Assessing Binary Measurement Systems

by

Oana Mihaela Danila

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Binary measurement systems (BMS) are widely used in both manufacturing industry and medicine. In industry, a BMS is often used to measure various characteristics of parts and then classify them as pass or fail, according to some quality standards. 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. In medicine, BMSs are known as diagnostic or screening tests, and are used to detect a target condition in subjects, thus classifying them as positive or negative. Assessing the performance of a medical test is essential in quantifying the costs due to misclassification of patients, and in the future prevention of these errors.

In both industry and medicine, the most commonly used characteristics to quantify the performance a BMS are the two misclassification rates, defined as the chance of passing a non-conforming (non-diseased) unit, called the consumer's risk (false positive), and the chance of failing a conforming (diseased) unit, called the producer's risk (false negative). In most assessment studies, it is also of interest to estimate the conforming (prevalence) rate, i.e. probability that a randomly selected unit is conforming (diseased).

There are two main approaches for assessing the performance of a BMS. Both approaches involve measuring a number of units one or more times with the BMS. The first one, called the "gold standard" approach, requires the use of a gold-standard measurement system that can

determine the state of units with no classification errors. When a gold standard does not exist, is too expensive or time-consuming, another option is to repeatedly measure units with the BMS, and then use a latent class approach to estimate the parameters of interest. In industry, for both approaches, the standard sampling plan involves randomly selecting parts from the population of manufactured parts.

In this thesis, we focus on a specific context commonly found in the manufacturing industry. First, the BMS under study is nondestructive. Second, the BMS is used for 100% inspection or any kind of systematic inspection of the production yield. In this context, we are likely to have available a large number of previously passed and failed parts. 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, separate from the assessment study. Finally, we assume that during the time of the evaluation, the process is under statistical control and the BMS is stable.

Our main goal is to investigate the effect of using sampling plans that involve random selection of parts from the available populations of previously passed and failed parts, i.e. conditional selection, on the estimation procedure and the main characteristics of the estimators. Also, we demonstrate the value of combining the additional information provided by the baseline data with those collected in the assessment study, in improving the overall estimation procedure. We also examine how the availability of baseline data and using a conditional selection sampling plan affect recommendations on the design of the assessment study.

In Chapter 2, we give a summary of the existing estimation methods and sampling plans for a BMS assessment study in both industrial and medical settings, that are relevant in our context. In Chapters 3 and 4, we investigate the assessment of a BMS in the case where we assume that the misclassification rates are common for all conforming/nonconforming parts and that repeated measurements on the same part are independent, conditional on the true state of the part, i.e. conditional independence. We call models using these assumptions fixed-effects models. In Chapter 3, we look at the case where a gold standard is available, whereas in Chapter 4, we investigate the "no gold standard" case. In both cases, we show that using a conditional selection plan, along with the baseline information, substantially improves the accuracy and precision of

the estimators, compared to the standard sampling plan.

In Chapters 5 and 6, we investigate the case where we allow for possible variation in the misclassification rates within conforming and nonconforming parts, by proposing some new random-effects models. These models relax the fixed-effects model assumptions regarding constant misclassification rates and conditional independence. As in the previous chapters, we focus on investigating the effect of using conditional selection and baseline information on the properties of the estimators, and give study design recommendations based on our findings.

In Chapter 7, we discuss other potential applications of the conditional selection plan, where the study data are augmented with the baseline information on the pass rate, especially in the context where there are multiple BMSs under investigation.

Acknowledgments

First and foremost, I would like to thank my supervisors Dr. Jock MacKay and Dr. Stefan Steiner for their wonderful guidance and constant support. Throughout my years at the University of Waterloo they have been most generous in sharing their knowledge and experience. I have learned so much from them and I am truly grateful to have had both of them as my supervisors. Our collaboration has been fun and intellectually stimulating due to their enthusiasm about our work. I am grateful also that they have given of their time to help me, whether it was teaching, consulting or working on other endeavors. I cannot overstate their profound influence on my work.

Many other professors from the department have made my time at the University of Waterloo a great learning experience. Special thanks to Dr. Richard Cook for his insightful advice on my thesis and for his guidance during my internship. Also, thank you to Dr. Jeanette O'Hara Hines for her generous support and guidance during my consulting work. I was very fortunate to have Dr. Mary Thompson and Dr. Cyntha Struthers as my instructors. They are outstanding teachers and I am grateful for everything I learned from them.

Many thanks to my very supportive friends Anita, Dasha and Audrey, and special thanks to Adrian. To my warm, wonderful family thank you for always believing in me and supporting my goals.

Dedication

To my wonderful mother Catrina.

Contents

List of Tables

List of Figures

Chapter 1

Background: Problems, Definitions and Notation

Binary measurement systems (BMS) are widely used in both industry and medicine, where they are known as diagnostic or screening tests. In industry, a BMS is used to measure various characteristics of parts and then classify them as conforming or nonconforming according to some quality standards. In the medical context, a BMS is used to detect a target condition in subjects, thus classifying them as positive or negative.

In practice, there are two kinds of BMSs. There are systems whose classification output is based on measuring a dichotomous variable, that is, the BMS is actually measuring the presence or absence of some characteristic, and there are systems that measure one or more continuous or ordinal characteristics and then classify a part or subject based on established threshold values.

In industry, BMSs can be either destructive, when the measurement system changes the characteristics of the parts during the inspection, or nondestructive. This thesis focuses on nondestructive BMSs, which in industry are part of the larger category of nondestructive tests. Nondestructive evaluation methods are widely applied in industry for process monitoring and sampling or exhaustive inspection of the production yield. For example, liquid-penetrant inspection is used to detect surface cracks or other surface flaws in a part (Olin and Meeker, 1996). The test consists of applying a special fluorescent liquid on the surface of an object that is then dried and cleaned. Next, the object is examined by a human inspector under ultraviolet light, and is accepted or rejected based on whether a flaw is visible or not. In this case, the output variable is binary, as the size of the crack or flaw is not considered when making the decision (Olin and Meeker, 1996).

Another example is the inspection of blank credit cards that are passed or rejected by an automated visual inspection system. The 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 pre-specified range, the card fails the inspection. In this case, although each measurement is continuous and there is a threshold for each of them, the fact that the decision to pass or fail the part is based on all these measurements makes the final output binary and the measurement system is considered a BMS.

In medical field, diagnostic and screening tests are used to identify the presence or absence of a certain condition and they include, for example, bacterial cultures, radiographic images and biochemical tests. As mentioned before, some tests are dichotomous by their nature, for example a test for myocardial infarct diagnosis that identifies the presence of new Q-waves in the ECG or the presence of elevated Creatine Kinase - MB Fraction, over a certain period of time (Rindskopf and Rindskopf, 1986).

Diagnostic tests are used in medical practice for identifying the presence or absence of a target disease in subjects that have signs or symptoms and are suspected of having the condition. Screening tests, on the other hand, are usually conducted within a population of healthy people that are at risk of developing the target condition. For example, yearly breast cancer screening with mammography of women over 50 years of age and cervical cancer screening with Pap smear are common medical practice. Diagnostic and screening tests differ in a few important ways.

First, as mentioned earlier, diagnostic tests are usually applied only to subjects suspected of having the condition, while screening tests are used on healthy subjects; second, a positive screening test is usually followed by a definitive or more accurate test and not directly by the treatment. Usually, screening tests are non-invasive and inexpensive, so that they can be applied on a large scale to the population at risk.

From the point of view of the statistical methodology used to evaluate the performance of these tests, diagnostic and screening tests are not different. This is why in the medical literature pertaining to this topic they are generically called "medical" tests (Pepe, 2003).

In industry, there are two main aspects of the performance of a measurement system: accuracy and precision. For continuous measurements, accuracy (bias) is defined as the expected measurement error while precision is the variation in repeated measurements on the same unit by the same operator (repeatability) and by different operators (reproducibility). In the case of a BMS, the classic definitions of accuracy and precision are not applicable, since the mean and variance are functionally related. Instead, the terms "performance" or "effectiveness" measures are usually used for quantifying the quality of the measurement system.

In the medical literature, the term accuracy is widely used for characterizing the performance of a medical test. In this thesis, we use a consistent terminology for both industrial and medical fields. Thus, we use the term "performance measures" when we refer to the quality of a BMS.

There are two main approaches for assessing the performance of a BMS. The first one, called the "gold standard" approach (Pepe, 2003; Farnum, 1994), requires the use of a definitive or gold-standard measurement system that can determine the state of parts or subjects with no classification errors. The approach then compares the outcomes of the gold standard with the ones from the BMS under study. Performance measures express how the results of the BMS agree with the outcomes of the gold-standard test. For example, in a coronary artery surgery study, subjects who were suspected to have coronary heart disease underwent an exercise stress test and also had their chest pain history taken (Weiner et al., 1979). These were the two BMSs under investigation and the true disease status was determined by arteriography, the gold-standard test in this case. In the industrial context, for the credit card example, a human inspector can compare each blank credit card with its pre-specified layout and then fail or pass the card. Although the gold-standard approach is considered the "standard" method in both industrial and medical contexts (AIAG, 2002; Pepe, 2003), there are cases where the gold standard is too expensive, time consuming or invasive to be used on a routine basis (Boyles, 2001; Walter and Irwig, 1988).

In other situations, a gold standard does not exist. For example, in the medical context, some conditions such as migraines cannot be determined based on histological or biochemical changes. In that case, the condition is defined by a combination of symptoms and signs and the determination is not error-free. Assessment methods have also been developed for such "no gold standard" situations (Rutjes et al., 2007; Pepe, 2003; Boyles, 2001; Van Wieringen and Van der Heuvel, 2005). These methods involve using an imperfect reference measurement system with known characteristics, or using a latent class approach, where no reference test is used. A more detailed review of these methods can be found in Chapter 2.

In the first two chapters of this thesis, we use a unified terminology and notation for both industrial and medical fields. Therefore, we use the term "units" for both parts and subjects. When a part passes the BMS inspection or a subject tests positive, we say the unit passes the inspection. Similarly, when a part fails the inspection or a subject tests negative, we say the unit fails the inspection. Also, if a part is conforming to the quality standards or a subject is diseased we say the unit is conforming, and similarly for the nonconforming and non-diseased case.

With BMSs considered here, each unit has a "true" quality or disease state, conforming (diseased) or nonconforming (non-diseased), also called the measurand. For unit i, we denote:

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

Each unit can be classified by the BMS as pass (positive) or fail (negative) and the outcome of the test is denoted by:

$$Y_i = \begin{cases} 1, & \text{if unit } i \text{ passes the inspection} \\ 0, & \text{if unit } i \text{ fails the inspection} \end{cases}$$

There are two parameters related to these variables: the conforming rate or the prevalence of the disease, π_C, i.e. the chance that a randomly selected unit is conforming (diseased), and the pass rate, π_P, i.e. the chance that a randomly selected unit passes the inspection (positive). There are many ways to quantify the performance of a BMS. Here we introduce the most common ones in both industrial and medical contexts.

Misclassification rates, defined as the chance of passing a nonconforming unit, called the false positive (consumer's risk), and the chance of failing a conforming unit, called the false negative (producer's risk), are commonly used in quantifying the performance of a BMS in industry (AIAG, 2002; Johnson et al., 1991; Boyles, 2001) and medicine (Hui and Walter, 1980; Walter and Irwig, 1988; Pepe, 2003).

We denote the two misclassification probabilities by:

$$\alpha = \Pr(Y_i = 1 \mid X_i = 0), \quad \text{the false positive} \tag{1.1}$$

and

$$\beta = \Pr(Y_i = 0 \mid X_i = 1), \quad \text{the false negative} \tag{1.2}$$

A gold standard system has $\alpha = \beta = 0$.

The four parameters, α, β, π_C and π_P, must lie between 0 and 1 and are constrained by the identity:

$$\pi_P = \Pr(Y_i = 1) = \Pr(Y_i = 1, X_i = 1) + \Pr(Y_i = 1, X_i = 0) = (1-\beta)\pi_C + \alpha(1-\pi_C) \tag{1.3}$$

Solving for π_C, we have

$$\pi_C = \frac{\pi_P - \alpha}{1-\beta-\alpha}$$

For a reliable BMS we expect low misclassification probabilities, α andβ. Also, it is usually the case that the chance of passing a conforming unit is higher than the chance of passing a non-conforming unit, i.e. $1-\beta > \alpha$. This assumption is also valid in the medical context, where "a process is to be called a test if and only if it selects diseased persons with higher probability than it does non-diseased persons" (Rogan and Gladen, 1978). This is equivalent to saying that a group of people with positive tests will be expected to have higher disease prevalence than the population.

In the above definitions, we have implicitly assumed that the misclassification probability is the same for each conforming/nonconforming unit. This may not be true, as in the liquid-penetrant inspection system (Olin and Meeker, 1996), where parts with large cracks are more easily classified. Later we will relax the assumption of constant misclassification probabilities.

In both industry and medicine, there are two main goals for conducting a BMS assessment study. First, the performance and the costs associated with misclassification by a BMS should be known before using the system for routine testing. The second is related to statistical inference, including estimation and hypotheses testing, on different parameters characterizing the production process or the population of interest, when the BMS is used to measure the variables of interest. Next, we explain in more detail the importance of these goals for both the industrial and medical contexts.

In industry, parts passed by the BMS are sent to the customer while failed parts are scrapped or re-worked. As a result, it is important to assess the rates associated with a wrong decision. The implications of the false positive classification are usually more serious than those of the false negative. Making a false positive error can lead to an increase in quality complaints and a decrease in customer satisfaction. On the other hand, false negative errors lead to unnecessary rework or scrapping of parts with additional costs for the producer. As there are different costs associated with each incorrect decision, we estimate both α and β so that the overall cost can be quantified. Also, knowing the misclassifications rates has an impact on how to improve a manufacturing process, when a study is conducted to identify whether the cause of output variation is related to the measurement system or to the rest of the process. For continuous outcomes, the overall variation can be partitioned into components due to the measurement system and the manufacturing process. In the binary outcome case, we can equivalently partition the pass rate into components due to the BMS, i.e. the misclassification rates α and β, and to the process, i.e. π_C. If the misclassification rates are high, the measurement system is not adequate and it must be improved before conducting further investigations.

Also, in a production process there are situations when the process is in statistical control and is being monitored with a fraction non-conforming p-chart (Burke et al., 1995). In this context, it is often mistakenly thought that the centre line of the control chart provides an estimate of the fraction nonconforming for the process. However, this estimate is based on measurements by a BMS. Thus, if the false positive and false negative rates are not accounted for, using the pass rate as an estimate of the conforming rate can lead to serious bias. Therefore, there is a need to estimate the misclassification probabilities and incorporating them in the estimation of the nonconforming rate.

In medical research, it is common to assess the performance of diagnostic test using the misclassification rates, α and β, or their complements, $1-\alpha$, called specificity and $1-\beta$, called sensitivity (Fleiss, 1981; Walter and Irwig, 1988; Pepe, 2003). The first goal of assessing the misclassification rates is similar to the one in the industrial setting: quantifying the costs due to misclassification of patients. Classifying a subject as non-diseased, when the subject is actually diseased, can prevent or delay the appropriate treatment that can cause aggravation of the disease or even death. When a non-diseased subject tests positive, unnecessary treatments that can be both invasive and costly are applied to the subject, causing discomfort, trauma and increased health care costs.

Aside from knowing the performance characteristics of a medical test for decision-making purposes, it is also important to use this information in studies where the prevalence or different measures of association are estimated. Medical tests are used to measure not only a target condition, but also exposure variables or risk factors, confounders, etc., and when the tests are not error-free, information about their performance should be included in the estimation of prevalence and incidence rates or indices of association, such as relative risk. Quade et al.

(1980), Rogan and Gladen (1978) and Yanagawa and Gladen (1984) show that inferences based on medical tests can lead to seriously biased estimators of prevalence, incidence and remission rates when information about the classification errors is not included. They also prove that the power of the test for comparing rates is also affected by errors due to the BMS. They propose methods to account for these misclassification errors, assuming that the sensitivity and specificity of the medical test are known or estimable. Barron (1977) shows the implications of not accounting for the misclassification errors on the statistical inference of the relative risk and suggests a way to account for these errors.

There are other measures of performance that are sometimes estimated in a BMS assessment study. For example, in the medical field, the predictive values, i.e. the probability that a subject that tests positive is actually diseased, called the positive predictive value (PPV), and the probability that a subject that test negative is actually non-diseased, called the negative predictive value (NPV), are sometimes used. These predictive values are expressed as follows:

$$PPV = \Pr(X_i = 1 \mid Y_i = 1) = \frac{(1-\beta)\pi_C}{(1-\beta)\pi_C + \alpha\pi_C},$$

$$NPV = \Pr(X_i = 0 \mid Y_i = 0) = \frac{(1-\alpha)\pi_C}{(1-\alpha)\pi_C + \beta\pi_C}.$$

We note that these two measures depend on both the misclassification rates, α and β, and the prevalence, π_C. They are not used to quantify the inherent performance of the BMS, but how well the test reflects the true state of a subject; therefore they quantify the clinical value of the test (Pepe, 2003).

Other measures of performance, used mostly for medical tests, are the positive and negative likelihood ratios, which quantify the change in the odds of the disease when given the result of the diagnostic test (Pepe, 2003). The likelihood ratios are expressed as follows:

$$PLR = \frac{\Pr(Y_i = 1 \mid X_i = 1)}{\Pr(Y_i = 1 \mid X_i = 0)} = \frac{1-\beta}{\alpha},$$

$$NLR = \frac{\Pr(Y_i = 0 \mid X_i = 1)}{\Pr(Y_i = 0 \mid X_i = 0)} = \frac{\beta}{1-\alpha}.$$

We note that the likelihood ratios are simple functions of the misclassification probabilities, α and β, and therefore summarize the performance of the BMS.

This thesis focuses on a particular industrial setting that can also be found in the medical field. First, the BMS under study is nondestructive and the quality variable X is a dichotomy that represents the presence or absence of a characteristic. Second, the study is conducted as a routine evaluation of the performance of the BMS, after it has been implemented and in use for a period of time. Third, the BMS is used for 100% inspection or any kind of systematic inspection of the production yield. Fourth, at the time of the evaluation, the process is under statistical control and the BMS is stable. In this context, we can obtain a good estimate of the pass rate, π_P, using a large number of measured units, called baseline measurements. For example, in the credit card case when the automated visual system is used, thousands of cards are measured every day and we can record the number of passes and fails and obtain an estimate of the proportion of passes. Therefore, we could assume that the pass rate is known or well estimated before the BMS assessment study is conducted. It would be statistically inefficient to ignore this information. In this thesis, we quantify how the assumed knowledge of π_P improves the estimation of α, β and π_C. As we will see in Chapters 3, 4, 5, and 6, having baseline data about π_P also allows us to choose sampling plans that otherwise we would not be able to use. In the medical context, screening tests that are already in use represent a similar setting, where we may have a good estimate of the rate of testing positive.

Our discussion will mainly focus on the case where only one BMS is assessed, but we will also discuss possible generalizations. Also, as we focus on an industrial context, most of the terminology we use here is from the manufacturing industry.

In Chapter 2, we give a summary of the existing sampling plans and estimation methods for a BMS assessment study that are relevant in our setting. In Chapters 3 and 4, we propose new procedures that improve and adapt the current ones to our case, when we assume that the misclassification rates α and β are constant within the nonconforming/conforming units. In Chapters 5 and 6, we propose new models for relaxing the assumption of constant misclassification rates. In all these new methods, we use the baseline information about the pass rate π_P, along with a proposed sampling scheme made possible by the availability of these baseline measurements. We evaluate the gain in the accuracy and precision of the estimators corresponding to these new procedures in comparison to the ones given by the standard (current) ones. In Chapter 7, we discuss possible extensions and future work, such as assessing multiple (possibly parallel) binary measurement systems, and assessing a BMS in the context where a reference or "anchored" system is present and we have baseline measurements available.

Chapter 2

Current Methods for Assessing a Binary Measurement System

While there is an extensive literature on the assessment of continuous measurement systems in industry (AIAG, 2002; Wheeler and Lyday, 1989), much less research has addressed assessing a BMS. AIAG (2002) provides a method that assumes there is a known underlying continuous measure that has been discretized and a threshold is used for classification. However, as mentioned before, our goal is to focus on cases where the measurand X is dichotomous.

In medical research, the performance of medical tests with positive/negative outcomes has been extensively studied (Walter and Irwig, 1988; Zhou et al., 2002; Rutjes et al., 2007). In this chapter, we give a summary of the methods developed in both industrial and medical fields, and provide a unified view on the inference methods and sampling schemes used in such assessment studies. All these proposed methods consider only the case where there is no prior information about the pass rate, π_P, of the studied BMS.

As mentioned before, there are two major approaches for assessing a BMS. The first one is used when a gold standard is available and we can compare the results of the BMS to the classification provided by the gold standard. The second approach is used when there is no gold standard or using a gold standard is not feasible.

2.1 The “Gold Standard Approach”

Methods using this approach require the existence and use of a gold standard, a test or measurement system that classifies units with no classification errors. When it is possible to use the gold standard on all units selected for the assessment study, we have “complete verification”’ with the gold standard. When the gold standard is too expensive or invasive to be used on the whole sample and only a sub-sample of units are tested with the gold standard, we have “partial verification”.

2.1.1 Complete Verification with the Gold-standard System

This method requires the use of the gold standard on all selected units, and it is considered the standard method for assessing a BMS in both the industrial and medical fields.

The method assumes that a total of n units are selected and then classified by both the BMS and the gold standard. We can summarize the results of classifying n units by each measurement system using Table 2.1. To make the notation easier to follow, we use the terms “pass (P)” for both “passed parts” and “positive subjects” and “fail $(\bar{P})$” for both “failed parts” and “negative subjects”, as determined by the BMS. Also, the terms “conforming (C)” is used for “conforming parts” and “diseased subjects”, and “nonconforming $(\bar{C})$” is used for both “nonconforming parts” and “non-diseased subjects”, as determined by the gold standard. The usual assumptions for the

Table 2.1 Data from a BMS Performance Study with Complete Verification by the Gold Standard

	Conform (C)	Not conform $(\bar{C})$	Total units
Pass (P)	n_{PC}	$n_{P\bar{C}}$	n_P
Fail $(\bar{P})$	$n_{\bar{P}C}$	$n_{\bar{P}\bar{C}}$	$n_{\bar{P}}$
Total units	n_C	$n_{\bar{C}}$	n

standard assessment study are:

- All units have the same probability of conforming, π_C;
- Nonconforming units have a common probability of passing the inspection, α;
- Conforming units have a common probability of failing the inspection, β;
- Measurements of different units are independent.

In practice, there are two sampling schemes used in selecting the units for an assessment study: Farnum's or the case-control study sampling plan, and the random selection or the cohort study sampling plan.

Farnum's or Case-control Sampling Plan

In the industrial context, Farnum (1994) suggested an assessment method for a BMS where two equal-size independent samples of conforming and nonconforming items are selected ($n_C = n_{\bar{C}}$) and then evaluated by the BMS. In Table 2.1, n_C, $n_{\bar{C}}$ and n are fixed and all the other quantities are random.

This sampling scheme is equivalent to the sampling protocol of a case-control study for assessing the performance of a medical test, where pre-specified numbers of diseased and non-diseased subjects, as determined by the gold standard, are selected (Pepe, 2003).

With this sampling scheme, both misclassification errors, α and β, can be directly estimated as:

$$\hat{\alpha} = \frac{n_{P\bar{C}}}{n_{\bar{C}}} \tag{2.1}$$

$$\hat{\beta} = \frac{n_{\bar{P}C}}{n_C} \tag{2.2}$$

These estimators are the Maximum Likelihood (ML) estimators, are unbiased and their variances are:

$$\text{Var}(\hat{\alpha}) = \frac{\alpha(1-\alpha)}{n_{\bar{C}}} \quad \text{Var}(\hat{\beta}) = \frac{\beta(1-\beta)}{n_C}.$$

We notice that the conforming rate, π_C, cannot be estimated when this sampling scheme is used. This is one of the major drawbacks of this method. We also notice that the predictive probabilities, PPV and NPV, cannot be estimated, unless prior information about the conforming rate is known. The predictive values are important, especially in the medical field, and, because for many conditions an estimate of the prevalence is known a priori, they can still be estimated in a case-control study.

The main advantage of Farnum's sampling scheme is the direct estimation of α and β. In an industrial setting, when we are dealing with a high-performance process, the main disadvantage of this method is related to the extensive use of the gold standard. Usually, the gold standard is expensive or time consuming and finding a large number of nonconforming parts through the use of the gold standard can be unreasonably costly or impractical. In the medical field, even in the case of low-prevalence diseases, this is not necessarily a problem, as usually there are enough people with the disease status already identified that can be enrolled in the BMS assessment study.

Random Selection or Cohort Study Sampling Plan

Another sampling scheme involves randomly selecting a sample of n units from a target population, and then measuring each unit using both the BMS and the gold standard. In this sampling scheme, in Table 2.1, only the total, n, is fixed and all the other quantities are random. This is the sampling scheme used in cohort studies in the medical field (Pepe, 2003).

We notice that with this sampling scheme all parameters of interest can be directly estimated. The ML estimates for the misclassification probabilities are the same as in the case-control sampling, Eq. (2.1) and (2.2). Also, we can estimate the conforming rate:

$$\hat{\pi}_C = \frac{n_C}{n} \tag{2.3}$$

The predictive values can also be obtained directly from Table 2.1. We note that, although the formulas for the estimates are the same as in a case-control study, the properties of the estimators, including precision, are different. In a cohort study, the quantities in the denominator of the estimators of α and β are now random, whereas in the case-control situation they are fixed. Confidence intervals for α and β can be derived using exact methods or asymptotic methods and the logistic transformation (Pepe, 2003).

Burke et al. (1995) propose a different testing protocol for use in industry. The method involves randomly selecting n units, and then classifying each unit by the gold standard, which in their example is a panel of knowledgeable experts that classify parts by consensus. The studied BMS is a single inspector involved in the routine classification of the manufactured parts. After the true state of a unit is determined, the inspector measures the unit r times. The estimates for the two misclassification probabilities are expressed as "the average values of $\hat{\alpha}$ and $\hat{\beta}$" for the n parts. The authors do not explain what "the average values of $\hat{\alpha}$ and $\hat{\beta}$" actually mean, and also do not specify any assumption about the independence of measurements. Therefore, we were not able to fully understand their estimation procedure. In Chapter 3, we will explore this testing protocol in greater detail and propose a new sampling plan. We will also make some appropriate assumptions about the measurements on the same part and on different parts, and derive the MLEs of α, β and π_C.

2.1.2 Partial Verification by the Gold Standard

Here we consider the case when there is an available gold standard, but for a variety of reasons it cannot be applied on a large sample of units. In the medical context, some gold-standard tests involve invasive procedures, such as surgery, biopsy, etc., and it is considered unethical to apply it on subjects that do not show signs of the target condition, that is, they test negative on the BMS. One example of such a gold standard is angiography which is used for the detection of pulmonary embolism. Angiography can cause serious complications in the tested subjects; therefore, it is considered unethical to perform this test on people with negative D-dimer results, the studied BMS in this case (Rutjes et al., 2007).

In industry, some gold-standard systems can be time-consuming, expensive or even destructive. In the credit card example, the gold standard was a human inspector that checks the cards for various defects. Although we assume the inspector classifies parts with no error, the inspection process is slow and costly when a large number of parts have to be tested. In both contexts, a compromise is to classify a small sample of units using both the gold standard and the BMS. The usual testing protocol involves selecting a large sample of t units and then measuring them all with the BMS. Next, a smaller sample of size n is selected from the initial one and then measured with the gold standard. The data can be summarized as in Table 2.2. This methodology is known as a two-phase or double sampling in the sample survey and industrial literature (Särndal and Swensson, 1987; Johnson et al., 1986). In the medical literature, it is called partial or incomplete verification (Rutjes et al., 2007).

Table 2.2 Data from a BMS Performance Study with Partial Verification by the Gold Standard

	Pass (P)	Fail ($\bar{P}$)	Total units
Conform (C)	n_{PC}	$n_{\bar{P}C}$	n_C
Not conform ($\bar{C}$)	$n_{P\bar{C}}$	$n_{\bar{P}\bar{C}}$	$n_{\bar{C}}$
Units verified by the BMS and the Gold Standard	n_P	$n_{\bar{P}}$	n
Units not verified by the Gold Standard	$t_P - n_P$	$t_{\bar{P}} - n_{\bar{P}}$	$t - n$
Total units	t_P	$t_{\bar{P}}$	t

The assumptions for a partial verification study are the same as in a standard one, i.e. units have the same conforming probability, π_C, nonconforming units have common probability of passing, α, conforming units have common probability of failing, β, and the classification outcomes are independent for different units.

This methodology was designed and proposed for two practical applications. For the first one, the goal is to estimate the conforming rate or prevalence. Boss (1954) shows that using the pass rate, π_P, as an estimate for the true conforming rate, π_C, can lead to serious bias, unless the BMS is perfect. The standard method for estimating the conforming rate is to use a gold standard on a large sample of parts, as discussed in Section 2.1.1. When this is not possible, a compromise is to conduct a two-phase study and estimate the conforming rate using the data at hand, accounting for the misclassification probabilities. The two-phase methodology was proposed in this context by Tenenbein (1970), (1971), and (1972).

The second purpose, which has been discussed in the medical context, is related to the estimation of the misclassification probabilities, α and β. Here, many assessment studies involve testing the subjects with the BMS, and then, due to the invasive character of the gold standard, only a proportion of the positive-tested persons are subjected to the gold standard. As Zhou (1998) mentions, in many such studies the estimates of the misclassification probabilities are based on the "complete-cases" data, i.e. the data related to subjects tested by both the BMS and the gold standard. Therefore, although the design is not standard, the data analysis is conducted as if it were. These naïve estimators are biased and usually lead to an overestimation of the performance of the BMS, in terms of sensitivity and specificity (Zhou, 1998).

There are various methods for selecting the second sample in a two-phase study. The two most relevant ones are random selection, when the second sample is randomly selected from the initial one, and conditional re-sampling, when pre-determined numbers of units are randomly selected from two strata: one made up of units that passed the BMS, and the other by those that failed. For random selection in phase two, except for n and t, in Table 2.2, all quantities are random. In the conditional re-sampling case, n, t, n_P and $n_{\bar{P}}$ are fixed, and all other quantities are random.

In the two-phase sampling with random selection, Tenenbein (1970) shows the ML estimators for the conforming rate and misclassification probabilities are:

$$\hat{\pi}_C = \frac{t_P}{t}\frac{n_{PC}}{n_P} + \frac{t_{\bar{P}}}{t}\frac{n_{\bar{P}C}}{n_{\bar{P}}} = \hat{\pi}_P\frac{n_{PC}}{n_P} + (1-\hat{\pi}_P)\frac{n_{\bar{P}C}}{n_{\bar{P}}} \tag{2.4}$$

$$\hat{\alpha} = \frac{\frac{t_P}{t}\frac{n_{P\bar{C}}}{n_P}}{1-\hat{\pi}_C} = \frac{\hat{\pi}_P\frac{n_{P\bar{C}}}{n_P}}{\hat{\pi}_P\frac{n_{P\bar{C}}}{n_P} + (1-\hat{\pi}_P)\frac{n_{\bar{P}\bar{C}}}{n_{\bar{P}}}} \tag{2.5}$$

$$\hat{\beta} = \frac{\frac{t_{\bar{P}}}{t}\frac{n_{\bar{P}C}}{n_{\bar{P}}}}{\hat{\pi}_C} = \frac{(1-\hat{\pi}_P)\frac{n_{\bar{P}C}}{n_{\bar{P}}}}{\hat{\pi}_P\frac{n_{PC}}{n_P} + (1-\hat{\pi}_P)\frac{n_{\bar{P}C}}{n_{\bar{P}}}} \tag{2.6}$$

We note that these estimates are undefined if either n_P or $n_{\bar{P}}$ is zero.

Tenenbein (1970) also derives the asymptotic variance of $\hat{\pi}_C$:

$$\text{Var}(\hat{\pi}_C) \simeq \frac{\pi_C(1-\pi_C)}{n}\left[1 - \frac{\pi_C(1-\pi_C)(1-\alpha-\beta)^2}{\pi_P(1-\pi_P)}\right] + \frac{\pi_C^2(1-\pi_C)^2(1-\alpha-\beta)^2}{m\pi_P(1-\pi_P)},$$

and provides sample size determinations for both stages when there is a fixed budget and we want to minimize the variance of $\hat{\pi}_C$, or when we desire a certain precision of $\hat{\pi}_C$ and we want to minimize the measurement cost.

For the case of conditional re-sampling, parallel work has been conducted on the estimation procedure for π_C, α and β (Haitovsky and Rapp, 1992; Begg and Greenes, 1983; Zhou, 1998). In the medical field, there was a need for estimation methods for studies with partial verification. We mentioned before that the common practice is to analyze the data from this kind of study as if they were multinomial data, and that this results in seriously biased estimators. A way to approach this problem is to look at it as a "missing-data" case, where data is missing on the gold-standard measurements for some units. When the probability of selection in the second phase does not depend on the result of the BMS test, i.e. in the random selection case mentioned above, the situation is called "Missing Completely At Random" (MCAR). When the probability of selecting a unit in the second phase depends on the result of the BMS test but not on its true state, we have a “Missing at Random” (MAR) case. In the MAR context, one estimation method proposed by Begg and Greenes (1983) uses quantities that can be directly derived from the data, i.e. the estimates of the predictive values PPV and NPV and the pass rate π_P, and then uses Bayes’ theorem to derive estimators for π_C, α and β. Begg and Greenes’s procedure is also known as a “correction” method (Rutjes et al., 2007). Another proposed analysis procedure involves imputation and is usually conducted in two phases. In the imputation phase, each missing value is replaced, and in the analysis phase, the estimates of the parameters are computed based on the complete data, using the standard method (Pepe, 2003; Rutjes et al., 2007). In the simplest imputation method (Pepe, 2003), n_{PC} and $n_{P\bar{C}}$ in Table 2.2 are multiplied by the inverse of the probability that a BMS-positive subject is selected in the second-phase sample, while $n_{\bar{P}C}$ and $n_{\bar{P}\bar{C}}$ are multiplied by the probability that a BMS-negative subject is selected. Then the data are analyzed as in a standard random BMS assessment study.

In the industrial context, Haitovsky and Rapp (1992) extended and improved Tenenbein’s double-sampling method by proposing a conditional re-sampling protocol. They derived the MLEs of π_C, α and β. It turns out that the estimates are the same as in the random selection case, but the properties of the estimators, such as their precision, are different, as now n_P, $n_{\bar{P}}$ are fixed. Zhou (1998) proved that when the MAR assumption holds, the MLE, the Begg and Greenes’ method and the simple imputation method provide the exact same estimators for π_C, α and β as given by Eq. (2.4), (2.5), (2.6).

The advantage of the conditional re-sampling method is that it controls the number of passed and failed parts in the sample, therefore avoiding the undefined estimates problem mentioned earlier. It also provides more precise estimators for some of the parameters.

2.2 The "No Gold Standard" Approach

As mentioned before, in the "no gold standard" situation, the gold standard is too expensive, invasive or time consuming to be used in an assessment study, or there is no available measurement system or test that can classify units without error.

2.2.1 Using a Reference System with Known Performance Characteristics – Anchored Method

In both industrial and medical contexts, there are cases where the gold-standard system is not available for the BMS assessment study. Instead, the best available testing method is used, usually called the reference test or the alloyed gold standard. This test is not error-free, but its performance characteristics, i.e. misclassification probabilities or their complements, sensitivity and specificity, are assumed known. Treating this test as a gold standard and then conducting a standard data analysis results in seriously biased estimators of α, β and π_C (Hadgu et al., 2005).

To overcome this issue, a new methodology was developed, where the known characteristics of the reference test are incorporated in the estimation procedure. With this method we classify a sample of units with both the reference test and the BMS, and then obtain the MLEs of α, β and π_C. This method was also recommended as an alternative to another assessment method, called discrepant analysis. Discrepant analysis was initially proposed as an attempt to solve the problem of biased estimators in the case where only an imperfect reference test is available (Hadgu et al., 2005). It involves the use of the BMS and the reference test and the cases with discordant results between the two tests are re-tested using an ancillary test or resolver, which is also imperfect. This method was severely criticized by many researchers as "biased and unscientific" (Hadgu et al., 2005), and is not considered a solution to the biasness problem. Another alternative is the latent class analysis which we will discuss later in this chapter.

Boyles (2001) also discusses using a reference system and calls the statistical model used to accommodate the data the "anchored model". From now on, we will use the term "anchored" method whenever a reference-standard system with known performance characteristics is used.

It is common in the medical field for a new BMS to be assessed by comparing its classification outcomes to those from a reference test. For example, new Nucleic Acid Amplification tests for detecting infectious conditions such as Chlamydia trachomatis infection are assessed by comparing their classification results with the results of a cell culturing test. The selected subjects are usually tested only once with each procedure.

In industry and some medical contexts, the selected units can be repeatedly measured by both the BMS and the reference test. We give the notation, assumptions and estimation procedure for the repeated-measurements case, as it is a generalization of the single-measurement case discussed above.

For unit i, we denote the outcome of the kth inspection with the reference test by:

$$Z_{ik} = \left\{ \begin{array}{ll} 1, & \text{if unit } i \text{ passes the } k\text{th inspection with reference test} \\ 0, & \text{otherwise} \end{array} \right\}, i = 1, \ldots, n, k = 1, \ldots, p$$

Also, the outcome of the jth inspection by the BMS for unit i is denoted by:

$$Y_{ij} = \left\{ \begin{array}{ll} 1, & \text{if unit } i \text{ passes the } j\text{th inspection with BMS} \\ 0, & \text{otherwise} \end{array} \right\}, i = 1, \ldots, n, j = 1, \ldots, r$$

The total number of passes for unit i with the reference test is denoted by $U_i = \sum_{k=1}^{p} Z_{ik}$, whereas the total number of passes with the BMS is denoted by $S_i = \sum_{j=1}^{r} Y_{ij}$.

The true state of a unit is denoted by X_i, consistent with the notation from Chapter 1. As before, the misclassification rates of the BMS are denoted by α and β.

The known performance characteristics of the reference test are:

$$\alpha_R = \Pr(Z_{ik} = 1 \mid X_i = 0) \quad \text{and} \quad \beta_R = \Pr(Z_{ik} = 0 \mid X_i = 1).$$

There are four main assumptions for the anchored model:

- The probability of passing the BMS inspection is common for all nonconforming units. The same is true for the probability of failing any conforming unit. The same assumption is made for the reference test.
- The BMS and the reference test do not change the units in any way, so the misclassification probabilities do not change from one measurement to the next, on the same unit;
- Given the true state of a unit, the repeated measurement, whether done by the BMS or the reference test, are independent. That is:

$$\Pr(Y_{i1}, Y_{i2}, \ldots, Y_{ir}, Z_{i1}, Z_{i2}, \ldots, Z_{ip} \mid X_i) = \prod_{j=1}^{r} \Pr(Y_{ij} \mid X_i) \prod_{k=1}^{p} \Pr(Z_{ik} \mid X_i) \tag{2.7}$$

  This is also known as “conditional independence”;
- Measurements on different units are independent.

Using these assumptions and notation, the conditional distribution of the total number of times unit i passes the BMS inspection given the part is conforming is $S_i \mid (X_i = 1) \sim \text{Binomial}(r, 1-\beta)$, and given it is nonconforming is $S_i \mid (X_i = 0) \sim \text{Binomial}(r, \alpha)$. For the reference test, $U_i \mid (X_i = 1) \sim \text{Binomial}(p, 1-\beta_R)$, and $U_i \mid (X_i = 0) \sim \text{Binomial}(p, \alpha_R)$.

When the units are randomly selected from a study population, the likelihood function can be expressed as:

$$L(\alpha, \beta, \pi_C \mid s_i, r_i) \propto \prod_{i=1}^{n} [(1-\beta)^{s_i} \beta^{r-s_i} (1-\beta_R)^{u_i} \beta_R^{p-u_i} \pi_C + \alpha^{s_i} (1-\alpha)^{r-s_i} \alpha_R^{u_i} (1-\alpha_R)^{p-u_i} (1-\pi_C)] \tag{2.8}$$

We notice that the score equations of the likelihood function do not have a closed form. Boyles (2001) and other authors use the EM algorithm to estimate the parameters. More details about this optimization procedure are given in Chapter 4.

