of the true state of the measured items. In his approach, each unit is measured several times and he assumes that the measurements are conditionally independent. The parameters are identifiable if we reasonably assume that $1 - \beta > \alpha$, that is, conforming items are more likely to be passed than the nonconforming ones. Van Wieringen and van der Heuvel (2005) provide an overview and comparison of Boyle's approach and other methods such as Cohen's kappa, intraclass-correlation coefficient and log-linear models. They conclude the latent-class model is preferred.

In this article, we propose two new plans for assessing a BMS under the following assumptions:

- the pass rate, π_p, is known, and
- the true class of each unit can be determined.

The first assumption is reasonable in cases where the BMS is currently in use. For example, in the credit-card application, thousands of cards are classified each hour so the pass rate is well known. We wish to use this information to help us assess the binary measurement system. The second assumption requires the existence of a "gold-standard" measurement device that can determine whether an item is either conforming or nonconforming. In the example, credit cards can be classified as conforming or nonconforming by a human operator. We assume that the human inspector determines the true status of the cards with no classification error. Because human inspectors are slow and expensive, they are not used in regular production.

The assumption that π_p is known imposes restrictions on the conditional probabilities, α and β, and π_c. Recall that π_p is a function of the performance of both the measurement and production processes. We have $\pi_p = (1 - \beta)\pi_c + \alpha(1 - \pi_c)$, and solving for π_c, we get $\pi_c = (\pi_p - \alpha)/(1 - \beta - \alpha)$. Because $0 \leq \pi_c \leq 1$, we know $0 \leq (\pi_p - \alpha)/(1 - \beta - \alpha) \leq 1$. It is reasonable to assume that $1 - \beta - \alpha > 0$, as the misclassification rates α and β are usually small. Alternatively, this is equivalent to Boyles' (2001) assumption, that is, conforming items are more likely to be passed than the nonconforming ones. Therefore, we have two restrictions, $\alpha \leq \pi_p$ and $\beta \leq 1 - \pi_p$.

Assuming π_p is known allows the indirect estimation of the misclassification rates α and β (see the next section for details) with a more convenient plan than Farnum's and allows us to estimate π_c, i.e., the performance of the production process. By estimating π_c as well as α and β, we can assess the relative contribution of the measurement system and production processes to the pass rate. This mirrors the calculation of the gauge R&R percentage in the continuous measurement case (AIAG (2002)).

In the methods section, we describe the two new plans and give the corresponding maximum likelihood estimates (MLEs) for α, β, and π_c. Next, we compare the two plans using both theoretical and simulation methods and discuss the advantages and disadvantages by comparing the efficiency of the estimators, the costs, and the practicality of each method.

Methods

In Table 1, we summarize the results of a BMS assessment plan where we classify N items using both the BMS under study and the gold-standard measurement device. The notation is straightforward; for

TABLE 1. Summary Data from Assessing a BMS

	Conform (c)	Not conform ($\bar{c}$)	Total Items
Pass (p)	n_{pc}	$n_{p\bar{c}}$	n_p
Reject ($\bar{p}$)	$n_{\bar{p}c}$	$n_{\bar{p}\bar{c}}$	$n_{\bar{p}}$
Total Items	n_c	$n_{\bar{c}}$	N

instance, n_{pc} represents the observed number of conforming items that passed. Note that, with Farnum's method, n_c, $n_{\bar{c}}$, and N are fixed and all the other quantities are random.

Plan I (Binomial)

In the binomial plan, we select two independent samples, one containing n_p items from the population of passed items and another of size $n_{\bar{p}}$ from the population of rejected items. Then, the true state—conforming or nonconforming—is determined for all selected items. Thus, in Table 1, n_p, $n_{\bar{p}}$, and are fixed and all the other quantities are random.

With this plan, we cannot directly estimate the two misclassification probabilities α and β. Instead, we start with the MLEs: $\widehat{\Pr}(\bar{c}/p) = n_{p\bar{c}}/n_p$ and $\widehat{\Pr}(\bar{c}/\bar{p}) = n_{\bar{p}\bar{c}}/n_{\bar{p}}$. Then, by Bayes' Rule,

$$\alpha = \frac{\Pr(p\cap\bar{c})}{\Pr(\bar{c})} = \frac{\Pr(\bar{c}/p)\Pr(p)}{\Pr(\bar{c}/p)\Pr(p) + \Pr(\bar{c}/\bar{p})\Pr(\bar{p})}$$

and

$$\beta = \frac{\Pr(\bar{p}\cap c)}{\Pr(\bar{c})} = \frac{\Pr(c/\bar{p})\Pr(\bar{p})}{\Pr(c/\bar{p})\Pr(\bar{p}) + \Pr(c/p)\Pr(p)}.$$

