Assessing Binary Measurement Systems Using Targeted Verification with a Gold Standard

by

Daniel Ernest Severn

A thesis

presented to the University Of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Statistics

Waterloo, Ontario, Canada, 2017

© Daniel Ernest Severn 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Steven Rigdon Professor (Saint Louis University)
Supervisor(s)	Stefan Steiner Professor
Internal Member	Joel Dubin Associate Professor
Internal Member	Peisong Han Assistant Professor
Internal-external Member	Gordon Savage Professor (Systems Design Engineering)
Other Member(s)	Jock MacKay Professor Emeritus

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the 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

Statement of Contributions

An article based upon the findings of Chapter 2 of this thesis has been published in the Journal of Quality Technology (Vol. 48, No. 2, April 2016, p. 128-138). I was the first author on this paper, with Stefan Steiner and Jock MacKay as co-authors. With supervision from Stefan and Jock, I developed the novel concepts, coded and performed the simulation studies, and wrote the contents of Chapter 2. The published article references this thesis and derives from the work done for this thesis.

Abstract

Binary Measurement Systems (BMS) are used to classify objects into two categories. Sometimes the categories represent some intrinsically dichotomous characteristic of the object, but sometimes continuous or even multidimensional characteristics are simplified into a dichotomy. In medicine, pregnancy is the typical example of a truly dichotomous characteristic; whereas Alzheimer's disease may be a continuous or multidimensional characteristic that one may none-the-less wish to simplify into a dichotomy in diagnosis. In both cases BMS are used to classify the patient into two categories, pregnant or not pregnant, diseased or non-diseased. Most BMS are not inerrant, they misclassify patients and these misclassifications can have very damaging consequences for the patients' health. Therefore in the search to understand and improve the BMS being used or developed, there needs to be a formalized way of studying and judging the merits of a BMS.

While BMS are used throughout society, the two main areas where they are formalized in this way are medicine and manufacturing. Medical BMS are designed to determine the presence of a disease or other medical condition. Manufacturing BMS are designed to determine whether a manufactured item meets a specified quality standard. This abstract will use language and examples typical in the medical application because this is easier to understand and relate to for most people. However most of the thesis was written with an eye to publication in journals for quality improvement and thus typically is written for that audience.

There are two primary attributes of BMS that are used to judge their quality: when measuring a subject once with the BMS what is the probability of a false positive diagnosis, and what is the probability of a false negative diagnosis. In the standard statistical framework (PPDAC – Problem, Plan, Data, Analysis, and Conclusion), the problem this thesis tries to address is determining these two quantities for a BMS. It develops new plans and estimation techniques for this purpose.

These plans assume that a perfect "gold standard" measurement system is available. It also assumes that it is possible to repeatedly measure a subject, and one measurement does not affect other measurements. The plans in this thesis consider reducing the number of gold standard measurements needed for a given level of precision as a primary goal. The context usually implies that there is some difficulty in using the gold standard measurement system in practice; were this not the case the gold standard could be used instead of the BMS being assessed. For example some gold standard measurement systems can only be performed on a dead patient while, the BMS being assessed is intended for a living patient. Alternately the gold standard could be very expensive because no errors are permitted.

The thesis considers two scenarios; one assessing a new BMS where no information is available prior to the study and where only sampling directly from the population of subjects is possible. The second, assessing a BMS that is currently in use where some information is available prior to the study and where subjects previously classified by the BMS are available to sample from. Chapters 2 and 3 consider the first scenario, while Chapters 4 and 5 consider the second scenario. Chapter 1 gives an introduction to the assessment of BMS and a review of the academic literature relevant to this thesis.

Chapter 2 considers a sequential statistical plan for assessing a BMS that introduces a new innovative design concept called Targeted Verification. Targeted Verification refers to targeting specific parts to "verify" with the gold standard based on the outcome of previous phases in the sequential plan. This plan can dramatically reduce the number of patients that need to be verified while attaining performance similar to that of plans that verify all patients and avoiding the pitfalls of plans that verify no patients. Chapter 3 develops a set of closed form estimates that avoid making subjective assumptions and thus have relevant theoretical properties but retain competitive empirical performance.

Chapter 4 takes the Targeted Verification concept and adapts it to the second scenario where a BMS is currently in use. It incorporates the information that is previously available about the BMS and takes advantage of the availability of patients previously categorized by the BMS in sampling. It shows that the Targeted Verification concept is much more efficient than similar plans that would verify all subjects, and much more reliable than plans than do not use a gold standard. Chapter 5 develops a set of estimates with a design philosophy the same as that of Chapter 3. To incorporate the design elements of Chapter 4, the new estimates are no longer closed form, but still avoid making subjective assumptions. The estimates have relevant theoretical properties and competitive empirical performance.

Chapter 6 summarizes and discusses the findings of the thesis. It also provides directions for future work that make use of the Targeted Verification concept.