In the case of a single measurement per unit by the BMS and the reference test, the resulting data can be summarized in a 2x2 table. After some re-parameterization of the likelihood function, the MLEs of α and β can be expressed in terms of the data, z_{ik} and y_{ij}, and the known misclassification rates of the reference test, α_R and β_R (Staquet et al., 1981).

The validity of the estimators given by the anchored model depends on the model assumptions. The conditional independence assumption is considered the most difficult to justify in many practical situations, especially in the medical field. If two medical tests are based on the same physiologic phenomenon, then, given the true state of a subject, there is likely a positive correlation between the outcomes of the two tests, and the conditional independence assumption is thus violated. For example, if both tests are based on a particular antibody reaction, something that inhibits the reaction for one test may have a similar effect on the other. Another example is the detection of lumbar herniation if both tests are imaging tests, such as MRI and radiography, as both of them focus on the detection of abnormalities of the discus (Rutjes et al., 2007).

If the anchored model is used when the outcomes of the two tests are correlated, the estimators of α and β can be seriously biased. In the next section, we introduce latent class analysis, and discuss some of the methods developed to account for the possible conditional dependence that can also be applied in the anchored model case.

2.2.2 Latent Class Analysis

Latent class (LC) analysis methods have been known for decades (Lazarsfeld and Henry, 1968) and they have been applied in many areas of research, such as psychology and sociology, and more recently in assessment studies for medical tests (Hui and Walter, 1980; Walter and Irwig, 1988) and BMSs used in industry (Boyles, 2001; Van Wieringen and Van der Heuvel, 2005; Van Wieringen and De Mast, 2008).

LC analysis methods acknowledge that there is no available gold standard and therefore, that the true state of a unit—conforming or nonconforming—cannot be observed. The true state is considered a “latent” variable and we only observe the outcomes of the BMSs measurements. The main assumption of the LC analysis is that, conditioning on the true state, the measurements from different BMSs or repeated measurements by the same BMS are independent. As noted above, this is called the conditional independence assumption (Eq. (2.7)).

In a manufacturing context, it is of interest to assess the performance of a measurement system used for routine inspection, for example an automated visual system, when no gold or reference standard is available. This case was considered by Boyles (2001), who proposes a testing procedure that consists of repeatedly measuring randomly selected parts with the studied BMS. Then, he uses the LC model for one BMS. The model assumptions are similar to the ones for the anchored model:

- The probability of passing the BMS inspection is common for all nonconforming units, i.e. α is constant. The same is true for the probability of failing any conforming unit, i.e. β is constant;

- The BMS does not change the units in any way, so the misclassification probabilities do not change from one measurement to the other;

- Conditioning on the true state of a unit, measurements by the BMS are independent, i.e.

$$\Pr(Y_{i1}, Y_{i2}, \ldots, Y_{ir} \mid X_i) = \prod_{j=1}^{r} \Pr(Y_{ij} \mid X_i), \text{ for each } i = 1, \ldots, n;$$

- Measurements on different units are independent.

Assuming the above conditions hold, the likelihood function is:

$$L(\alpha, \beta, \pi_C \mid s_i) \propto \prod_{i=1}^{n} [(1-\beta)^{s_i} \beta^{r-s_i} \pi_C + \alpha^{s_i} (1-\alpha)^{r-s_i} (1-\pi_C)] \tag{2.9}$$

which is a mixture of two Binomial distributions. As this likelihood function is based on the observed measurement outputs of the BMS, it is also called the observed-data likelihood. To make the parameters identifiable, we have to assume that $\alpha + \beta < 1$, which is a reasonable condition as we mentioned in Chapter 1, and that there are at least three measurements per unit, i.e. $r \geq 3$ (Boyles, 2001; Van Wieringen and Van der Heuvel, 2005).

To find the maximum likelihood estimates for α, β and π_C, Boyles (2001) uses the EM algorithm. The algorithm uses the likelihood function for the complete-data which include the number of passes, s_i, and the (unobserved) true state of the part, x_i, as follows:

$$L_C(\alpha,\beta,\pi_C \mid (s_i,x_i)) \propto \prod_{i=1}^{n}[(1-\beta)^{s_i}\beta^{r-s_i}\pi_C]^{x_i}[\alpha^{s_i}(1-\alpha)^{r-s_i}(1-\pi_C)]^{1-x_i} \tag{2.10}$$

Boyles derives likelihood-based 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 is the asymptotic variance-covariance matrix of the ML estimators. More details about this method, including estimator precision, sample size determination and sampling plan assessment are discussed in Chapter 4.

Boyles (2001) and Van Wieringen and Van der Heuvel (2005) also apply the LC analysis to the case where selected parts are repeatedly measured by several human inspectors. In this case, inspectors are considered different BMSs, and their performance parameters are estimated using a more general LC model. Suppose there are t BMSs under study and each randomly selected unit is measured r_j times by the jth BMS, $j = 1,\ldots,t$. The model assumptions are the same as in the case of a single BMS, with the addition that given the true state of a unit, the measurements by different BMSs are independent.

In this case, the likelihood function is:

$$L(\alpha_j,\beta_j,\pi_C \mid s_{ij}) \propto \prod_{i=1}^{n}[\pi_C\prod_{j=1}^{t}(1-\beta_j)^{s_{ij}}\beta_j^{r_j-s_{ij}} + (1-\pi_C)\prod_{j=1}^{t}\alpha_j^{s_{ij}}(1-\alpha_j)^{r_j-s_{ij}}],$$

where s_{ij} represents the observed number of times unit i passes the inspections of the jth BMS, and α_j and β_j are the misclassification probabilities specific to the jth BMS.

For the general LC model, the parameters are identifiable when:

- $0 < \pi_C < 1$;

- $0 \le \alpha_j < 1-\beta_j \le 1$. As mentioned earlier, this is a very reasonable assumption and it is actual the definition of a "test" in the medical context;

- As derived by Van Wieringen (2005)

$$\prod_{j=1}^{t}(r_j+1)-1 \geq 2t+1 \tag{2.11}$$

In the medical field, the LC analysis is used in a variety of situations when no gold standard is available and subjects are tested simultaneously or in sequence with different BMSs or repeatedly tested with the same BMS. Some examples include the case where several diagnosticians independently assess subjects for the same condition; for example, several anaesthetists independently conduct a pre-operative assessment on subjects and decide whether or not the subject is fit to undergo a general anaesthetic (Dawid and Skene, 1979). Another example is when subjects are tested with different diagnostic methods for detecting the same condition, such as Mantoux, tine, imotest and “monovacc” tests for tuberculin sensitivity (Gutjahr et al., 1982). The last category is when subjects are repeatedly tested with the same diagnostic test, such as the sequence of six stool guaiac tests for colon cancer (Walter and Irwig, 1988) and the repeated biopsies following cardiac transplantation (Spiegelhalter and Stovin, 1983).

In some situations, repeatedly testing a subject with the same BMS is not acceptable for ethical or economic reasons, such as the case where the BMS is invasive or poses some health risks to the subject, e.g. surgery or biopsy. One possibility is to include other less invasive BMSs in the assessment study and then use the LC model to estimate all performance parameters. The problem is that in some situations, there is only one alternate diagnostic test and in that case, the identifiability condition (2.11) is not met. This condition requires that in the case of a single population of subjects (one prevalence, π_C) and no repeated measurements by the BMSs on the same unit, at least three tests have to be included in the study to make all parameters estimable. To see this, we can also use a more informal argument for the identifiability condition (2.11) that can be found in the medical literature, based on comparing the number of parameters to be estimated with the number of degrees of freedom in the data (Walter and Irwig, 1988; Pepe, 2003). For instance, in the case of three tests and no repeated measurements, there are seven parameters and seven degrees of freedom, as there is a total of eight possible combinations of test results.

Much attention has been given to estimating the performance of a diagnostic test when there is only one alternate test, and one proposed solution is to impose restrictions on the parameters. For example, Hui and Walter (1980) prove that in the case where only two tests are available for identifying a certain condition, the parameters are still estimable if subjects are selected from two populations with different prevalence rates, and it is assumed that the two tests have the same performance within the two populations. Another option is to use a Bayesian approach to estimate the performance parameters, α and β, (Joseph et al., 1995).

For the LC model, ML estimation is usually carried out using either the EM algorithm (Dawid and Skene, 1979; Boyles, 2001) or using direct optimization techniques (Torrance-Rynard and Walter, 1997; Fujisawa and Izumi, 2000), as in general, the estimates have to be derived numerically. However, in the case where subjects are selected from two populations and two BMSs are studied, there is a closed-form solution for the ML estimates as shown by Hui and Walter (1980).

Based on the large-sample properties of the ML estimators, the asymptotic variance-covariance matrix of the estimators is derived using either the observed or Fisher information matrix (Hui and Walter, 1980; Pepe, 2003). When the EM algorithm is used, the observed information matrix for the incomplete-data likelihood is obtained using the missing information principle (Meng and Rubin, 1991; McLachlan and Krishnan, 1997).

As in the "anchored" model case, the validity of the LC analysis depends on whether the model assumptions hold. While the identifiability conditions and the assumption of independent measurements for different units can be guaranteed by the study design, the conditional independence and constant error rates assumptions depend on the nature of the true quality/disease status. Many authors (Pepe, 2003; Qu et al., 1996; Vacek, 1983; Torrance-Rynard and Walter, 1997; Van Wieringen and De Mast, 2008) consider the conditional independence assumption unrealistic in many medical and industrial situations. In the medical field, we can imagine that if the tests included in the study are based on the same biological phenomenon, e.g.

blood sample or imaging methods, then multiple test results on a person with a certain disease status are likely to be correlated. Also, if the condition has different degrees of severity then subjects at a more advanced stage of the disease are more likely to test positive. In the industrial context, in the case where parts are inspected for different surface flaws such as scratches or cracks, it is easier for the visual BMS to detect flaws of a larger size, although the BMS does not measure the actual size. In all these cases, as we will see later in Chapters 5 and 6, we could assume the presence of an underlying latent variable whose value is specific for each unit and affects the probability that the unit passes or fails the inspection. In other words, α and β are not constant within the nonconforming/conforming units. In the presence of this intermediate latent variable, such as the severity of an parasitic infection or the size of a surface flaw, the detectability of the measurand by the BMS varies from unit to unit. Also, the measurements on the same unit given that the unit is conforming/nonconforming are not independent anymore. That is, in these cases, the conditional independence does not hold.

Vacek (1983) investigates the effects of the conditional dependence on the estimators given by the LC model, when two BMSs are used on two different populations of subjects. She proves that when conditional dependence exists and it is not accounted for in the model, the resulting estimators are biased and she provides expressions of the corresponding biases. Diagnostic tests have been proposed for checking the independence assumption, including χ^2 goodness-of-fit test when there are sufficient degrees of freedom in the data (Rindskopf and Rindskopf, 1986).

There are several proposed methods for accounting for the conditional dependence of the repeated measurements. One approach is to include in the LC model some parameters that quantify the conditional correlation between measurements (Vacek, 1983; Torrance-Rynard and Walter, 1997). Vacek considers the case where there are two diagnostic tests and two populations and proposes a model that includes the covariance between the measurements of the two BMSs, conditioning on the true state of a unit, i.e. conditional covariance. This method does not model the variability of the misclassification rates, but rather the conditional dependence of repeated measurements that is induced by a possible intermediate latent variable.

Other authors (Qu et al., 1996; Fujisawa and Izumi, 2000) propose the use of a random-effects model, where it is assumed that subject-specific random effects account for the varying α and β, and, therefore, for the conditional dependence between repeated measurements on a unit. Another approach is to use a Bayesian random effects model (Dendukuri and Joseph, 2001), where the misclassification probabilities, α and β, are considered random variables and Beta prior distributions are assumed. Then, point estimates and credibility intervals for α and β are derived. Rutjes et al. (2007) note that the Bayesian approach is sensitive to the chosen prior distribution and different priors can lead to differences in estimates. They also note that the Bayesian approach is helpful in situations where the number of parameters is large relative to the available degrees of freedom.

In Chapters 5 and 6, we address the issue of varying misclassification rates (and conditional dependence) by proposing a new random-effects model where we assume a certain distribution for α and β. We explore the properties of the estimators given by this random-effects model when a new sampling scheme is used and baseline information about the pass rate π_P is included in the estimation.

Chapter 3

New Methods for Assessing a Binary Measurement System with Constant Misclassification Rates – Gold Standard Available

3.0.3 Introduction

In this chapter, we focus on the situation where a gold-standard system is available for the assessment study, the BMS has been in use for a while, and we plan to conduct a routine assessment of its performance characteristics. In such cases, parts previously measured with the BMS are available for re-measuring, especially the rejected ones, as they are not immediately shipped to the customers. Therefore, in this case, we can sample parts conditionally on the previous (baseline) measurement, i.e. we can use a conditional selection (CS) plan. With the CS, we select two independent samples of parts from the collections of previously passed and rejected parts. The proportion of previously passed parts in the selected sample can be set prior to the assessment study, and here we investigate the effect of changing this proportion on the estimation procedure. Additionally, a BMS used for any systematic inspection in high volume processes typically tracks the number of parts previously passed and rejected over a certain period of time. Therefore, we often have baseline data that can be used to assess the current pass rate π_P, separate from the assessment study. We propose to improve the overall estimation procedure by using conditional sampling and augmenting the measurement assessment data with the available baseline data. We compare the characteristics of the estimators given by different CS plans with the ones given by the standard sampling plan (SP) commonly used in industrial practice. That involves a random selection of parts from the collection of manufactured parts. This design does not incorporate any additional information regarding the collection of parts we sample from, except that it is a representative sample of parts produced over a certain period of time. The standard plan is appropriate when, for example, we conduct an assessment study before the BMS is used for regular inspection. However, in current industrial practice (Boyles, 2001; Burke et al., 1995), the SP is also used in cases where the BMS has been used for routine inspection and we have additional free information that can be used in the assessment study.

In this context, we consider two scenarios for assessing a BMS that also uses results from a gold standard. First, we discuss the case where we select n parts for the assessment study, and then measure them once with the gold standard and r times with the BMS. We investigate the properties of the estimators for the misclassification rates α and β, and the conforming rate π_C, when we use the standard plan and the conditional selection augmented with baseline data. In the context of repeated measurements by the BMS, the SP, which involves random selection of parts from the population of manufactured parts, was proposed by Burke et al. (1995). In Section 3.1, we compare the accuracy and precision of the estimators obtained with SP and CS sampling schemes. In Section 3.2, we investigate the testing protocol where parts are measured once with the gold standard and once with the BMS. This is a special case of the previous scenario with $r = 1$. We again look at the two sampling schemes, the standard and the conditional selection plans, and compare the accuracy and precision of the corresponding estimators.

3.1 Using the Gold Standard and Repeated Measurements by the BMS

In this section, we investigate an assessment procedure that involves the use of the gold standard and repeated measurements by the BMS. This idea is similar to the one proposed by Burke et al. (1995), which we described in Chapter 2. Here, we suppose there is one BMS under study and consider two sampling schemes. The standard plan (SP), where parts are randomly selected from the population of manufactured parts and no additional information about the pass rate is available, and the conditional selection (CS), where parts are randomly selected from the populations of previously passed and failed and we have m baseline measurements. In the CS case, the proportion of passed parts in the sample, f, is pre-determined during the design stage of the study.

For all these plans, we assume the following:

- The probability of passing the BMS inspection is common for all nonconforming parts, i.e. α is constant. The same is true for the probability of failing any conforming parts, i.e. β is constant;

- The BMS does not change the parts in any way, so the misclassification probabilities do not change from one measurement to the another;

- Conditioning on the true state of a part, X, measurements by the BMS are independent;

- Measurements on different parts are independent;

- The BMS performance does not change during the assessment study; similarly, the manufacturing process is under statistical control.

Additionally, for the CS plan, we assume that the baseline measurements come from a time period where the BMS and process parameters are not different from the time of the assessment study.

3.1.1 Standard Plan

In the standard sampling plan, n units are randomly selected and their true state, $x_i, i = 1, \ldots, n$, is determined by the gold standard. Then, each unit is measured r times and the number of passes, $s_i, i = 1, \ldots, n$, is recorded.

With this plan the likelihood function is:

$$L_{SP}(\alpha, \beta, \pi_C \mid (s_i, x_i), i = 1 \ldots n) \propto \prod_{i=1}^{n} [(1-\beta)^{s_i} \beta^{r-s_i} \pi_C]^{x_i} [\alpha^{s_i} (1-\alpha)^{r-s_i} (1-\pi_C)]^{1-x_i} \tag{3.1}$$

We notice that this function is what we called the "complete-data" likelihood (Eq. (2.10)) in the latent class analysis in Chapter 2, when the EM algorithm was used. In the case discussed here, we observe the x_i's, the true state of parts.

From (3.1), the ML estimates have a closed form as follows:

$$\hat{\alpha}(SP) = \frac{\sum_{i=1}^{n} (1-x_i) s_i}{r \sum_{i=1}^{n} (1-x_i)} \tag{3.2}$$

$$\hat{\beta}(SP) = \frac{\sum_{i=1}^{n} x_i (r - s_i)}{r \sum_{i=1}^{n} x_i} \tag{3.3}$$

$$\hat{\pi}_C(SP) = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.4}$$

We note that the estimate of α is not defined when there are no nonconforming parts in the sample, and the same is true for the estimate of β when there are no conforming parts. In the case of a high-performance process, i.e. π_C close to 1, and when the parts are randomly sampled, there is a substantial risk of selecting a sample with no nonconforming units, in which case α is not estimable. This suggests that a sampling plan that increases the chance of selecting nonconforming units is desirable.

The variance-covariance matrix of the estimators can be approximated using the Fisher

(Expected) information matrix, as given by (3.5):

$$J = \mathrm{E}(I) = \mathrm{diag}\left(\frac{nr(1-\pi_C)}{\alpha(1-\alpha)}, \frac{nr\pi_C}{\beta(1-\beta)}, \frac{n}{\pi_C(1-\pi_C)}\right) \tag{3.5}$$

We note that the estimators are asymptotically uncorrelated.

In this case, we can also think of the observed data as the realization of two binomial random variables. When all the assumptions hold and the true state of a part is first determined by the gold standard, two strata of conforming and nonconforming parts are generated. Within each stratum, parts have common probabilities of passing the BMS inspections; therefore, we can think of the total number of passes within the conforming/nonconforming stratum as realizations from (3.6) or (3.7), respectively.

$$\sum_{i=1}^{N_C} S_i \sim \mathrm{Binomial}(rN_C, 1-\beta) \tag{3.6}$$

$$\sum_{i=1}^{N_{\bar{C}}} S_i \sim \mathrm{Binomial}(rN_{\bar{C}}, \alpha) \tag{3.7}$$

where N_C and $N_{\bar{C}}$ are the random variables corresponding to the number of conforming/conforming parts in the sample. The MLEs of the parameters can also be expressed in terms of N_C and $N_{\bar{C}}$, and by conditioning on the values of these variables, we can see that $\hat{\alpha}$ and $\hat{\beta}$ are unbiased. It is also easy to see that the estimator of π_C is unbiased.

We will refer to model (3.6)–(3.7) again in Chapter 4, when we discuss the latent class analysis. When using model (3.6), the variance of $\hat{\alpha}$ is:

$$\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{N_{\bar{C}}} S_i}{rN_{\bar{C}}}\right) = \mathrm{Var}\left[\mathrm{E}\left(\frac{\sum_{i=1}^{n_{\bar{C}}} S_i}{rn_{\bar{C}}} \mid N_{\bar{C}} = n_{\bar{C}}\right)\right] + \mathrm{E}\left[\mathrm{Var}\left(\frac{\sum_{i=1}^{n_{\bar{C}}} S_i}{rn_{\bar{C}}} \mid N_{\bar{C}} = n_{\bar{C}}\right)\right],$$

Now,

$$\mathrm{E}\left(\frac{\sum_{i=1}^{n_{\bar{C}}} S_i}{rn_{\bar{C}}} \mid N_{\bar{C}} = n_{\bar{C}}\right) = \alpha, \text{ if } n_{\bar{C}} \neq 0. \text{ Therefore, } \mathrm{Var}\left[\mathrm{E}\left(\frac{\sum_{i=1}^{n_{\bar{C}}} S_i}{rn_{\bar{C}}} \mid N_{\bar{C}} = n_{\bar{C}}\right)\right] = 0.$$

Also,

$$\text{Var}\left(\frac{\sum_{i-1}^{n_{\bar{C}}} S_i}{r n_{\bar{C}}} \mid N_{\bar{C}} = n_{\bar{C}}\right) = \frac{\alpha(1-\alpha)}{r n_{\bar{C}}}, \text{ if } n_{\bar{C}} \neq 0, \text{ and } \text{E}\left[\text{Var}\left(\frac{\sum_{i-1}^{n_{\bar{C}}} S_i}{r n_{\bar{C}}} \mid N_{\bar{C}} = n_{\bar{C}}\right)\right] \simeq \frac{\alpha(1-\alpha)}{r\,\text{E}(N_{\bar{C}})}.$$

Therefore,

$$\text{Var}(\hat{\alpha}) \simeq \frac{\alpha(1-\alpha)}{r\,\text{E}(N_{\bar{C}})} \tag{3.8}$$

The approximated variance for the estimator of β is derived in a similar way from (3.7):

$$\text{Var}(\hat{\beta}) \simeq \frac{\beta(1-\beta)}{r\,\text{E}(N_C)} \tag{3.9}$$

These approximations show that when we first use the gold standard and then repeatedly measure the units with the BMS, the precision of the estimators of α and β depends on the true values of the parameters, the number of repeated measurements and the expected numbers of nonconforming and conforming units in the sample, respectively. In the case of the standard plan, $\text{E}(N_C) = n\pi_C$ and $\text{E}(N_{\bar{C}}) = n(1-\pi_C)$, and the approximations are the ones given by the Fisher information matrix, Eq. (3.5). For the SP, since π_C is usually large, $\text{E}(N_C)$ is much larger than $\text{E}(N_{\bar{C}})$, and β is better estimated than α.

The expressions for the variance approximations suggest that we can achieve a certain precision of the estimators by controlling the expected number of conforming and nonconforming parts in the sample. One approach is to allocate equal resources for estimating α and β, and then the goal would be to get a more balanced sample, i.e. to make $\text{E}(N_C)$ and $\text{E}(N_{\bar{C}})$ approximately equal. Also, for the SP, we note that the variances of $\hat{\alpha}$ and $\hat{\beta}$ depend on the values of α and β, respectively, of π_C and on the total number of measurements $n \times r$. Therefore, in the case where a gold standard is available and we use a SP, the allocation of n and r does not have an effect on the precision of $\hat{\alpha}$ and $\hat{\beta}$, for the same total number of measurements. However, as the estimate of π_C is based on the total number of parts n in the study, increasing r and therefore decreasing n leads to a loss in precision for $\hat{\pi}_C$.

Next, we look at the precision of the parameter estimators $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}_C$, i.e. the size of the

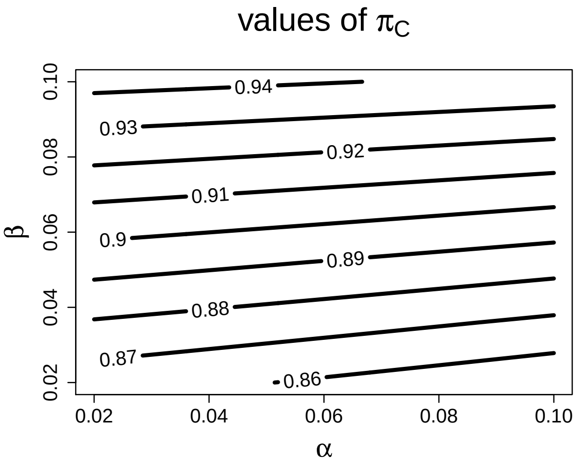

Figure 3.1 Contours of π_C for a Grid of Values for α, β, when $\pi_P = 0.85$.

standard deviations as derived from Equations (3.8) and (3.9). We focus on the case where the BMS has good performance characteristics and we have a high-quality manufacturing process. Therefore, we consider cases where the actual misclassification probabilities are small (e.g. $\alpha, \beta \leq 0.1$), and the conforming rate is large (e.g. $0.85 \leq \pi_C$). These cases are commonly found in the manufacturing industry. In our investigations, we construct a grid of parameter values by keeping π_P constant, and then varying α and β within the above limits. As π_P is related to the other three parameters by Eq. (1.3), π_C takes on different values across the grid. For example, when $\pi_P = 0.85$ and $0.02 \leq \alpha, \beta \leq 0.1$, π_C varies from 0.85 to 0.94, as shown in Figure 3.1. Note that π_C is insensitive to changes in α. We first look at small sample sizes $n = 100$ and $r = 2$ number of repeated measurements and conclude that, although the estimators are unbiased, their precision is very poor (results not shown here). Next, we increase the sample size to $n = 200$ and $r = 5$. Figure 3.2 shows the asymptotic standard errors of the estimates for a total number of measurements $n \times r = 1{,}000$. We note that the precision of $\hat{\alpha}$ varies mostly with α and it is small enough to be useful. For a few cases where α is small and π_C large, the size of the standard error is almost as large as the true parameter value α. The other two parameters are well estimated across the whole grid of values. Note that the precision of $\hat{\beta}$ varies mostly with β, and the precision of $\hat{\pi}_C$ with π_C. We expect similar results for larger values of $n \times r$ and other

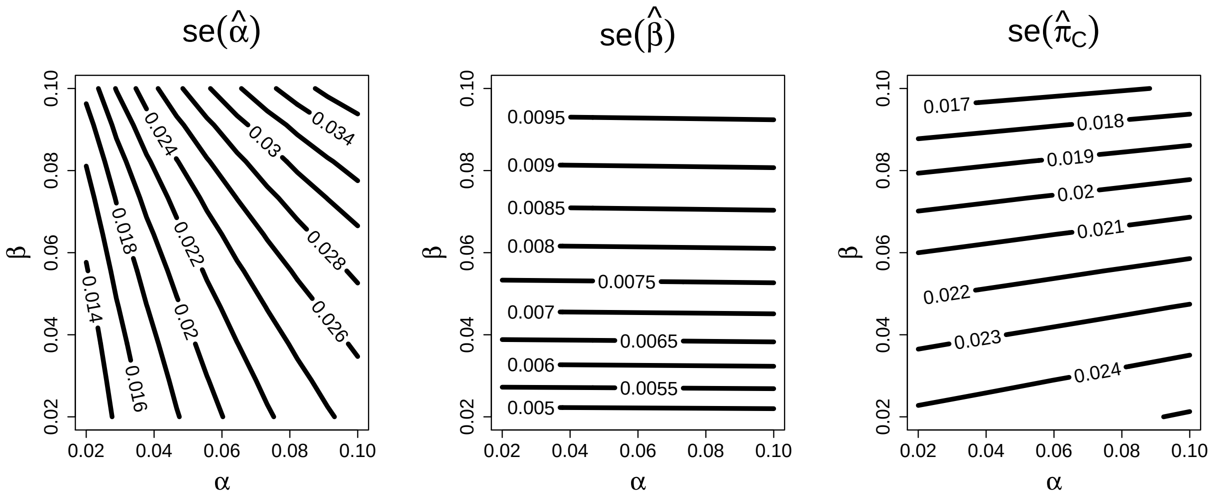

Figure 3.2 Asymptotic Standard Errors – Standard Plan with $n = 200$ and $r = 5$, when $\pi_P = 0.85$

values of π_P.

3.1.2 Conditional Selection (CS) without Baseline Data

In this section, we consider the sampling plan where predetermined numbers of previously passed and failed parts are sampled for the assessment study. We mentioned before that, in our context, large populations of passed and failed parts are readily available and we can sample conditionally on the result of a previous (baseline) BMS measurement. Here, we investigate the case where we use a conditional selection (CS) plan, but we do not include the baseline information about the pass rate π_P in the estimation. Although, in practice, this information is available and it should be used, here we are interested in looking at the behavior of the estimators when we have no prior information about π_P.

With the CS plan, we control the number of passed and failed parts in the sample, and therefore we can increase the chance of having some nonconforming parts in the sample, since we expect $\Pr(\bar{C} \mid \bar{P}) > \Pr(\bar{C})$. As noted before, if there are no nonconforming parts in the sample, the estimate of α is not defined.

Also, we notice that in Equations (3.6)–(3.7), the precision of the estimator of α depends on $\mathrm{E}(N_{\bar{C}})$, the expected number of nonconforming units, whereas the precision of the estimator of β depends on $\mathrm{E}(N_C)$. If we focus on estimating α with high precision, we have to choose a sampling plan that gives a large expected number of nonconforming units. If we focus on allocating equal resources for estimating α and β, the sampling plan should give approximately equal $\mathrm{E}(N_{\bar{C}})$ and $\mathrm{E}(N_C)$. When sampling conditionally on the result of a previous measurement by the BMS, we can control both $\mathrm{E}(N_{\bar{C}})$ and $\mathrm{E}(N_C)$ by choosing the appropriate proportion of passed units, f. The following expression shows how the expected number of nonconforming units varies with f:

$$\mathrm{E}(N_{\bar{C}}) = n\left[\frac{\alpha(1-\pi_C)}{(1-\beta)\pi_C+\alpha(1-\pi_C)}f + \frac{(1-\alpha)(1-\pi_C)}{\beta\pi_C+(1-\alpha)(1-\pi_C)}(1-f)\right] \tag{3.10}$$

Therefore, the CS offers more flexibility with regard to the goal of the assessment study, as we can choose f according to this goal.

The likelihood function for the CS sampling plan is:

$$\begin{aligned} L_{CS}(\alpha,\beta,\pi_C \mid (s_i,x_i), i=1\ldots n) &= \prod_{i=1}^{n_P} \frac{[(1-\beta)^{s_i+1}\beta^{r-s_i}\pi_C]^{x_i}[\alpha^{s_i+1}(1-\alpha)^{r-s_i}(1-\pi_C)]^{1-x_i}}{(1-\beta)\pi_C+\alpha(1-\pi_C)} \\ &\times \prod_{i=1}^{n_{\bar{P}}} \frac{[(1-\beta)^{s_i}\beta^{r-s_i+1}\pi_C]^{x_i}[\alpha^{s_i}(1-\alpha)^{r-s_i+1}(1-\pi_C)]^{1-x_i}}{\beta\pi_C+(1-\alpha)(1-\pi_C)} \end{aligned} \tag{3.11}$$

Now, we are interested in comparing the precision of the estimators given by the SP with the ones given by the CS plan without baseline. We derive the asymptotic standard deviations for the CS plan using the Fisher information matrix. We do not include here the expression for the information matrix, as it has a complicated form, but we provide R-code (R Development Core Team, 2010) for obtaining the asymptotic standard errors. Next, we compare the asymptotic standard errors for the CS with $f = 0.5$ with the ones from the SP plan. Figure 3.3 shows the ratios of the asymptotic standard deviations for the estimators of α, β and π_C, for the SP and CS plans,

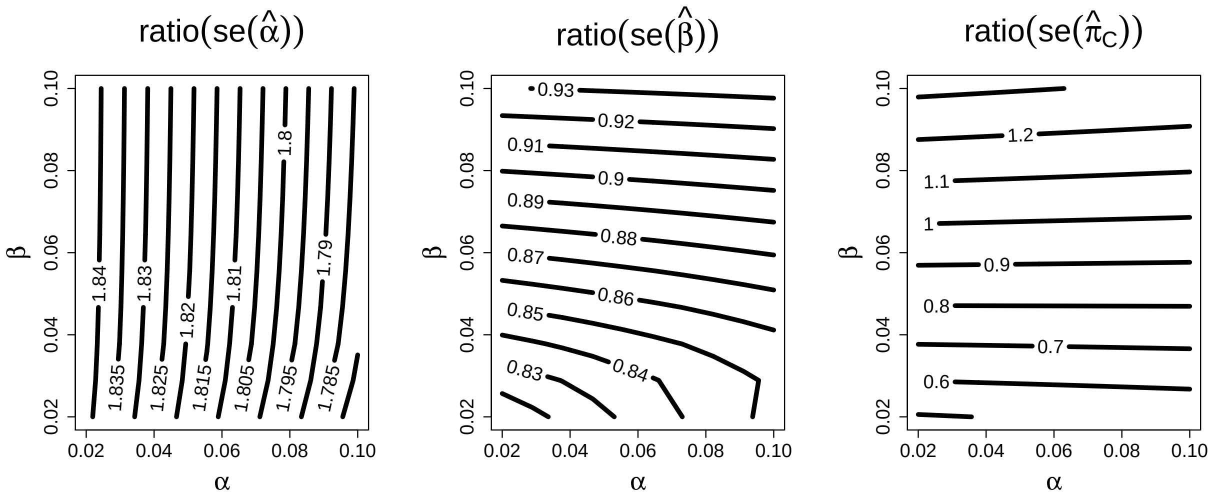

Figure 3.3 Ratios of Asymptotic Standard Errors, $se(SP)/se(CS, f = 0.5)$, when $n = 200$, $r = 5$, and $\pi_P = 0.85$

when $f = 0.5$ and $r = 5$.

We notice that the CS plan gives a more precise estimator of α and the SP a more precise estimator of β, a result that is consistent with (3.6)–(3.7), and the fact that in the CS plan the expected number of nonconforming parts is larger than in the SP case. For example, when $\alpha = 0.02$, $\beta = 0.02$, $\pi_C = 0.9$ and $n = 200$, for the standard plan, the expected number of nonconforming parts in the sample $\mathrm{E}(N_{\bar{C}}) = 20$, whereas for the conditional selection plan, $\mathrm{E}(N_{\bar{C}}) = 85$. Another option is to select only from the population of previously failed units, i.e. $f = 0$, in which case $\mathrm{E}(N_{\bar{C}}) = 169$. Table 3.1 shows the expected proportion of nonconforming parts in the sample for the grid of values considered in Figure 3.1.

Table 3.1 Expected Proportion of Nonconforming Parts in the Sample, when $\pi_P = 0.85$

α	β	π_C	$E_{SP}(N_{\bar{C}}/n)$	$E_{CS_{f=0.5}}(N_{\bar{C}}/n)$	$E_{CS_{f=0}}(N_{\bar{C}}/n)$
0.02	0.02	0.865	0.1354	0.4440	0.8847
0.04	0.02	0.862	0.1377	0.4454	0.8850
0.05	0.02	0.860	0.1401	0.4470	0.8853
0.07	0.02	0.857	0.1425	0.4485	0.8857
0.08	0.02	0.855	0.1451	0.4502	0.8860
0.10	0.02	0.852	0.1477	0.4519	0.8864
0.02	0.04	0.879	0.1208	0.3959	0.7890
0.04	0.04	0.877	0.1228	0.3973	0.7895
0.05	0.04	0.875	0.1250	0.3988	0.7900
0.07	0.04	0.873	0.1272	0.4004	0.7905
0.08	0.04	0.870	0.1295	0.4019	0.7911
0.10	0.04	0.868	0.1319	0.4036	0.7917
0.02	0.05	0.894	0.1056	0.3462	0.6899
0.04	0.05	0.893	0.1075	0.3476	0.6906
0.05	0.05	0.891	0.1094	0.3490	0.6912
0.07	0.05	0.889	0.1114	0.3504	0.6919
0.08	0.05	0.887	0.1134	0.3519	0.6927
0.10	0.05	0.884	0.1156	0.3535	0.6934
0.02	0.07	0.910	0.0899	0.2948	0.5874
0.04	0.07	0.908	0.0915	0.2960	0.5882
0.05	0.07	0.907	0.0932	0.2973	0.5889
0.07	0.07	0.905	0.0949	0.2986	0.5897
0.08	0.07	0.903	0.0967	0.3000	0.5905
0.10	0.07	0.901	0.0986	0.3015	0.5913
0.02	0.08	0.926	0.0737	0.2415	0.4813
0.04	0.08	0.925	0.0750	0.2426	0.4820
0.05	0.08	0.924	0.0764	0.2437	0.4828
0.07	0.08	0.922	0.0778	0.2449	0.4836
0.08	0.08	0.921	0.0793	0.2461	0.4844
0.10	0.08	0.919	0.0809	0.2474	0.4853
0.02	0.10	0.943	0.0568	0.1863	0.3712
0.04	0.10	0.942	0.0579	0.1872	0.3719
0.05	0.10	0.941	0.0590	0.1881	0.3726
0.07	0.10	0.940	0.0601	0.1891	0.3734
0.08	0.10	0.939	0.0613	0.1901	0.3742
0.10	0.10	0.938	0.0625	0.1912	0.3750

3.1.3 Conditional Selection (CS) with Baseline Data

Model Formulation

Suppose we have a baseline population of m parts, each measured once for inspection purposes. That is, we record the output of the routine inspection, i.e. passed or failed, over a certain period of time prior to the assessment study. Therefore, we know that out of m parts, m_P passed the BMS inspection. Here, we assume that the performance of the BMS and the process characteristics do not change between the time we collect the baseline data and the time of the BMS assessment. The n parts sampled from the previously passed and failed for the assessment study do not necessarily have to be selected from the baseline once-measured parts m. That is, the baseline measurements can be independent of the study measurements. For high volume processes, m is typically large. For the m parts measured once only, the likelihood is:

$$L_b(\pi_P) \propto \pi_P^{m_P}(1-\pi_P)^{m-m_P} \tag{3.12}$$

where m_P is the number of passed parts in this group. We call m the size of the baseline sample. Note that $L_b(\pi_P)$ can be rewritten in terms of α, β, and π_C, using the constraint $\pi_P = (1-\beta)\pi_C + \alpha(1-\pi_C)$, though it is not possible to separately estimate α, β, and π_C using $L_b(\pi_P)$ alone.

The overall likelihood for a CS plan with baseline data is proportional to:

$$L_b(\pi_P) \times L_{CS}(\alpha, \beta, \pi_C | (s_i, x_i), i = 1 \dots n)$$

where L_{CS} is the likelihood in Equation (3.11) and s_i is the number of times part i passed in the r repeated measurements.

Bias and Precision of the Estimators

We investigate the properties of the maximum likelihood estimators for different CS plans with a baseline of size m, by conducting several simulation studies, where we focus on the same range of parameters values as illustrated by Figure 3.1. In our simulations, we run 500 repeats for each combination of parameters values from the grid. For each repeat, we first generate n_P random samples from Binomial$(1, \Pr(X = 1 \mid Y_0 = 1))$ and $n_{\bar{P}}$ from Binomial$(1, \Pr(X = 1 \mid Y_0 = 0))$, where Y_0 denotes the random variable representing the previous (baseline) measurement. Y_0 is 1 for a previously passed part and 0 for a previously failed. Then, for each conforming part in the whole sample, i.e. $x = 1$, we generate the number of passes, s, a realization of $S \mid (X = 1)$, which is distributed Binomial$(r, 1-\beta)$. Similarly, for each nonconforming part, we generate the total number of passes s from Binomial(r, α). The estimation of the parameters is based on the number of passes for each part, s, the information regarding the initial measurement y_0, and the true state x. As in the CS case the MLEs do not have a closed form, we use constrained optimization routines in the R environment (R Development Core Team, 2010) for the parameters estimation, such as Nelder-Mead algorithm (Nelder and Mead, 1965). We optimize the likelihood function in Eq. (3.11) under the constraints $0 < \alpha, \beta, \pi_C < 1$. For each combination of parameters values, we get the sample errors and standard deviations as approximations of the biases and standard deviations of the estimators. We fit separate local polynomial regression (loess) smoothing models (Cleveland et al., 1992) to biases and standard errors, and then get predictions over the grid.

Also, for each combination of parameters values, we derive the asymptotic standard errors based on the Fisher (expected) information matrix. We note that for the CS plan, the estimators are not asymptotically uncorrelated anymore, as they are in the SP case. We start our investigation with small sample sizes, numbers of repeated measurements, and baseline size (e.g. $n = 200$, $r = 5$, and $m = 1{,}000$).