We also have $\pi_c = \Pr(c\cap p) + \Pr(c\cap\bar{p}) = \Pr(c/p)\cdot \Pr(p) + \Pr(c/\bar{p})\Pr(\bar{p})$. Using the invariance property of the MLEs, we obtain the following estimates for α, β, and π_c from Plan I:

$$\hat{\alpha}(I) = \frac{\pi_p \dfrac{n_{p\bar{c}}}{n_p}}{\pi_p \dfrac{n_{p\bar{c}}}{n_p} + (1-\pi_p)\dfrac{n_{\bar{p}\bar{c}}}{n_{\bar{p}}}} = \frac{\pi_p n_{p\bar{c}} n_{\bar{p}}}{\pi_p n_{p\bar{c}} n_{\bar{p}} + (1-\pi_p) n_{\bar{p}\bar{c}} n_p}$$

$$\hat{\beta}(I) = \frac{(1-\pi_p)\dfrac{n_{\bar{p}c}}{n_{\bar{p}}}}{(1-\pi_p)\dfrac{n_{\bar{p}c}}{n_{\bar{p}}} + \pi_p \dfrac{n_{pc}}{n_p}} = \frac{(1-\pi_p) n_{\bar{p}c} n_p}{(1-\pi_p) n_{\bar{p}c} n_p + \pi_p n_{pc} n_{\bar{p}}}$$

$$\hat{\pi}_c(I) = \pi_p \frac{n_{pc}}{n_p} + (1-\pi_p)\frac{n_{\bar{p}c}}{n_{\bar{p}}}. \tag{1}$$

With Plan I, $\hat{\pi}_c(I)$ is an unbiased estimator and the biases for both $\hat{\alpha}(I)$ and $\hat{\beta}(I)$ are of order $1/N$. We can also derive approximations for the variances of the estimators corresponding to $\hat{\alpha}(I)$ and $\hat{\beta}(I)$ in terms of α, β, and π_p (known), while the variance of the estimator corresponding to $\hat{\pi}_c(I)$ can be directly derived. See Danila et al. (2006) for detailed results and derivations. All approximations in this article are derived using the δ *method* (Casella and Berger (2002)). We obtain the results

$$\mathrm{Var}(\hat{\alpha}(I)) \simeq \frac{\alpha(1-\alpha)(\pi_p-\alpha)}{1-\beta-\pi_p} \times \left[\frac{1-\alpha-\beta+\alpha\beta}{n_p} + \frac{\alpha\beta}{n_{\bar{p}}}\right]$$

$$\mathrm{Var}(\hat{\beta}(I)) \simeq \frac{\beta(1-\beta)(1-\beta-\pi_p)}{\pi_p-\alpha} \times \left[\frac{\alpha\beta}{n_p} + \frac{1-\beta-\alpha+\alpha\beta}{n_{\bar{p}}}\right]$$

$$\mathrm{Var}(\hat{\pi}_c(I)) = \frac{(1-\beta-\pi_p)(\pi_p-\alpha)}{(1-\alpha-\beta)^2} \times \left[\frac{\alpha(1-\beta)}{n_p} + \frac{\beta(1-\alpha)}{n_{\bar{p}}}\right].$$

Figures 1 and 2 illustrate how the standard deviations of the estimators vary with the proportion of the passed items, i.e., n_p/N, for some specific values of α, β, and π_p.

In Figure 1, we notice that $\mathrm{Std}(\hat{\alpha}(I))\sqrt{N}$ decreases slowly from 0.86 to 0.57 over the interval $[0.4, 1)$, the function being close to flat over this range. Therefore, to estimate α, selecting the percentage of passed items anywhere from 40% to 99% gives roughly the same results. $\mathrm{Std}(\hat{\beta}(I))\sqrt{N}$ increases very slowly over the entire interval for the proportion of passed items. In Figure 2, $\mathrm{Std}(\hat{\pi}_c(I))\sqrt{N}$ varies from 0.047 to 0.057 over the interval $[0.2, 0.8]$ for the proportion of passed items. The function is almost flat from 0.25 to 0.75.