Acknowledgements

I would like to thank my supervisor Stefan Steiner as well as Jock MacKay for their guidance in writing this thesis and in my studies at the University of Waterloo. I would like to thank the Ontario government and the University of Waterloo for providing the funding and environment for my PhD studies. In connection with this I would also like to thank members of the Statistics and Actuarial Science department for their personal contributions to making the department a good place to study.

I would like to thank my parents, Bob and Arlene Severn, for their financial and personal support throughout my education. I would like to thank my fiancée Victoria Chan for her love and care during my studies. I would also like to thank Andy Carducci for his spiritual support during my studies. Lastly and most importantly I would like to thank Jesus Christ, King of Kings and Lord of Lords, my Lord and Saviour. All good things come from you and I pray I would have a heart of thankfulness toward you and to all through which your gifts come.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Binary classification is ubiquitous in society. There are many instances when this classification task becomes so important it needs to be formalized and assessed. This thesis refers to any system that performs binary classification as a Binary Measurement System (BMS). The two main areas where a formal statistical assessment of a BMS may occur are manufacturing and medicine. In the manufacturing industry a BMS is used to test whether parts conform to some specification to verify the quality of products delivered to the customer. In the medical industry a BMS is used to diagnose various diseases and other medical conditions so that patients can receive the appropriate medical treatment. This thesis considers statistical plans and analysis for rigorously assessing a BMS.

Here I introduce the basic notation used in this thesis and some terminology used in the research area. A unit of a study is either a part in the manufacturing setting or a patient in the medical setting. For simple plans let units be indexed by $i \in \{1, 2, ..., n\}$. For each part let Y_i be the outcome of a single binary measurement, and X_i be the true value of the measurand. Capital letters represent the random variables for each quantity while lower case letters represent data or specific values. This thesis will focus on the industrial application where $X_i = 1, 0$ indicates a conforming and non-conforming part respectively and $Y_i = 1, 0$ represents a passed and failed inspection respectively. The quantities of interest for a BMS are defined in terms of these random variables. The misclassification probability for non-conforming parts is defined as $\mu_A = P(Y_i = 1 | \mathrm{X}_i = 0)$. The misclassification probability for conforming parts is defined as $\mu_B = P(Y_i = 0 | \mathrm{X}_i = 1)$. The conforming probability is defined as $\pi_C = P(X_i = 1)$.

When considering medical diagnosis let $X_i = 1, 0$ represent non-diseased and diseased respectively and $Y_i = 1, 0$ represent a negative and positive test result respectively. These definitions are counter-conventional to medical diagnosis notation where disease status is defined as $D_i = 1 - X_i$ and the test result as $T_i = 1 - Y_i$. In medical diagnosis there are three alternative quantities of interest which are defined as follows: prevalence $P(D_i = 1)$, sensitivity $P(T_i = 1 | D_i = 1)$, and specificity $P(T_i = 0 | D_i = 0)$. Under the above definitions for X_i and Y_i these quantities are equal to $1 - \pi_C$, $1 - \mu_A$ and $1 - \mu_B$ respectively.

A common element of assessment plans for measurement systems is the use of repeated measurements, which means repeatedly measuring a part in the same manner with the measurement system being assessed. Let S_i denote the sum of a set of r repeated measurements on a single part. Since S_i is the sum of binary measurements, $S_i \in \{0, 1, 2, ..., r\}$. S_i is not used in the definition of the quantities of interest but its expectation is related to those quantities and thus it can be used to estimate those quantities. Because the repeated measurements are conducted in the exact same manner and are thus interchangeable, they are combined into the S_i statistics. Given the support of the random variables the binomial distribution springs to mind. However, in practice all parts do not have the same probability of passing inspection and thus S_i are not iid binomial random variables. However, it is possible to model S_i with binomial distributions with different probabilities of passing inspection for each part; this will be discussed in Chapter 2.

The properties of each part Y_i, S_i and X_i, are independent of the properties of all other parts, that is, from part-to-part, all data are considered independent and identically distributed.

1.1 History

Medical Literature

The earliest works on assessing a BMS are in the medical literature. Yerushalmy (1947) studied the diagnosis of tuberculosis, introducing the measures sensitivity and specificity. Neyman (1947) studied the problem of diagnosis more generally. Their papers are published in the same issue of Public Health Reports side-by-side and the two seemed to have collaborated. Another early influential work is written by Bross (1954) and looks at the effect of misclassification in 2x2 tables on various statistical techniques. Cochran looked at many forms of measurement error and touched on binary measurement error in Cochran (1968). His student Aaron Tenenbein studied the problem of binary measurement extensively within his doctoral thesis and three subsequent research papers Tenenbein (1969, 1970, 1971, 1972). Tenenbein's work formalized the idea of using a gold standard reference measurement system to objectively assess another flawed, but potentially useful, BMS.