First, we consider the case of an equal number of passed and failed items, i.e. $f = 0.5$. Figure 3.4 shows the smoothed biases of the estimators and we notice that all estimators are virtually

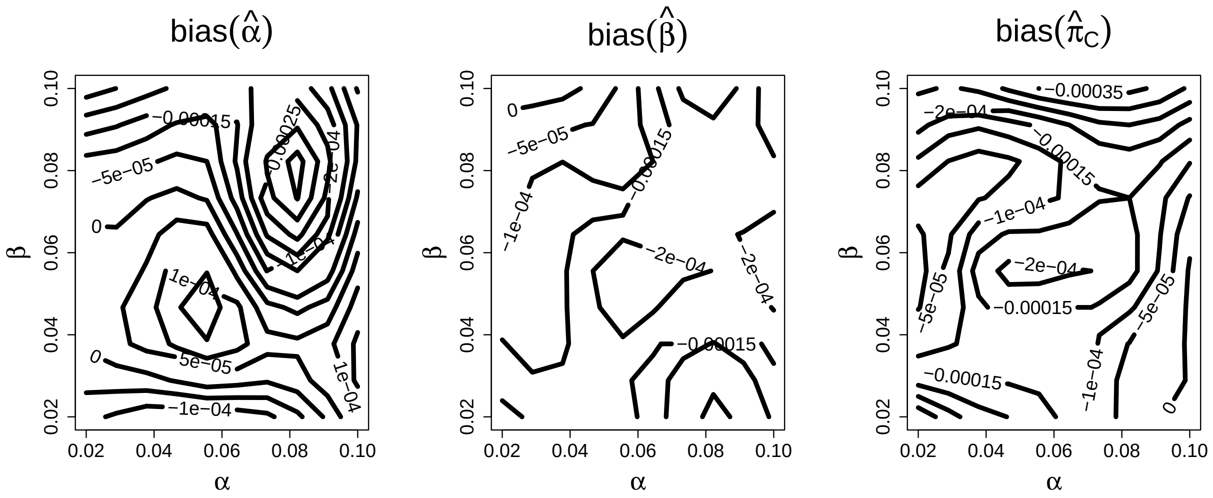

Figure 3.4 Smoothed Biases – Conditional Selection with $n = 200$, $r = 5$, $m = 1{,}000$, and $f = 0.5$.

unbiased. Next, we compare the simulation-based and asymptotic standard errors and conclude that they are very close (results not shown here). Figure 3.5 shows the asymptotic standard deviations for the CS plan with $f = 0.5$ and $m = 1{,}000$. We note that all parameters are estimated with good precision, including α. For larger values of n and r, we expect the same conclusions.

Now, we are interested in comparing the precision of the estimators given by the SP plan with the ones given by the CS with $f = 0.5$ and baseline $m = 1{,}000$. The ratios of asymptotic standard deviations are shown in Figure 3.6, and we notice that the CS plan gives uniformly better estimators, especially for α and π_C. The gains in precision are substantial when we use the CS plan compared to the SP. As in Section 3.1.2, where we do not use the baseline information, we expect to get more nonconforming parts in the sample when using the CS plan with $f = 0.5$ than with the SP, for the same sample size n (see Table 3.1). Therefore, we expect more efficient estimators for α. However, adding the baseline information about the pass rate π_P leads to surprising results regarding the precision of $\hat{\beta}$. That is, β is better estimated with the CS plan, although there are fewer conforming parts in the sample than in the SP case. The baseline provides an estimate of $\pi_P = (1-\beta)\pi_C + \alpha(1-\pi_C)$, a function of α, β, and π_C. Since we are considering situations where π_C is large, i.e. high-quality manufacturing processes, and α is

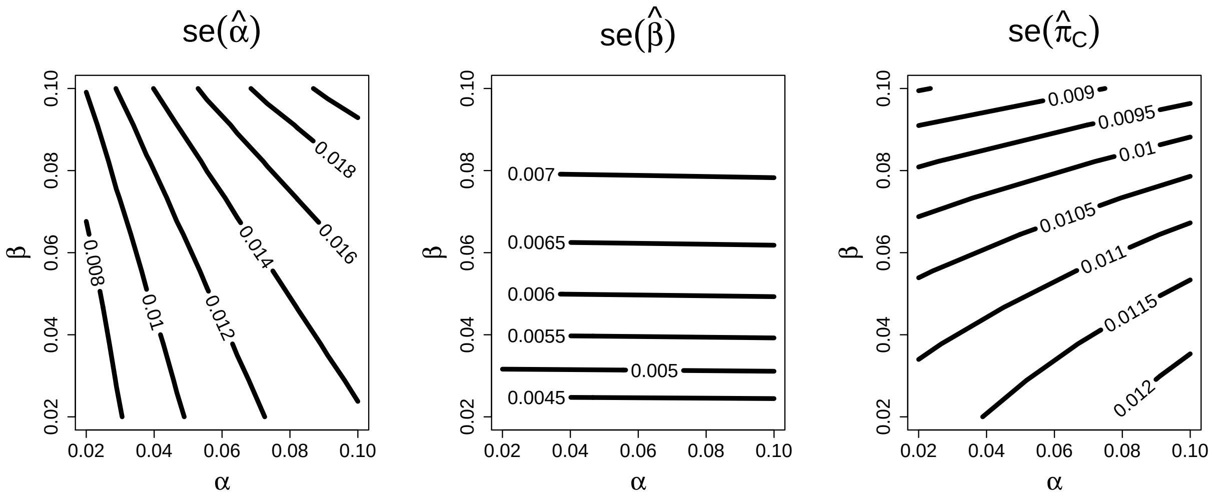

Figure 3.5 Asymptotic Standard Errors – Conditional Selection Plan with $n = 200$, $r = 5$, $m = 1{,}000$, and $f = 0.5$.

small, π_P is strongly influenced by β and π_C. In this case, the additional information about β and π_C from the baseline outweighs the lost information due to fewer conforming parts in the sample.

Next, we look at a CS plan with parts sampled only from the population of failed parts, i.e. $f = 0$, and then compare the standard deviations given by the CS when $f = 0.5$ with the ones given by the CS with $f = 0$, as in Figure 3.7. We would expect the CS plan with $f = 0$ to give smaller precision for $\hat{\beta}$, when compared to the CS with $f = 0.5$, for the same baseline size $m = 1{,}000$. CS with $f = 0$ yields a smaller expected number of conforming than CS with $f = 0.5$, and the estimator of β is mostly influenced by the number of conforming parts in the sample. However, we note that the CS with $f = 0$ gives better precision for the estimators of $\hat{\alpha}$ and $\hat{\pi}_C$, and similar precision for $\hat{\beta}$. As we cannot explain this unusual result, we look at other values of n and r, and conclude that for $n = 200$ and $r = 10$ the standard errors of $\hat{\beta}$ are smaller for the CS plan with $f = 0.5$ than for $f = 0$.

Therefore, using a conditional selection plan, aside from decreasing the chance of not having nonconforming units in the sample, also gives more efficient estimators. We note that

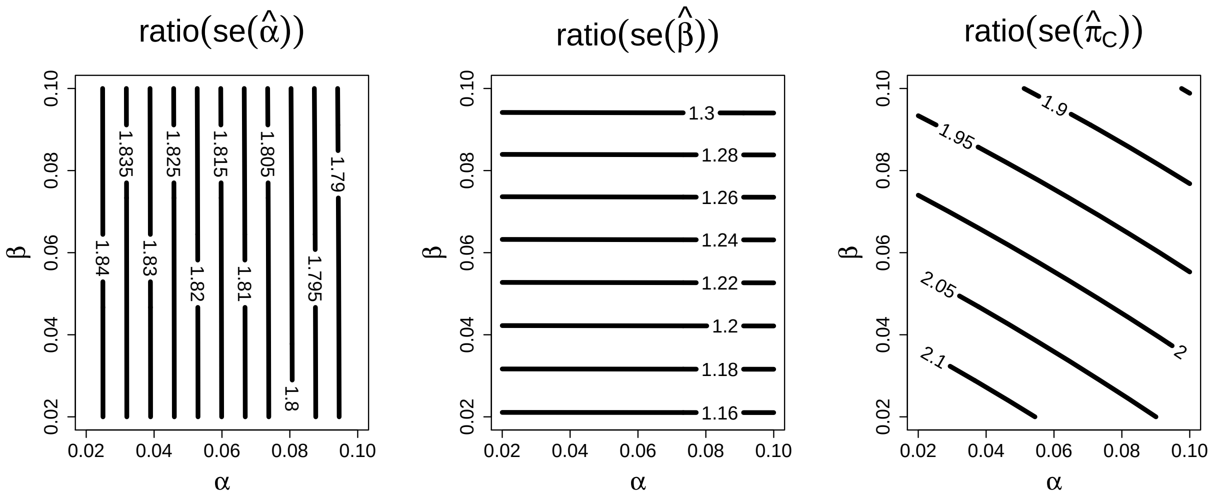

Figure 3.6 Ratios of Asymptotic Standard Errors – $se(SP)/se(CS, f = 0.5, m = 1{,}000)$, when $n = 200$, $r = 5$, and $\pi_P = 0.85$.

when we use the CS plan with baseline, the results do not follow the logic in model (3.6)–(3.7), where the approximate precision of the estimator of β is a function of the expected number of conforming parts in the sample. The CS plan with $f = 0$ and baseline gives a more precise estimator of β than the SP, even though the expected number of conforming is higher with the latter plan.

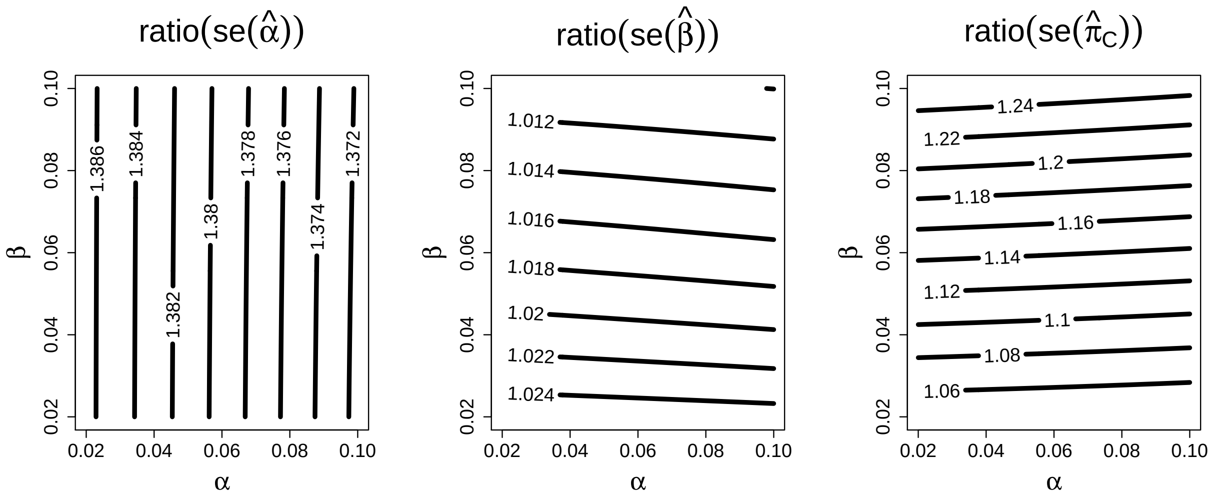

Figure 3.7 Ratios of Asymptotic Standard Errors – $se(CS, f = 0.5)/se(CS, f = 0)$, when $m = 1{,}000$, $n = 200$, $r = 5$, and $\pi_P = 0.85$.

3.2 Using the Gold Standard and a Single Measurement by the BMS

In this section, we focus on an assessment procedure that involves one measurement with the studied BMS and one with the gold-standard system. We first look at a sampling plan commonly used in both manufacturing (Boyles, 2001) and medical (Pepe, 2003) contexts, the random selection or cohort study design described in Chapter 2, Section 2.1.1. With this plan, we randomly select parts from the population of manufactured parts and then each part is measured once with the BMS and once with the gold standard. This assessment procedure is a special case of the standard plan (SP) discussed in Section 3.1.1, with $r = 1$. Next, we investigate a study design where parts are selected from the collections of previous passed and failed, that is, we use a conditional selection (CS) plan, and then measured once with the gold standard. This design corresponds to a special case of the CS plan discussed in 3.1.3, with $r = 0$. That is, during the assessment study we only measure the sampled parts with the gold standard, knowing for each part the outcome of the previous (baseline) measurement by the BMS. Additionally, we assume we have baseline information about the pass rate π_P. Although the industrial context considered here is different, this special case of a CS plan is similar to the partial verification with conditional re-sampling design described in Chapter 2 and proposed by Tenenbein (1971), and Haitovsky and Rapp (1992). Contrary to the partial verification with conditional re-sampling case, in our context, we do not assume that the sampled parts come from the collection of parts that make up the baseline measurements. However, here we look at a special case that is not investigated in the literature related to the partial verification design, that is, we look at the limiting case where the baseline size $m \to \infty$, or equivalently π_P is known.

3.2.1 Standard Plan

When parts are measured once with the gold standard and once with the BMS, we can organize the study data in a 2 × 2 table as in Table 3.2.

When parts are randomly selected from the population of manufactured parts and then measured once with the BMS and once with the gold standard (we use a standard plan) all quantities except n are random in Table 3.2. The maximum likelihood estimates for the parameters of interest are:

$$\begin{aligned}\hat{\alpha}(SP) &= \frac{n_{P\bar{C}}}{n_{\bar{C}}}\\ \hat{\beta}(SP) &= \frac{n_{\bar{P}C}}{n_C}\\ \hat{\pi}_C(SP) &= \frac{n_C}{n}\end{aligned} \tag{3.13}$$

The estimators are unbiased and their asymptotic variances are obtained using standard theory related to the multinomial distribution (Casella and Berger, 2002).

We look at the precision of the estimators for the whole grid of parameters values as in Figure 3.1. As noted in Section 3.1.1, we need a relatively large total number of measurements to get the estimators with good precision. Therefore, we start our investigation with a sample size

Table 3.2 Data from a BMS Performance Study with Complete Verification by the Gold Standard

	Conform (C)	Not conform ($\bar{C}$)	Total parts
Pass (P)	n_{PC}	$n_{P\bar{C}}$	n_P
Fail ($\bar{P}$)	$n_{\bar{P}C}$	$n_{\bar{P}\bar{C}}$	$n_{\bar{P}}$
Total parts	n_C	$n_{\bar{C}}$	n

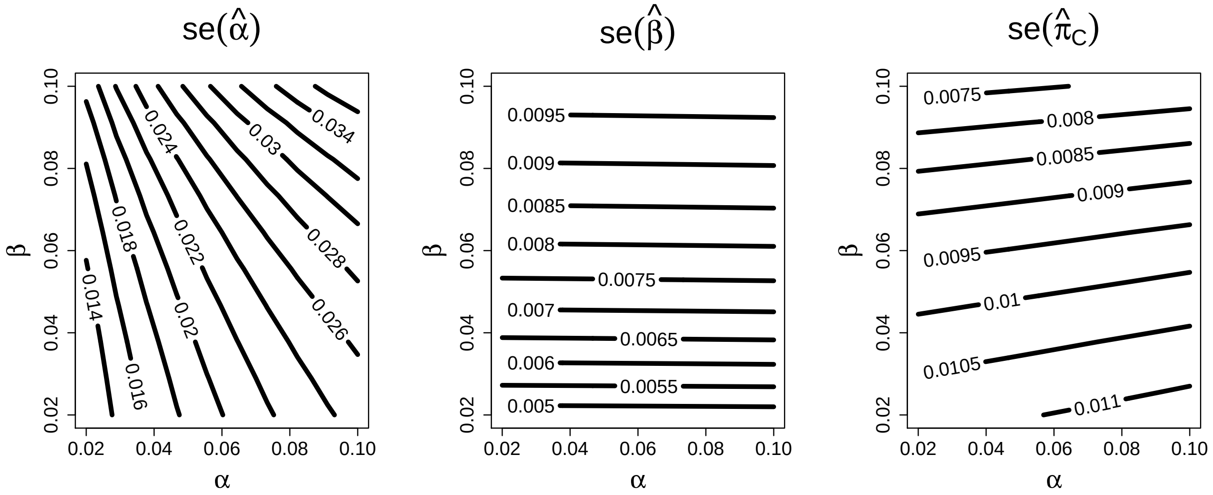

Figure 3.8 Asymptotic Standard Errors – Standard Plan with $n = 1000$, when $\pi_P = 0.85$

of $n = 1{,}000$.

Figure 3.8 shows the asymptotic standard deviations of the estimators and we note that, except for the case where $\alpha = 0.02$ and π_C is large, the estimators have good precision. Actually, the precision of $\hat{\alpha}$ and $\hat{\beta}$ in Figure 3.8 are identical with the ones given in Figure 3.2, as the total number of measurements is the same for the two designs and the allocation of $n \times r$ does not have an effect on these standard errors.

3.2.2 Conditional Selection (CS) Plan with π_P Known

When large collections of passed and failed units are readily available, we propose a sampling plan that involves selecting two independent samples from these populations, where the numbers of selected passed and failed units, n_P and $n_{\bar{P}}$, are pre-determined. Then, all the selected parts are inspected using the gold-standard system, and their true state is determined. With the CS plan, in Table 3.2, n_P, $n_{\bar{P}}$, and n are fixed, and all the other quantities are random. We also assume that the pass rate π_P is known prior to the assessment study.

With the CS plan, the two misclassification probabilities, α and β, cannot be directly estimated. Instead, we start with the ML estimates:

$$\hat{\Pr}(\bar{C} \mid P) = \frac{n_{P\bar{C}}}{n_P} \quad \text{and} \quad \hat{\Pr}(\bar{C} \mid \bar{P}) = \frac{n_{\bar{P}\bar{C}}}{n_{\bar{P}}}.$$

Then, by Bayes' Rule:

$$\alpha = \frac{\Pr(P \cap \bar{C})}{\Pr(\bar{C})} = \frac{\Pr(\bar{C} \mid P)\Pr(P)}{\Pr(\bar{C} \mid P)\Pr(P) + \Pr(\bar{C} \mid \bar{P})\Pr(\bar{P})} \tag{3.14}$$

and

$$\beta = \frac{\Pr(\bar{P} \cap C)}{\Pr(C)} = \frac{\Pr(C \mid \bar{P})\Pr(\bar{P})}{\Pr(C \mid \bar{P})\Pr(\bar{P}) + \Pr(C \mid P)\Pr(P)} \tag{3.15}$$

We also have:

$$\pi_C = \Pr(C \cap P) + \Pr(C \cap \bar{P}) = \Pr(C \mid P)\Pr(P) + \Pr(C \mid \bar{P})\Pr(\bar{P}) \tag{3.16}$$

We note that all parameters are estimable only if the pass rate, $\Pr(P) = \pi_P$, is known.

We know that $N_{P\bar{C}} \sim \text{Binomial}(n_P, \Pr(\bar{C} \mid P))$ and $N_{\bar{P}\bar{C}} \sim \text{Binomial}(n_{\bar{P}}, \Pr(\bar{C} \mid \bar{P}))$, where $N_{P\bar{C}}$ and $N_{\bar{P}\bar{C}}$ are the random variables whose realisations are $n_{P\bar{C}}$ and $n_{\bar{P}\bar{C}}$.

Using the invariance property of the ML estimators and the known π_P, we obtain the

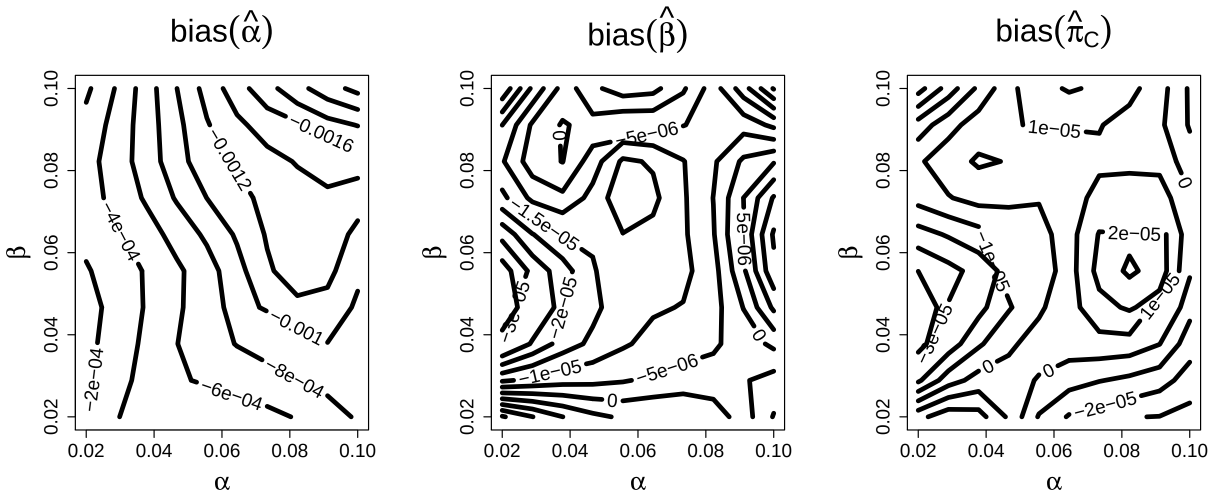

Figure 3.9 Smoothed Biases – Conditional Plan with $f = 0.5$ and $n = 1{,}000$, when $\pi_P = 0.85$.

following ML estimates:

$$\hat{\alpha}(CS) = \frac{\pi_P \dfrac{n_{P\bar{C}}}{n_P}}{\pi_P \dfrac{n_{P\bar{C}}}{n_P} + (1-\pi_P)\dfrac{n_{\bar{P}\bar{C}}}{n_{\bar{P}}}} \tag{3.17}$$

$$\hat{\beta}(CS) = \frac{(1-\pi_P)\dfrac{n_{\bar{P}C}}{n_{\bar{P}}}}{\pi_P \dfrac{n_{PC}}{n_P} + (1-\pi_P)\dfrac{n_{\bar{P}C}}{n_{\bar{P}}}} \tag{3.18}$$

$$\hat{\pi}_C(CS) = \pi_P \frac{n_{PC}}{n_P} + (1-\pi_P)\frac{n_{\bar{P}C}}{n_{\bar{P}}} \tag{3.19}$$

We also obtain approximations for the variances of the estimators in terms of α, β and π_P

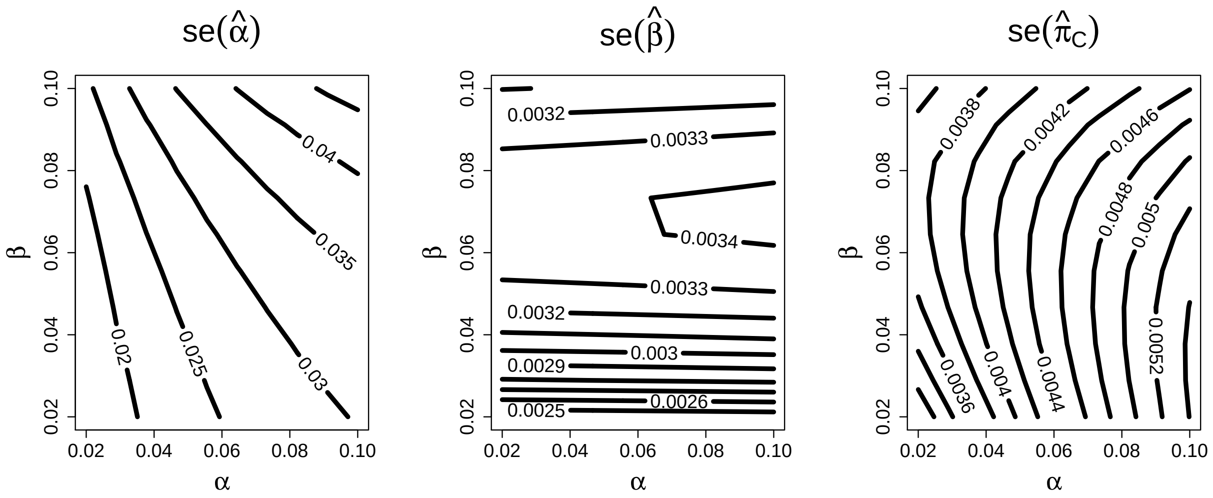

Figure 3.10 Asymptotic Standard Errors – Conditional Plan with $f = 0.5$ and $n = 1{,}000$, when $\pi_P = 0.85$.

(known) using the δ-method (Casella and Berger, 2002):

$$\text{Var}(\hat{\alpha}(CS)) \simeq \frac{\alpha(1-\alpha)(\pi_P-\alpha)}{1-\beta-\pi_P}\left(\frac{1-\alpha-\beta+\alpha\beta}{n_P}+\frac{\alpha\beta}{n_{\bar{P}}}\right)$$

$$\text{Var}(\hat{\beta}(CS)) \simeq \frac{\beta(1-\beta)(1-\beta-\pi_P)}{\pi_P-\alpha}\left(\frac{\alpha\beta}{n_P}+\frac{1-\beta-\alpha+\alpha\beta}{n_{\bar{P}}}\right)$$

The variance of $\hat{\pi}_C(CS)$ can be directly derived as:

$$\text{Var}(\hat{\pi}_C(CS)) = \frac{(1-\beta-\pi_P)(\pi_P-\alpha)}{(1-\alpha-\beta)^2}\left(\frac{\alpha(1-\beta)}{n_P}+\frac{\beta(1-\alpha)}{n_{\bar{P}}}\right)$$

First, we look at the properties of the estimators by simulating study data using the R environment (R Development Core Team, 2010), over a grid of values as in Figure 3.1, for a CS plan with $f = 0.5$ and $n = 1{,}000$. For each simulation run, we obtain the ML estimates using Equations (3.17)–(3.19). Then, for each combination of parameters values, we get the sample errors and standard deviations as approximations for the biases and standard deviations of the estimators. In Figure 3.9, we note small biases for the estimator of α, for large values of α and β. All the other estimators are virtually unbiased. The simulation-based standard errors are close to the asymptotic ones

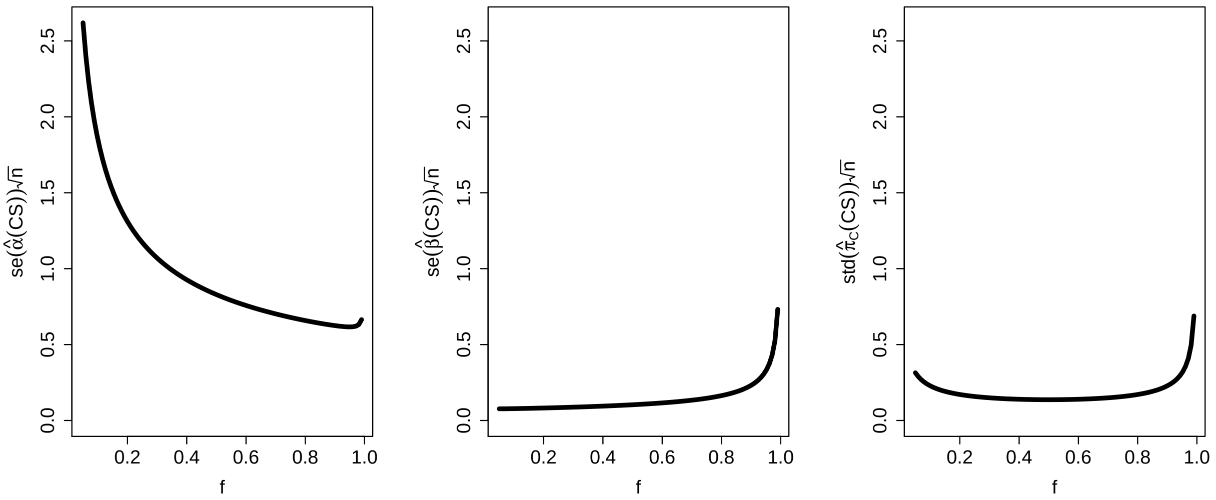

Figure 3.11 sd($\hat{\alpha}(CS)$)$\sqrt{n}$, sd($\hat{\beta}(CS)$)$\sqrt{n}$, and sd($\hat{\pi}_C(CS)$)$\sqrt{n}$ as functions of the proportion of passed items in the sample, $\alpha = 0.05$, $\beta = 0.05$, $\pi_C = 0.9$

(results not shown here). Figure 3.10 shows the asymptotic standard errors and we note that $\hat{\beta}$ and $\hat{\pi}_C$ have very good precision. However, the estimator of α is less efficient, with the standard errors close to the size of the parameter when α is small and π_C large.

Next, we look at how the precision of the estimators varies with the proportion of passed parts in the sample. Figure 3.11 illustrates how the asymptotic standard deviations of the estimators multiplied by the square root of the sample size n vary with the proportion of passed items in the sample, $f = n_P/n$, for some specific values of α, β and π_C. Note that with the CS plan, we control this proportion.

We notice that sd($\hat{\alpha}(CS)$)$\sqrt{n}$ decreases with the proportion of passed, with a big drop from 0.1 to 0.4, and then slowly over the interval $[0.6, 1)$. Therefore, to estimate α, selecting f anywhere between 0.6 and 0.9 gives roughly the same results.

We also note that sd($\hat{\beta}(CS)$)$\sqrt{n}$ increases very slowly over the entire interval for f. If the main goal of the study is to estimate both α and β with good precision, we can select as many as 80% passed parts and 20% failed parts. Also, sd($\hat{\pi}_C(CS)$)$\sqrt{n}$ varies very little over the interval $f \in [0.2, 0.8]$. Therefore, for this combination of parameters values, a CS plan with $f = 0.8$ seems

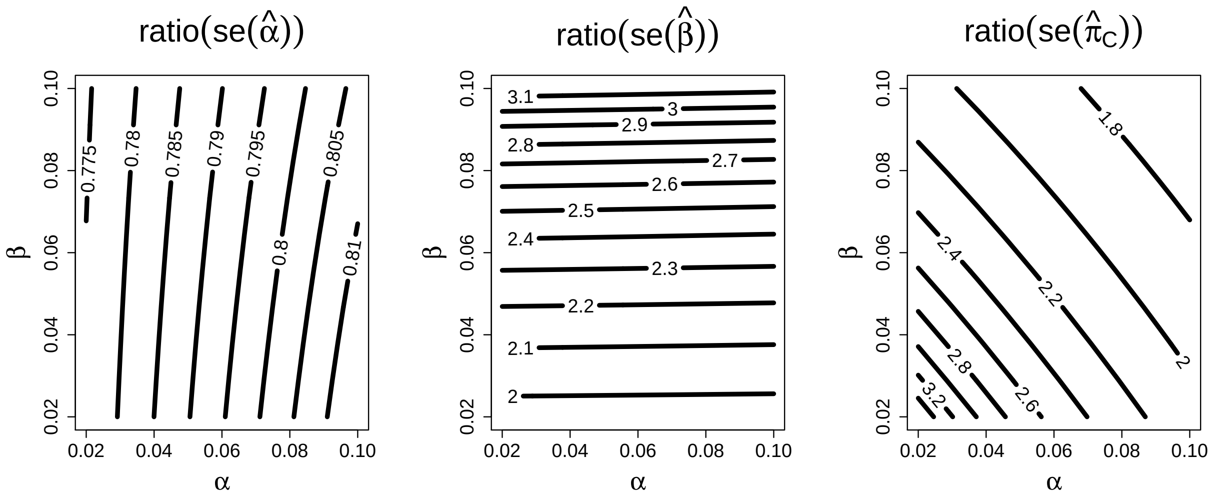

Figure 3.12 Ratios of Asymptotic Standard Errors – $se(SP)/se(CS, f = 0.5, \pi_P$ known) – when n = 1,000 and $\pi_P = 0.85$.

an efficient design.

We find the above results unusual, as we expect that sampling heavily from the population of previously failed (i.e. f close to 0) and, therefore, getting more nonconforming parts in the sample, would improve the precision of $\hat{\alpha}$ and decrease the precision of $\hat{\beta}$. This is the case in Section 3.1, where parts are repeatedly measured with the BMS. However, as we can see in Equations (3.14) – (3.16), in the case where only the baseline measurement is included, the estimation of α, β, and π_C is based on two functions of the data, $\hat{\Pr}(\bar{C} \mid P)$ and $\hat{\Pr}(\bar{C} \mid \bar{P})$, and the known values of π_P. Therefore, it is difficult to intuitively see how the precision of the estimators changes with the proportion of passed f.

Next, we compared the efficiencies of the standard plan and the conditional selection plan with $f = 0.5$, by looking at the ratios of asymptotic standard errors for $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}_C$. In Figure 3.12, we notice that SP gives a consistently better estimator for α over all values of α and β, and π_C. The parameter β has a smaller influence than α on the ratio. However, the CS plan gives a more precise estimator of β than the SP, as the ratio of the standard deviations is uniformly larger than 1 over the whole range of α, β, and π_C. Also, the ratio increases with the value of β, and is insensitive to changes in α.

It is interesting to note that these results are counter-intuitive and differ from the ones we found in the case where the parts are repeatedly measured with the BMS (Section 3.1). As with that case, we would expect a higher precision for the estimator of α given by the CS plan, as α is estimated based on the number of nonconforming units in the sample and we expect a larger number of nonconforming units with the CS plan.

We also notice that the CS plan gives a more precise estimator for π_C, for the whole grid of parameters values.

3.2.3 Sample Size Calculation

So far we have concentrated on the analysis and properties of the estimators. Now, suppose we are in the planning stage and are interested in designing a plan to get a prescribed precision for an estimator. Then, the objective is to find the minimum sample size that achieves this precision. The functions that give the sample size for the CS plan, when we want a certain precision as specified by the standard deviation and assume certain values for α, β and known π_P are given below:

$$n_0(\hat{\alpha}(CS)) = \frac{(\pi_P - \alpha)(1-\alpha)\alpha(1-f-\beta+\beta f-\alpha+\alpha f+\alpha\beta)}{[\mathrm{sd}_0(\hat{\alpha})]^2 f(1-f)(1-\beta-\pi_P)} \tag{3.20}$$

$$n_0(\hat{\beta}(CS)) = \frac{(\alpha\beta+f-\beta f-\alpha f)\beta(1-\beta)(1-\beta-\pi_P)}{[\mathrm{sd}_0(\hat{\beta})]^2 f(1-f)(\pi_P-\alpha)} \tag{3.21}$$

$$n_0(\hat{\pi}_C(CS)) = \frac{(\alpha+\beta f-\alpha\beta-\alpha f)(1-\beta-\pi_P)}{[\mathrm{sd}_0(\hat{\pi}_C)]^2 f(1-f)(1-\beta-\alpha)^2} \tag{3.22}$$

where $f = \frac{n_P}{n}$, the proportion of the passed items in the total sample size.

If we are interested to find the sample sizes for a standard plan, we can use an indirect method by looking at the contour plots in Fig. 3.12 and find the ratio of the standard deviations for the assumed α and β values and the known π_P. Once we have the ratio, we can derive the corresponding standard error given by the CS plan, and finally get the total sample size, n_0.

For example, suppose that our goal is to estimate α with a precision of 0.0145 (i.e. $\mathrm{sd}_0(\hat{\alpha}) = 0.0145$), and we assume $\alpha = 0.03$, $\beta = 0.04$, and $\pi_P = 0.85$. The standard plan gives the best estimator for α and now we are interested in finding the minimum total sample size, n_0, for this method. We look at Figure 3.12 and find that, for the given values of α, β and π_P, the ratio of the standard deviations is 0.78. From this ratio we can compute $\mathrm{sd}_0(\hat{\alpha}(CS)) = 0.0186$ and the minimum sample size to achieve that, using Eq. (3.20), is $n_0(0.0186) = 1171$.

More details about the assessment of a BMS when there is only one measurement by the BMS and one measurement by the gold standard can be found in the paper "Assessing a Binary Measurement System" by Danila, Steiner, and MacKay (2008), published in the "Journal of Quality Technology".

3.3 Summary

In this chapter, we investigate two assessment procedures in the case where a gold-standard system is available, and we use two different sampling schemes, the standard and the conditional selection plans. We focus on a context in the manufacturing industry where parts coming from a high-volume process are routinely measured with the studied BMS, and large collections of previously passed and failed parts are readily available for the BMS assessment study. Also, we have baseline information about the pass rate, π_P, over a period of time in which the process and the properties of the BMS are stable.

The first assessment method involves measuring parts once with the gold standard and repeatedly with the BMS. The standard plan currently used in industrial practice involves random selection of parts from the population of manufactured parts. We look at the properties of the estimators of α, β, and π_C, under this sampling plan and we conclude that, for the range of parameters values commonly found in industry, i.e. small α and β, and large π_C, we need at least $n \times r = 1{,}000$ total number of measurements in order to get efficient estimators. We also note that the way the total number of measurements is allocated does not have an effect on the precision of $\hat{\alpha}$ and $\hat{\beta}$, whereas choosing a larger sample size n at the expense of the number of repeated measurements r gives better precision for the estimator of π_C.

Next, we investigate a sampling design that uses information available from the routine inspection of parts. With this plan, called the conditional selection plan, parts are randomly selected from collections of previously passed and failed parts. In many cases, a BMS can keep track of the number of passed parts during routine inspection of parts, and we propose using the baseline information about the pass rate in the BMS assessment study. We demonstrate that in the case where parts are repeatedly measured with the BMS, the CS plan augmented with baseline data gives uniformly better estimators than the SP. Also, we find that sampling heavily from the collection of previously failed parts gives estimators with better precision than other CS plans and the SP.

The second assessment method requires one measurement with the BMS and one with the gold standard. We look at both the standard plan and the conditional selection plan with $m \to \infty$ (i.e. π_P known). Contrary to the case where we repeatedly measure the sampled parts with the BMS, we conclude that the SP gives better precision that the CS plan with π_P known for the estimator of α, whereas $\hat{\beta}$ and $\hat{\pi}_C$ are better estimated with the CS plan. Also, we conclude that a CS plan with parts heavily sampled from the population of previously passed parts, e.g. $f = 0.8$, gives uniformly better estimators than other CS plans.

Chapter 4

New Methods for Assessing a Binary Measurement System with Constant Misclassification Rates – No Gold Standard Available

In Chapter 2, we mention that in both industrial and medical contexts, there are cases where the gold standard is too expensive, time consuming or invasive to be used on a routine basis. In other situations, a gold standard does not exist. There are two major approaches used for assessing the performance of a BMS when no gold standard is available. The first one involves the use of an imperfect measurement system with known performance characteristics, where an "anchored" model is used for parameter estimation (Boyles, 2001). In the second approach, sampled parts are either repeatedly measured by the BMS of interest or measured by several BMSs and a latent class (LC) analysis is used for estimation (Boyles, 2001; Van Wieringen and De Mast, 2008). In Chapter 4, we investigate the effect of different sampling plans on the estimation procedure for α, β and π_C, when there is no gold standard. In Section 4.1, we give a brief review of current methods used in industrial practice, in the case where the gold standard is not available and we repeatedly measure parts with the BMS. In Section 4.2, we investigate the standard plan (SP) where parts are randomly selected from the collection of manufactured parts. In Section 4.3, we look at the properties of the estimators given by a conditional selection (CS) plan, where parts are randomly selected from the collections of previously passed and failed parts, augmented with baseline data. Next, we compare the two sampling schemes with respect to their precision and make study design recommendations.

4.1 Current Selection Methods and Models

In Chapter 2, we introduce the latent class (LC) model for the specific case where the units are repeatedly measured by the studied BMS, a testing procedure that is commonly used in the industrial setting (Boyles, 2001; Van Wieringen and Van der Heuvel, 2005; Van Wieringen and De Mast, 2008).

Boyles (2001) proposes selecting a random sample of n parts from the population of parts, and measuring each part r times with the BMS, i.e. the standard plan (SP). Note that the SP can be used in cases where the BMS has not yet been used for regular inspection and we assume no prior information.

The results of the repeated measurements are summarized by the total number of passes for each part, $s_i = \sum_{j=1}^{r} y_{ij}, i = 1, \ldots, n$, which is a sufficient statistic for the distribution of $Y_{i1} \ldots Y_{ir}, i = 1, \ldots, n$. Assuming conditional independence and common misclassification probabilities for all parts, the conditional distribution of S_i, given the part is conforming or nonconforming, is:

$$S_i \mid (X_i = 1) \sim \text{Binomial}(r, 1-\beta) \quad \text{or} \quad S_i \mid (X_i = 0) \sim \text{Binomial}(r, \alpha)$$

Under the assumption of independent measurements on different units, the likelihood function for the SP is a mixture of two binomial distributions as follows:

$$L_{SP}(\alpha,\beta,\pi_C \mid s_i, i = 1\dots n) \propto \prod_{i=1}^{n} \left[(1-\beta)^{s_i} \beta^{r-s_i} \pi_C + \alpha^{s_i} (1-\alpha)^{r-s_i} (1-\pi_C) \right] \tag{4.1}$$

To make the parameters identifiable, there should be at least three repeated measurements, i.e. $r \geq 3$, and $1-\beta > \alpha$ (Van Wieringen and Van der Heuvel, 2005; Van Wieringen and De Mast, 2008). The assumption $1-\beta > \alpha$ is reasonable, since 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, we expect both α and β to be relatively small.

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 and β. For sample size calculations during the planning stage, Boyles (2001) 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.

Although the proposed model “treats the units in the study as a random sample from some population” (Boyles, 2001, pp. 223), the author 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 take this point further, but the idea of a conditional selection is later pursued by Van Wieringen and De Mast (2008).