In summary, if the main goal of the study is to estimate α and β, we can select as many as 75% passed items and 25% rejected items and have almost

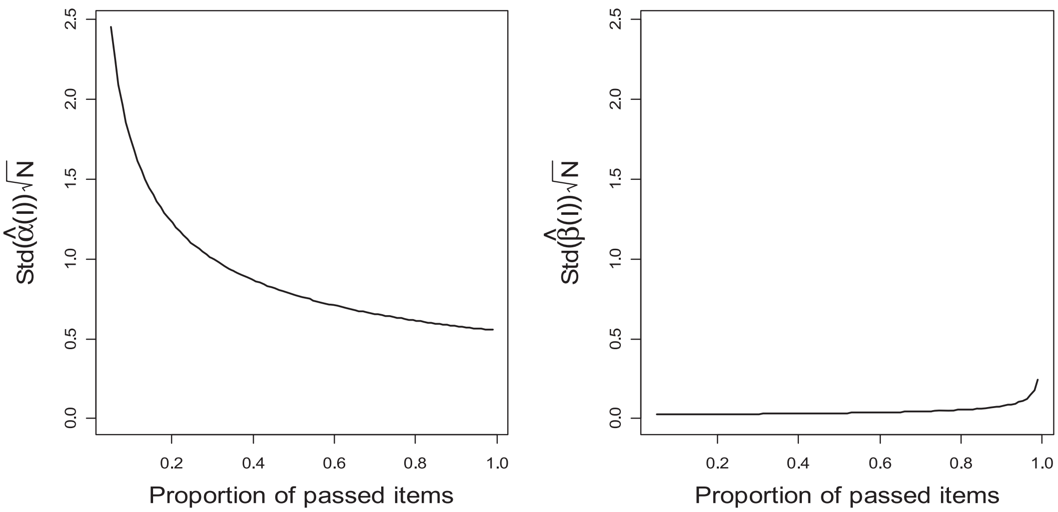

FIGURE 1. Std$(\hat{\alpha}(I))\sqrt{I}$ and Std$(\hat{\beta}(I))\sqrt{I}$ as Functions of the Proportion of Passed Items. Total sample size is N and $\alpha = 0.01$, $\beta = 0.02$, $\pi_p = 0.95$.

the same precision for the estimates as when selecting 50% passed and 50% rejected. This result makes Plan I appealing, as it may be difficult to select a large number of rejected items, $1 - \pi_p$ usually being small.

Plan I is much easier to implement than Farnum's plan. Because we assume the measurement system has been operating for some time, selecting n_p passed items and $n_{\bar{p}}$ rejected items is straightforward and only N $(= n_p + n_{\bar{p}})$ items have to be classified by the gold-standard measurement system.

Plan II (Multinomial)

In Plan II (called the multinomial method), N items are selected, then classified as conforming or nonconforming by the gold-standard system and then measured with the BMS. Thus, with Plan II, in Table 1, only N is fixed and all other quantities are random.

Following a similar derivation as with Plan I, the multinomial plan gives the following MLEs:

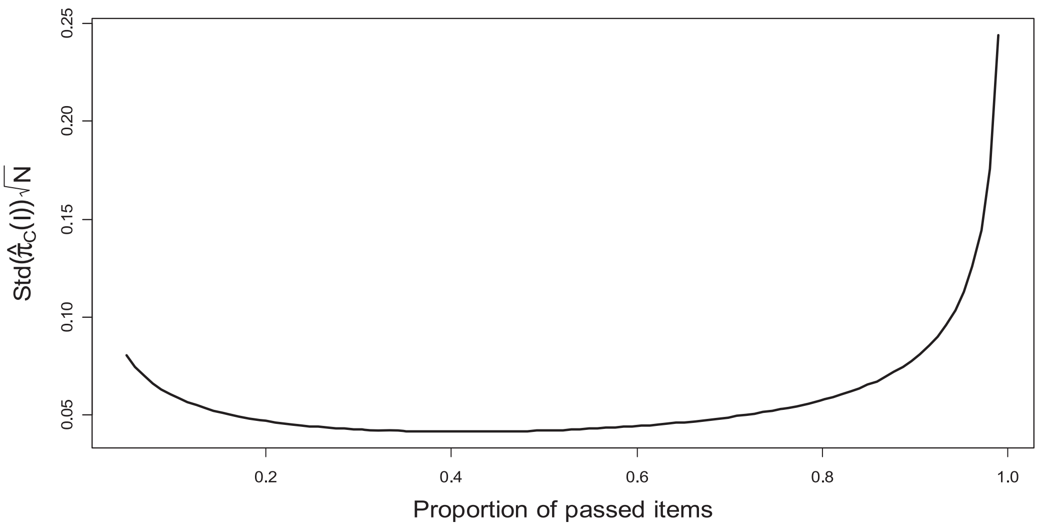

FIGURE 2. Std$(\hat{\pi}_C(I))\sqrt{N}$ as a Function of the Proportion of Passed Items. Sample size in the total sample size, N, when $\alpha = 0.01$, $\beta = 0.02$, $\pi_p = 0.95$.

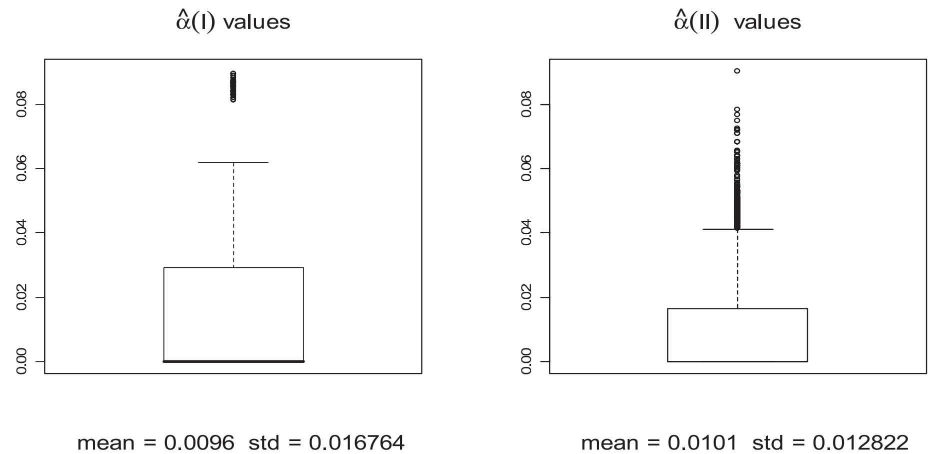

FIGURE 3. Simulated Distributions of the Estimate for α Using Plan I and Plan II, $\alpha = 0.01$, $\beta = 0.02$, $\pi_p = 0.95$, and the Total Sample Size Is 2,000 Items.

$$\begin{aligned}\hat{\alpha}(II) &= \pi_p n_{p\bar{c}}(n_{\bar{p}c} + n_{\bar{p}\bar{c}}) \\ &\quad \div [\pi_p n_{p\bar{c}}(n_{\bar{p}c} + n_{\bar{p}\bar{c}}) \\ &\qquad + (1-\pi_p) n_{\bar{p}\bar{c}}(n_{pc} + n_{p\bar{c}})] \\ \hat{\beta}(II) &= (1-\pi_p) n_{\bar{p}c}(n_{pc} + n_{p\bar{c}}) \\ &\quad \div [(1-\pi_p) n_{\bar{p}c}(n_{pc} + n_{p\bar{c}}) \\ &\qquad + \pi_p n_{pc}(n_{\bar{p}c} + n_{\bar{p}\bar{c}})] \\ \hat{\pi}_c(II) &= \frac{\pi_p n_{pc}}{n_{pc} + n_{p\bar{c}}} + \frac{(1-\pi_p) n_{\bar{p}c}}{n_{\bar{p}c} + n_{\bar{p}\bar{c}}}.\end{aligned}$$

Note that with Plan II, in contrast with Plan I, the proportions of passed and rejected items are random.

The estimators for α and β given by the multinomial method have biases of order $1/N$, and the estimator of π_c is unbiased. The detailed derivations of the MLEs using Lagrange multipliers and of approximations for the variances and biases of the estimators can be found in Danila et al. (2006).

Comparison of Plans I and II

In this section, both simulation and theoretical results are used to compare Plans I and II. If not otherwise stated, we use the special case of Plan I where the number of passed items equals the number of rejected items, i.e., $n_p = n_{\bar{p}} = N/2$. Figures 3, 4, and 5 show simulated distributions of the estimators for α, β, and π_p for some specific plausible values of α, β, and n_p.