Van Wieringen and De Mast (2008) use the latent class model in a similar context, where they do not assume parts are randomly selected from the population of manufactured parts. Therefore, they define the likelihood function in terms of the two misclassification probabilities, α and β, and another parameter corresponding to the proportion of conforming parts in the sample, π_S, that does not represent the conforming rate of the production process (unless the sample is selected at random from the population of all parts). The likelihood function has the same form as in Eq.(4.1), with the population-based conforming rate π_C replaced by π_S. π_S depends on the chosen sampling plan and when, for example, SP is used, $\pi_S = \pi_C$. The authors also suggest and compare two estimation methods, maximum likelihood using the EM algorithm and an application of the method of moments.

Based on simulation results and comparison of the two estimation methods, Van Wieringen and De Mast (2008) give planning recommendations, including selection method and sample size. In particular, they suggest trying to obtain a sample that has an equal number of conforming and nonconforming parts. To achieve such a balanced sample, in their example, they suggest selecting two random samples of parts from the population of previously passed and rejected parts. We call this plan the conditional selection (CS) plan, as it involves sampling parts conditionally on a previous (baseline) measurement. The authors propose selecting equal numbers of previously passed and rejected parts.

There are several issues related to the model proposed by Van Wieringen and De Mast (2008). First, the conforming rate representing the production process is not estimable, unless we understand the sampling mechanism used to select the sample. The model is useful in the case where there is no information related to the selection method of the sample (for example, parts are just "grabbed" and then included in the assessment study). In that case, we can estimate the parameters characterizing the BMS, i.e. α and β, whereas π_S is considered a nuisance parameter that does not characterize the manufacturing process. Therefore, π_S is not a population-based parameter, as it does not characterize the population from which the parts are randomly sampled. In the case where the parts are selected conditionally on the baseline measurement, the probability of having a conforming part is not the same for previously passed and rejected parts. Therefore, we suggest a different LC model based on a conditional likelihood which includes α, β, and π_C, and also incorporates the baseline measurement of parts.

Also, Van Wieringen and De Mast (2008) recommend obtaining a balanced sample in terms of conforming and nonconforming parts. However, sampling equal numbers of previously passed and rejected parts does not necessarily yield a sample with an equal (expected) number of conforming and nonconforming parts, as seen in Table 3.1. We investigate this issue later in this chapter, by looking at CS plans with different proportions, f, of previously passed parts in the sample. We look at how the precision of the estimators change with the proportion of expected number of conforming parts, which is a function of f. We compare these plans to the SP and then with each other. We mentioned before that the SP is recommended when we need to assess the BMS before the system is used for regular inspection. However, in many practical situations, SP is used in cases where there is previous (baseline) information about the population of parts from which the sample is selected. That is, we usually know the proportion of passed parts from a large number of previously measured parts, as the BMS can keep track of that information. Our main goal is to show the value of using a CS plan with baseline information, and to find the value of f that gives the best plan in terms of the precision of the estimators.

4.2 Standard Plan – Accuracy and Precision of the Parameter Estimators

In this section, we further investigate the LC model in the SP case, with focus on accuracy and precision of the estimators, and the effect of the sample size and the number of repeated measurements on these characteristics.

As in the gold-standard case, we focus on assessing a BMS with reasonable performance characteristics, where we also assume the sampled parts come from a high-quality production process. In our investigations, we consider parameters values as in Chapter 3, where $0.02 \leq \alpha, \beta \leq 0.1$, and $0.85 \leq \pi_C \leq 0.94$.

To assess the bias and efficiency of the parameter estimators, we first simulate data for different n and r, and then obtain the ML estimates by maximizing the likelihood function (4.1), using the Nelder-Mead optimization algorithm (Nelder and Mead, 1965) in R (R Development Core Team, 2010). Other authors (Torrance-Rynard and Walter, 1997; Fujisawa and Izumi, 2000) also used direct optimization techniques rather than the EM algorithm. 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.

In our simulations, we run 500 repeats for each combination of parameter values from the grid in Figure 3.1. For each repeat, we first generate n realizations for the random variable X, the true state, which is distributed Binomial$(1, \pi_C)$. Then, for each conforming part, i.e. $x = 1$, we generate the number of passes, s, a realization of $S \mid (X = 1)$, which is distributed Binomial$(r, 1 - \beta)$. Similarly, for each nonconforming part, we generate the total number of passes s from Binomial(r, α). The estimation of the parameters is based on the number of passes for each part, s, as here we consider the case where the gold standard is not available.

When the sample size is small and the conforming rate π_C is close to 1, there might be no nonconforming parts in the sample and the optimization algorithm might have difficulty finding the global maximum. We add an additional step to the estimation procedure. That is, for each run, after the ML estimates are found using the Nelder-Mead algorithm, we compare the maximum log-likelihood value found by the algorithm with the maximum log-likelihood value at $\pi_C = 1$. The log-likelihood function with $\pi_C = 1$ contains only the parameter β and if the maximum value is larger than the one with $\pi_C < 1$, we skip to the next run. Another way to deal with such a case would be to consider $\hat{\pi}_C = 1$, $\hat{\beta} = (nr - \sum_{i=1}^{n} s_i)/(nr)$, and to admit there is not enough data to estimate α. Anyway, for our simulations, we decided to use the former solution. We first consider cases where the sample size is small ($n = 200$) and the number of repeated measurements is close to the identifiability condition ($r = 5$). We recall that these are the minimum values for n and r for which the SP gives useful standard errors for all estimates in the gold-standard case in Chapter 3. We run 500-repeat simulations at each combination of parameter values in the 36-value grid with $\pi_P = 0.85$, and then obtain the sample errors and standard deviations, as approximations for the biases and standard deviations of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\pi}_C$. We note that for these parameters values and sample size, there were no instances where the maximum log-likelihood as given by the optimization algorithm was smaller than the one with $\pi_C = 1$, which was checked with the additional step in the algorithm. However, as we will see in Section 4.3.3, this step is very useful in cases where π_C is close to 1, e.g. $\pi_C = 0.98$, and we look at small sample sizes.

We also derive asymptotic standard errors based on the Fisher (expected) information matrix based on the likelihood function (4.1). The second derivatives of the log-likelihood

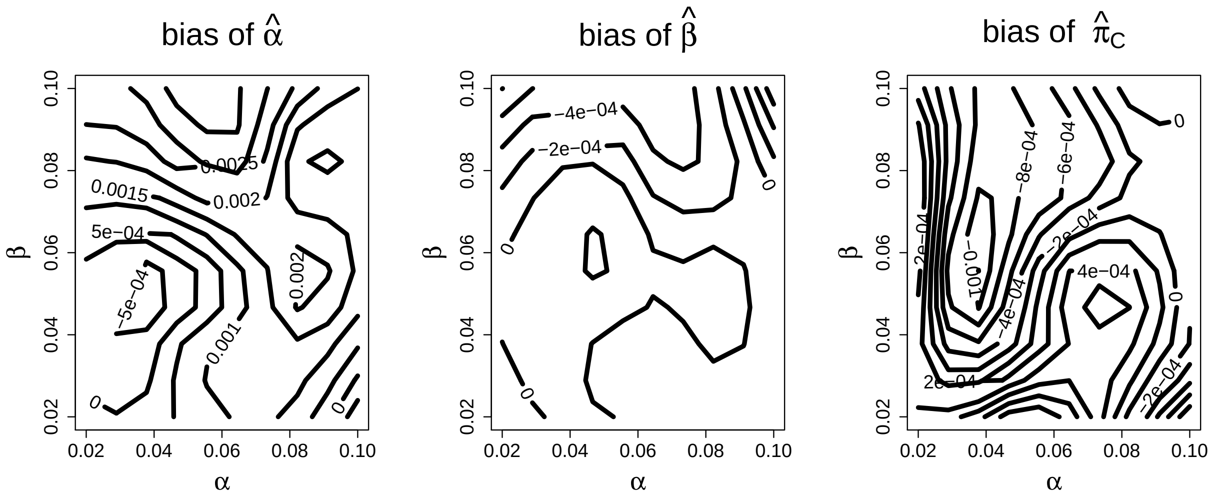

Figure 4.1 Contour Plots of Smoothed Biases – Standard Plan with $n = 200$ and $r = 5$, when $\pi_P = 0.85$

function are obtained using Maple software (Maplesoft, 2009) and they are too long to include here. The actual functions representing the elements of the information matrix are written in R (R Development Core Team, 2010). For any given parameter values, we obtain each element of the expected information matrix by summing the product of minus the value of the corresponding second derivative of the log-likelihood and the corresponding probability that a part passes the BMS inspection, over all possible values of s_i, i.e. $0,..,r$.

Next, we fit separate local polynomial regression (loess) smoothing models, (Cleveland et al., 1992) to biases and standard errors, and then get predictions over a 100-value grid. Figure 4.1 shows the biases of the estimators, and we note that all of them are virtually unbiased. We also look at the simulation-based and asymptotic standard errors and conclude they are close for all estimators, for most of the grid values (results not shown here). When α is small and π_C large, the simulation-based standard errors of $\hat{\alpha}$ are larger than the asymptotic ones (the ratio of the two standard errors is not larger than 1.3). Figure 4.2 shows the simulation-based standard errors and we note that the estimators of β and π_C have good precision for all values considered here, whereas $\hat{\alpha}$ has good precision for most of these values, except for small α and large π_C. In these

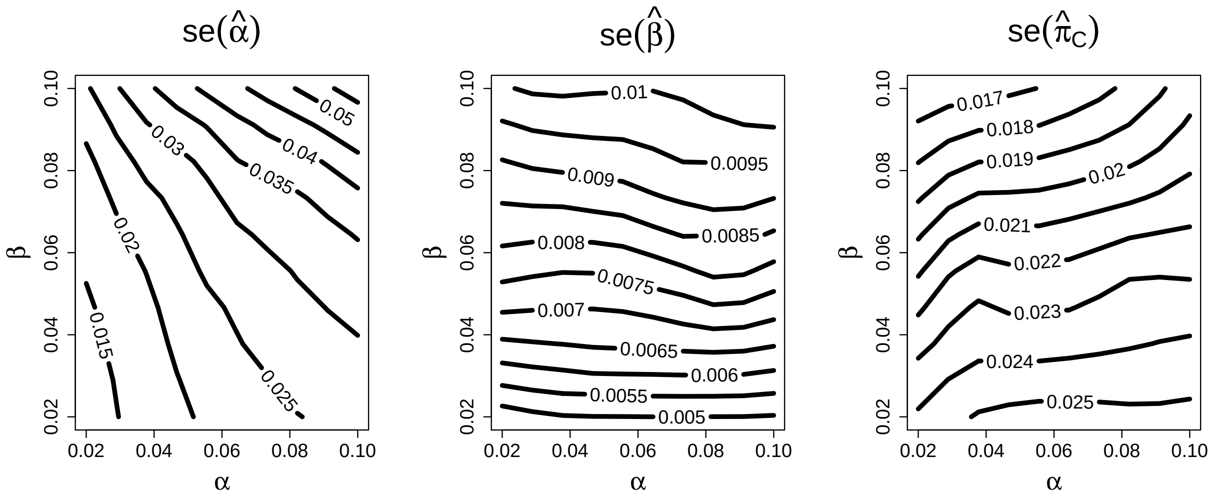

Figure 4.2 Contour Plots of Simulation-based Standard Errors – Standard Plan with $n = 200$ and $r = 5$, when $\pi_P = 0.85$

cases, the standard errors have the same size as the parameter α. We increase the number of repeated measurements to $r = 10$ and look again at the precision of the estimators. Our goal is to find the minimum sample size and number of repeated measurements so that all estimators have reasonable precision. In Figure 4.3, we see that the precision of the estimators, including the one for $\hat{\alpha}$, are reasonably good, with standard errors small enough to be useful, for the whole grid of parameters values.

Another option for finding the optimal combination of n and r, so that we achieve good precision for all estimators, is to increase the sample size from $n = 200$ to, for example, $n = 400$, and use the same number of repeated measurements $r = 5$. This way, we get an idea about how $n \times r = 2{,}000$ total measurements can be allocated so that we get a better efficiency for the parameter estimators. When we compare the standard errors in Figure 4.3 to the ones in Figure 4.4 we note that, for 2,000 total number of measurements, the plan with $r = 10$ gives slightly better estimators for α and β, whereas for π_C, the plan with $n = 400$ gives a more precise estimator. When we increase the sample size, we expect to get a more precise estimator for π_C, as $\hat{\pi}_C$ is mostly based on the (expected) number of conforming parts in the sample.

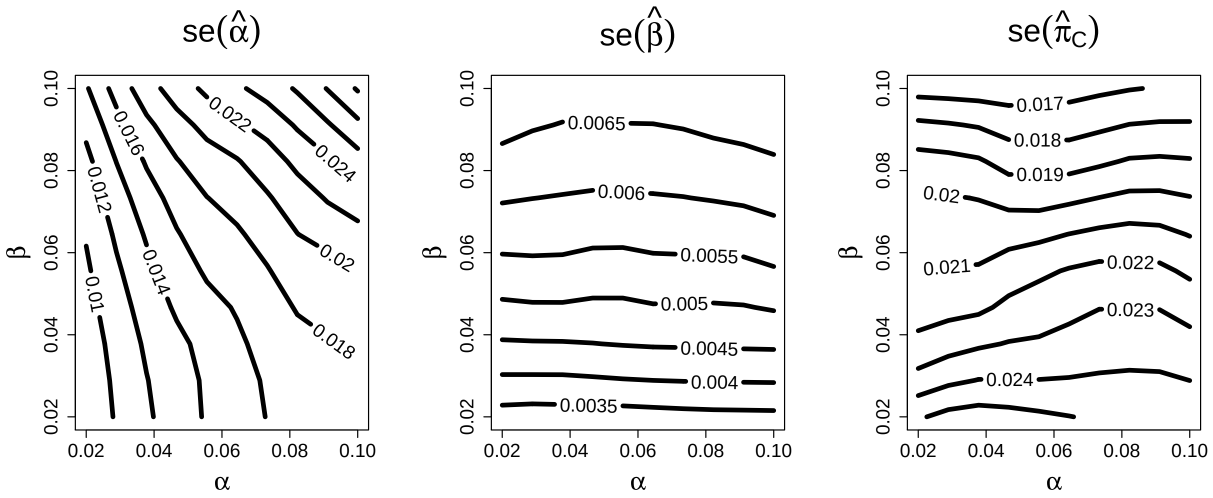

Figure 4.3 Contour Plots of Smoothed Simulation-based Standard Errors – Standard Plan with $n = 200$ and $r = 10$, when $\pi_P = 0.85$

We recall that in the gold-standard case in Chapter 3, the variances of $\hat{\alpha}$ and $\hat{\beta}$ depend on the values of the corresponding parameter, α or β, the value of π_C, and the total number of measurements $n \times r$, as given by Eq. (3.5). Therefore, for $\hat{\alpha}$ and $\hat{\beta}$, we get the same precision for the different allocations of $n \times r$, as shown in Figures 3.2 and 3.8, in Chapter 3, where $n = 200$, $r = 5$, and $n = 1{,}000$ and $r = 1$, respectively. Now, we want to see if these results also apply to the no gold-standard case. We plot the asymptotic standard errors for $\hat{\alpha}$ and $\hat{\beta}$, for eight combination of parameters values and $n \times r = 2{,}000$ total number of measurements. Here, we considered the minimum and maximum values for α, β, and π_C, from the grid in Figure 3.1, and three allocations of n and r. Figure 4.5 shows the curves of the standard errors for $\hat{\alpha}$ and $\hat{\beta}$, for all eight combinations of parameters values. For the standard error of $\hat{\alpha}$ (left panel), we note that the curves group by the values of α and π_C. This result is consistent with the variance expression in the gold-standard case. For small values of α and π_C, the curves of the standard errors are almost horizontal, which suggests that, in this case, the allocation of $n \times r$ does not effect the precision of $\hat{\alpha}$. However, for large values of α, we note a big drop in the value of the standard error when r increases from 5 to 10 (31% when $\alpha = 0.1$, $\beta = 0.1$, and $\pi_C = 0.95$), and a slight drop from $r = 10$ to $r = 20$. In the right panel of Figure 4.5, we note that the precision of $\hat{\beta}$ does not change with the

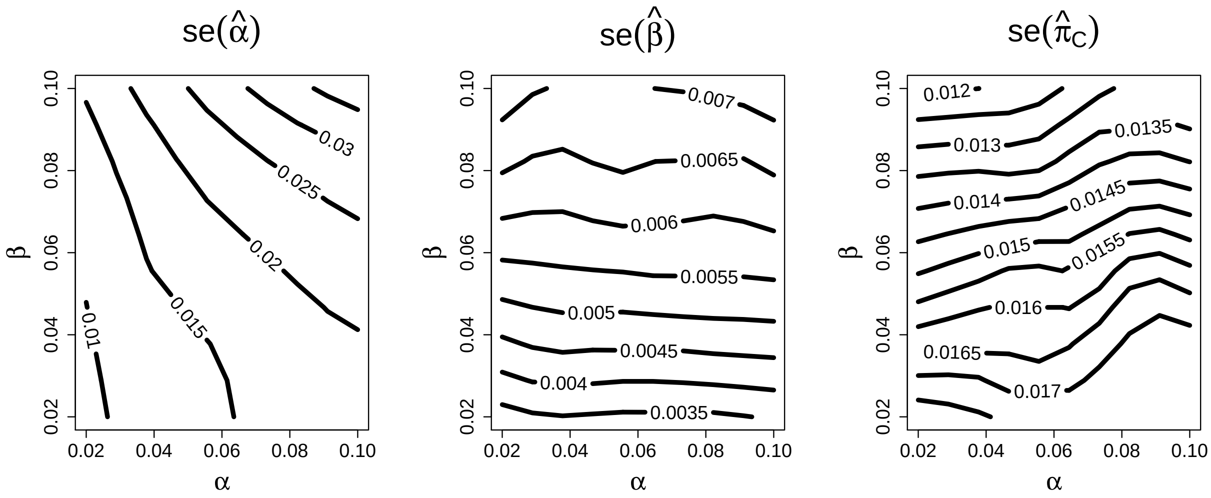

Figure 4.4 Contour Plots of Smoothed Simulation-based Standard Errors – Standard Plan with $n = 400$ and $r = 5$, when $\pi_P = 0.85$

allocation of n and r, for small values of β. For $\beta = 0.1$, $\alpha = 0.1$, and $\pi_C = 0.85$, the standard error of $\hat{\beta}$ decreases by 6%, when r goes from 5 to 10. In practice, the cost of sampling and measuring a new part is usually different than the cost of re-measuring an already selected part. Therefore, in some cases we might prefer increasing the number of measurements per part, whereas in other cases it might be less expensive to select a larger number of parts and re-measure them fewer times. However, as seen in Figure 4.5, if we increase the number of repeated measurements r at the expense of the sample size n, we get similar or better precision of the estimators of the consumer's and producer's risks. Therefore, if we focus on large values of r, for the range of parameters values $0.02 \leq \alpha, \beta \leq 0.1$ and $0.86 \leq \pi_C \leq 0.94$, we suggest a minimum sample size of $n = 200$ for $r = 10$ repeated measurements by the BMS.

Next, we look at the pattern of variation for the standard errors of the estimates in Figure 4.3. We note that for $\hat{\alpha}$, the standard error mostly increases with the values of α, for $\hat{\beta}$ with the values of β, whereas for π_C it mostly decreases with the values of β. This pattern can be explained by the idea of a "binomial model" approximation – Eq. (3.6) and (3.7) – for the variances of the estimators, proposed in Chapter 3. In the case where a gold-standard system is available, the

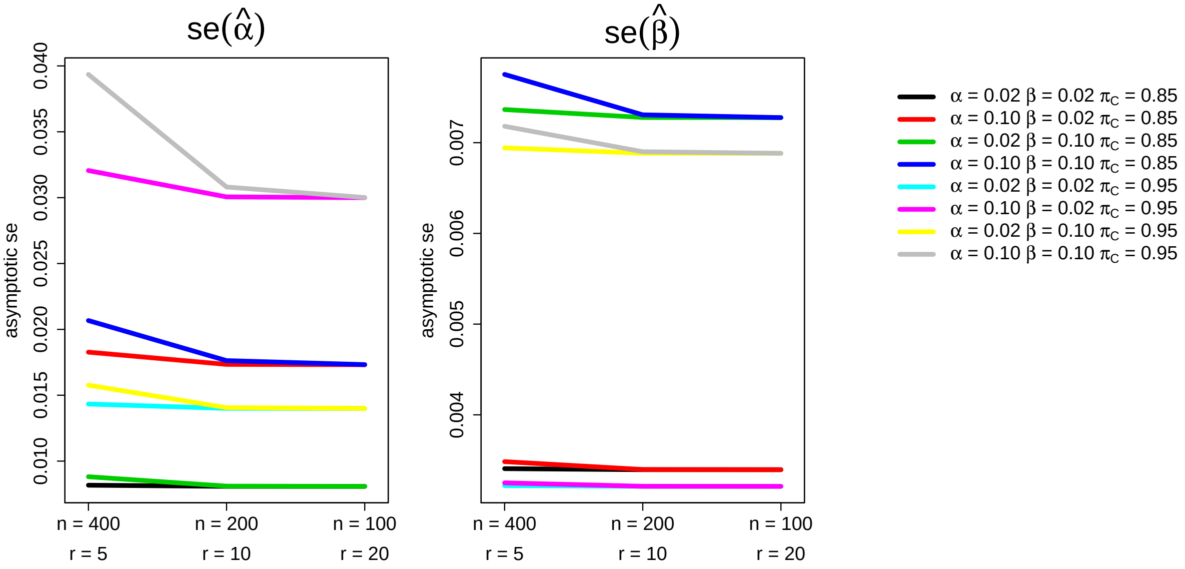

Figure 4.5 Plots of Asymptotic Standard Errors – Standard Plan with $n \times r = 2{,}000$ total measurements

assessment study involves selecting a sample of parts, measuring each part once with the gold standard, and then repeatedly measuring it with the BMS. Therefore, we can separate parts into conforming and nonconforming based on the gold-standard classification. The approximate variances of the estimators are given by Equations (3.8) – (3.9), and we note that for the estimator of α, the variance depends on the number of repeated measurements, the expected number of nonconforming parts in the sample, $E(N_{\bar{C}})$, and the true value of α; the variance of $\hat{\beta}$ depends on r, expected number of conforming parts in the sample, $E(N_C)$, and β. Finally, the variance of $\hat{\pi}_C$ varies with the sample size, n, and the value of π_C. Now, in the "no gold standard" case, the parts cannot actually be separated into conforming and nonconforming, as the true state is not known. Nevertheless, these approximations are still useful in explaining the pattern of variation for the precision of the estimators within a grid of values.

The "binomial" models (3.6)–(3.7) for $\hat{\alpha}$ and $\hat{\beta}$ can be extended to other sampling plans, as they only require that the parts are initially measured by the gold standard, and then separated into conforming and nonconforming. For a CS plan with the same number of repeated measurements and sample size as in an SP, the only quantity that changes is the expected number of (non)conforming parts in a sample. In the next section, we investigate how the change in the expected number of conforming parts affects the precision of the estimators for different CS plans, compared to the ones given by the SP.

4.3 Conditional Selection Plan

The conditional selection (CS) plan involves independently selecting two random samples of parts from the populations of the previously passed and rejected parts. With a CS plan, we can choose the proportion of previously passed parts in the sample, denoted by f.

The likelihood function for the CS plan is given by two mixtures of Binomial distributions, one for the previously passed and one for the previously rejected parts, as follows:

$$\begin{aligned} L_{CS}(\alpha,\beta,\pi_C|s_i, i=1\dots,n) \quad \propto \quad & \prod_{i=1}^{n_{\bar{P}}} \frac{\left(1-\beta\right)^{s_i}\beta^{r+1-s_i}\pi_C+\alpha^{s_i}\left(1-\alpha\right)^{r+1-s_i}\left(1-\pi_C\right)}{1-\pi_P} \times \\ & \prod_{i=1}^{n_P} \frac{\left(1-\beta\right)^{s_i+1}\beta^{r-s_i}\pi_C+\alpha^{s_i+1}\left(1-\alpha\right)^{r-s_i}\left(1-\pi_C\right)}{\pi_P} \end{aligned} \tag{4.2}$$

In Equation (4.2), $n_P = fn$ and $n_{\bar{P}} = (1-f)n$ represent the pre-determined number of previously passed and rejected parts in the sample. In practice, when a CS plan is possible, there is also available information about the pass rate, i.e. baseline data. Therefore, for practical reasons, we should consider CS plans where the assessment study data is augmented with the baseline data. As in Chapter 3, to separate the contributions of the CS and baseline data, we first consider the CS plan without baseline data, and compare the precision and bias of the estimators given by a CS plan with the ones given by the SP.

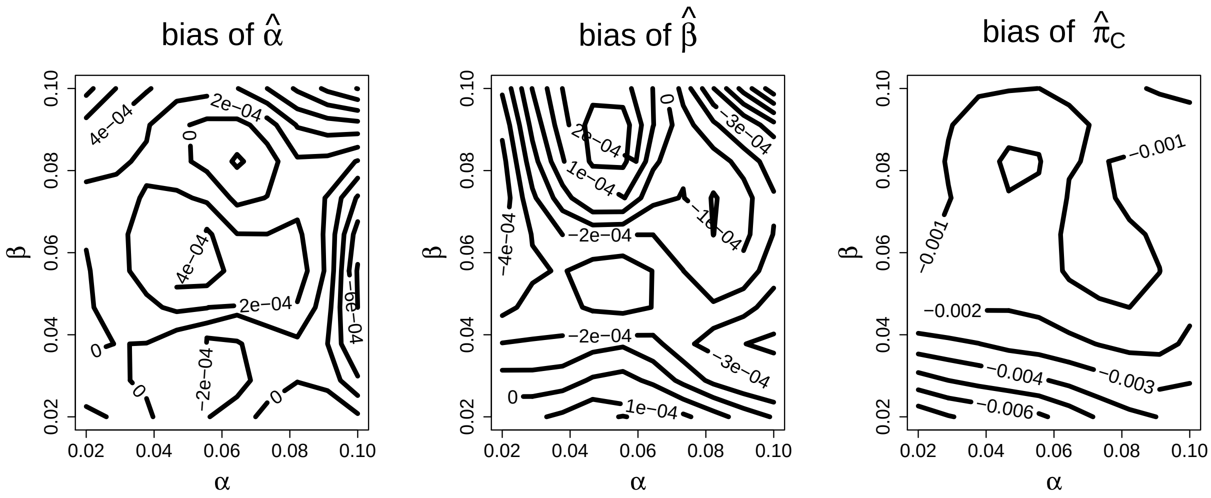

Figure 4.6 Contour Plots of Smoothed Biases – Conditional Selection Plan with $f = 0.5$, $m = 0$, $n = 200$ and $r = 5$, when $\pi_P = 0.85$

4.3.1 CS without Baseline Data

We start by looking at the biases of the simulation-based ML estimates, for $r = 10$, when parts are sampled in equal numbers from the population of previously passed and rejected parts. That is, in the design stage we choose $f = 0.5$. As in the SP case, we look at the biases of the estimators approximated by the sample errors from a simulation study. Figure 4.6 shows the smoothed (loess) biases for a 100-value grid with $\pi_P = 0.85$, and we note that all estimators are virtually unbiased. Next, we look at the comparison of the asymptotic standard deviations given by the SP to the ones given by CS with $f = 0.5$. Figure 4.7 shows the ratios of asymptotic standard deviations of the estimators given by the two plans, for a 100-value grid with $\pi_P = 0.85$. Note that these ratios do not depend on n. We note that the estimator of α given by the CS is more precise than the one from the SP, for the whole grid of values. For $\hat{\beta}$, we see the reversed situation, whereas the estimator of π_C is more precise for the CS, for most of the grid values. If we go back to the idea of a “binomial model” for the approximation of standard deviations of $\hat{\alpha}$ and $\hat{\beta}$, the results related to the precision of these estimators are not surprising. As we note in Table 3.1 in Chapter 3, the CS plans with $f = 0.5$ and $f = 0$ increase the expected proportion of

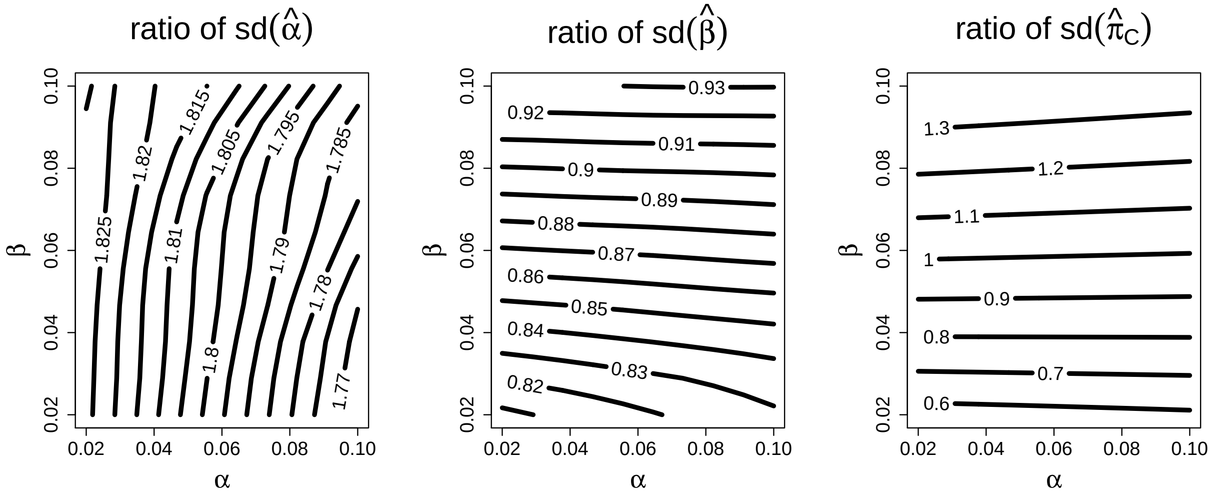

Figure 4.7 Contour Plots of Ratios of Asymptotic Standard Errors, $se(SP)/se(CS, f = 0, m = 0)$, when $\pi_P = 0.85$ and $r = 10$.

nonconforming parts in the sample, but decrease the expected proportion of conforming, over the whole grid of parameter values. Equations (3.8)–(3.9) show that the precision of $\hat{\alpha}$ increases with the expected number of nonconforming parts, whereas the precision of $\hat{\beta}$ decreases.

We conclude that there is a trade-off in terms of the precision of $\hat{\alpha}$ and $\hat{\beta}$, when we choose one selection method over the other. Now, this is the case when the CS plan is used without including the baseline information in the analysis. In the next section, we investigate the precision of the estimators given by the CS plan augmented by baseline data of various sizes, that is, we know the number of passed parts out of a certain (usually large) number of once measured parts, where this collection of previously inspected parts is a representative sample of the manufactured parts.

4.3.2 CS with Baseline Data

Suppose we have a baseline population of m parts, each measured once for inspection purposes. Also, we randomly select n_P parts from a large collection of previously passed, and $n_{\bar{P}}$ from the previously failed. Here we assume that the collections of passed and failed we sample from do not make up the m baseline measurements, so that the study and baseline measurements are independent. At the end of the assessment study, we have m parts measured once and n parts measured $r+1$ times. For high volume processes, m is typically large. For the m parts measured once only, the likelihood is:

$$L_b(\pi_P) \propto \pi_P^{m_P}(1-\pi_P)^{m-m_P} \tag{4.3}$$

where m_P is the number of passed parts in this group, and m is the size of the baseline data. As in Chapter 3, we note that $L_b(\pi_P)$ can be rewritten in terms of α, β, and π_C, using the constraint $\pi_P = (1-\beta)\pi_C + \alpha(1-\pi_C)$, though it is not possible to separately estimate α, β, and π_C using $L_b(\pi_P)$ alone. The overall likelihood for a CS plan with baseline data is proportional to: $L_b(\pi_P) \times L_{CS}(\alpha,\beta,\pi_C|s_1,\ldots,s_n)$, where L_{CS} is the likelihood in Equation (4.2), and s_i is the number of times part i passed in the r repeated measurements.

We expect the results regarding biases for the CS without baseline to also hold for the CS with baseline, as adding more information can only improve the estimation procedure. We check this assumption and conclude that, for CS with $f = 0.5$ and $f = 0$, sample size $n = 200$ and number of repeated measurements $r = 10$, all estimators are virtually unbiased and have reasonable precision (results not shown here). We also look at the case where $r = 5$ and $n = 200$, and find that the CS plans with $f = 0.5$ and $f = 0$, and $m = 1{,}000$ give estimators with good accuracy and precision for all parameters. Also, for these (and larger) values of n and r, the asymptotic standard deviations agree with the standard errors based on simulated data. Therefore, we can proceed with comparing the asymptotic standard deviations of the estimators given by the SP to the ones given by CS with baseline data and $f = 0.5$ and $f = 0$, respectively. We start with a small baseline sample $m = 1{,}000$, and compare the asymptotic standard deviations given by the SP with the ones given by a CS plan with $f = 0.5$ (Figure 4.8). When comparing Figure 4.7 with Figure 4.8, we note that adding the baseline data improves the precision given by the CS, for all estimators. The ratios of standard deviations for the estimator of β are now larger than 1, for the whole grid of parameter values. Note that the estimators of α and π_C given by the CS plan with $f = 0.5$ are almost twice as efficient as the ones given by the SP. Therefore, when we

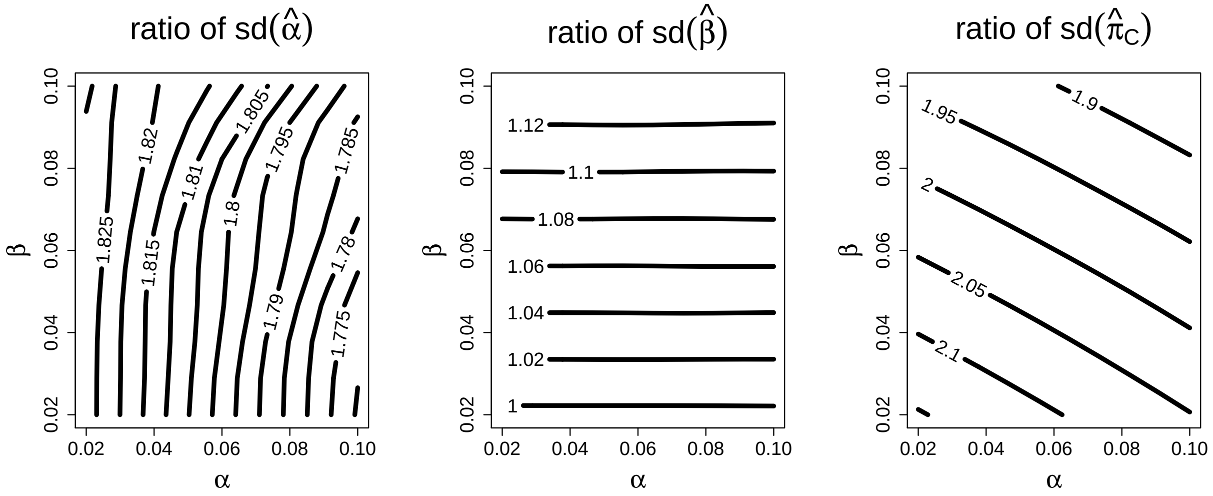

Figure 4.8 Contour Plots of Ratios of Asymptotic Standard Errors, $se(SP)/se(CS, f = 0.5)$, when $\pi_P = 0.85$, $n = 200$, $r = 10$ and $m = 1{,}000$.

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 4.9 and 4.10, we compare the precision of each estimator for CS plans with $f = 0.5$ and $f = 0$, for baseline sample sizes $m = 1{,}000$ and $m = 10{,}000$. These figures suggest that, as the size of the baseline data increases, a CS plan with $f = 0$ becomes uniformly more efficient than a CS plan with $f = 0.5$, especially when β is large. Also, for larger values of the pass rate ($\pi_P \geq 0.85$), 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 (results not shown here). We expect to get more information about α with conditional sampling and $f = 0$, since with this scheme we will likely select more nonconforming parts. The increased precision for the estimator of β is perhaps surprising. Here it is the baseline measurements that help. The baseline data provide an estimate of $\pi_P = (1-\beta)\pi_C + \alpha(1-\pi_C)$, a function of α, β, and π_C. Since we are considering situations where π_C is large and α is relatively small, π_P is strongly influenced by β and π_C. In this case, the additional information about β and π_C from the baseline data outweighs the lost information due to fewer conforming parts in the sample.

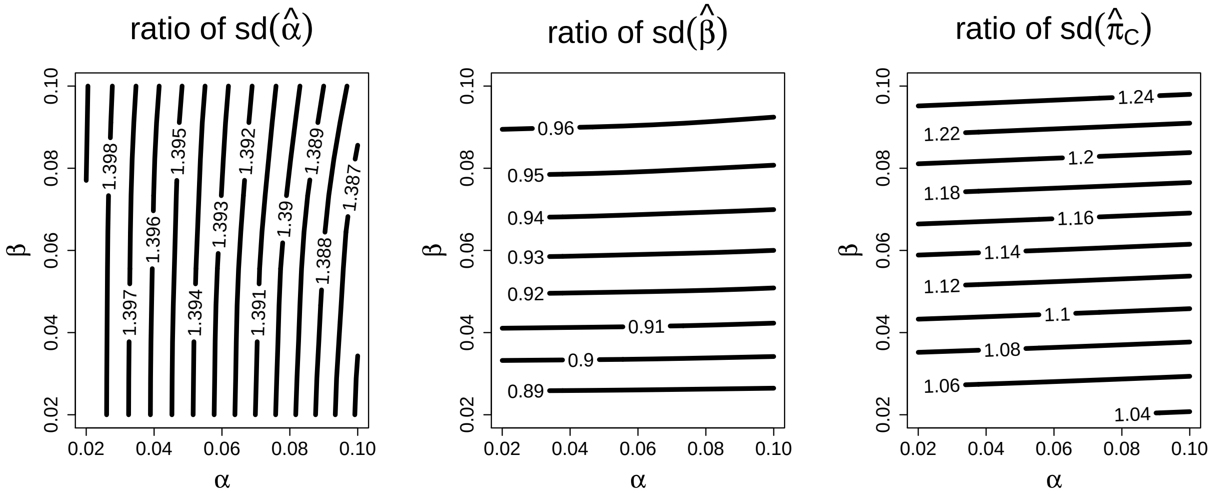

Figure 4.9 Contour Plots of Ratios of Asymptotic Standard Errors, $se(CS, f = 0.5)/se(CS, f = 0)$, when $\pi_P = 0.85$, $n = 200$, $r = 10$ and $m = 1{,}000$.

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 rejected parts. 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 4.11, 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. The estimators of α given by the CS with $f = 0$ are twice times more efficient than the ones given by the SP. The estimator of β is slightly more precise for the CS plan, for smaller values of β (and π_C), and is one and a half times more precise, for larger values of β. We see the biggest gain in precision for the estimator of π_C, as the ratio of standard derivation varies from 3.5 to 7 over the grid of values with $\pi_P = 0.85$. We see similar large gains for larger values of π_P (results not shown here).

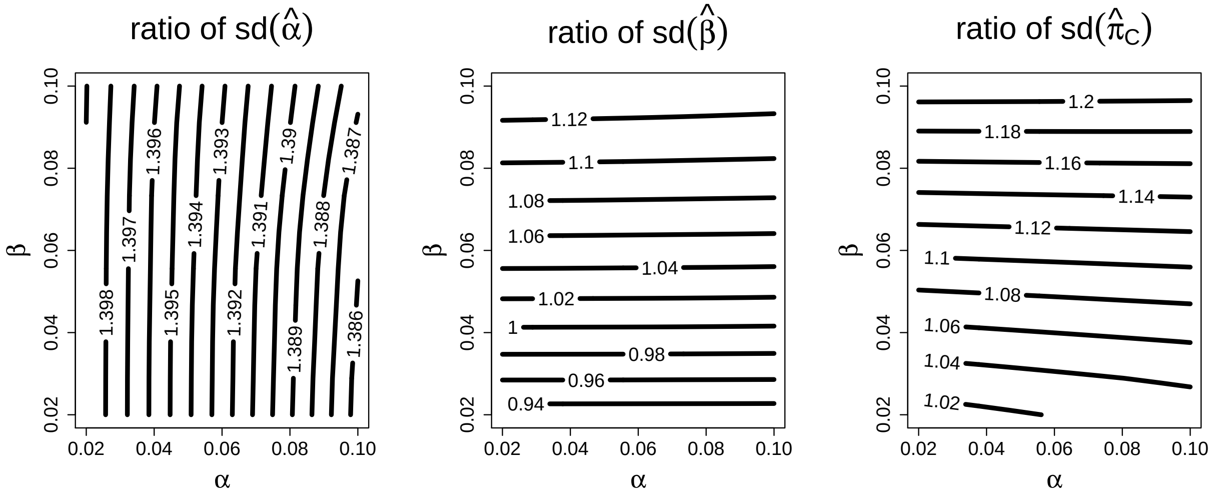

Figure 4.10 Contour Plots of Ratios of Asymptotic Standard Errors, $se(CS, f = 0.5)/se(CS, f = 0)$, when $\pi_P = 0.85$, $n = 200$, $r = 10$ and $m = 10{,}000$.

4.3.3 Special Cases – large conforming rate and small n and r

We also look at cases where the conforming rate π_C is very close to 1, e.g. $\pi_C = 0.98$, and we investigate the properties of the estimators when we design a study with a small total number of repeated measurements $n \times r$, using the standard and conditional selection plans. For these cases, we run several simulation studies, as in Sections 4.2 and 4.3, and obtain the biases and standard errors of the estimates. We first look at the case where $\alpha = 0.1$, $\beta = 0.09$, and $\pi_C = 0.988$, and we randomly select $n = 100$ parts, i.e. we use the SP, and measure them $r = 4$ times with the BMS. We find the estimators of α and π_C highly biased, with many cases where the estimates of α are very large (close to 1). This can be explained by the fact that, when π_C is close to 1, the expected proportion of nonconforming parts in the sample is small, and when, in addition, the sample size is small, some samples may contain very few or no nonconforming parts. For example, for the case considered above, $E_{SP}(N_{\bar{C}}/n) = 0.0123$, and when $n = 100$, we expect to have one nonconforming part in the sample. We further investigate samples that in our example give unreasonably large values for $\hat{\alpha}$, i.e. $\hat{\alpha} > 0.5$, and small values for $\hat{\pi}_C$, i.e. $\hat{\pi}_C < 0.5$. One such sample in the simulation contains no nonconforming parts and there are 5 parts out of 100 that

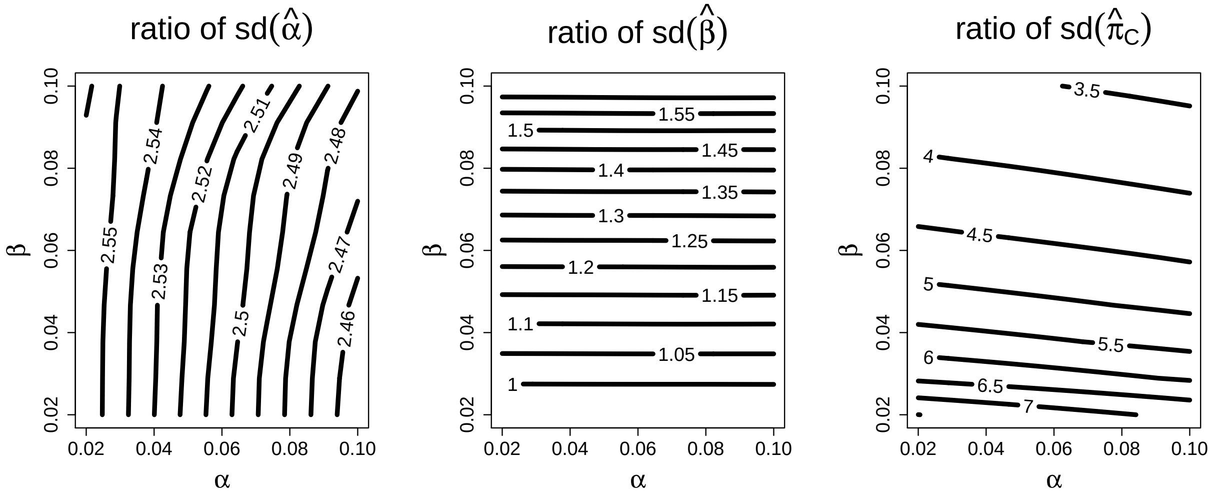

Figure 4.11 Contour Plots of Ratios of Asymptotic Standard Errors, $se(SP)/se(CS, f = 0)$, when $\pi_P = 0.85$, $n = 200$, $r = 10$, and $m = \infty$ (π_p known).

pass the inspection twice, 25 parts pass it three times, and 70 pass it four times. For this sample, $\hat{\alpha} = 0.89$, $\hat{\beta} = 0.99$, and $\hat{\pi}_c = 0.14$. After visually checking the profile likelihood functions, we conclude that the optimization algorithm does find the global maximum. When compared to the log-likelihood value at $\pi_c = 1$, the additional step described in Section 4.2, the maximum value found by the algorithm is larger. We encounter similar situations in 17% of the simulated samples, and some of these samples do include one nonconforming part. It seems that, when the number of nonconforming parts in the sample is smaller than 2 and the number of repeated measurements r is small, it is difficult for the LC model to distinguish between conforming and nonconforming parts. For these cases, we admit that there is not enough data to reliably estimate α and π_C.

Next, we increase the value of r to 10 and we note that increasing the number of repeated measurements substantially reduces the bias of the estimators, for the same number of selected parts. We compare the case discussed earlier, where $\pi_c = 0.988$, $\alpha = 0.1$, $\beta = 0.09$, $n = 100$, and $r = 4$, to the case where we have the same parameter values and sample size, but $r = 10$. The number of times when the log-likelihood value at $\pi_c = 1$ is larger than the maximum value found by the algorithm is similar for the two cases (around 25%). The difference lies in the number of times the estimates for α are unreasonably large (i.e. larger than 0.5). When $r = 10$, 3% of the simulated samples yield these results, compared to 17% in the $r = 4$ case. That is, when the number of repeated measurements is large, in most cases, the LC model is able to distinguish between conforming and nonconforming parts, even when the sample size is small, and thus, the number of nonconforming parts in the sample is small.

Now, as we can see in Table 3.1, for the same parameter values, the expected proportion of nonconforming parts in the sample is larger for the CS plans compared to the SP. Also, the CS plan with $f = 0$ gives a larger expected proportion of nonconforming parts than the CS plan with $f = 0.5$, over the range of parameter values considered here. Therefore, when we select parts conditionally on their baseline measurement and $f > 0.5$, the expected number of nonconforming parts is larger than in the SP, for the same sample size. Thus, with such a CS plan, we need a smaller number of parts than in the SP in order to get unbiased estimators. We check this using simulated data and, for the case considered above, we conclude that the biases for all CS estimators are negligible.

4.4 Conditional Selection with Baseline Data - Study Design

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 BMS performance does not change and the process is stable. To make sure that this is true, we recommend examining the stability of the baseline data using statistical process control techniques (Montgomery, 1996). Note that we only need 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 failed the initial inspection. Rejected parts are typically set aside in any case to be repaired or scrapped.

We choose n and r using an algorithm coded in the R environment (R Development Core Team, 2010) that provides feasible combinations that 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 $\hat{\alpha}$ and $\hat{\beta}$ derived from the expected information matrix using the likelihood of Equations (4.2) and (4.3). 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, β, and π_C, 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 expected number of nonconforming parts in the sample.

To find feasible values for n and r, the algorithm uses a simple search strategy. It starts with a minimum number of repeated measurements $r = 5$ and a minimum of parts $n = 10$ 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 select $f = 0$ and we have an additional $m = 10{,}000$ previously measured parts. We also assume that the true (unknown) parameter values are $\alpha = 0.02$, $\beta = 0.02$, and $\pi_C = 0.86$ ($\pi_P = 0.85$). Suppose also that the desired precision for the estimators of α and β are $se(\hat{\alpha}) = 0.005$ and $se(\hat{\beta}) = 0.005$. The corresponding sample size n, the number of repeated measurements r, the total number of measurements $r \times n$, the resulting asymptotic standard deviations as provided by the algorithm, and the expected number of nonconforming parts $E(N_{\bar{C}})$ in the sample are given in Table 4.1.

To choose the best combination of n and r, we can select the combination that results in the fewest total number of measurements $n \times r$, in this case $n = 111$ and $r = 8$, or some other combination that takes into account the relative costs of measuring and sampling a part. Note that, in Table 4.1, the plans with r between 6 and 8 all have roughly the same total number of measurements.

Table 4.1 Recommended Sample Sizes, $\alpha = 0.02$, $\beta = 0.02$, $\pi_P = 0.85$, $sd(\hat{\alpha}) = 0.005$, $sd(\hat{\beta}) = 0.005$, $m = 10{,}000$

n	r	$n \times r$	$sd(\hat{\alpha})$	$sd(\hat{\beta})$	$sd(\hat{\pi}_C)$	$E(N_{\bar{C}})$
179	5	895	0.0050	0.0039	0.0048	158
148	6	888	0.0050	0.0043	0.0051	131
127	7	889	0.0050	0.0046	0.0053	112
111	8	888	0.0050	0.0049	0.0055	98
103	9	927	0.0049	0.0050	0.0056	91
102	10	1020	0.0047	0.0050	0.0056	90
100	11	1100	0.0045	0.0050	0.0056	88
99	12	1188	0.0043	0.0050	0.0056	88
97	13	1261	0.0042	0.0050	0.0056	86
96	14	1344	0.0041	0.0050	0.0056	85
95	15	1425	0.0039	0.0050	0.0056	84

4.5 Summary

In this chapter, we investigate methods for assessing a BMS in the case where a gold-standard system is not available. First, we give a brief review of methods currently used in industrial practice. Next, we discuss the standard plan, where parts are randomly selected from the population of manufactured parts, and then each part is repeatedly measured with the BMS. A latent class model (LC) is used for parameter estimation. We look at the properties of the estimators from the LC model with the SP, by simulating data for a whole grid of parameter values, and different values of the sample size n and the number of repeated measurements r. We conclude that, when the BMS has good performance and the manufacturing process is high-quality, we need fairly large total number of measurements (e.g. $n \times r = 2{,}000$, for the range of parameters values considered here) in order to estimate the parameters with reasonable accuracy and precision. We demonstrate that the way we allocate n and r for the total number of measurements has an effect on the precision of $\hat{\alpha}$ and $\hat{\beta}$; that is, when using an SP design with a larger number of repeated measurements increases the efficiency of the estimators of α and β than in the case where the sample size is larger. This is different than in the case where a gold-standard system is available, where we get constant precision for $\hat{\alpha}$ and $\hat{\beta}$, for the same $n \times r$.

Next, we explore the conditional selection plan, where parts are randomly selected from the populations of previously passed and failed parts. We propose a new LC model for parameter estimation, which is based on the conditional probability of passing the inspection, given the initial (baseline) measurement of a part. More importantly, we also propose augmenting the study data obtained from repeatedly measuring the sample parts, with baseline data that are readily available when the BMS has been in used for routine inspection. We demonstrate that the CS with $f = 0$, i.e. we sample only from the previously failed parts, supplemented with baseline measurements provides more precise estimators for all model parameters, when compared to other CS plans and the SP. For one choice of n and r, using CS with $f = 0$ gives the largest possible expected proportion of nonconforming parts in the sample. That is, we expect to have the most information about α. Also, by including the baseline measurements in the estimation we get an estimate of the pass rate π_P, and therefore additional information about β and π_C.

We also give planning recommendations, where we suggest using a CS plan with $f = 0$ and as much baseline data as possible. We provide an algorithm for sample size determination, where the input is the desired precision for the estimators of α and β, and the output represents different combinations of n and r so that we achieve the desired precision. We include one example in this chapter and provide the R-code for the algorithm at www.bisrg.uwaterloo.ca.

The main results presented in this chapter are also included in the paper “Assessing a Binary Measurement System in Current Use” by Danila, Steiner, and MacKay (2010), published in the Journal of Quality Technology.

Chapter 5

Random-Effects Model – Gold-Standard System Available

5.1 Introduction

In Chapters 3 and 4, we discuss statistical models that assume conditional independence of repeated measurements given the true state of the part, and constant misclassification probabilities within conforming/nonconforming parts. These assumptions have been widely criticized in both medical (Hui and Walter, 1980; Walter and Irwig, 1988; Pepe, 2003; Torrance-Rynard and Walter, 1997; Qu et al., 1996; Fujisawa and Izumi, 2000) and industrial (Van Wieringen and De Mast, 2008; De Mast et al., 2011) contexts.

As mentioned before, our main goal is to estimate the misclassification errors of a measurement system in the context where the "true state" of a part denoted X is binary, and the output of the system classification is binary, that is, we use a "pass/fail" inspection. In most cases, the binary variable X, also called the measurand (De Mast et al., 2011), represents the presence or absence of a certain characteristic, such as the presence of a certain infection in a patient's organism or the fact that a person suffered a myocardial infarction (Rindskopf and Rindskopf, 1986), in the medical context, or the presence of a scratch on the surface of a part from an injection molding process (Van Wieringen and De Mast, 2008), in the industrial field. The measurement system (medical test) is supposed to detect the presence/absence of this characteristic and classify the unit as pass/fail (positive/negative). In this context, there are two distinct cases that can often occur in both diagnostic testing and manufacturing industry. In the first case, the measurement system detects the presence or absence of the characteristic of interest X in units with the same probability, within the sub-populations of conforming/nonconforming (positive/negative). That is, there is no characteristic other than the true state X that influences the chance the BMS correctly classifies a unit. This is a strong assumption that, while a reasonable approximation in some cases, will not hold generally. In the second case, which is more prevalent in both industrial and medical contexts, there exists another variable (possibly more than one) that influences the chance that, for example, a nonconforming part is rejected or a diseased subject tests positive. One example is the inspection of parts for surface cracks using the liquid-penetrant method (Olin and Meeker, 1996). If the crack does exists, its detectability will depend on the size of the crack. Similarly, in the medical context, in the case of breast cancer screening tests, it is likely that some breast cancer lesions are more easily detected by the testing procedure than others (Shen et al., 2001). Also, in the case where subjects are tested for a certain parasitic disease, in a severely diseased case, there is a larger concentration of parasites, making it easier to detect (Dendukuri and Joseph, 2001). In all these cases, the measurand X is still a binary variable, but the probability of detecting the true state depends now on another characteristic that varies within the nonconforming parts (diseased subjects).

Therefore, in many cases, it may be unreasonable to assume that the misclassification rates α and β are the same for all conforming/nonconforming parts. Some parts may be harder to correctly classify than others. Suppose that there is another undetermined characteristic of the part denoted Z that affects the misclassification rates so that $\Pr(Y = 0|X = x, Z = z)$ is not equal to $\Pr(Y = 0|X = x)$. That is, the probability of failing the inspection depends on the value of z, as well as the true conforming/nonconforming state, x. If we assume that repeated measurements on a single part are independent, given $X = x$ and $Z = z$, then it is easy to prove that the repeated measurements on a part, given $X = x$, are now dependent, contrary to the basic assumptions in Chapters 3 and 4. For independence to hold, we require that $\Pr(Y = 0|X = x, Z = z)$ does not depend on z for any latent variable. That is, given $X = x$, no other characteristic of the parts affects the properties of the BMS.

There are different approaches to address the issue of varying values of α and β over different units. Fujisawa and Izumi (2000) use a random-effects model where they specify a value for both α and β for each unit, and assume that their joint distribution is Dirichlet. In our opinion, assuming that each unit has a value associated with both α and β is not reasonable, as a unit is either conforming or nonconforming. Also, with the Dirichlet model, they assume a common variability parameter for the two misclassification rates. This assumption is difficult to check and it is likely not reasonable in many applications.

Qu et al. (1996) construct a random-effects model to specify the joint distribution of $Y_{i1},\dots,Y_{ir}$, for part $i = 1,\dots,n$. At the first level, given $X = x$ and a latent variable $Z \sim N(0,1)$, they assume $Y_{i1},\dots,Y_{ir}$ are (conditionally) independent with $\Pr(Y_{ij}|X = x, Z = z) = \Phi(a_{ix}+b_{ix}z)$, where Φ is the cumulative distribution function of a standard normal variable. Dendukuri and Joseph (2001) use a fully Bayesian extension of the random-effects model of Qu et al. (1996). All these models are proposed in the context where a gold-standard system is not available.

In this thesis, we adopt a random-effects model to relax the assumption that α and β are constant for nonconforming and conforming parts, respectively. That is, we suppose for any randomly selected nonconforming part, the consumer's risk α has probability density function $f(\alpha;\theta_\alpha)$, $0 < \alpha < 1$, and, given $X = 0$ and α, repeated measurements on a part are independent so that:

$$\Pr(Y_{i1} = y_{i1},\dots,Y_{ir} = y_{ir} \mid X = 0,\alpha) = \alpha^{s_i}(1-\alpha)^{r-s_i}$$

Similarly, for any conforming part, we assume producer's risk β has density $f(\beta;\theta_\beta)$, $0 < \beta < 1$, and, given $X = 1$ and β, repeated measurements on part i are independent so that:

$$\Pr(Y_{i1} = y_{i1},\dots,Y_{ir} = y_{ir} \mid X = 1,\beta) = (1-\beta)^{s_i}\beta^{r-s_i}$$

The random-effects model explicitly allows for variation in the producer's and consumer's risks within the set of conforming and nonconforming parts. In a sense, the random variables for α and β model the effects of all latent variables on the properties of the BMS, given the value of X.

As with the model with constant misclassification rates in Chapters 3 and 4, which from now on we call the "fixed-effects" model, we can question the conditional independence assumption for the repeated measurements in the random-effects model. Here, the conditioning has moved one level deeper in the hierarchy. It is certainly possible to increase the number of levels but eventually we see no option but to assume conditional independence of the repeated measurements. Each additional level adds more unknown parameters to the model and we see little value in further increasing the complexity.

In Chapters 5 and 6, we explore the use of a Beta-binomial random-effects model. In Chapter 5, we consider the case of an available gold-standard system, while in Chapter 6, we look at the situation where a gold standard is not available. As we discuss later, the model formulation we propose, where α and β follow Beta distributions, allows more flexibility than the Dirichlet model proposed by Fujisawa and Izumi (2000).

In this chapter, we first investigate the properties of the estimators from the random-effects model in the case where parts are randomly selected from the population of manufactured parts (i.e. we use the standard sampling plan). Next, we look at the properties of the fixed-effects model estimators (3.2-3.4) when fitted to data generated from a random-effects model. We also investigate the case where data are generated from a fixed-effects model as in Chapter 3, and we fit a random-effects model. Next, we look at the properties of the estimators from the random-effects model, when parts are selected from the population of previously rejected parts, i.e. we use a conditional selection (CS) with $f = 0$, augmented by m baseline measurements. We demonstrate the advantage of using the CS plan with $f = 0$ and baseline data, in terms of accuracy and precision of the estimators. Also, we show that this design requires a smaller total number of measurements $n \times r$ in order to get useful estimates for the random-effects model parameters. In the last section, we investigate the effect of the design parameters n, r, and m on the precision of the estimators from a random-effects model when we use CS with baseline data.

5.2 Random-Effects Model – Standard Plan

5.2.1 Model Formulation

We first look at the case where n parts are randomly selected from the population of manufactured parts, and each of them is measured r times with the BMS and once with the gold standard. As in the fixed-effects model case, we call this sampling plan the standard plan (SP). The results of the repeated measurements are summarized by the total number of passes for each part, $s_i = \sum_{j=1}^{r} y_{ij}, i = 1, \ldots, n$. We also know the true state for each part $x_i, i = 1, \ldots, n$.

In this chapter, we adopt a random-effects model based on the Beta distribution to relax the assumption that α and β are constant for nonconforming and conforming parts, respectively. Therefore, for any randomly selected nonconforming part i, we assume that the probability of passing the inspection has distribution $A_i|X_i = 0$, with density:

$$f(a) = \frac{a^{g_A-1}(1-a)^{h_A-1}}{Beta(g_A, h_A)}, \quad 0 < a < 1, \tag{5.1}$$

where $Beta(g_A, h_A) = \int_0^1 t^{g_A-1}(1-t)^{h_A-1} dt$ is the Beta function. We also assume that given $X_i = 0$ and $A_i = \alpha$, repeated measurements on the part are independent, so that:

$$\Pr(Y_{i1} = y_{i1}, \ldots, Y_{ir} = y_{ir} \mid X_i = 0, A_i = \alpha) = \alpha^{s_i}(1-\alpha)^{r-s_i} \tag{5.2}$$

Similarly, for any conforming part i, we assume the probability of failing the inspection has distribution $B_i|X_i = 1$, with density:

$$f(b) = \frac{b^{g_B-1}(1-b)^{h_B-1}}{Beta(g_B, h_B)}, \quad 0 < b < 1, \tag{5.3}$$

and given $X_i = 1$ and $B_i = \beta$, repeated measurements on the part are independent, so that:

$$\Pr(Y_{i1} = y_{i1}, \ldots, Y_{ir} = y_{ir} \mid X_i = 1, B_i = \beta) = (1-\beta)^{s_i} \beta^{r-s_i} \tag{5.4}$$

Models (5.1), (5.2), and (5.3), (5.4) are beta-binomial models as described by Griffiths (1973).

The mean and variance of A are:

$$E(A) = \mu_A = \frac{g_A}{g_A + h_A}, \quad \text{and} \quad Var(A) = \frac{\mu_A(1-\mu_A)}{g_A + h_A + 1} = \phi_A \mu_A (1-\mu_A),$$

where $\phi_A = 1/(g_A + h_A + 1)$. Also, the mean and variance of B are:

$$E(B) = \mu_B = \frac{g_B}{g_B + h_B}, \quad \text{and} \quad Var(B) = \frac{\mu_B(1-\mu_B)}{g_B + h_B + 1} = \phi_B \mu_B (1-\mu_B),$$

where $\phi_B = 1/(g_B + h_B + 1)$. Parameters μ_A and μ_B are interpreted as the average consumer's and producer's risks and are the primary parameters of interest. As the measures of variability of the risks ϕ_A and ϕ_B approach zero, we recover the corresponding fixed-effects model (3.1) from Chapter 3.

Note that $0 < \mu_A, \mu_B, \phi_A, \phi_B < 1$. The parameters g_A, h_A, g_B, and h_B can be expressed in terms of μ_A, μ_B, ϕ_A and ϕ_B as follows:

$$g_A = \mu_A \frac{1-\phi_A}{\phi_A}, \quad h_A = (1-\mu_A)\frac{1-\phi_A}{\phi_A}, \quad g_B = \mu_B \frac{1-\phi_B}{\phi_B} \quad h_B = (1-\mu_B)\frac{1-\phi_B}{\phi_B} \tag{5.5}$$

Now, with these assumptions, for any nonconforming part i, with s_i passes in r repeated measurements ($s_i = \sum_i^r y_{ij}$), we have:

$$\begin{aligned}
\Pr(S_i = s_i, X_i = 0) &= \Pr(S_i = s_i | X_i = 0)(1-\pi_C) \\
&= (1-\pi_C)\int_0^1 \Pr(\sum_{j=1}^{r} Y_{ij} = s_i | X_i = 0, A = \alpha) f(\alpha; g_A, h_A) d\alpha \\
&= (1-\pi_C)\int_0^1 \binom{r}{s_i} \alpha^{s_i}(1-\alpha)^{r-s_i} \frac{\alpha^{g_A-1}(1-\alpha)^{g_A-1}}{Beta(g_A, h_A)} d\alpha \\
&= (1-\pi_C)\binom{r}{s_i}\frac{Beta(s_i+g_A, r-s_i+h_A)}{Beta(g_A, h_A)} \qquad (5.6)
\end{aligned}$$

Also, for any conforming part i that passes the inspection s_i times, we have:

$$\begin{aligned}
\Pr(S_i = s_i, X_i = 1) &= \Pr(S_i = s_i | X_i = 1)\pi_C \\
&= \pi_C \binom{r}{s_i}\frac{Beta(r-s_i+g_B, s_i+h_B)}{Beta(g_B, h_B)} \qquad (5.7)
\end{aligned}$$

Note that in this model, the repeated measurements $Y_{i1}...Y_{ir}$ given X_i are not independent. The covariance between two repeated measurements Y_{ij} and Y_{ik} given part i is conforming is:

$$\begin{aligned}
Cov(Y_{ij}, Y_{ik} | X_i = 1) &= E(Y_{ij}Y_{ik} | X_i = 1) - E(Y_{ij} | X_i = 1)E(Y_{ik} | X_i = 1) \\
&= \Pr(Y_{ij} = 1, Y_{ik} = 1 | X_i = 1) - \Pr(Y_{ij} = 1 | X_i = 1)\Pr(Y_{ik} = 1 | X_i = 1) \\
&= (1-\mu_B)(1-\mu_B+\mu_B\phi_B) - (1-\mu_B)^2 \\
&= \phi_B\mu_B(1-\mu_B) = Var(B) \qquad (5.8)
\end{aligned}$$

Similarly, the covariance between two repeated measurements given the part is nonconforming part is:

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}|X_i = 0) &= E(Y_{ij}Y_{ik}|X_i = 0) - E(Y_{ij}|X_i = 0)E(Y_{ik}|X_i = 0) \\ &= \Pr(Y_{ij} = 1, Y_{ik} = 1|X_i = 0) - \Pr(Y_{ij} = 1|X_i = 0)\Pr(Y_{ik} = 1|X_i = 0) \\ &= \phi_A\mu_A(1-\mu_A) = Var(A) \qquad (5.9) \end{aligned}$$

Therefore, the dependence between two repeated measurements on the same part is driven by the variability of the corresponding misclassification probability. The higher the variability, the stronger the dependence between repeated measurements.

The above random-effects models (5.1), (5.2), and (5.3) (5.4) explicitly allow for variation in the consumer's and producer's risks, within the set of nonconforming and conforming parts. In a sense, the random variables A and B model the effects of all latent variables on the properties of the BMS, given the value of X. That is, we are assuming for any latent variable Z_i, that $\Pr(Y_{i1} = y_{i1}, \ldots, Y_{ir} = y_{ir}|X_i = 0, A_i = \alpha, Z_i = z)$ and $\Pr(Y_{i1} = y_{i1}, \ldots, Y_{ir} = y_{ir}|X_i = 1, B_i = \beta, Z_i = z)$ do not depend on z.

Now, combining expressions (5.5), (5.6), and (5.7) gives the likelihood function:

$$\begin{aligned} L(\mu_A, \phi_A, \mu_B, \phi_B, \pi_C|(x_i, s_i)) \quad &\propto \quad \prod_{i=1}^{n} \left[\frac{Beta(r - s_i + \mu_B\frac{1-\phi_B}{\phi_B}, s_i + (1-\mu_B)\frac{1-\phi_B}{\phi_B})}{Beta(\mu_B\frac{1-\phi_B}{\phi_B}, (1-\mu_B)\frac{1-\phi_B}{\phi_B})}\pi_C \right]^{x_i} \times \\ &\times \left[\frac{Beta(s_i + \mu_A\frac{1-\phi_A}{\phi_A}, r - s_i + (1-\mu_A)\frac{1-\phi_A}{\phi_A})}{Beta(\mu_A\frac{1-\phi_A}{\phi_A}, (1-\mu_A)\frac{1-\phi_A}{\phi_A})}(1-\pi_C) \right]^{1-x_i} \end{aligned} \qquad (5.10)$$

Suppose that in a sample of n units, we use the gold standard to determine that there are n_C conforming parts with $x_i = 1$. Using (5.10), we can write the log-likelihood as a sum:

$$l(\mu_A, \phi_A, \mu_B, \phi_B, \pi_C) = l_A(\mu_A, \phi_A) + l_B(\mu_B, \phi_B) + n_C log(\pi_C) + (n - n_C)log(1-\pi_C) \qquad (5.11)$$

where $l_A(\mu_A,\phi_A)$ and $l_B(\mu_B,\phi_B)$ are beta-binomial log-likelihood functions corresponding to the $n-n_C$ nonconforming and n_C conforming parts in the sample, respectively. For example, $l_A(\mu_A,\phi_A)$ is given by:

$$\sum_{i=1}^{n-n_C}\left[lbeta(s_i+\mu_A\frac{1-\phi_A}{\phi_A},r-s_i+(1-\mu_A)\frac{1-\phi_A}{\phi_A})-lbeta(\mu_A\frac{1-\phi_A}{\phi_A},(1-\mu_A)\frac{1-\phi_A}{\phi_A})\right]$$

where $lbeta() = log[Beta()]$ is the log of the Beta function, and the expression for $l_B(\mu_B,\phi_B)$ is similar.

All parameters are estimable for $r>1$. In the case where $r=1$, that is, the selected parts are measured once with the gold standard and once with the BMS, the model collapses to the fixed-effects model (3.1) with $r=1$ from Chapter 3, with α replaced by μ_A, and β by μ_B. To see this, as mentioned in Chapter 3, in the $r=1$ case, the data can be summarized as in the 2×2 Table 2.1, where we count the number of conforming and nonconforming parts that pass and fail the BMS inspection. We denote these quantities by n_{CP}, $n_{\bar{C}P}$, $n_{C\bar{P}}$, and $n_{\bar{C}\bar{P}}$, respectively. We can then write the probability that a conforming part passes as:

$$\begin{aligned}
\Pr(Y_i=1|X_i=1) &= \int_0^1 \Pr(Y_i=1|X_i=1,B_i=\beta)f(\beta;g_B,h_B)d\beta \\
&= \int_0^1 (1-\beta)f(\beta;g_B,h_B)d\beta \\
&= 1-\mu_B \qquad (5.12)
\end{aligned}$$

The other conditional probabilities can be derived similarly as in (5.12), and the likelihood for the random effects in the $r=1$ case is:

$$\begin{aligned}
L(\mu_A,\mu_B,\phi_A,\phi_B,\pi_C|n_{CP},n_{\bar{C}P},n_{C\bar{P}},n_{\bar{C}\bar{P}}) &= [(1-\mu_B)\pi_C]^{n_{CP}}\times[\mu_A(1-\pi_C)]^{n_{\bar{C}P}}\times \\
&\times\ [\mu_B\pi_C]^{n_{C\bar{P}}}\times[(1-\mu_A)(1-\pi_C)]^{n_{\bar{C}\bar{P}}}
\end{aligned}$$

$$(5.13)$$

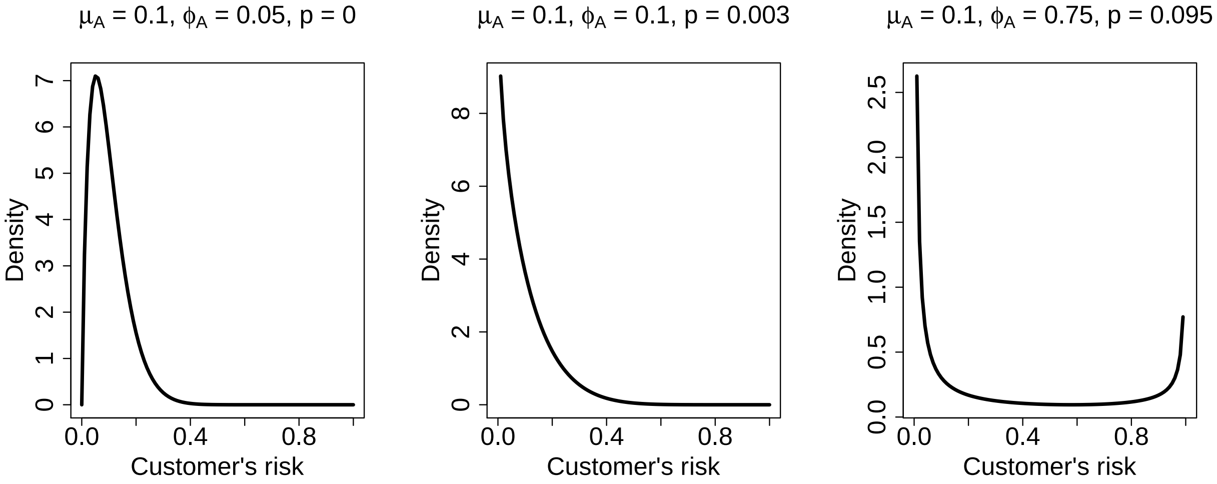

Figure 5.1 Densities Functions for the Beta Distributions - Consumer's Risk

Note that Eq. (5.13) does not include the variability parameters ϕ_A and ϕ_B. Therefore, we are not able to estimate ϕ_A and ϕ_B when $r = 1$. For the rest of the chapter, we focus on the assessment of a BMS when $r \geq 2$. For the case $r = 1$, all the results found in Chapter 3 apply.

As with the constant α and β case, we are generally interested in binary measurement systems with reasonable performance characteristics. In the varying misclassification rates case, a good BMS will have small average consumer's and producer's risks (e.g. μ_A and μ_B smaller than 0.1), and the variation of these rates will not be too extreme. Therefore, in this thesis, we consider only Beta distributions with densities that decrease to zero as α and β get larger. We therefore exclude the u-shaped Beta distribution. This leads to the constraints $h_A \geq 1$ and $h_B \geq 1$, or equivalently, $\mu_A + 2\phi_A - \mu_A\phi_A \leq 1$ and $\mu_B + 2\phi_B - \mu_B\phi_B \leq 1$. Additionally, we are interested in parameter values where the chance of a nonconforming part with a value of A larger than 0.5 is small. In Figure 5.1, we look at different Beta distributions as in (5.1) where the expected values are 0.1, for different values of the variability parameter ϕ_A. We note that for $\phi_A = 0.75$, which is equivalent to $h_A = 0.3$, the variation of A is large and that in 9.5% of the cases, the consumer's risk is larger than 0.5 ($p = \Pr(A > 0.5) = 0.095$). In our subsequent investigations of the random-effects models with assumed Beta distributions for the consumer's and producer's

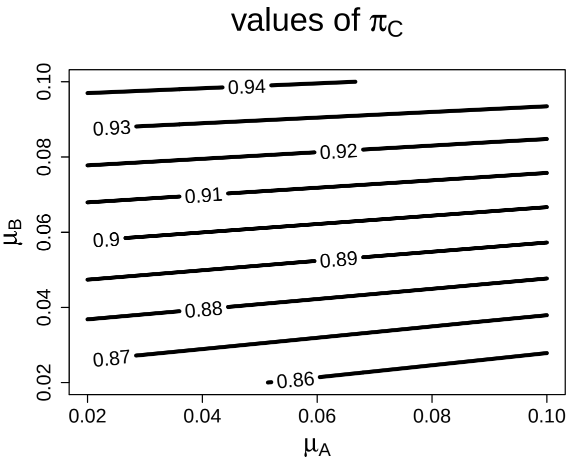

Figure 5.2 Contours of π_C for a Grid of Values for μ_A, μ_B, when $\pi_P = 0.85$.

risks, we will look at average values in the range $0.02 \leq \mu_A, \mu_B \leq 0.1$ and variability parameters in the range $0.01 \leq \phi_A, \phi_B \leq 0.1$.

5.2.2 Bias and Precision of the Estimators

As mentioned above, here we focus on assessing a BMS with reasonable performance characteristics, and we assume that the manufacturing process has a high conforming rate, i.e. π_C is large. Therefore, we focus on the ranges $\pi_C \geq 0.85$ and $\mu_A, \mu_B \leq 0.1$, and we construct a grid of parameter values with π_P constant and μ_A, μ_B, and π_C varying . For most of the examples given in this chapter, we use a grid of values with $\pi_P = 0.85$, where μ_A and μ_B vary from 0.02 to 0.1, and π_C varies from 0.85 to 0.94, as shown in Figure 5.2.

We expect that when we use the SP and the conforming rate is large, we will need large samples of parts (n large) to get a sufficient number of nonconforming parts in the sample. Also, since μ_A and μ_B are assumed small, we need a large total number of measurements ($n \times r$) to get standard errors that are small enough to be useful.

In order to investigate the properties of the maximum likelihood (ML) estimators, we simulate data from the beta-binomial random-effects model (5.10), for different n and r. We conduct a simulation study with a 2^5 structure, where we select two levels for each of the model parameters, $\mu_A = 0.02, 0.1$, $\mu_B = 0.02, 0.1$, $\phi_A = 0.01, 0.1$, $\phi_B = 0.01, 0.1$, and $\pi_C = 0.85, 0.95$. The ranges of values considered here for μ_A, μ_B, and π_C are the same as the ones in Figure 5.2. For each of the 32 combinations of the model parameters values, we simulate 500 random samples from the random-effects model. We expect that when the conforming rate is close to 1 and the sample size is small, there might be no nonconforming parts in the sample. Therefore, for each simulation run, we count the number of nonconforming parts in the sample and whenever this number is zero, we skip to the next run. For each accepted sample, we obtain the ML estimates corresponding to the fixed- and random-effects models, using the Nelder-Mead optimization algorithm (Nelder and Mead, 1965) in the R environment (R Development Core Team, 2010). For estimating the parameters, we only use the constraints $0 < \mu_A, \phi_A, \mu_B, \phi_B, \pi_C < 1$. We look at the sample error and the sample standard deviation of the estimates from both models.

In this section, we focus on the properties of the estimators from a random-effects model fitted to data simulated from the beta-binomial model (5.10). In the next subsection, we look at the properties of the estimators corresponding to a fixed-effects model fitted to the same sets of data.

We first consider cases where the sample size is small ($n = 200$) and there are $r = 5$ repeated measurements on each part. Figure 5.3 shows the simulation-based biases for the estimators of all the random-effects model parameters, for all 32 combinations of parameter values considered in the simulation study. We note negligible biases for $\hat{\mu}_A$, $\hat{\mu}_B$, and $\hat{\pi}_C$, small biases for $\hat{\phi}_B$, and unreasonably large biases, in some cases, for $\hat{\phi}_A$. The bias of $\hat{\phi}_A$ is larger than 0.1 when $\mu_A = 0.02$, $\phi_A = 0.1$, and $\pi_C = 0.95$. Recall that with the standard plan parts are randomly selected from the population of manufactured parts, and thus, for small sample sizes and large conforming rates, the expected number of nonconforming parts is small. Therefore, we expect difficulties estimating $\hat{\mu}_A$ and $\hat{\phi}_A$, as the estimation of these quantities is based on the number of times nonconforming parts pass the BMS inspection. Furthermore, for small numbers of repeated measurements, we expect to encounter additional problems estimating $\hat{\phi}_A$ and $\hat{\phi}_B$. It seems

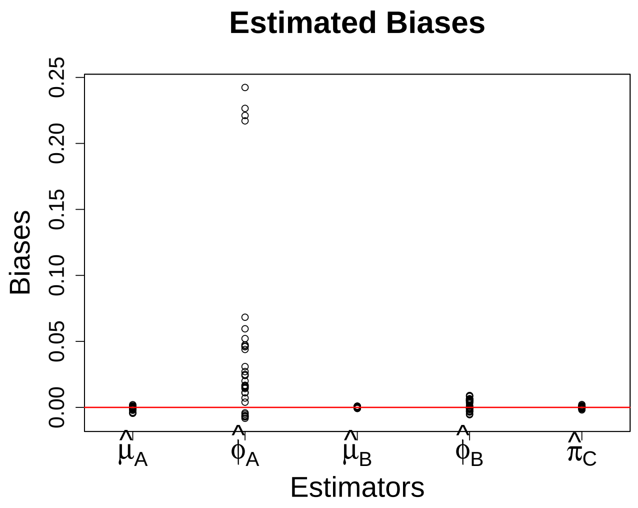

Figure 5.3 Plots of Biases – Random-Effects Model with Standard Plan, $n = 200$, $r = 5$.

that for $r = 5$ repeated measurements and a sample size of $n = 200$ we are able to estimate with reasonable accuracy all parameters except ϕ_A and ϕ_B. For $n = 500$ and $r = 5$ (results not shown here), all parameters except for ϕ_A are estimated with good accuracy.

Next, we increase the number of repeated measurements to $r = 10$, for a sample size $n = 200$. Although for $r = 10$ the bias of $\hat{\phi}_A$, for cases where μ_A is small, and ϕ_A and π_C are large, decreases by half compared to the $r = 5$ case, we still do not gain enough accuracy to make the estimator of ϕ_A useful. However, for $r = 10$ and $n = 200$, the estimators of μ_A, μ_B, ϕ_B, and π_C are virtually unbiased (results not shown here). After running more simulation studies for larger sample sizes, we found that for a sample of size $n = 2{,}000$ and $r = 10$ repeated measurements we can estimate with good accuracy all parameters including ϕ_A (results not shown here).