From Figures 3, 4, and 5, we conclude that Plan II

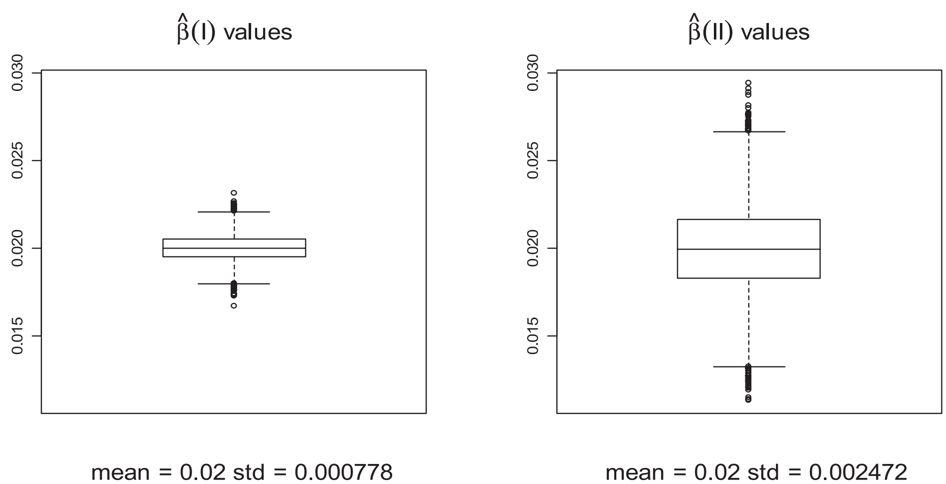

FIGURE 4. Simulated Distributions of the Estimate for β Using Plan I and Plan II, $\alpha = 0.01$, $\beta = 0.02$, $\pi_p = 0.95$, and the Total Sample Size Is 2,000 Items.

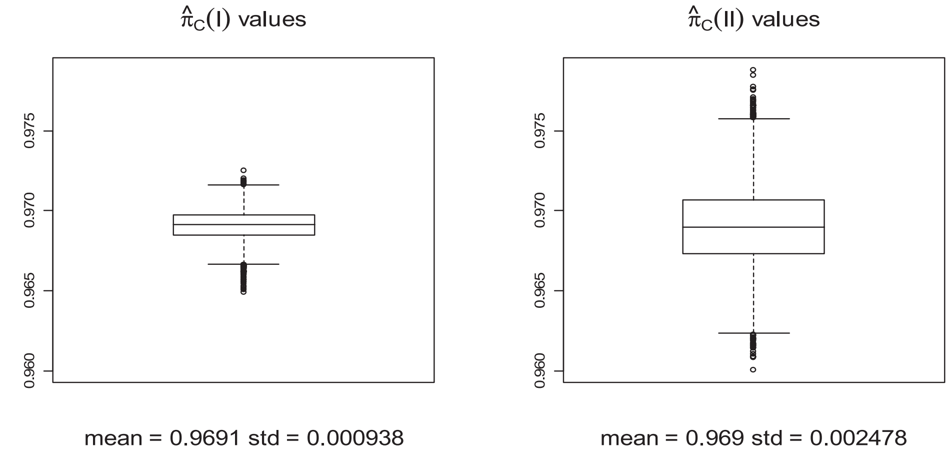

FIGURE 5. Simulated Distributions of the Estimate for π_c Using Plan I and Plan II, $\alpha = 0.01$, $\beta = 0.02$, $\pi_p = 0.9$, and the Total Sample Size Is 2,000 Items.

is better at estimating α than Plan I, while Plan I is better for estimating β and π_c. Also, in general, the theoretical approximated standard deviations based on the delta method and the simulated standard deviations agree closely, differing by at most 3.3% for the conditions considered here. As a result, for the remaining comparisons, we use the theoretical approximations.

Next, we compare the standard deviations of the estimators over a large range of α and β values for different values of π_p. Figures 6–8 show contours for the ratio of the estimator standard deviations from Plan I and II for α, β, and π_c, respectively, where we assume with Plan I an equal number of passed and failed items are selected.

Figure 6 shows that Plan II gives a consistently better estimate for α over all values of α and β and different values of π_p. Also, as π_p gets larger, the ratio gets larger for the same values of α, and β has a smaller influence on the ratio. From Figure 7, we see that Plan I gives better estimates of β than Plan II over the whole range of α, β, and π_p. Also, the ratio varies slightly over the α and β ranges and the ratio becomes less sensitive to β for larger values of π_p (e.g., 0.95). In Figure 8, we notice that Plan I gives a more precise estimate for π_p for small values of π_p and α. When $\pi_p = 0.95$, β has little influence on the ratio, especially for α less than 0.05.

To implement one of the proposed plans, it may be useful to determine the sample size required to