For all simulation studies, we also look at the standard errors of the estimates. Figure 5.4 shows the simulation-based standard errors for all estimates except for $\hat{\phi}_A$, for $n = 500$ and $r = 10$. For smaller sample sizes (e.g. $n = 200$) and $r = 10$ repeated measurements, the standard errors of the estimates of μ_A and ϕ_B are too large to be useful. For example, for one of the cases where $\mu_A = 0.02$, the standard error of $\hat{\mu}_A$ is 0.02. We also looked at the case where $n = 500$ and

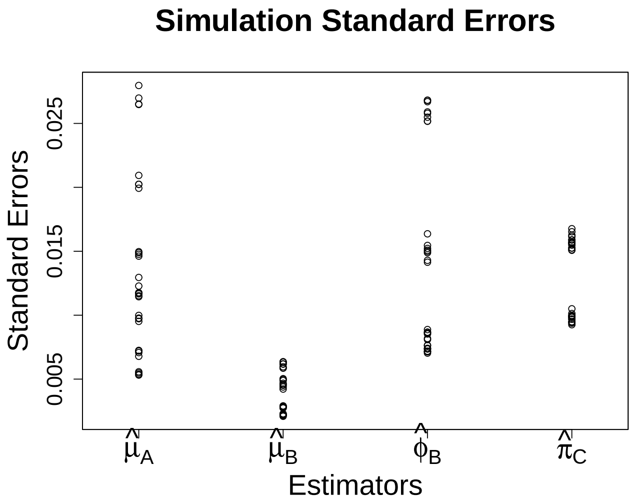

Figure 5.4 Plots of Simulation-based Standard Errors – Random-Effects Model with Standard Plan, $n = 500$, $r = 10$.

$r = 5$ and found that the standard errors are reasonably small for the estimates of μ_A, μ_B, ϕ_B, and π_C, except when $\phi_B = 0.01$. There, the standard errors of $\hat{\phi}_B$ are as large as the true value of parameter. From now on, we focus on the case where all four parameters μ_A, μ_B, ϕ_B, and π_C are well estimated in terms of accuracy and precision (i.e. $n = 500$, $r = 10$). In Figure 5.4, we don't include the standard errors for $\hat{\phi}_A$, as for these values of n and r, and some values of the model parameters, this estimator is highly biased. All four estimators have reasonable precision, with $\hat{\mu}_B$ having the smallest standard errors for the parameter values considered here.

Next, we compare the standard errors of the estimates obtained in the simulation studies with the asymptotic ones. We are interested in finding the minimum combination of design parameters so that the simulation-based standard errors match the asymptotic ones. We obtain the asymptotic standard errors from the Fisher information matrix corresponding to the random-effects likelihood function (5.10). The second derivatives of the log-likelihood function are obtained using Maple software (Maplesoft, 2009) and they are too long to include here. The actual functions representing the elements of the information matrix are written in the R environment (R Development Core Team, 2010). For any given parameter values, we obtain each element of the

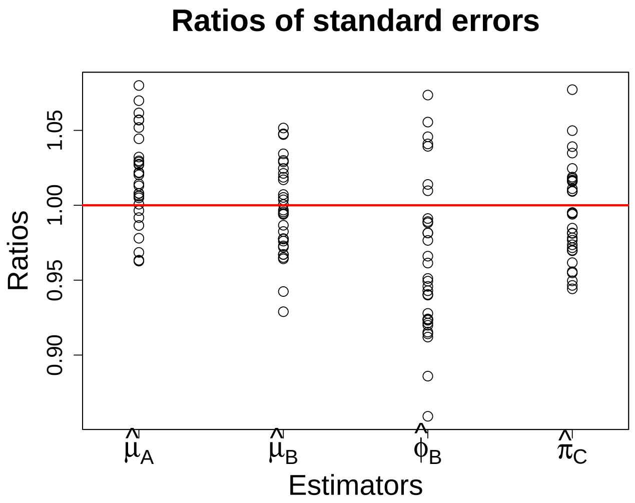

Figure 5.5 Plots of Ratios Standard Errors – Simulation-based over Asymptotic - Random-Effects Model with Standard Plan, $n = 500$, $r = 10$.

expected information matrix by summing the product of minus the value of the corresponding second derivative of the log-likelihood and the corresponding probability that a part passes the BMS inspection, over all possible values of s_i, i.e. $0,..,r$. For sample sizes as small as 500 and ten repeated measurements ($r = 10$), the ratios of the simulation-based standard error over the asymptotic one are close to 1 for all four parameter estimators, for most combinations of model parameter values (Figure 5.5). The two cases where the asymptotic standard errors of $\hat{\phi}_B$ are not close to the simulation-based ones (i.e. ratio is less than 0.9) happen when both ϕ_A and ϕ_B are small (i.e. 0.01).

5.2.3 Fixed-Effects Model Estimators – Bias and Precision

For each sample simulated from the beta-binomial model (5.10), we also obtain the ML estimates for α, β, and π_C from the fixed-effects model (3.1), and we look at their biases and standard errors. For each simulation run, the estimate for α is almost the same as the one for μ_A from the random-effects model. We see similar results for β and μ_B, and for the two estimators of

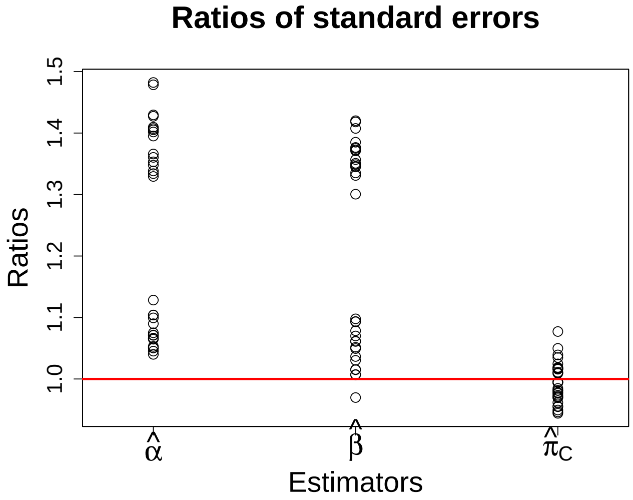

Figure 5.6 Plots of Ratios of Standard Errors – Simulation-based over Asymptotic – Fixed-Effects Model Estimates with Standard Plan, $n = 500$, $r = 10$.

π_C (results not shown here). Therefore, the fixed-effects estimators are almost unbiased for reasonable sample sizes and number of repeated measurements (e.g. $n = 500$ and $r = 10$), and their simulation-based standard errors are very close to the ones from the random-effects model. There is one problem related to the precision of the fixed-effects model estimators. The asymptotic-based standard errors for α and β derived from the fixed-effects model (3.1) seriously underestimate the simulation-based standard errors. Figure 5.6 shows the ratio of simulation-based over asymptotic standard errors based on the fixed-effects likelihood for α and β to demonstrate the inadequacy of the fixed-effects model asymptotics in this situation.

5.2.4 Fixed-Effects Model Data

Our next step focuses on the case where the data come from a model with constant misclassification probabilities α and β, i.e. a fixed-effects model as in Eq. (3.1) or random-effects model with $\phi_A = \phi_B = 0$, and we fit both the fixed- and random-effects models. We are interested in comparing the properties of the estimators from the two models, with emphasis on the precision.

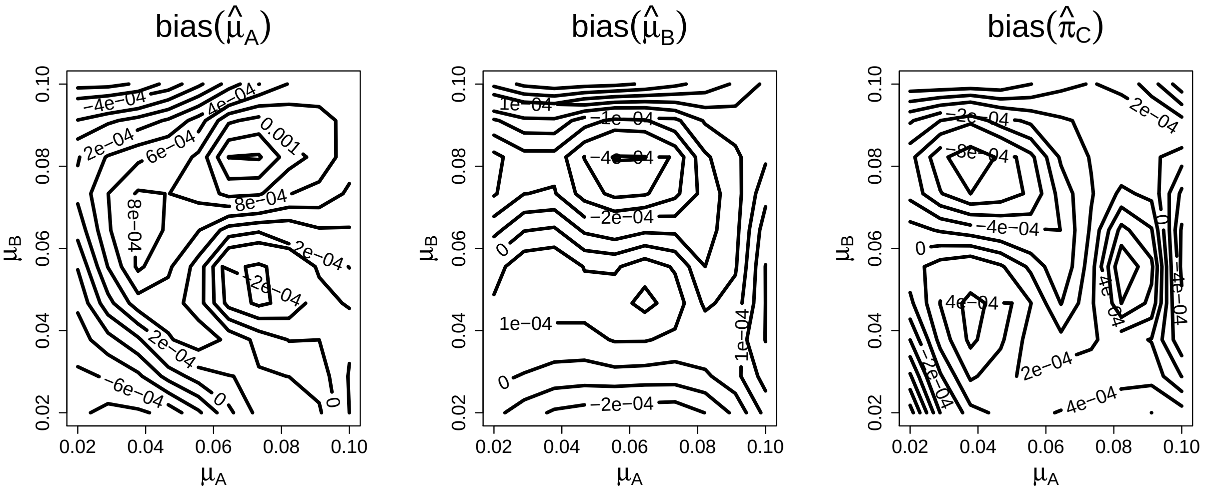

Figure 5.7 Contour Plots of Smoothed Biases – Random-Effects Estimators, when $n = 200$, $r = 5$, $\phi_A = 0$, $\phi_B = 0$, and $\pi_P = 0.85$.

We simulated 500 samples for each combination of $\alpha(\mu_A)$ and $\beta(\mu_B)$ shown in Figure 5.2, where $n = 200$ and $r = 5$. As we expected, the fixed-effects estimators are unbiased (results not shown here). Furthermore, the corresponding estimators from the random-effects model are also unbiased (Figure 5.7) and they are as precise as the ones from the fixed-effects model (Figure 5.8). We see the same results for larger values of the design parameters and π_P (results not shown here).

Therefore, we conclude that, when we have repeated measurements by the BMS ($r \geq 2$), and we have reasons to suspect that there is variation in the misclassification rates, the random-effects model should be used to get the appropriate measures of precision for the estimates of μ_A and μ_B and to provide an idea of how variable the misclassification rates are. On the other hand, if it happens that α and β do not vary from part to part, we can still safely use the random-effects model, as its estimators are unbiased and there is no loss of precision when we are fitting this more complicated model.

We can also examine the issue of varying misclassification probabilities by separately testing the hypotheses $H_0 : \phi_A = 0; H_A : \phi_A \neq 0$ and $H_0 : \phi_B = 0; H_A : \phi_B \neq 0$. The likelihood ratio test

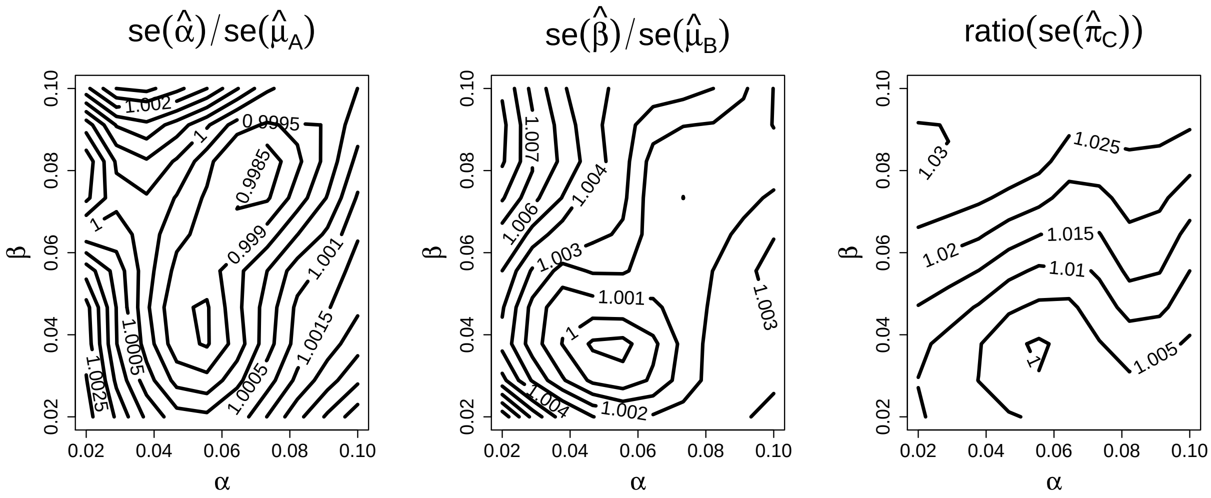

Figure 5.8 Contour Plots of Ratios of Smoothed Standard Errors – Fixed-Effects over Random-Effects Estimates, when $n = 200$, $r = 5$, $\phi_A = 0$, $\phi_B = 0$, and $\pi_P = 0.85$.

statistic for the hypothesis $H_0 : \phi_A = 0$ is given by $-2[l_A(\hat{\mu}_A, 0) - l_A(\hat{\mu}_A, \hat{\phi}_A)]$, with $l_A(\hat{\mu}_A, \hat{\phi}_A)$ defined as in (5.11). Because ϕ_A is on the boundary of the parameter space, using the results of Self and Liang (1987), the distribution of the test statistic under the null hypothesis is approximately an equal mixture of a discrete random variable with probability 1 at the origin and a χ^2_1 random variable. Note that the numerator in the likelihood ratio statistic is the maximum value of the log-likelihood for the fixed effects model (3.1), with μ_A replaced by α. Similar results apply for testing the hypothesis $H_0 : \phi_B = 0$.

In conclusion, when we randomly sample parts from the population of manufactured parts and the conforming rate is large, we need large sample sizes and number of repeated measurements to estimate all random-effects model parameters with reasonable accuracy and precision. ϕ_A is the hardest to estimate, as in the case of SP and π_C close to 1, the expected number of nonconforming parts in the sample can be unreasonably small. Also, we conclude that the fixed-effects estimators have similar proprieties as the random-effects ones, except for the fact that the asymptotic standard errors based on the fixed-effects model can substantially underestimate the asymptotic standard deviations. When data come from a model with constant misclassification rates, i.e. $\phi_A = \phi_B = 0$, the random-effects model estimators have similar properties as the fixed-effects ones, with virtually no loss of precision.

5.3 Random-Effects Model – Conditional Selection Plan with Baseline Data

As discussed in Chapters 3 and 4, in industrial practice, there are many examples of high-volume production processes where parts are systematically inspected by a BMS, especially in the cases where parts are visually inspected by an automated BMS. In these situations, passed and failed parts are segregated after measurement by the BMS. As well, the pass rate is recorded by hour, shift or other fixed-time period. Therefore, we can independently select two random samples from the available populations of previously passed and failed parts, i.e. use the conditional selection plan, and then measure each part with the gold standard and then repeatedly with the BMS. Additionally, we can augment the study data with the baseline information regarding the number of passed parts out of a (large) number of parts routinely measured with the BMS.

In Chapter 3, we investigated the advantages of using a CS plan augmented by baseline data, when we measure parts once with a gold-standard system and once or repeatedly with the BMS. We demonstrated the gain in the precision of the estimators given by the CS plan with baseline, in the context where we assumed constant misclassification probabilities α and β for all nonconforming/conforming parts. Now, we are interested in investigating the effect of using the CS plan with baseline on the characteristics of the estimators, when we assume α and β vary within nonconforming/conforming parts. That is, we look at the bias and precision of a random-effects model estimators when the CS plan with baseline data is used. Additionally, since with a CS plan we have control over the proportion of previously passed parts in the sample, f, using a CS can help increase the expected number of nonconforming parts in the sample, and this way we might be able to produce useful estimates with smaller sample sizes than those required by the SP.

In the next subsection, we give the formulation of a beta-binomial random-effects model when a CS plan with baseline is used. Then, we investigate the properties of the corresponding estimators, with focus on accuracy and precision, and assess the effect of changing the design parameters n, r, and the baseline size m, on the precision of the estimators. Finally, we compare the precision of the random-effects model estimators given by a CS plan with baseline with the ones given by the SP, for the same sample size and number of repeated measurements.

5.3.1 Model Formulation

Let $Y_{i0} = y_{i0}$ denote the initial routine measurement for part i. We randomly sample n_P parts from the population of previously passed and $n_{\bar{P}}$ from the population of previously failed. Then we measure each part once with the GS and r times with the BMS.

The contribution to the likelihood of any part i that passes s_i times in the assessment study is:

$$\Pr(S_i = s_i, X_i = x_i \mid Y_{i0} = y_{i0}) = \frac{\Pr(S_i = s_i, Y_{i0} = y_{i0} \mid X_i = x_i)\Pr(X_i = x_i)}{\Pr(Y_{i0} = y_{i0})} \tag{5.14}$$

Now, for a previously passed part, i.e. $Y_{i0} = 1$, we can write the contribution to the likelihood (5.14) as:

$$\left[\frac{\Pr(S_i = s_i, Y_{i0} = 1 \mid X_i = 1)\Pr(X_i = 1)}{\Pr(Y_{i0} = 1)}\right]^{x_i} \left[\frac{\Pr(S_i = s_i, Y_{i0} = 1 \mid X_i = 0)\Pr(X_i = 0)}{\Pr(Y_{i0} = 1)}\right]^{1-x_i}$$

Using the distributions (5.1) and (5.3) for A and B, respectively, we can further write (5.14) as:

$$\frac{1}{\pi_p}\left[\frac{Beta(r - s_i + \mu_B\frac{1-\phi_B}{\phi_B}, s_i + 1 + (1-\mu_B)\frac{1-\phi_B}{\phi_B})}{Beta(\mu_B\frac{1-\phi_B}{\phi_B}, (1-\mu_B)\frac{1-\phi_B}{\phi_B})}\pi_C\right]^{x_i} \times$$

$$\left[\frac{Beta(s_i + 1 + \mu_A\frac{1-\phi_A}{\phi_A}, r - s_i + (1-\mu_A)\frac{1-\phi_A}{\phi_A})}{Beta(\mu_A\frac{1-\phi_A}{\phi_A}, (1-\mu_A)\frac{1-\phi_A}{\phi_A})}(1-\pi_C)\right]^{1-x_i} \tag{5.15}$$

Similarly, for a previously failed part the contribution to the likelihood is:

$$\frac{1}{1-\pi_p}\left[\frac{Beta(r-s_i+1+\mu_B\frac{1-\phi_B}{\phi_B},s_i+(1-\mu_B)\frac{1-\phi_B}{\phi_B})}{Beta(\mu_B\frac{1-\phi_B}{\phi_B},(1-\mu_B)\frac{1-\phi_B}{\phi_B})}\pi_C\right]^{x_i}\times$$
$$\left[\frac{Beta(s_i+\mu_A\frac{1-\phi_A}{\phi_A},r-s_i+1+(1-\mu_A)\frac{1-\phi_A}{\phi_A})}{Beta(\mu_A\frac{1-\phi_A}{\phi_A},(1-\mu_A)\frac{1-\phi_A}{\phi_A})}(1-\pi_C)\right]^{1-x_i} \tag{5.16}$$

where $\pi_P = (1-\mu_B)\pi_C + \mu_A(1-\pi_C)$. There are n_P contributions as in (5.15) and $n_{\bar{P}}$ as in (5.16), in the likelihood function. In addition to these contributions, for the CS plans considered here, we also have the contribution from the baseline measurements:

$$L_b(\pi_P) \propto \pi_P^{m_P}(1-\pi_P)^{m-m_P}$$

where m_P is the number of parts passing the initial inspection out of m parts.

As in the standard plan case, we maximize the overall likelihood function using Nelder-Mead with the constraints as defined in the previous section.

5.3.2 Bias and Precision of the Estimators

To investigate the properties of the CS sampling plan, we simulate data from the beta-binomial random-effects model, where parts are selected using a CS plan augmented with baseline data of size m. We focus on the case where we only sample from the previously failed parts, i.e. $f = 0$ and thus $n_P = 0$, $n_{\bar{P}} = n$. If π_C is large, as expected, then sampling from failed parts increases the expected number of nonconforming parts in the sample. For a given sample size and model parameter values, sampling parts from the population of previously failed gives the highest expected proportion of nonconforming parts in the sample. In practice, it is convenient to use failed parts in the assessment study as these parts are typically retained as scrap or for re-work. However, depending on the parameter values, we may end up with more than half the parts being nonconforming which suggests that μ_B and ϕ_B would become now the more difficult

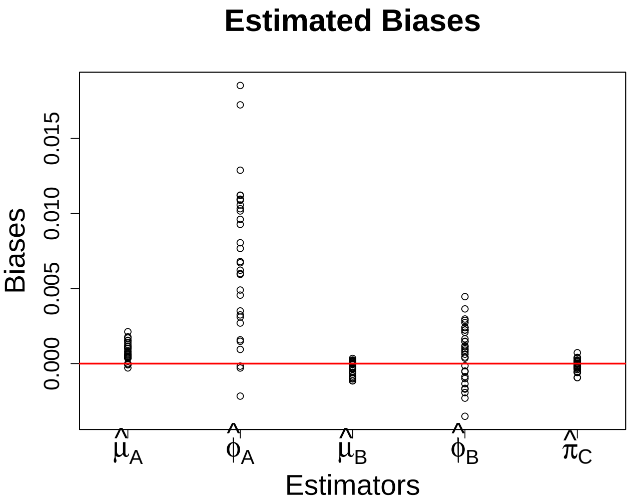

Figure 5.9 Plots of Biases – Random-Effects Model with Conditional Selection Plan, $n = 200$, $r = 5$, $m = 1{,}000$.

parameters to estimate. However, the baseline data provide significantly more information about μ_B than μ_A. Therefore, similar to the fixed-effects model case in Chapters 3 and 4, when using a CS plan with baseline we are better off with $f = 0$ in terms of precision for all estimators, even when this results in more nonconforming than conforming parts in the sample.

We start with small sample sizes ($n = 200$), numbers of repeated measurements ($r = 5$), and baseline size ($m = 1{,}000$). We expect that increasing any of the design parameters n, r or m will improve the performance of the estimates. For assessing the bias and precision of the estimators, we look at pairs of small and large values of the model parameters, i.e. $\mu_A, \mu_B = 0.02, 0.1$, $\phi_A, \phi_B = 0.01, 0.1$, and $\pi_C = 0.85, 0.95$, and conduct a 2^5 factorial simulation study. For each combination of the parameters, we determine the bias and standard error for each estimate, as well as the ratio of the simulated over the asymptotic approximation of the standard errors. The results are summarized in Figures 5.9, 5.10, and 5.11.

In Figure 5.9 we see there is negligible bias in any of the estimates except for $\hat{\phi}_A$. However, we note that the biases for the estimator of ϕ_A are generally much smaller than in the SP case, for the same values of the design parameters (see Figure 5.3).

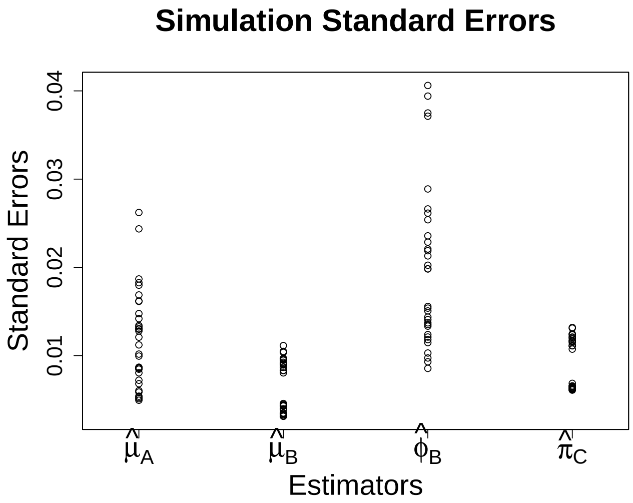

Figure 5.10 Plots of Simulation Standard Errors – Random-Effects Model with Conditional Selection Plan, $n = 200$, $r = 5$, $m = 1{,}000$.

The standard errors of the estimators of μ_A, μ_B, ϕ_B, and π_C are reasonably small, as shown in Figure 5.10, and are close to the asymptotic ones (Figure 5.11), for most cases except for a few cases where the ratios for $\hat{\mu}_A$ are larger than 1.1, and some cases where the ratio for $\hat{\phi}_B$ are smaller than 0.8. Therefore, for design parameters as large or larger than the ones used in the experiment, we can safely use the asymptotic approximation as a measure of the precision of the estimators, when the true values are close to those tested.

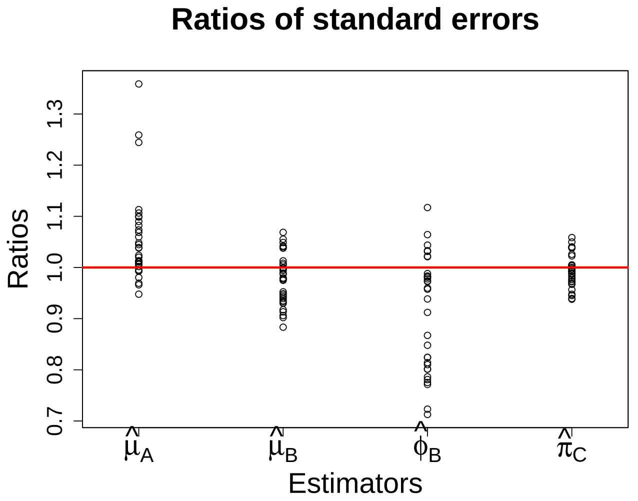

Figure 5.11 Plots of Ratios of Standard Errors – Simulation over Asymptotic – Random-Effects Model with Conditional Selection, $n = 200$, $r = 5$, $m = 1{,}000$.

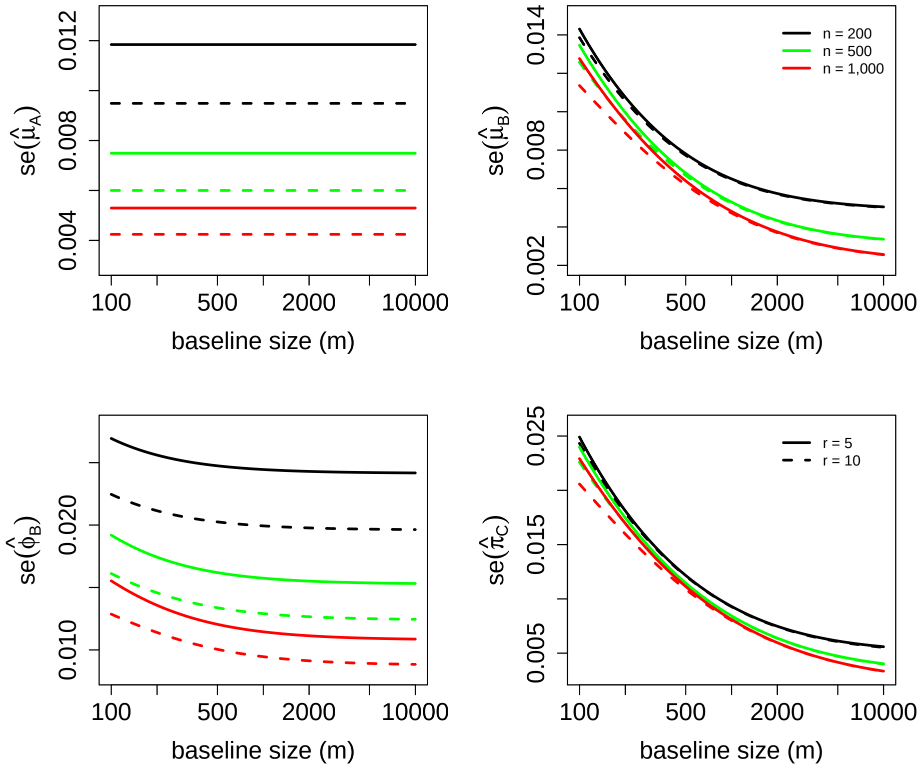

Figure 5.12 Effect of Changing the Baseline Size m on Standard Error of the Estimators – Random-Effects Model with Conditional Selection, $\mu_A = 0.05$, $\phi_A = 0.1$, $\mu_B = 0.05$, $\phi_B = 0.1$, and $\pi_C = 0.9$.

5.3.3 Effect of Changing the Design Parameters

With a conditional selection plan augmented with baseline data, the design parameters that can be determined prior to the assessment study are the number of parts used for the study n, the number of repeated measurements r, and the total number of parts routinely measured with the BMS, i.e. the baseline size m. We are interesting in assessing the effect of changing these design parameters on the precision of the estimators from the random-effects model. We focus on the case where we choose only previously failed parts, i.e. $f = 0$.

Based on the results from the previous section, in our investigation, we look at parameter values for which the asymptotic standard errors represent good approximations. Also, we look at sample sizes and numbers of repeated measurements as large or larger than the ones in the experiment, i.e. $n \geq 200$ and $r \geq 5$.

Figure 5.12 shows the change in the asymptotic standard errors of $\hat{\mu}_A$, $\hat{\mu}_B$, $\hat{\phi}_B$, and $\hat{\pi}_C$, as the baseline size increases from 100 to 10,000, for different sample sizes and numbers of repeated measurements. The precision of $\hat{\phi}_A$ is not included, as we already know that we need larger sample sizes to make the asymptotic approximation work for this estimator. In the plots, the baseline size m is represented on the horizontal axis on a logarithm scale. First, we note that increasing the baseline size does not affect the precision of $\hat{\mu}_A$, for any sample size or number of repeated measurements considered here. For the estimate of μ_B, there is a dramatic increase in precision as the baseline size increases, with a large drop (75%) in standard errors when m increases from 100 to 2,000. We see a similar result for the precision of $\hat{\pi}_C$. Increasing the baseline size has a small impact on the precision of $\hat{\phi}_B$, with the largest gain when m increases from 100 to 2,000. We also note that for all estimates, the standard errors approach a limit as m gets larger, i.e. $m \geq 10{,}000$. Letting m go to infinity corresponds to the situation where the pass rate π_P is known.

In Figure 5.13, we look at the effect of changing the sample size n on the standard errors of the estimates, when r and m are held constant. For the estimate of μ_A, we note that the standard errors are identical for different values of m (solid, dashed and dotted lines of the same color in the first panel of Figure 5.13 overlap), and they are changing at a $1/\sqrt{n}$ rate. For the estimates of μ_B and π_C, when $m = 100$, the effect of changing n depends on the number of repeated measurements r (see top black and green solid lines in the second and fourth panels of Figure 5.13). When m is larger, the curves of the standard errors are identical for $r = 5$ and $r = 10$. That is, the effect of changing n does not depend on r anymore. In the case of $\hat{\phi}_B$, the relative decrease in standard errors is similar for different values of r and m, with a slightly steeper slope for $r = 5$ and large values of m (see black dotted and dashed lines in the third panel of Figure 5.13).

Regarding the effect of changing the number of repeated measurements r, we note in Figure 5.14 that the curves for the standard errors of $\hat{\mu}_A$ for different sample sizes n have similar slopes, with a slightly larger relative decrease for smaller sample sizes ($n = 200$). Again, we see no

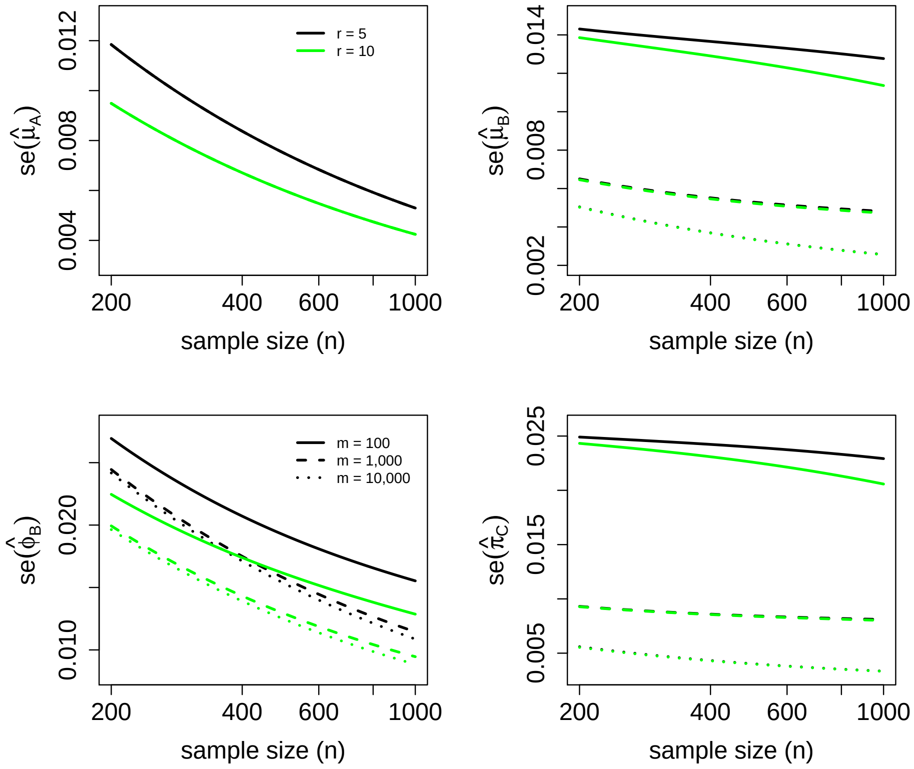

Figure 5.13 Effect of Changing the Sample Size n on Standard Error of the Estimators – Random-Effects Model with Conditional Selection, $\mu_A = 0.05$, $\phi_A = 0.1$, $\mu_B = 0.05$, $\phi_B = 0.1$, and $\pi_C = 0.9$.

difference in the curves for different baseline sizes (solid, dotted and dashed lines of the same color in the first panel of Figure 5.14 overlap). For the estimates of μ_B and π_C, when the baseline size is small, i.e. $m = 100$, there is a small effect of changing r on the standard errors, especially for large sample sizes, i.e. $n = 1{,}000$ (see top solid lines in the second and fourth panels of Figure 5.14). When the baseline size is large, increasing the number of repeated measurements does not affect the precision of $\hat{\mu}_B$ and $\hat{\pi}_C$. For the estimator of ϕ_B, the curves have similar slopes, with a slightly steeper slope for smaller sample sizes, i.e. $n = 200$ (see black solid, dashed and dotted lines in the third panel of Figure 5.14).

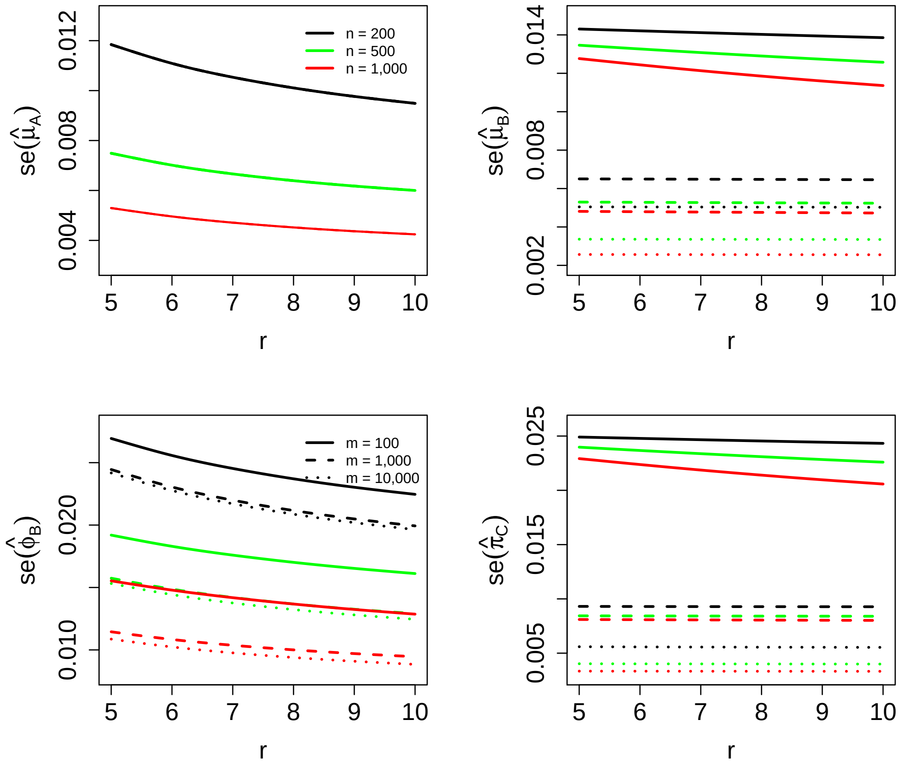

Figure 5.14 Effect of Changing the Number of Repeated Measurements r on Standard Error of the Estimators – Random-Effects Model with Conditional Selection, $\mu_A = 0.05$, $\phi_A = 0.1$, $\mu_B = 0.05$, $\phi_B = 0.1$, and $\pi_C = 0.9$.

5.4 Comparison of Standard Plan and Conditional Selection Plan with Baseline

In this section, we compare the precision of the estimators from a beta-binomial model with conditional selection plan with baseline data to the ones given by the same random-effects model with a standard plan. For the conditional selection plan, we sample parts only from the population of previously failed parts, i.e. $f = 0$, and augment the study data with $m = 1{,}000$ baseline data observations. Since we assume the baseline measurements are available from routine inspection, the conditional and standard sampling plans have the same total number of measurements $n \times r$. The following comparisons are made for the case where both sampling schemes yield virtually unbiased estimators. That is, we look at relatively large sample sizes

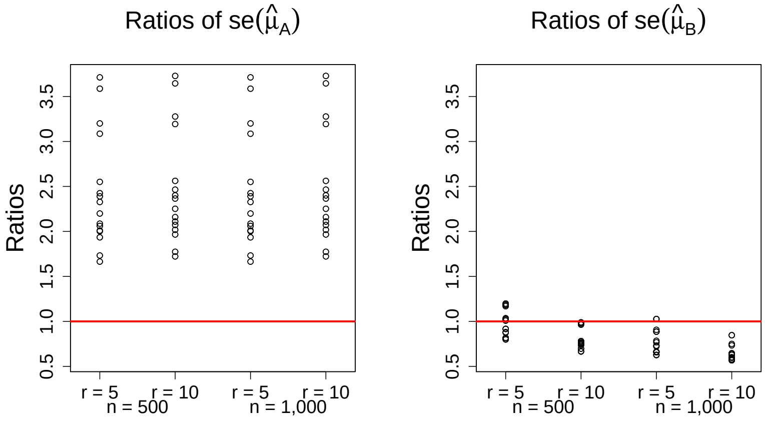

Figure 5.15 Plots of Ratios of Asymptotic Standard Deviations – Standard Plan over Conditional Selection with $m = 1{,}000$ and $f = 0$, for the estimators of μ_A and μ_B.

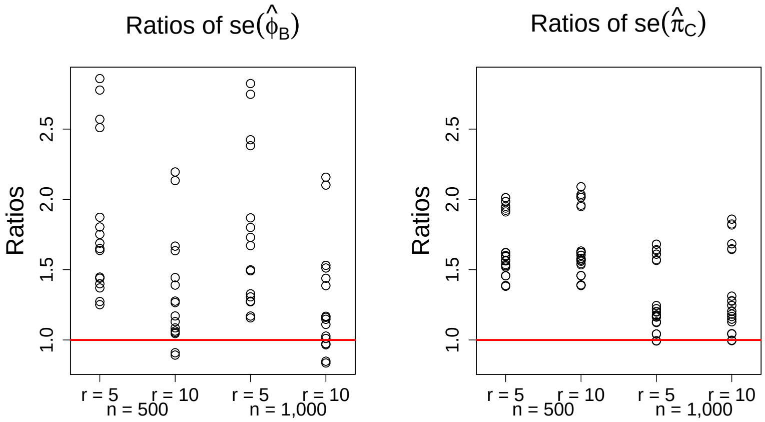

Figure 5.16 Plots of Ratios of Asymptotic Standard Deviations – Standard Plan over Conditional Selection with $m = 1{,}000$ and $f = 0$, for the estimators of ϕ_B and π_B.

(e.g. $n = 500$), so that the SP estimators are unbiased. For the same combination of model parameter values, we need smaller n to obtain unbiased estimators with the CS plan. However, for "scientific" reasons, we compare the precision of the estimators from the two sampling plans, for sample sizes large enough to make both sets of estimators unbiased.

For these comparisons, we again use a factorial structure, where we look at all possible

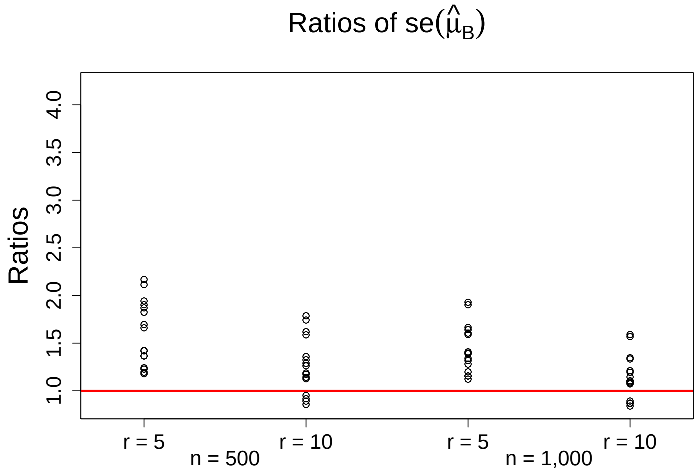

Figure 5.17 Plots of Ratios of Asymptotic Standard Deviations – Standard Plan over Conditional Selection with $m = 10{,}000$ and $f = 0$, for the estimators of μ_B.

combinations of $n = 500, 1{,}000$, $r = 5, 10$, and $\mu_A = 0.02, 0.1$, $\mu_B = 0.02, 0.1$, $\phi_A = 0.01, 0.1$, $\phi_B = 0.01, 0.1$, and $\pi_C = 0.85, 0.95$. Also, we use the asymptotic standard errors of the estimates of μ_A, μ_B, ϕ_B, and π_C, as we have seen in the previous sections, for these design parameters and parameter values, the asymptotic approximations are reasonable.

Figures 5.15 and 5.16 show the ratios of asymptotic standard errors of $\hat{\mu}_A$, $\hat{\mu}_B$, $\hat{\phi}_B$, and $\hat{\pi}_C$, respectively, for the standard plan compared to the conditional selection plan with $f = 0$ and $m = 1{,}000$. Ratios are larger than 1 for the cases where the CS plan estimators are more precise than the SP ones. We note that the CS gives better estimators for μ_A and π_C, for all combinations of model parameter values and choices of r and n. Furthermore, for $r = 5$ and $n = 500$, the ratios for the standard errors of $\hat{\pi}_C$ are larger than 1.5, whereas for $\hat{\mu}_A$, the ratios are larger than 1.5 for all values of r and n. For $\hat{\phi}_B$, the CS plan gives more efficient estimators for all cases where the number of repeated measurements is small $(r = 5)$, and in most cases for $r = 10$. However, as shown in Figure 5.15, for $\hat{\mu}_B$, the ratios of asymptotic standard errors are larger than 1 only for some cases where the sample size and the number of repeated measurements are small , i.e. $n = 500$ and $r = 5$. For the estimator of μ_B, all the cases where the CS estimators are more precise than the SP ones happen when the value of μ_B is small (i.e. 0.02).

We expect to get more information about μ_A with conditional sampling since with this scheme we will likely select more nonconforming parts. The increased precision for the estimators of μ_B and ϕ_B, for some model and design parameters, is perhaps surprising. Here it is the baseline measurements that help, as shown in Figures 5.13 and 5.12. The baseline data provide an estimate of $\pi_P = (1-\mu_B)\pi_C + \mu_A(1-\pi_C)$, a function of μ_A, μ_B, and π_C. Since we are considering situations where π_C is large and μ_A is relatively small, π_P is strongly influenced by μ_B and π_C. In this case, the additional information about μ_B and π_C from the baseline data outweighs the lost information due to having fewer conforming parts in the sample. To further illustrate this argument, we obtained the ratios of asymptotic standard errors for the estimator of μ_B, for the case where we augment the CS plan with $m = 10{,}000$ baseline measurements. In Figure 5.17, we note that the CS plan with $f = 0$ and $m = 10{,}000$ gives estimators for μ_B with similar or better precision than the SP.

5.5 Summary

In this chapter, we introduce the idea of varying misclassification rates of the binary measurement system that leads to inadequacy of conditional independence assumption. We give a brief overview of proposed methods that allow for variation in consumer's and producer's risks within nonconforming/conforming parts. Next, we propose a new random-effects model that assumes Beta distributions for the misclassification rates α and β. We first investigate the properties of the model parameters in the case where parts are randomly selected from the population of manufactured parts, i.e. we use the standard plan. We obtain biases and standard errors from simulation studies and conclude that, for the parameter values considered here, we need large sample sizes (e.g. $n = 2{,}000$) and numbers of repeated measurements (e.g. $r = 10$) to get unbiased estimators for all model parameters, including the variation parameters ϕ_A and ϕ_B. However, we only need $n = 500$ and $r = 5$ to estimate with good accuracy and precision all the main parameters μ_A, μ_B, and π_C.

For the same simulated data, we also obtain the MLEs for the fixed-effects model parameters and conclude that the fixed-effects estimates for α and β are almost identical to the ones for μ_A, μ_B. We find that, although the estimates from the two models are very similar, the asymptotic standard errors given by the fixed-effects model are not good approximations. Therefore, we recommend fitting the random-effects model as it also provides information about the variation parameters ϕ_A and ϕ_B. In the case where data are generated from a fixed-effects model, fitting the random-effects model leads to unbiased estimators and virtually no loss of precision when compared to the fixed-effects estimators.

Next, we investigate the random-effects model for the case where parts are randomly selected from the population of previously failed parts (i.e. the conditional selection plan with $f = 0$) and we supplement the study data with $m = 1{,}000$ baseline measurements. We find that, for relatively small sample sizes ($n = 200$) and numbers of repeated measurements ($r = 5$), the CS gives estimators with reasonable accuracy and precision for all model parameters except for ϕ_A. When we compare the precision of the estimators from a random-effects model with CS with the ones given by the SP, we conclude that the CS plan yields more precise estimators for μ_A, ϕ_B, and π_C. When we increase the baseline size to $m = 10{,}000$, the CS plan estimators are uniformly better than the SP ones. Therefore, we recommend using a conditional selection plan with parts sampled from the population of previously failed parts augmented with at least $m = 1{,}000$ parts. Using a larger number of baseline measurements increases the precision of the estimator of μ_B, although, even for smaller baseline sizes, the large gain in the precision of $\hat{\mu}_A$ may outweigh the small loss in the precision of $\hat{\mu}_B$.

We also look at the effect of the design parameters n, r, and m on the precision of the estimators from a random-effects model with CS and baseline measurements.

The main results presented in this chapter are also included in the paper “Assessing a Binary Measurement System with Varying Misclassification Rates when a Gold Standard is Available” by Danila, Steiner, and MacKay (2011), submitted to “Technometrics” in August 2011.

Chapter 6

Random-Effects Model – No Gold-Standard System Available

In the case where a gold-standard system is not available, the true state of a part, i.e. conforming or nonconforming, is unknowable. When parts are randomly selected from the population of manufactured parts, we can write the probability that the outcomes of measuring part i r times with the BMS are $y_{i1},\ldots,y_{ir}$ as:

$$\begin{aligned}\Pr(Y_{i1}=y_{i1},\ldots,Y_{ir}=y_{ir}) \quad &= \quad \Pr(Y_{i1}=y_{i1},\ldots,Y_{ir}=y_{ir} \mid X_i=1)\Pr(X_i=1)+\\ &\quad\quad \Pr(Y_{i1}=y_{i1},\ldots,Y_{ir}=y_{ir} \mid X_i=0)\Pr(X_i=0)\end{aligned}$$

If we further assume that the misclassification rates α and β are constant for all nonconforming/conforming parts, we use can a latent class model to estimate the parameters of interest (see Eq. (4.1)).

In the case where the misclassification rates vary from part to part, we can adopt a random-effects model using the assumptions (5.3), (5.4), and (5.1), (5.2) from Chapter 5, and then get the two probabilities $\Pr(Y_{i1}=y_{i1},\ldots,Y_{ir}=y_{ir} \mid X_i=1)$ and $\Pr(Y_{i1}=y_{i1},\ldots,Y_{ir}=y_{ir} \mid X_i=0)$

by integrating over the possible values of β and α, respectively. In this chapter, we investigate the properties of the estimators from a random-effects model when there is no available gold standard and parts are repeatedly measured only with the BMS. We first look at the case where parts are randomly selected, i.e. we use the standard plan (SP). We compare the properties of the estimators from both the fixed-effects model (4.1) and the random-effects model, in the case where misclassification rates are not constant. Then, we investigate the case where α and β are constant and we fit the random-effects model. Next, we use the conditional selection (CS) plan augmented with baseline data and look at the properties of the estimators of the random-effects model and compare them to the ones given by a standard plan. We look at the effect of changing the design parameters n, r, and m on the precision of the estimators from a CS plan and then give planning recommendations.

6.1 Random-Effects Model – Standard Plan

6.1.1 Model Formulation

When n parts are randomly selected from the population of manufactured parts, each part is measured r times with the BMS and we record the total number of times each part passes the inspection as $s_i = \sum_{j=1}^{r} y_{ij}$, $i = 1,\ldots,n$, we can write the likelihood function for the random-effects model as:

$$\begin{aligned} L(\mu_A,\mu_B,\phi_A,\phi_B,\pi_C|s_1,\ldots,s_n) &\propto \prod_{i=1}^{n} \Pr(S_i = s_i) \\ &= \prod_{i=1}^{n} \Bigg[\frac{Beta(r - s_i + \mu_B \frac{1-\phi_B}{\phi_B}, s_i + (1-\mu_B)\frac{1-\phi_B}{\phi_B})}{Beta(\mu_B \frac{1-\phi_B}{\phi_B}, (1-\mu_B)\frac{1-\phi_B}{\phi_B})} \pi_C + \\ &\quad + \frac{Beta(s_i + \mu_A \frac{1-\phi_A}{\phi_A}, r - s_i + (1-\mu_A)\frac{1-\phi_A}{\phi_A})}{Beta(\mu_A \frac{1-\phi_A}{\phi_A}, (1-\mu_A)\frac{1-\phi_A}{\phi_A})} (1-\pi_C) \Bigg] \end{aligned} \tag{6.1}$$

The random-effects model (6.1) depends only on the number of passes $s_1,\ldots,s_n$ in the r repeated measurements. There are $r+1$ possible values for each s_i, and the probabilities associated with each possible value must add to 1. The model has five parameters (μ_A, μ_B, ϕ_A, ϕ_B, π_C) and to be identifiable we require $r \geq 5$. If $r = 4$, there is an infinite number of parameter values that give the same distribution for $(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$. Also, from the likelihood, it is clear that for identifiably we need the sensible constraint $\mu_A < 1 - \mu_B$, similar to that required for the fixed effect model, i.e. $\alpha < 1 - \beta$. That is, we assume that the average pass rate for nonconforming parts is less than the average pass rate for conforming parts. Van Wieringen and Van den Heuvel (2005) investigate the parameters identifiability issue for the fixed-effects model and obtain the constraint $\alpha < 1 - \beta$ and the minimum number of measurements r for different cases, e.g. single or multiple BMSs under study, etc. Similar rigorous derivations can be attempted for the random-effects model and further research on this topic is needed.

Also, similarly to the case where a gold-standard system is available, we note that if we hold μ_A and μ_B fixed and let ϕ_A and ϕ_B approach zero, $\Pr(S_i = s_i)$ in (6.1) approaches $\binom{n}{r}[(1-\beta)^{s_i}\beta^{r-s_i}\pi_C + \alpha^{s_i}(1-\alpha)^{r-s_i}(1-\pi_C)]$, the fixed-effects (latent class) model (4.1), with $\alpha = \mu_A$ and $\beta = \mu_B$. That is, the fixed-effects is a limiting case of the random-effects model.

We can estimate the five parameters by maximizing the log-likelihood using a standard approach such as Nelder-Mead (Nelder and Mead, 1965).

6.1.2 Bias and Precision

We investigate the properties of the estimators from a random-effects model through a simulation study with a 2^5 factorial structure. We use two values for each model parameter, with $\mu_A, \mu_B = 0.02, 0.1$, $\phi_A, \phi_B = 0.01, 0.1$, and $\pi_C = 0.85, 0.95$. For μ_A, μ_B, and π_C, the values considered in the simulation study are the minimum and maximum values from the grid in Figure 5.2.

As with the fixed-effects model in Chapter 4, we expect that when we use the SP and the conforming rate π_C is large, we need a large sample size n and number of repeated measurements r to get unbiased estimators with useful precision. In the case of the random-effects model, we expect to need even larger total number of measurements $n \times r$, as this model has a higher degree of complexity than the fixed-effects model. Therefore, in the simulation study, we focus on cases where the number of repeated measurements is large (e.g. $r = 10$) and we have different sample sizes (e.g. $n = 250, 500, 1{,}000$).

For a given sample size and for each combination of parameter values, we run 500-repeat simulations. For each run, we get the ML estimates for both the fixed- and random-effects models parameters, using the Nelder-Mead optimization algorithm in the R Environment (R Development Core Team, 2010). For the random-effects model, we maximize the likelihood function subject to the constraints $\mu_A + 2\phi_A - \mu_A\phi_A \leq 1$ and $\mu_B + 2\phi_B - \mu_B\phi_B \leq 1$, needed to avoid the u-shaped Beta distribution, and $\mu_A + \mu_B < 1$, for identifiability. For the fixed-effects model, we maximize the likelihood under the constraint $\alpha + \beta < 1$. As in Chapter 4, we use an additional step in the optimization algorithm that checks whether the maximum log-likelihood value found by the Nelder-Mead algorithm when $\pi_C < 1$ is larger than the one when $\pi_C = 1$, both values being found by the Nelder-Mead algorithm. For the random-effects model, the log-likelihood with $\pi_C = 1$ contains only two parameters, μ_B and ϕ_B, and if the maximum value is larger than the one with $\pi_C < 1$, we skip to the next run. This additional step avoids cases where the global optimum is not found by the algorithm as a result of not having sufficient nonconforming parts in the sample. This occasionally happened when the sample size is small, π_C is close to the upper boundary 1, and we sample parts randomly from the population of manufactured parts, i.e. we use the standard plan (SP).

Also, we derive the asymptotic standard deviations of the estimators using the expected Fisher information matrix. The second derivatives of the log-likelihood functions are obtained using Maple (Maplesoft, 2009) and they are too long to be included here. The actual functions representing the elements of the information matrix are written in the R Environment (R Development Core Team, 2010). For given parameter values, we obtain each element of the information matrix by summing up the product of minus the value of the corresponding second derivative of the log-likelihood and the corresponding probability that a part passes the BMS inspection, over all possible values of s_i, i.e. $0,..,r$.

In this section, we focus on estimating the accuracy and precision of the random-effects model estimators, and also on comparing the standard errors as given by the simulation study to the asymptotic ones. We are interested in finding the minimum sample size for which the biases are negligible, the standard errors are small enough to be useful, and the simulation-based standard errors match the asymptotic ones.

We start by looking at the results of the simulation study for $n = 250$. We summarize the results by plotting the biases (Figure 6.1), simulation-based standard errors (Figure 6.2) and ratios of simulation versus asymptotic standard errors (Figure 6.3), for all combinations of parameter values considered in the study.

We note relatively large biases for $\hat{\mu}_A$ and unreasonably large biases for $\hat{\phi}_A$. Most cases when $\hat{\mu}_A$ is biased (e.g bias larger than 0.01) arise when ϕ_B and π_C are large. Also, the simulation-based standard errors for $\hat{\mu}_A$ are as large as 0.1, whereas for $\hat{\phi}_A$ standard errors are as large as 0.2. These standard errors as too large to be useful, considering the small parameter values used in the study ($\mu_A = 0.02, 0.1$ and $\phi_A = 0.01, 0.1$). More positively, $\hat{\mu}_B$ and $\hat{\pi}_C$ are virtually unbiased, with small simulation-based standard errors. $\hat{\phi}_B$ has small biases and relatively small standard errors. In Figure 6.3, we note that the simulation-based standard errors match the asymptotic ones for

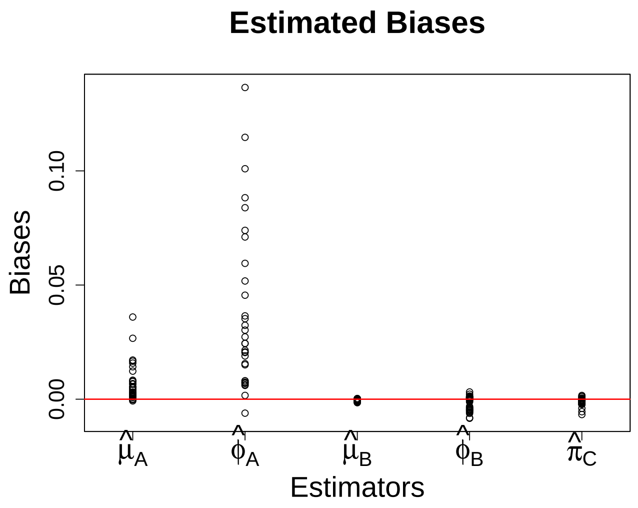

Figure 6.1 Plots of Biases – Random-Effects Model with Standard Plan, $n = 250$, $r = 10$.

$\hat{\mu}_B$, $\hat{\phi}_B$ and $\hat{\pi}_C$, whereas for $\hat{\mu}_A$ and $\hat{\phi}_A$ the two values are quite different, especially for $\hat{\phi}_A$.

Next, we increase the sample size and run simulation studies for the same combinations of parameter values. We still find large biases and standard errors for $\hat{\mu}_A$ and $\hat{\phi}_A$, for sample sizes $n = 500, 1{,}000$ (results not included here). Figure 6.4 shows the biases of the estimators for $n = 2{,}000$ and $r = 10$. Although the biases of $\hat{\mu}_A$ and $\hat{\phi}_A$ reduce drastically compared to the case where $n = 250$, we note there are still some large values, especially for the estimator of ϕ_A (bias up to 0.085). For $\hat{\mu}_A$, the large biases arise when the values of μ_B, ϕ_A, ϕ_B, and π_C are large, i.e. 0.1, 0.1, 0.1, and 0.95, respectively.

Regarding the simulation-based standard errors for the case where $n = 2{,}000$, when we compare Figure 6.5 with Figure 6.2, we note that there is a substantial gain in precision for all estimators, except for $\hat{\phi}_A$. When we increase the sample size from 250 to 2,000, the largest standard error for $\hat{\phi}_A$ only decreases from 0.2 to 0.17. The ratios of simulation-based to asymptotic standard errors for $\hat{\mu}_A$ are close to 1 for most combinations of parameter values, although there are still few cases where the ratio is as high as 1.7 (results not shown here).

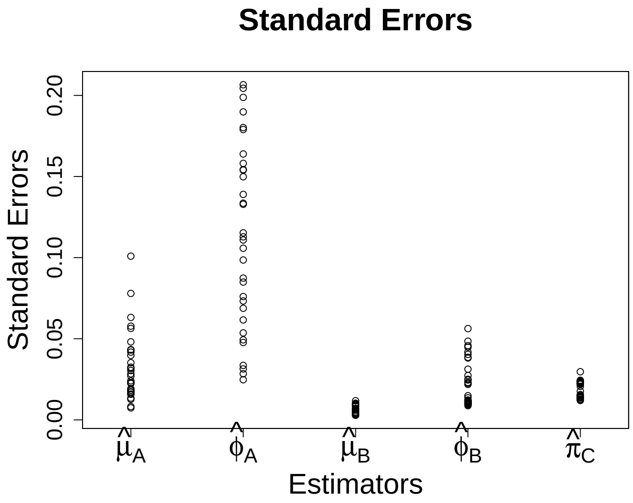

Figure 6.2 Plots of Simulation-based Standard Errors – Random-Effects Model with Standard Plan, $n = 250$, $r = 10$.

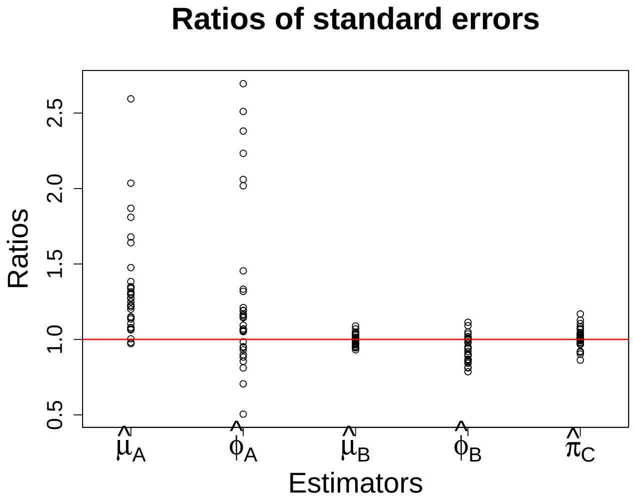

Figure 6.3 Plots of Ratios of Standard Errors – Simulation over Asymptotic – Random-Effects Model with Standard Plan, $n = 250$, $r = 10$.

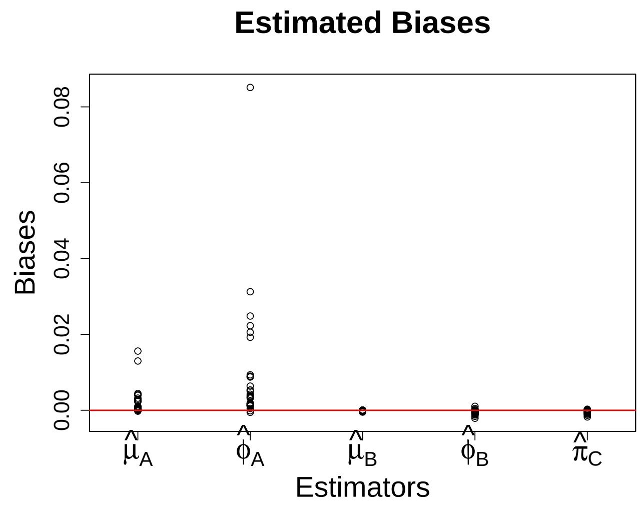

Figure 6.4 Plots of Biases – Random-Effects Model with Standard Plan, $n = 2{,}000$, $r = 10$.

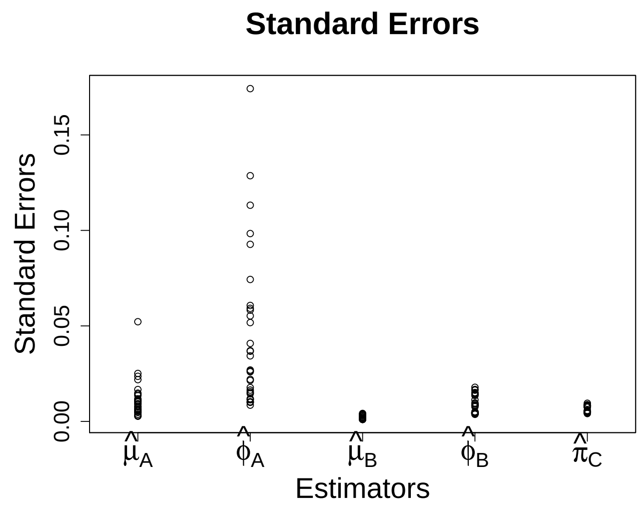

Figure 6.5 Plots of Simulation-based Standard Errors – Random-Effects Model with Standard Plan, $n = 2{,}000$, $r = 10$.

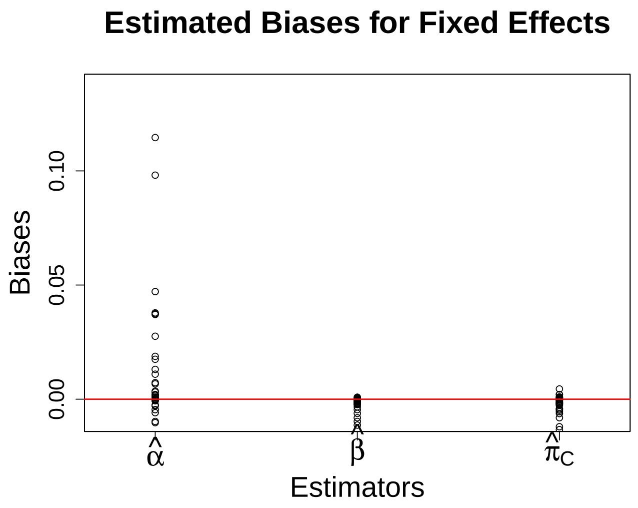

Figure 6.6 Plots of Biases – Fixed-Effects Model with Standard Plan, $n = 250$, $r = 10$.

6.1.3 Fitting the Fixed-Effects Model to Random Effects Data

For each run in the simulation studies discussed above, aside from the random-effects model parameters, we also estimate the parameters from the fixed-effects model (4.1). The results are quite different than the ones in Chapter 5, where a gold-standard system is used in the assessment study, and the corresponding fixed-effects estimators are unbiased. In the case where a gold standard is not available, the fixed-effects estimators are highly biased, as shown in Figure 6.6. Unlike the case of random-effects model estimators, the biases of the fixed-effects estimators do not substantially reduce when we increase the sample size to 2,000, especially for $\hat{\alpha}$ (results not shown here).

6.1.4 Fitting the Random-Effects Model to Fixed Effects Data

Next, we investigate the properties of the random-effects model estimators, when we fit the random-effects model to data from an assessment study where the misclassification rates are constant for conforming/nonconforming parts. As in Chapter 5, we are interested to see whether

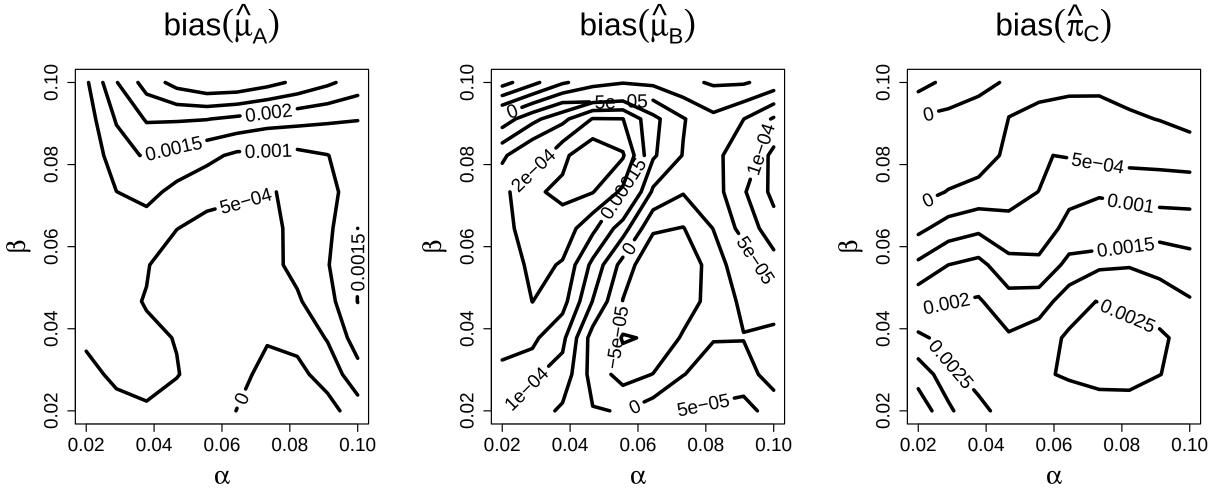

Figure 6.7 Plots of Biases – Fixed-Effects Model with Standard Plan, n = 250, r = 10, ϕ_A = 0, $\phi_B = 0$.

fitting a random-effects model produces estimates with the same accuracy and precision as the fixed-effects model. We first look at a case with a relatively small sample size (n = 250) and a large number of repeated measurements (r = 10). We simulate data from a fixed-effects model over a grid of values $\mu_A = \alpha = 0.02, 0.1$, $\mu_B = \beta = 0.02, 0.1$, and $\pi_C = 0.84, 0.94$, as in Figure 5.2. We fit both fixed- and random-effects models and find the ML estimates using the Nelder-Mead algorithm in the R environment, with the additional step where we check the value of the likelihood function at $\hat{\pi}_C = 1$. From Figure 6.7 that shows the bias results, we note that the random-effects model gives unbiased estimators for μ_B, and there is at most a small bias for μ_A and π_C. Also, as shown in Figure 6.8, the simulation-based standard errors for the random-effects estimators are generally close to the ones given by the fixed-effects model. The only exception is for $\hat{\mu}_A$ and $\hat{\pi}_C$, when μ_B and π_C are large (upper part of the grid). For these cases, fitting the random-effects model leads to some loss of precision in estimating $\mu_A = \alpha$ and surprisingly some gain in estimating π_C. We ran similar simulation studies for larger sample sizes and found out that the bias and loss in precision for the estimators given by the random-effects model disappear for sample sizes larger than 1,000. These results are different than those in Chapter 5, where a gold-standard system is available. There, fitting the random-effects model generated

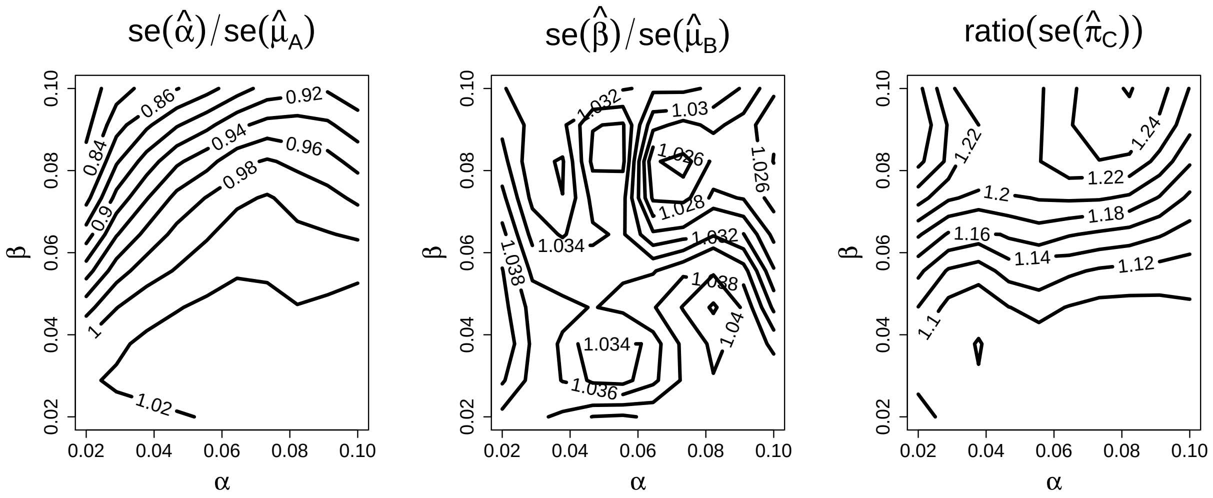

Figure 6.8 Plots of Ratios of Standard Errors - Fixed- vs. Random-Effects Model with Standard Plan, $n = 250$, $r = 10$, $\phi_A = 0$, $\phi_B = 0$.

estimators with the same accuracy and precision as the ones from a fixed-effects model, even for small sample sizes (i.e. $n = 200$). We mentioned before that the fixed-effects model is a limiting case of the random-effects model, with ϕ_A and ϕ_B approaching 0. In order to test whether the misclassification rates are constant, we need to develop a statistical test for the hypothesis $H_0 : \phi_A = \phi_B = 0$. Self and Liang (1987) discuss several likelihood-ratio tests when the true parameter is on the boundary of the parameter space. However, unlike in the case where a gold standard is available and we have two separate beta-binomial distributions, their results are not suitable for the "no gold standard" case. More work has to be done on designing a formal statistical test for the hypothesis $H_0 : \phi_A = \phi_B = 0$, in the case where a gold-standard system is not available.

After investigating the properties of the random- and fixed-effects estimators from a standard plan in the context where the misclassification rates are not constant, we conclude that, for high quality processes, i.e. π_C close to one, the standard assessment plan is not practical. It requires very large sample sizes to produce useful and reliable estimates of the primary parameter μ_A. The problem occurs because we need n very large in order to get sufficient nonconforming parts in the sample. Also, we conclude that using the fixed-effects model with the standard plan can produce badly biased estimators of the consumer's and producer's risks if the misclassification rates vary from part to part. In the case where the misclassification rates are constant and the sample sizes are small, when we fit the random-effects model there are some small biases for the estimators of μ_A (or α in this case) and π_C, for large values of conforming rate π_C and β.

6.2 Random-Effects Model – Conditional Selection with Baseline Data

To address the requirement of extreme sample sizes with the standard plan when π_C is close to one, we explore conditional sampling where we randomly select parts from the populations of previously passed and failed parts. In particular, to obtain more nonconforming parts in the sample, we over-sample from the population of previously failed parts. We also augment the assessment study with baseline data, which are available from routine inspection of parts by the studied BMS.

6.2.1 Model Formulation

With a conditional selection plan, we randomly sample n_P parts from the population of previously passed and $n_{\bar{P}}$ from the population of previously failed. Then, we measure each part r times with the BMS.

For a previously passed part, i.e. $Y_{i0} = 1$, we can write the contribution to the likelihood as:

$$\frac{\Pr(S_i = s_i, Y_{i0} = 1 \mid X_i = 1)\Pr(X_i = 1) + \Pr(S_i = s_i, Y_{i0} = 1 \mid X_i = 0)\Pr(X_i = 0)}{\Pr(Y_{i0} = 1)} \tag{6.2}$$

Using the two Beta distributions (5.1) and (5.3) for A and B, respectively, we can further write 6.2

as:

$$\frac{\binom{n}{k}}{\pi_P}\left[\frac{Beta(r-s_i+\mu_B\frac{1-\phi_B}{\phi_B},s_i+1+(1-\mu_B)\frac{1-\phi_B}{\phi_B})}{Beta(\mu_B\frac{1-\phi_B}{\phi_B},(1-\mu_B)\frac{1-\phi_B}{\phi_B})}\pi_C+\right.$$
$$\left.\frac{Beta(s_i+1+\mu_A\frac{1-\phi_A}{\phi_A},r-s_i+(1-\mu_A)\frac{1-\phi_A}{\phi_A})}{Beta(\mu_A\frac{1-\phi_A}{\phi_A},(1-\mu_A)\frac{1-\phi_A}{\phi_A})}(1-\pi_C)\right] \tag{6.3}$$

Similarly, for a previously failed part the contribution to the likelihood is:

$$\frac{\binom{n}{k}}{1-\pi_P}\left[\frac{Beta(r-s_i+1+\mu_B\frac{1-\phi_B}{\phi_B},s_i+(1-\mu_B)\frac{1-\phi_B}{\phi_B})}{Beta(\mu_B\frac{1-\phi_B}{\phi_B},(1-\mu_B)\frac{1-\phi_B}{\phi_B})}\pi_C+\right.$$
$$\left.\frac{Beta(s_i+\mu_A\frac{1-\phi_A}{\phi_A},r-s_i+1+(1-\mu_A)\frac{1-\phi_A}{\phi_A})}{Beta(\mu_A\frac{1-\phi_A}{\phi_A},(1-\mu_A)\frac{1-\phi_A}{\phi_A})}(1-\pi_C)\right] \tag{6.4}$$

For the CS plans augmented with baseline data, we also have the contribution:

$$L_b(\pi_P)\propto \pi_P^{m_P}(1-\pi_P)^{m-m_P}$$

where m_P is the number of parts passing the initial inspection out of m, and $\pi_P=(1-\mu_B)\pi_C+\mu_A(1-\pi_C)$. The overall likelihood function is a product of n_P terms as in (6.3), $n_{\bar{P}}$ terms as in (6.4), and the baseline likelihood function.

6.2.2 Bias and Precision of the Estimators

We conduct simulation studies for different CS plans with baseline data, similar to the ones for the SP. Each study has a 2^5 factorial structure, where we vary the model parameters $\mu_A=0.02,0.1$, $\mu_B=0.02,0.1$, $\phi_A=0.01,0.1$, $\phi_B=0.01,0.1$, and $\pi_C=0.85,0.95$. We look at cases where we randomly select parts only from the population of previously failed parts, i.e. $f=0$, and we measure each part $r=10$ times with the BMS. We add baseline data of size $m=1{,}000$ and look

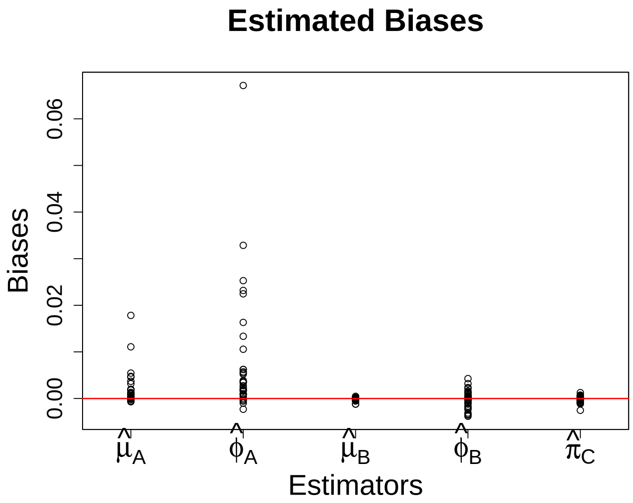

Figure 6.9 Plots of Biases – Random-Effects Model with Conditional Selection Plan, $n = 250$, $r = 10$, $m = 1{,}000$.

at sample sizes of $n = 250, 500, 1{,}000$. We maximize the overall likelihood function numerically, using a Nelder-Mead algorithm under the constraints $\mu_A + \mu_B < 1$, $\mu_A + 2\phi_A - \mu_A\phi_A < 1$, and $\mu_B + 2\phi_B - \mu_B\phi_B < 1$. We use 500 simulation runs for each combination of parameter values, for a given sample size, $r = 10$ and $m = 1{,}000$. Then, we obtain the ML estimates for both fixed- and random-effects models parameters. For each combination of parameter values, we estimate the biases and get the standard errors of the estimates.

Figure 6.9 shows the estimated biases for all random-effects estimators, for $n = 250$. We note that there is some bias for $\hat{\mu}_A$ (up to 0.018) and large bias for $\hat{\phi}_A$ (up to 0.067) that arise when all parameters are at their highest value. The estimators of μ_B and π_C are unbiased, whereas $\hat{\phi}_B$ has some small bias. We also look at the fixed-effects estimators (results not shown here), and find very large biases for the estimator of μ_B (up to 0.05) and some bias for the estimators of μ_A and π_C.

However, the random-effects estimators for μ_A and ϕ_A when we use a CS plan with $f = 0$ and $m = 1{,}000$ are overall much more accurate than the ones from a SP, for the same sample size $n = 250$. Recall that for the SP with $n = 250$, biases for $\hat{\mu}_A$ were as high as 0.036, and for $\hat{\phi}_A$ as

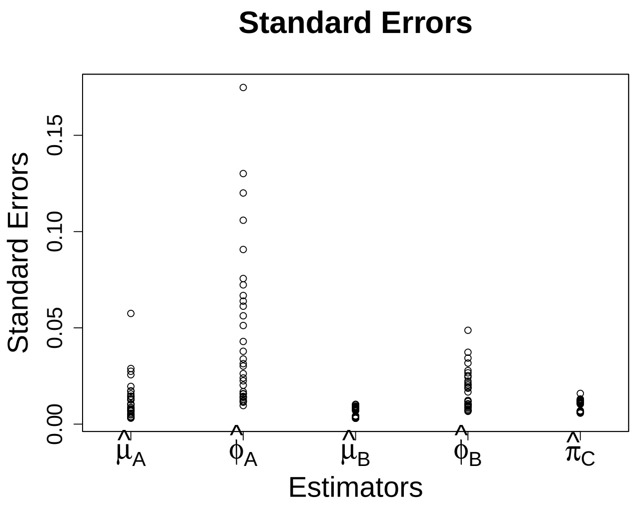

Figure 6.10 Plots of Simulation-based Standard Errors – Random-Effects Model with Conditional Selection Plan, $n = 250$, $r = 10$, $m = 1{,}000$.

high as 0.136 (see Figures 6.1 and 6.9). Using the CS plan leads to a decrease in bias of almost 50%.

Using a CS plan with $f = 0$ and baseline of size $m = 1{,}000$ gives estimators with reasonable precision, including $\hat{\mu}_A$. Except for the cases where all parameters are at their highest values, the simulation-based standard errors for $\hat{\mu}_A$ are less than 0.025, which is a substantial improvement compared to the SP, where standard errors go as high as 0.1 (Figure 6.2).