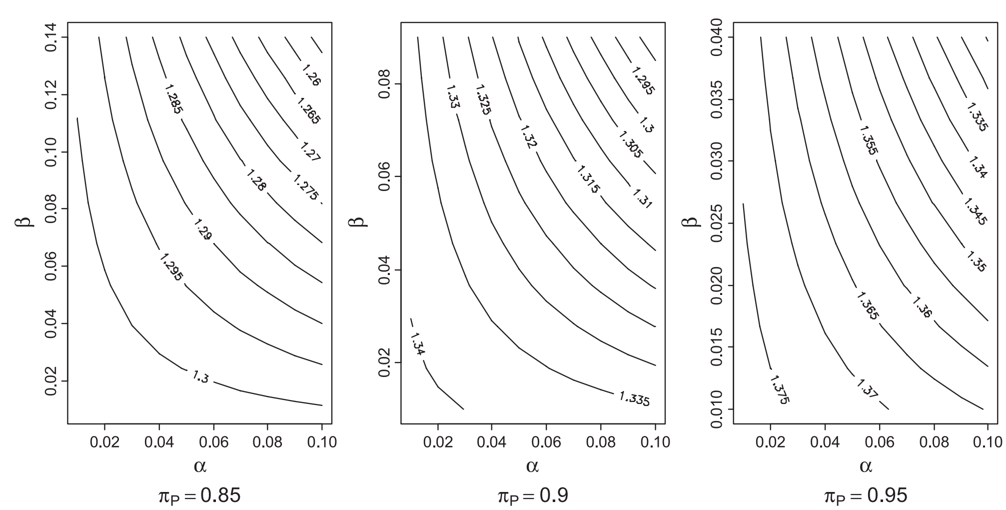

FIGURE 6. Contour Plots for Std($\hat{\alpha}(I)$)/Std($\hat{\alpha}(II)$).

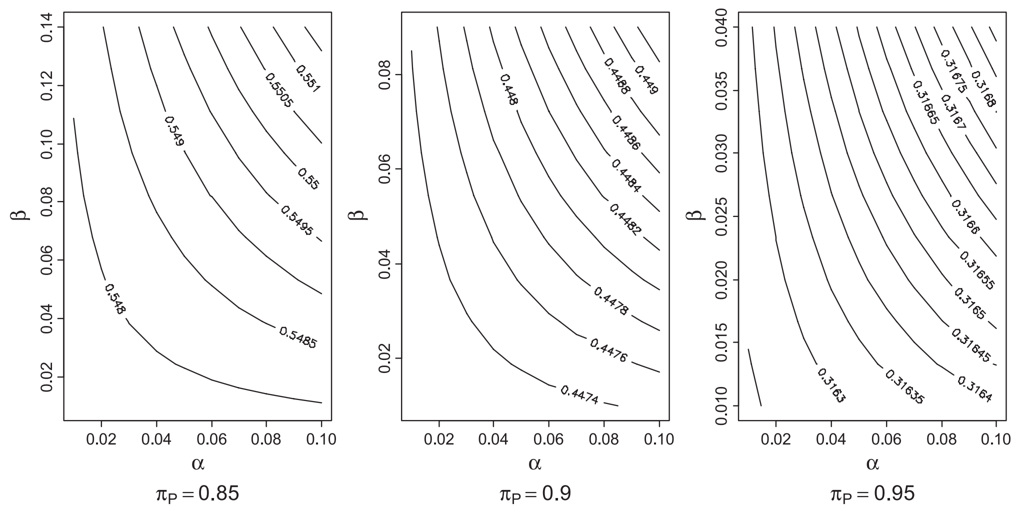

FIGURE 7. Contour Plots for Std($\hat{\beta}(I)$)/Std($\hat{\beta}(II)$).

attain a certain precision in estimating α, β, or π_c. As in most sample-size determination problems, general results are not possible unless we assume the parameters are known. The results below give the minimum sample size for Plan I that can achieve the desired precision, denoted std_0:

$$N_0(\alpha(I)) = \frac{(\pi_p - \alpha)\alpha(1-\alpha)[(1-f)(1-\beta-\alpha)+\alpha\beta]}{[\mathrm{std}_0(\hat{\alpha}(I))]^2 f(1-f)(1-\beta-\pi_p)}$$

$$N_0(\beta(I)) = \frac{(\alpha\beta + f - \beta f - \alpha f)\beta(1-\beta)(1-\beta-\pi_p)}{[\mathrm{std}_0(\hat{\beta}(I))]^2 f(1-f)(\pi_p-\alpha)}$$

$$N_0(\pi_c(I)) = \frac{(\alpha+\beta f - \alpha\beta - \alpha f)(1-\beta-\pi_p)(\pi_p-\alpha)}{[\mathrm{std}_0(\hat{\pi}_c(I))]^2 f(1-f)(1-\beta-\alpha)^2}, \tag{2}$$

where f is the selected proportion of the passed items in the total sample size.

If we are interested in using Plan II, the complicated functions that give the required sample sizes similar to Equation (2) can be found in Danila et al. (2006). Alternatively, we may use an indirect method based on Equations (2) and the contour plots of Figures 6–8. The contour plots give the relative precision

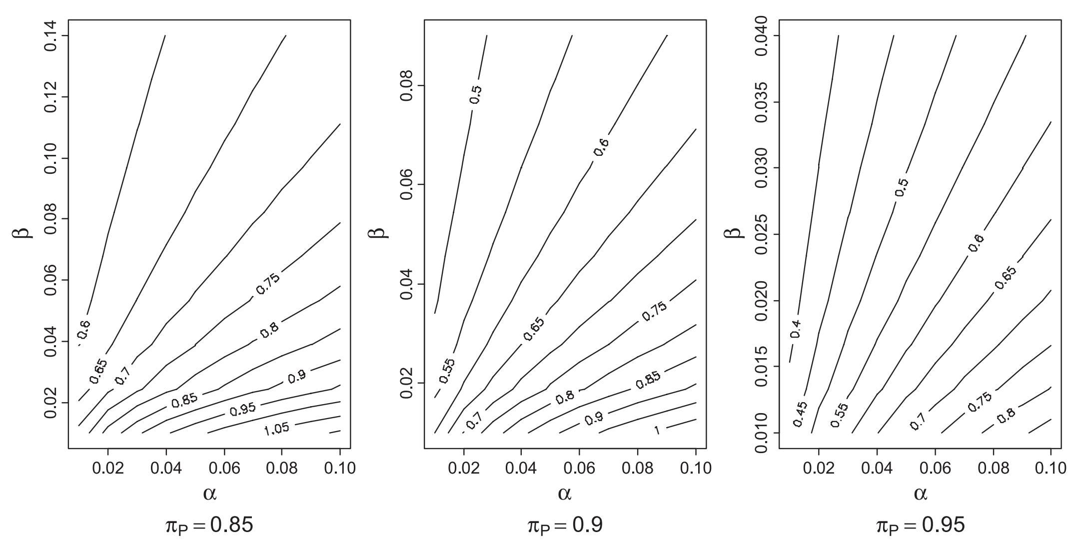

FIGURE 8. Contour Plots for Std($\hat{\pi}_c(I)$)/Std($\hat{\pi}_c(II)$).

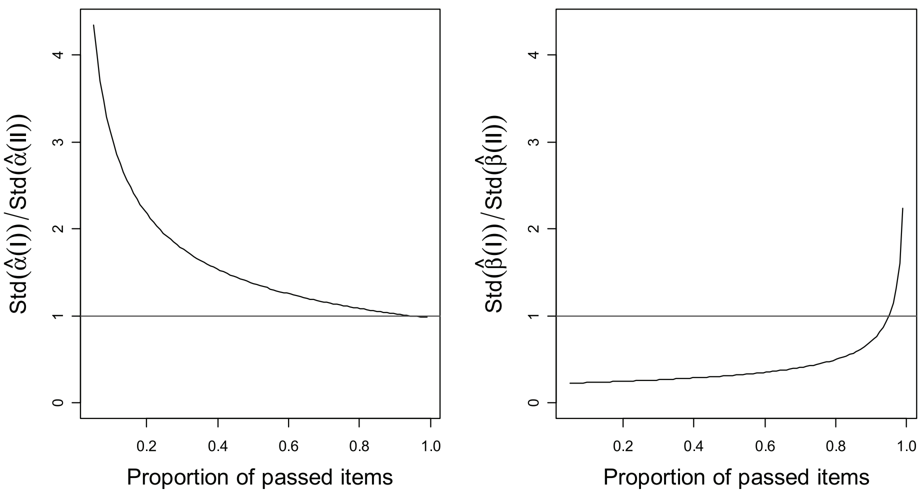

FIGURE 9. Ratio of the Standard Deviations for the α and β Estimators $\alpha = 0.01$, $\beta = 0.02$, and $\pi_p = 0.95$.

of the two plans when the sample size is equal. Once we have the ratio, we can derive the corresponding standard deviation given by Plan I and use Equation (2) to get the required sample size for Plan II.

For example, suppose our goal is to use Plan II to estimate α with a standard deviation of at most 0.02 (i.e., $\text{std}_0(\hat{\alpha}(II)) = 0.02$) and we assume $\alpha = 0.01$, $\beta = 0.02$, and $\pi_p = 0.95$. Figure 6 shows that, for the given values of α, β, and π_p, the ratio $\text{Std}_0(\hat{\alpha}(I))/\text{Std}_0(\hat{\alpha}(II))$ is roughly 1.375. From this ratio, we determine that using Plan I with the same sample size will give $\text{Std}_0(\hat{\alpha}(I)) = 0.0275$. Finally, setting $\text{std}_0(\hat{\alpha}(I)) = 0.275$ and $f = 0.5$ in (2), the minimum sample size for Plan II is roughly 800.