However, the estimators for μ_A given by the SP with $n = 250$ were highly biased for some combinations of the parameter values, and comparing their precision with the one given by the CS plan is not relevant. Therefore, we compared the precision of the estimators from SP and CS with $f = 0$ and $m = 1{,}000$ for larger sample sizes, i.e. $n = 1{,}000$, so that we have approximately unbiased estimators of μ_A, μ_B, ϕ_B, and π_C, for both plans. We made this comparison more for “scientific” reasons, as in practice we would always recommend using a selection plan that needs a smaller sample size in order to yield unbiased estimators, that is a CS plan with $f = 0$ and baseline. Figure 6.11 shows the ratios of the simulation-based standard errors for the SP estimates to the CS ones. We note that the CS plan substantially improves the precision of the estimator of μ_A for

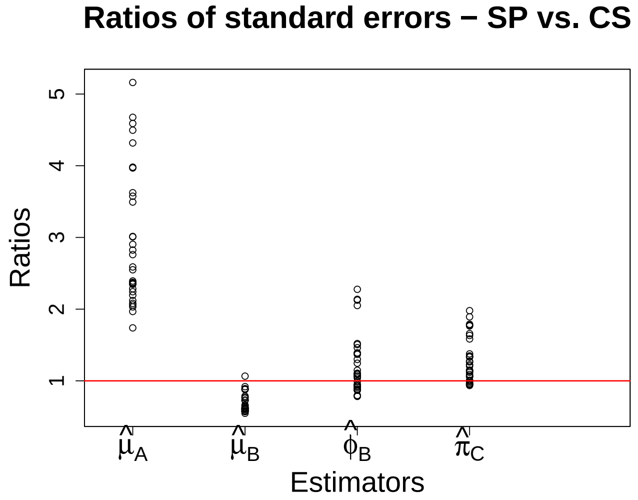

Figure 6.11 Plots of Ratios of Standard Errors – SP over CS with $f = 0$ and $m = 1{,}000$ – Random-Effects Model with $n = 1{,}000$, $r = 10$.

all combinations of parameter values . The CS plan also gives better precision than the SP for $\hat{\pi}_C$, except for few cases where the two plans have similar performance. The situation is different for the estimator of μ_B, which has a better precision when we use the SP, for almost all combinations of parameter values. When parts are selected only from the population of previously failed parts (CS plan with $f = 0$) the expected number of conforming parts in the sample is (much) smaller than in the case where parts are randomly sampled from the population of manufactured parts (SP). Therefore, we expect that $\hat{\mu}_B$ would have better precision with a SP plan. However, in the previous chapters, we noted that adding the information about the pass rate provided by the baseline measurements substantially improves the precision of $\hat{\mu}_B$ (or $\hat{\beta}$ in the fixed-effects model case). The baseline data provide an estimate of $\pi_P = (1-\mu_B)\pi_C + \mu_A(1-\pi_C)$. As in the cases considered here π_C is large and μ_A is small, thus having an estimate of π_P provides a constraint for the estimates of μ_B and π_C, and hence better precision for those parameters. Therefore, we increase the baseline size to $m = 5{,}000$ and compare the precision of $\hat{\mu}_B$ given by the SP versus the CS plan. There are fewer cases where the SP gives more precise estimators than the CS plan (results not shown here), and all of them happen when $\phi_B = 0.1$. Next, we increase

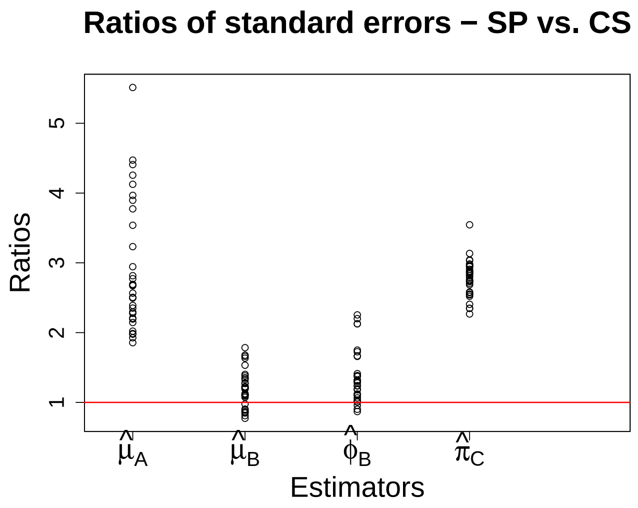

Figure 6.12 Plots of Ratios of Standard Errors – SP over CS with $f = 0$ and $m = 10{,}000$ – Random-Effects Model with $n = 1{,}000$, $r = 10$.

the baseline size to $m = 10{,}000$ and we note in Figure 6.12 that the CS plan gives more precise estimators for μ_B, for almost all combinations of parameter values. The same results apply to the estimator of ϕ_B. However, no matter the sample size, μ_B is generally better estimated than μ_A; therefore, for some combinations of parameter values, with the CS plan with $f = 0$ and baseline we might lose some precision for $\hat{\mu}_B$, but we gain a lot of precision for $\hat{\mu}_A$.

We also compare the simulation-based standard errors with the asymptotic ones. For small sample sizes ($n = 250$), the asymptotic results do not provide good approximations (results not shown here). We look at larger sample sizes, and although the asymptotic approximations work better, there are still some cases where they underestimate the standard errors of $\hat{\mu}_A$ and $\hat{\phi}_A$ (see Figure 6.13). In the case where we have to analyze data from a BMS assessment study and are interested in getting the ML estimates of the random-effects model parameters and their corresponding standard errors, especially for small sample and baseline sizes, we recommend using nonparametric or parametric bootstrapping techniques as proposed by Efron and Tibshirani (1994), and De Menezes (1999). In the nonparametric bootstrapping case, we generate new data by randomly drawing with replacement n parts from the initial sample and

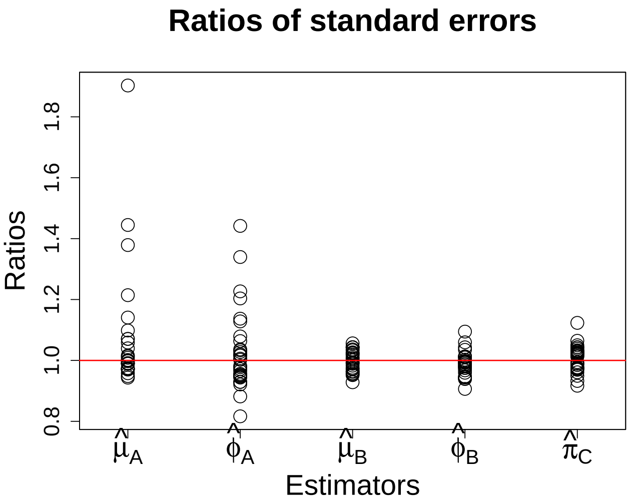

Figure 6.13 Plots of Ratios of Standard Errors – Simulation over Asymptotic – Random-Effects Model with Conditional Selection Plan, $n = 1{,}000$, $r = 10$, $m = 1{,}000$.

then get the ML estimates based on the new sample. Note that, for each re-sampled part i, we use the initial number of passes out of r measurements s_i. The procedure is repeated B times ($B \geq 500$). The standard errors of the estimates are given by the sample standard deviations based on the new B samples. In the parametric bootstrapping case, we estimate the random-effects model parameters from the original sample and then we generate a random sample of the same size as the original one, where both the true quality status x and the number of passes s are generated for each of the n parts. However, in the ML estimation we only use the total number of passes $s_i, i = 1, \ldots, n$.

We conclude that, for most combinations of parameter values, the CS plan with $f = 0$ and $m = 1{,}000$ works reasonably well even for small sample sizes ($n = 250$). This, along with the fact that the CS plan with $f = 0$ augmented with baseline data is easy to implement, gives a clear advantage to the CS plan compared to the standard plan, where we need (much) larger sample sizes to get unbiased estimators. This design also yields useful standard errors for the main parameters μ_A, μ_B, and π_C.

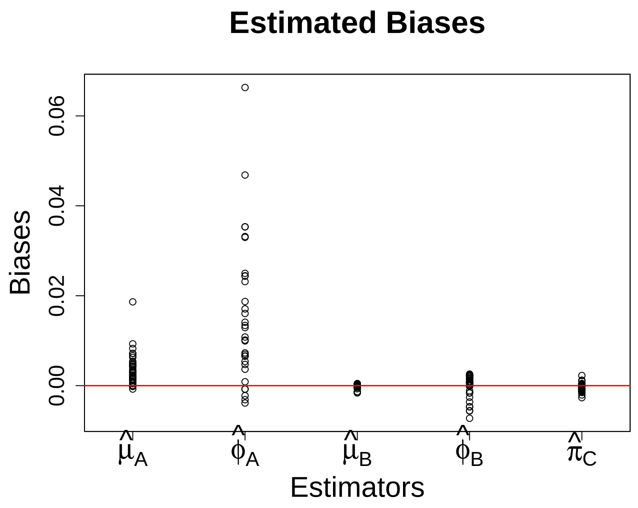

Figure 6.14 Plots of Biases – Random-Effects Model with Standard Plan Using the Ad Hoc Procedure, $n = 250$, $r = 10$.

6.2.3 Possible Solution to Large Biases of the Random-Effects Model Estimators – Constrained Random-Effects Model

In the previous sections, we concluded that the estimators of μ_A and ϕ_A given by a random-effects model with a standard plan and small sample sizes were highly biased, especially for large values of ϕ_B and π_C. Also, we found that the estimators given by a CS plan with baseline data were biased for small sample sizes, although the biases were smaller than in the SP case. For both selection plans, we looked at the cases where the estimates of both μ_A and ϕ_A were unreasonably large and found out that these estimates were actually on the boundary $\mu_A + 2\phi_A - \mu_A\phi_A = 1$. We propose to handle these cases where the estimates are on the boundary, by re-fitting the random-effects model with the further constraint that $\phi_A = \phi_B = \phi$. The result of this proposal is that, when needed, we borrow strength from the better preforming estimate of $\hat{\phi}_B$. Using a model with a common ϕ is similar to the so-called $2LCR1$ model described by Qu et al. (1996), and equivalent to the Dirichlet model of Fujisawa and Izumi (2000). We demonstrate the effect of this ad hoc procedure in Figure 6.14, for the SP, and in Figure 6.15, for the CS plan with $f = 0$, $m = 1{,}000$, both plans with a sample size of $n = 250$ and a number of repeated measurements

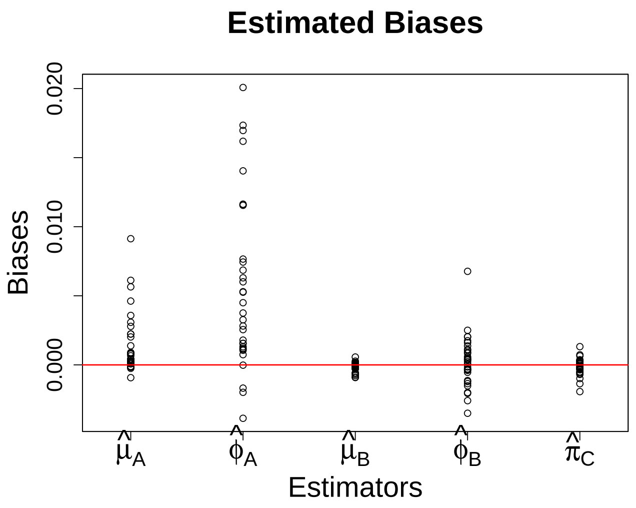

Figure 6.15 Plots of Biases – Random-Effects Model with Conditional Selection Plan Using the Ad Hoc Procedure, $n = 250$, $r = 10$, $m = 1{,}000$.

$r = 10$. For both SP and CS, we note that the biases of $\hat{\mu}_A$ are generally smaller than in the case where we only use the unconstrained model (see Figures 6.1 and 6.9). However, the CS with baseline gives better estimates than the SP, even when we use the additional ad hoc procedure. For $n = 250$ and $r = 10$, the CS with $m = 1{,}000$ provides fairly good estimates for all primary parameters μ_A, μ_B, and π_C. However, the biases and standard errors of $\hat{\phi}_A$ and $\hat{\phi}_B$ are too large for the estimates to be reliable and we still need large sample sizes to estimate these two parameters with good accuracy and precision.

There are other solutions to the problem of highly biased estimators of μ_A and ϕ_A. One possible approach is applicable only in cases where a gold-standard system is available, but otherwise too expensive to be used on a whole sample of parts. In that case, we can use the gold standard to measure the "problem" parts, that is, parts that pass the BMS inspection around $r/2$ times out of r. We believe that most estimation problems arise from the fact that, in these cases, the random-effects model with latent class is not able to distinguish between conforming and nonconforming parts. By using the gold standard, we can now determine the true quality state of these parts, i.e. conforming or nonconforming. Adding this piece of information to the

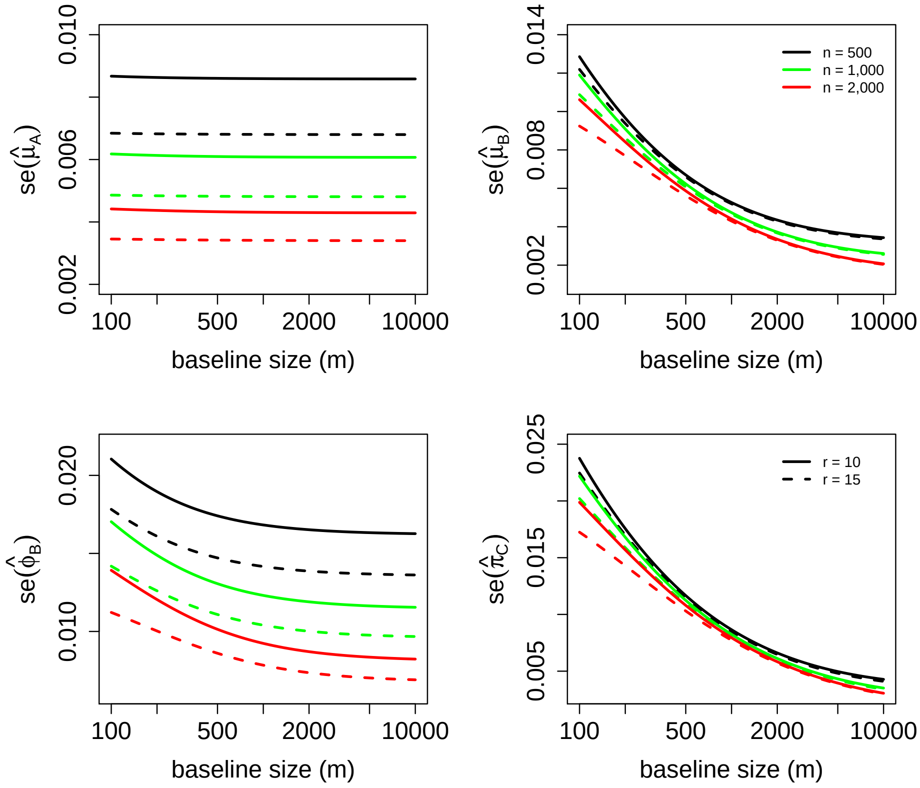

Figure 6.16 Effect of Changing the Baseline Size m on Standard Error of the Estimators – Random-Effects Model with Conditional Selection, $\mu_A = 0.05$, $\phi_A = 0.1$, $\mu_B = 0.05$, $\phi_B = 0.1$, and $\pi_C = 0.9$

likelihood function will eliminate the confusion that the random effects model with latent class usually faces. Another option for dealing with large biases would be to collect more data by increasing the total number of measurements $r \times n$ or the baseline size m.

6.2.4 Effect of Changing the Design Parameters on the Precision of Estimators

In this section, we investigate the effect of changing the design parameters m, n, and r, on the precision of the estimators from a random-effects model with conditional selection and baseline data. Since we are assuming that the conforming rate π_C is large, we look at the case $f = 0$, where we sample only from previously failed parts. We also assume the baseline size m is

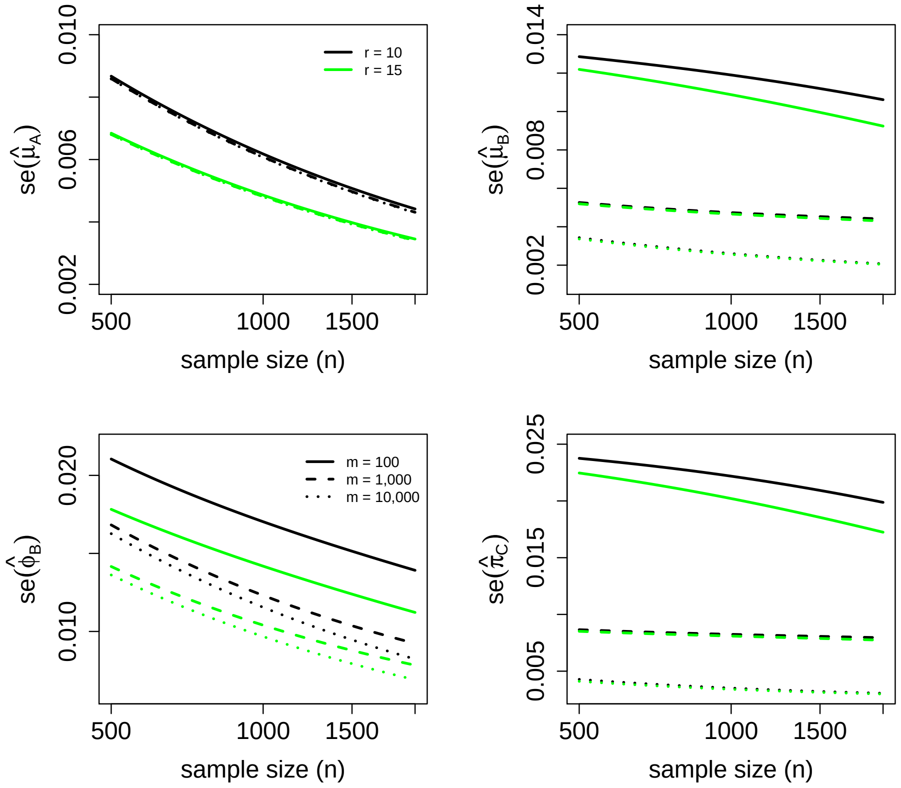

Figure 6.17 Effect of Changing the Sample Size n on Standard Error of the Estimators – Random-Effects Model with Conditional Selection, $\mu_A = 0.05$, $\phi_A = 0.1$, $\mu_B = 0.05$, $\phi_B = 0.1$, and $\pi_C = 0.9$

determined separately from the other plan parameters n and r. We start by comparing different choices for n, r, and m using the asymptotic standard errors from the Fisher information. As with the gold-standard case in Chapter 5, we look at design and model parameter values for which the asymptotic standard errors are good approximations. Therefore, we consider cases where $m = 100, 1{,}000, 10{,}000$, $r = 10, 15$, and $n = 500, 1{,}000, 2{,}000$, and moderate values for the model parameters, i.e. $\mu_A = 0.05$, $\phi_A = 0.1$, $\mu_B = 0.05$, $\phi_B = 0.1$, and $\pi_C = 0.9$. Note that for some combinations of n, r, and m, the total number of measurements $(n \times r + m)$ is unreasonably large. We actually look at these values only to get an idea about the effect of changing the design parameters on the precision of the estimators and these values are not necessarily recommended for study planning.

In Figure 6.16, we note that increasing the baseline size m has no effect on the precision of

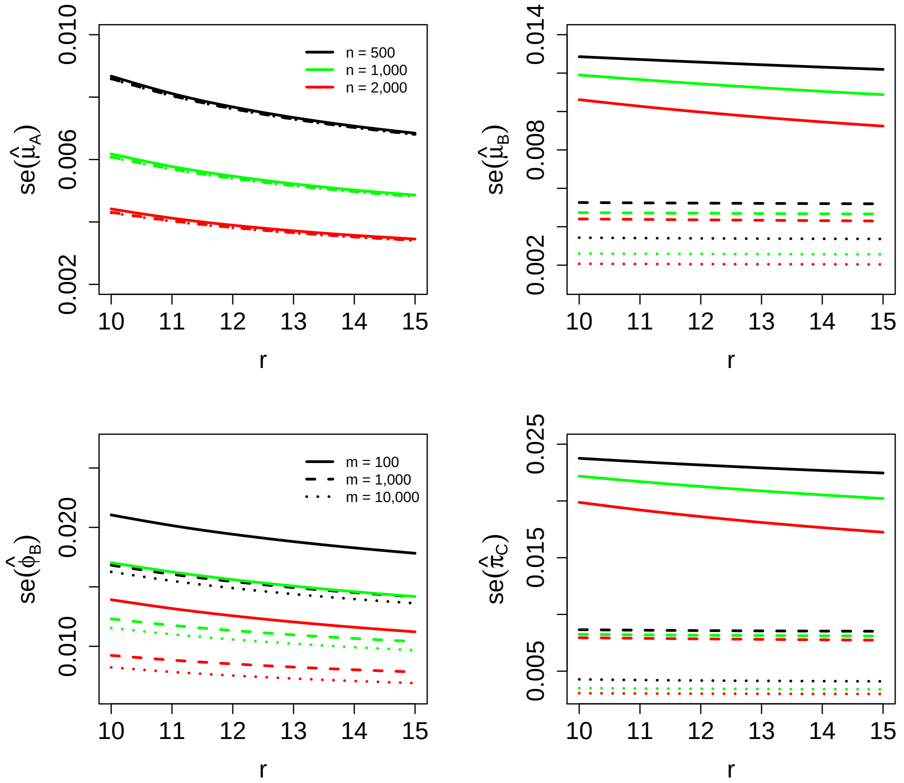

Figure 6.18 Effect of Changing the Number of Repeated Measurements r on Standard Error of the Estimators – Random-Effects Model with Conditional Selection, $\mu_A = 0.05$, $\phi_A = 0.1$, $\mu_B = 0.05$, $\phi_B = 0.1$, and $\pi_C = 0.9$

the estimator of μ_A. The curves for the asymptotic standard errors are all parallel with slope zero. The curves corresponding to $\hat{\mu}_B$ and $\hat{\pi}_C$ are very similar, with a dramatic increase in precision as m gets larger. For $n = 500$ and $r = 10$ (see black solid lines on the second and fourth panels of Figure 6.16), the standard errors of $\hat{\mu}_B$ and $\hat{\pi}_C$ decrease by 73% and 82%, respectively, as m goes from 100 to 10,000. When m is close to 500, the curves for $r = 10$ (solid lines) and $r = 15$ (dashed lines) overlap, for all sample sizes n considered in the study. Also, for both $\hat{\mu}_B$ and $\hat{\pi}_C$, the standard errors approach a limit as m gets larger. For $\hat{\phi}_B$, increasing m reduces the standard error, but with a smaller rate than for $\hat{\mu}_B$. For $n = 500$ and $r = 10$ (black solid line in third panel of Figure 6.16), there is a 22% decrease in the standard error of $\hat{\phi}_B$ when m goes from 100 to 2,000. For $m > 2{,}000$, the precision of $\hat{\phi}_B$ is not much affected by the baseline size.

Next, we investigate the effect of increasing the sample size n on the precision of the estimators. Figure 6.17 shows the curves of the asymptotic standard errors, for $m = 100, 1{,}000, 10{,}000$ and $r = 10, 15$, with the sample size varying from $n = 500$ to 2,000. We note that the curves for the standard error of $\hat{\mu}_A$ are identical for different values of m, which agrees with the results seen in Figure 6.16. For each value of r, the standard errors of $\hat{\mu}_A$ change at a $1/\sqrt{n}$ rate. For $\hat{\mu}_B$ and $\hat{\pi}_C$, we note some moderate effect of the sample size on the standard errors when the size of the baseline is small, i.e. $m = 100$ (see solid black and green solid lines in the second and fourth panels of Figure 6.16). For $m = 100$ and $r = 10$, the standard errors of $\hat{\mu}_B$ and $\hat{\pi}_C$ decrease by 17% and 16%, respectively, whereas for $r = 15$ by 24% and 23%. For larger values of m, the curves overlap for different values of r, with very small slopes. The effect of changing n on the precision of $\hat{\phi}_B$ is more dramatic and is similar for different values of r. For $m = 100$ and $r = 10$, the standard error of $\hat{\phi}_B$ decreases by 34% when n varies from 500 to 2,000.

Increasing the number of repeated measurements r, when n and m are held constant, has a moderate effect on the standard error of $\hat{\mu}_A$ (see Figure 6.18). For $n = 500$, the asymptotic standard error of $\hat{\mu}_A$ decreases by 21% when r varies from 10 to 15. The effect of increasing r is quite similar for $n = 1{,}000$ and $n = 2{,}000$. Also, increasing r leads to a moderate decrease in the standard errors of $\hat{\mu}_B$ and $\hat{\pi}_C$, for small baseline sizes (i.e. $m = 100$). For larger baseline sizes, the effect of increasing r is minimal, for all sample sizes considered here. For the estimator of ϕ_B, increasing r results in a decrease of the asymptotic standard error by 15%, for $n = 500$ and $m = 100$. Also, the standard error curves are similar for different baseline sizes, for the same sample size n.

We also looked at other values for the model parameters (results not shown here). We found out that, for larger values of the parameters, i.e. $\mu_A = 0.1$, $\mu_B = 0.1$, and $\pi_C = 0.95$, the effect of changing the design parameters on the asymptotic standard errors of the estimates is similar to that in the case discussed above. There were some differences in the effect of the baseline size m on the asymptotic standard errors of $\hat{\mu}_A$. That is, for $r = 10$, increasing m decreased the standard error of $\hat{\mu}_A$ by a small amount, whereas in the previous case it did not have any effect. Also, the standard errors of $\hat{\pi}_C$ did not converged to the same limit for all values of r and n. Regarding the effect of changing the sample size n, we noted the same patters as in the previous case, with the only difference that the effect of increasing the sample size on the standard errors of $\hat{\mu}_B$ and $\hat{\pi}_C$ is now larger. Also, the effect of increasing the number of repeated number of measurements r on the standard errors is stronger for all estimators.

For planning an assessment plan, we recommend using a CS plan with parts randomly sampled only from the population of previously failed parts, augmented with baseline data. This plan provides the largest expected number of nonconforming parts for a given sample size and combination of parameter values. We also recommend using as many baseline measurements as possible, with a minimum of $m = 1{,}000$. For high-quality manufacturing processes, i.e. the conforming rate π_C is large, we need relatively large sample sizes to get useful estimates, especially for μ_A and ϕ_A. We provide an algorithm coded in the R Environment (2005) that provides feasible combinations of n and r that achieve prespecified precision for the estimators of μ_A and μ_B, similar to the one for the fixed-effects model in Chapter 4.

6.3 Summary

In this chapter, we focus on the case where no gold-standard system is available and, therefore, the quality state of each part is unknown. Also, we consider the case where the misclassification rates are not constant within the conforming/nonconforming parts. We first investigate the properties of the random-effects model in the case where parts are randomly selected from the population of manufactured parts, i.e. we use the standard plan. We find out that, for realistic values of the model parameters and large number of repeated measurements ($r = 10$), the standard plan needs unreasonably large sample sizes ($n \geq 2{,}000$) in order to get unbiased estimators with useful standard errors. We also conclude that fitting the fixed-effects model when the misclassification rates are not constant yields biased estimators. Fitting the random-effects model to data with constant misclassification rates leads to almost no loss of precision for $n \geq 1{,}000$.

Next, we look at the properties of the random-effect estimators in the case where parts are randomly selected only from the population of previously failed parts, i.e use conditional selection with $f = 0$, and we use the information about the pass rate from the baseline data of size m. We recommend $f = 0$, since it increases the expected number of nonconforming parts in the sample and failed parts are usually readily available. We find out that the CS plan supplemented with baseline data gives estimators with reasonable accuracy and precision for the main parameters, μ_A, μ_B, and π_C, even for small sample sizes (i.e. $n = 250$). Using a standard plan with the same sample sizes and number of repeated measurements yields estimators with large biases, especially for μ_A. For samples where the estimates of μ_A and ϕ_A seem unreasonable, which occasionally occur when the values of μ_B, ϕ_B, and π_C are large, we propose a random-effects model that introduces the "common ϕ" constraint, that is, $\phi_A = \phi_B = \phi$. This method reduces the biases for $\hat{\mu}_A$, but not for $\hat{\phi}_A$ and $\hat{\phi}_B$, when we hold n and r constant. We also suggest other solutions, such as collecting more data or using the gold standard for parts that passed the BMS inspection close to half of the time.

We also look at the effect of changing the design parameters m, n, and r, on the asymptotic standard errors of the estimates from a random-effects model when we use a CS with $f = 0$ and baseline measurements. We also give some general planning recommendations for a BMS assessment study.

Regarding testing whether we need a random-effects model or not, that is checking whether the misclassification rates are constant within conforming/nonconforming parts, we need to develop a formal statistical test to check the hypothesis $\phi_A = \phi_B = 0$, similarly to the gold-standard case in Chapter 5. This is part of future research on the topic of random-effects models.

The main results presented in this chapter are also included in the paper "Assessing a Binary Measurement with Varying Misclassification Rates" by Danila, Steiner, and MacKay (2012), accepted for publication in the Journal of Quality Technology in December 2011.

Chapter 7

Conclusions and Future Work

7.1 Summary

In this thesis, we investigate the assessment of binary measurement systems (BMS) that are commonly found in both manufacturing industry and medical diagnostic testing. Here we adopt the manufacturing language and notation. We suppose the BMS has been in use for routine systematic inspection of parts. We consider assessment plans in two contexts. In one, we have available a gold-standard system that classifies parts without error. In the second, we do not have such a gold-standard system or it is too expensive or time consuming to use in the assessment study.

We suppose each part has a binary true status, i.e. it is conforming or nonconforming. The performance of the BMS is characterized by two misclassification rates, the false positive or the consumer's risk, α, and the false negative or the producer's risk, β, or by their complements, the specificity and sensitivity. The manufacturing process is characterized by the conforming rate, π_C. We concentrate on high-quality processes, i.e. we assume that π_C is close to 1 and α and β are small. The goal of a BMS assessment study is to estimate both misclassification rates and the conforming rate.

In some processes, the true status of each part is binary but the misclassification rates vary from part to part. For example, in the case where a visual BMS inspects parts for surface cracks, the actual size of the crack has an effect on the BMS being able to detect the flaw and classify the part accordingly. There may exist one or more latent variables, not measured during the assessment study or the routine use of the BMS, whose value influences the probability that the BMS misclassifies the part. In these situations, we adopt a model that allows the misclassification rates α and β to vary within the populations of nonconforming/conforming parts.

The main goal of this thesis is to investigate the assessment of a BMS in a situation commonly found in manufacturing processes. Large collections of measured parts are available prior to the BMS assessment study. Also, the BMS tracks the number of parts passed over a certain period of time, therefore providing baseline information about the pass rate, π_P. If we assume that, during the time the baseline measurements are recorded, the process is stable and the properties of the BMS do not change, then we can use the baseline information in the assessment study. We note here that the baseline measurements are not part of the BMS assessment study and come as free measurements. With these features in mind, we investigate a sampling scheme that involves random selection of parts from the previously passed and failed parts (i.e. conditional selection). Throughout this thesis, we look at the properties of the estimators given by the conditional selection plan augmented with baseline data and compare them with the corresponding estimators given by the current (standard) assessment plans.

In Chapter 1, we give an introduction to the general context of binary measurement systems in both manufacturing industry and medical diagnostic testing. In Chapter 2, we review various assessing methods proposed in the literature that are relevant to our context of interest.

In Chapters 3 and 4, we look at the cases where we assume that the misclassification rates α and β are constant within the nonconforming/conforming parts. Chapter 3 focuses on the case where a gold standard is available for the assessment study, whereas in Chapter 4, we look at the case where a gold standard is too expensive, time consuming or does not exist. In Chapters 5 and

6, we consider the same two cases but allow the misclassification rates to vary from part to part.

In the situation where π_C is close to 1 and α and β are small, we show that we require very large sample sizes with the standard plans to get useful estimates, i.e. estimates with standard errors materially smaller than the estimates themselves. In general, for the same sampling effort and cost, the use of freely available baseline data combined with conditional sampling produces substantially better estimates of the parameters of interest. Or, put another way, by using the recommended plans, we can get useful estimates with much smaller sample sizes.

To our knowledge, the ideas of a conditional selection plan augmented with baseline data have not been investigated in the literature nor implemented in practice. Since there is little added cost or complexity, we strongly recommend their application in routine BMS assessment studies.

7.2 Future Research and Extensions

7.2.1 Testing Various Hypotheses in the Case where the Misclassification Rates Vary

In Chapters 5 and 6, we investigate the properties of the random-effects model estimators and we conclude that we need very large sample sizes to get reasonable estimates for the variability parameters ϕ_A and ϕ_B. It would be useful to develop testing procedures for checking several hypotheses such as:

- $H_0 : \phi_A = \phi_B = 0$, i.e. both misclassification rates are constant
- $H_0 : \phi_A = 0$, i.e. the false positive rate is constant
- $H_0 : \phi_B = 0$, i.e. the false negative rate is constant
- $H_0 : \phi_A = \phi_B = \phi$, i.e. both the false positive and false negative rates are varying, but share the same variation parameter ϕ

Self and Liang (1987) give guidelines for developing likelihood-ratio testing procedures for the case where the parameters are at the boundary, which is the case with the first three hypotheses listed above. However, none of the cases discussed in this paper directly apply to our context where we use a conditional selection of parts. Therefore, more work has to be done on how to develop hypotheses tests in the case where we use the recommended plan, i.e. conditional selection, in the contexts where a gold standard is available or not.

7.2.2 Extension to Multiple Binary Measurement Systems

In industrial practice, there are cases where multiple BMSs work in parallel and routinely test manufactured parts. Therefore, we might be interested in assessing the performance of these systems at the same time. In medical diagnostic testing, there is an extensive literature on testing procedures of two or more tests, when a gold standard is not available (Pepe, 2003; Hui and Walter, 1980). However, for various reasons, including ethical considerations, subjects are not repeatedly measured by the same testing procedure. Therefore, it would be interesting to develop new methods for assessing the performance of several BMSs, where parts are repeatedly measured by these systems during the assessment study, especially if we allow for part-to-part varying misclassification rates.

Also, we can investigate how using a conditional selection plan augmented with baseline data can improve the estimation of the parameters. We can assume that the conforming rate is common, that is, the BMSs measure parts coming from the same manufacturing process. Also, we can think of two scenarios. First, during the routine measurement, parts are marked by the BMS that inspects them, so that we can get separate baseline information for the studied BMSs. The second scenario, parts are not marked by the BMSs and the baseline measurements give information about a common (mixed) pass rate.

7.2.3 Using Conditional Selection with Baseline in an "Anchored" Model

In Chapter 2, we discuss an approach for assessing a BMS when no gold standard is available that involves the use of a system with known performance characteristics. The statistical model used in this context is called the "anchored" model by Boyles (2001). This testing procedure has been investigated only in the context where parts are randomly sampled from the population , i.e. we use a standard plan. It would be interesting to investigate how using a conditional selection augmented with baseline measurements affects the properties of the parameter estimators. An added complication is that parts can be measured more than once with the anchored system.

Bibliography

AIAG (2002). *Measurement Systems Analysis* (3rd ed.). Southfield, MI: AIAG.

Barron, B. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics 33*, 414–418.

Begg, C. B. and R. A. Greenes (1983). Assessment of diagnostic test when disease verification is subject to selection bias. *Biometrics 39*, 207–215.

Boss, I. (1954). Misclassification in 2x2 tables. *Biometrics 10*, 474–486.

Boyles, R. A. (2001). Gage capability for pass-fail inspection. *Technometrics 43*, 223–229.

Burke, R. J., R. D. Davis, F. C. Kaminsky, and A. E. P. Roberts (1995). The effect of inspector errors on the true fraction non-conforming: an industrial experiment. *Quality Engineering 7*, 543–550.

Casella, G. and R. L. Berger (2002). *Statistical Inference.* Pacific Grove CA: Duxbury Press.

Cleveland, W. S., E. Grosse, and W. M. Shyu (1992). *Statistical Models in S.* Chapman & Hall. Chapter 8.

Danila, O., S. H. Steiner, and R. J. MacKay (2008). Assessing a binary measurement system. *Journal of Quality Technology 40*(3), 310–318.

Danila, O., S. H. Steiner, and R. J. MacKay (2010). Assessment of a binary measurement system in current use. *Journal of Quality Technology 42*(2), 152–164.

Danila, O., S. H. Steiner, and R. J. MacKay (2011). Assessing a binary measurement system with varying misclassification rates when a gold standard is available. *Technometrics*. Submitted August 2011.

Danila, O., S. H. Steiner, and R. J. MacKay (2012). Assessing a binary measurement system with varying misclassification rates. *Journal of Quality Technology*. Accepted December 2011.

Dawid, A. P. and A. M. Skene (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics 28*, 20–28.

De Mast, J., W. N. Van Wieringen, and T. P. Erdmann (2011). Measurement system analysis for binary inspection: Continuous versus dichotomous measurands. *Journal of Quality Technology 43*(2), 99–112.

De Menezes, L. M. (1999). On fitting latent class models for binary data: the estimation of standard errors. *British Journal of Mathematical and Statistical Psychology 52*, 149–168.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Dendukuri, N. and L. Joseph (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics 57*, 158–167.

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

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

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

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

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics 29*, 637–648.

Gutjahr, P. L., H. Jung, C. Herzog, and J. A. Lunn (1982). Detecting tuberculin sensitivity. *Lancet 768*, 768.

Hadgu, A., N. Dendukuri, and J. Hilden (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test. *A review of the statistical and epidemiologic issues. Epidemiology 16*, 604–612.

Haitovsky, Y. and J. Rapp (1992). Conditional resampling from misclassified multinomial data with application to sampling inspection. *Technometrics 34*, 473–483.

Hui, S. L. and S. D. Walter (1980). Estimating the error rates of diagnostic test. *Biometrics 36*, 167–171.

Johnson, N. L., S. Kotz, and R. N. Rodriguez (1986). Statistical effects of imperfect sampling: double sampling and link sampling. *Journal of Quality Technology 18*, 116–138.

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

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

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

Maplesoft (2009). *Maple 13.* Waterloo, ON.

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

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

Montgomery, D. (1996). *Introduction to Statistical Quality Control* (Third ed.). New York: Wiley.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *Computer Journal 7*, 308–313.

Olin, B. D. and W. Q. Meeker (1996). Applications of statistical methods to nondestructive evaluation. *Technometrics 38*, 95–112.

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

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

Quade, D., P. A. Lachenbruch, F. S. Whaley, D. K. McClish, and R. W. Haley (1980). Effects of misclassifications of statistical inference in epidemiology. *American Journal of Epidemiology 111*(5), 503–515.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available from: `http://www.R-project.org`.

Rindskopf, D. and W. Rindskopf (1986). The value of latent class analysis in medical diagnosis. *Statistics in Medicine 5*, 21–27.

Rogan, W. and B. Gladen (1978). Estimating prevalence from the results of screening test. *Am.Journal of Epidemiol. 107*, 71–76.

Rutjes, A. W. S., J. B. Reitsma, K. S. Coomarasamy, K. S. Khan, and P. M. M. Bossuyt (2007). Evaluation of diagnostic tests when there is no gold standard. a review of methods. *Health Technology Assessment 11*(50), 1–47.

Särndal, C. E. and B. Swensson (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *International statistical review 55*, 279–294.

Self, S. and K. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association 82*(2), 605–610.

Shen, Y., D. Wu, and M. Zelen (2001). Testing the independence of two diagnostic tests. *Biometrics 57*, 1009–1017.

Spiegelhalter, D. J. and P. G. I. Stovin (1983). An analysis of repeated biopsies following cardiac transplantation. *Statistics in Medicine 2*, 33–40.

Staquet, M., M. Rozencweig, Y. J. Lee, and F. M. Muggia (1981). Methodology for assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases 34*, 599–610.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of Am. Statist. Assoc. 65*, 1350–1361.

Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications: sample size determination. *Biometrics 27*, 935–944.

Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *and Technometrics 14*, 187–202.

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

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

Van Wieringen, W. N. (2005). On identifiability of certain latent class models, statistics and probability letters. *Technometrics 75*, 211–218.

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

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

Walter, S. D. and L. M. Irwig (1988). Estimation of test error rates, disease prevalence and relative risk for misclassified data: a review. *Journal of Clinical Epidemiology 41*, 923–937.

Weiner, D. A., T. J. Ryan, C. H. McCabe, J. W. Kennedy, M. Schloss, F. Tristani, B. R. Chaitman, and L. D. Fisher (1979). Correlations among history of angina, st-segment response and prevalence of coronary artery disease in the coronary artery surgery study. *New England Journal of Medicine 301*, 230–235.

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

Yanagawa, T. and B. Gladen (1984). Estimating disease rates from diagnostic tests. *Am.Journal of Epidemiol. 119*, 1015–1023.

Zhou, X. H. (1998). Correcting for verification bias in studies of a diagnostic test's accuracy. *Stat. Methods for Med. Research 7*, 337–353.

Zhou, X. H., D. K. McClish, and N. A. Obuchowski (2002). *Statistical methods in diagnostic medicine.* New York: Wiley.