It is also of interest to compare the precision of the estimators when we use different sample sizes for passed and failed items in Plan I, i.e., $n_p \neq n_{\bar{p}}$. Figures 9 and 10 illustrate how the ratio of the standard deviations for the estimators for the two plans varies with the proportion of the passed items. We see that Plan II gives a better estimate of α, regardless of the proportion of the passed items.

For estimating β, Plan I is better unless the proportion of passed items is very high. To estimate π_c, we notice that the ratio is smaller than unity unless the proportion of passed items is larger than about 0.9.

Conclusions

In this paper, we propose two new methods for estimating the misclassification probabilities that characterize the quality of a binary measurement sys-

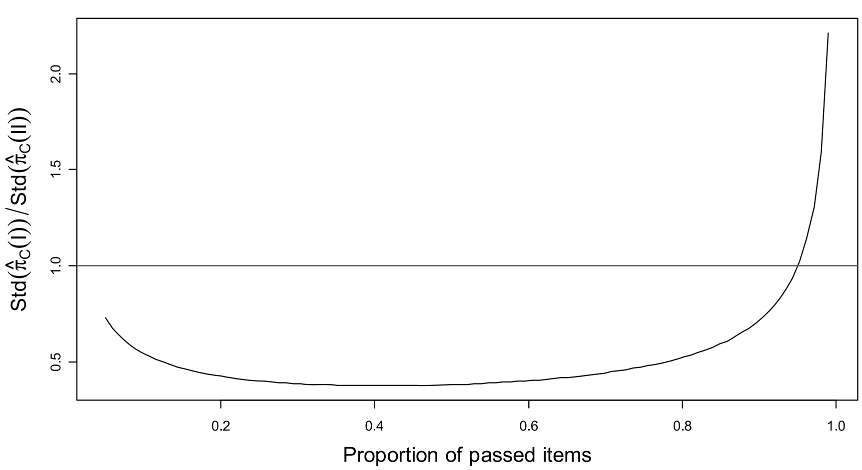

FIGURE 10. Ratio of Standard Deviations for π_c when $\alpha = 0.01$, $\beta = 0.02$, and $\pi_p = 0.95$.

tem and the probability of having a conforming item, which is a measure of the process quality. We consider the situation where we assumed that the passing rate, π_p, is known and the true class of any item can be determined.

Neither of the two methods was found to be consistently better than the other for estimating all the parameters. The best method depends on the objectives of the study. If we are primarily interested in estimating α, the probability of passing a nonconforming item, we recommend Plan II (multinomial). This method gives a more precise estimator than Plan I for a large range of α, β, and π_p values. On the other hand, if the study objective is to assess β, the probability of rejecting a conforming item, or to assess the process quality, i.e., π_c, it is better to use Plan I (Binomial). The standard deviation of the β estimator given by Plan I varies very slowly as we change the proportion of the passed items selected. Therefore, we can select unequal numbers of passed and rejected items. This makes Plan I easier to implement, as it is usually more difficult to obtain rejected as opposed to passed items. The standard deviation of π_c was almost constant when the proportion of the passed items sample size varied from 0.25 to 0.75. Therefore, we can select as many as 75% passed items and 25% rejected items and still have a precise estimate of π_c.

References

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

Barron, B. A. (1977). “The Effects of Misclassification on the Estimation of Relative Risk.” *Biometrics* 33, pp. 414–418.

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

Burdick, R. K.; Borror, C. M.; and Montgomery, D. C. (2005). *Design and Analysis of Gauge R&R Studies: Making Decisions with Confidence Intervals in Random and Mixed ANOVA Models*. ASA-SIAM Series on Statistics and Applied Probability, 17.

Casella G. and Berger, R. L. (2002). *Statistical Inference*, 2nd edition. Brooks/Cole Publishing Company, Pacific Grove, CA.

Danila, O.; Steiner, S. H.; and MacKay, R. J. (2006). “Assessing a Binary Measurement System.” Business and Industrial Statistics Research Group, Research Report 06-01, University of Waterloo (www.bisrg.uwaterloo.ca/archive/RR-06-01.pdf).

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

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edition. Wiley & Sons Inc., New York, NY.

Johnson, N. L.; Kotz, S.; and Wu, X. (1991). *Inspection Errors for Attributes in Quality Control*, 1st edition. Chapman and Hall, New York, NY.

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

Quade, D.; Lachenbruch, P. A.; Whaley, F. S., et al. (1980). “Effects of Misclassifications of Statistical Inference in Epidemiology.” *American Journal of Epidemiology* 111, pp. 503–515.

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.

Wheeler, D. J. and Lyday, R. W. (1989). *Evaluating the Measurement Process*, 2nd edition. SPC Press, Inc., Knoxville, TN.