The next major development relevant to this thesis was the "latent class" approach. This refers to assessing the properties of a BMS without ever measuring the x_i values. Rather these values are treated as latent variables, X_i, which are distributed according to a Bernoulli distribution with parameter π_C. It is not possible to estimate the properties of a BMS with data that only consists of single y_i measurements for each part. The first attempt to assess a BMS without observing the x_i values was

Gart & Buck (1966). This early paper measured each unit with the measurement system of interest as well as with a fallible BMS with known misclassification probabilities. The most influential approach did not come until much later with Hui et al. (1980). This paper measured patients with two tests in multiple populations with different prevalences. The misclassification rates are unknown but assumed to be equal for each population.

Another approach to assess a measurement system without observing the x_i values is to use “repeated measurements”. This was introduced for a more general categorical or polytomous measurement system by Dawid & Skene (1979). This paper used the then newly developed EM algorithm, Dempster et al. (1977), to calculate estimates without observing the true status or category with multiple raters and tests. This approach was more directly applied to assessing a BMS by Quade et al. (1980).

Spiegelhalter & Stovin (1983) modeled the results of multiple biopsies from single patients using similar methodology. The results of repeated measurements were considered conditionally independent of one another, given the true status, x_i. This assumption of conditional independence was challenged in Vacek (1985) who argued that tests based upon similar, or in the case of repeated measurements identical, physical principles often exhibit conditional dependence. The first attempt to model the conditional dependence between repeated measurements from the same test was by Qu et al. (1996) who introduced the Gaussian Random Effects Model (GRE). This model assumes misclassification rates vary patient to patient, and that the distribution of misclassification rates is different for diseases and non-diseased patients. It models the two distributions of misclassification rates separately using Normal distributions that have been transformed to $[0,1]$ by the cumulative distribution function of the standard Normal distribution.

The current medical literature for assessing BMS is very broad and extensive. For a view of the current techniques consider the textbooks of Pepe (2003) or Zhou et al. (2011). For a very extensive literature review of latent class models see Collins (2014). Not all of this broader scope of medical literature is relevant to this thesis because only a small subset of plans consider the use of repeated measurements with the same tests. Some medical studies make use of repeated testing however it is very rare that this actually refers to administering the test in the same way at the same time. Sometimes different doctors administering the same test on the same patient are treated as repeated measurements. More common are studies where the same test is applied to patients at different times, see Engel et al. (2010). This would allow for the disease status to change, which must be accounted for in the modeling. These forms of repeated testing are different from what I refer to as repeated measurements.

Quality Improvement Literature

The rarity of true repeated measurements in the medical literature is in stark contrast to the assessment of a BMS in the quality improvement literature where repeated measurements are considered necessary. This follows from the dominating influence of the Gage Repeatability & Reproducibility (Gage R&R) method for assessing continuous measurement systems based on the work of Mandel (1972). The first assessment plan for a BMS in quality improvement, proposed by McCaslin & Gruska (1972), was based on the repeatability principles found in the Gage R&R method. For more information on these methods and broad view of assessment methods in quality improvement please refer to AIAG guide of Measurement System Analysis (2010). Gage R&R also inspired Boyles (2001) to create a BMS assessment plan that calculated misclassification probabilities. Boyles implicitly assumed, however, that repeated measurements were conditionally independent given the true status, x_i. Similar to the development of medical literature, this assumption was questioned in Wieringen & De Mast (2008). The first paper in quality improvement literature to model the dependency between the repeated measurements was Danila et al. (2012). This paper assumes that misclassification rates vary from part-to-part (i.e. some parts are harder to classify than others) and models the varying misclassification rates with beta distributions.

This thesis uses the random effects model developed in Danila et al. (2012) and sometimes the model in Qu et al. (1996) for robustness comparisons. Both of these papers use their models for latent class analysis; that is, statistical plans for assessing a BMS that only collect s_i measurements for parts and no x_i measurements. Latent class plans are necessary when no gold standard is available, however some papers have demonstrated significant flaws in the approach. Albert & Dodd (2004) showed that when dependence between test results is misspecified, the estimates can be significantly biased. Furthermore, they showed that likelihood ratio tests or other model comparison techniques are not very effective for determining the appropriate dependence structure when the number of tests involved is limited. Van Wieringen (2005) addressed some identifiability concerns of latent class models. Akkerhuis (2016) thoroughly investigated the difficulties and limitation of implementing the latent class approach. For further discussion of latent class plans please see Sections 2.2 and 2.10.

1.2 Targeted Verification

One of the primary novel contributions of this thesis is the development of a design element of BMS assessment plans called Targeted Verification. Verification refers to measuring the x_i values for parts, that is verifying the true conforming/non-conforming status of a part. A full verification plan measures the x_i values for all parts in the study. A no verification or latent class plan measures the x_i values for none of the parts in the study. A partial verification plan measures the x_i values for a subset of parts in the experiment. This thesis considers partial verification plans where parts are selected for verification in an intentional "targeted" manner that improves efficiency. The implementation of such targeted verification plans will be discussed in Chapter 2. The motivation for considering the partial verification plans over full verification plans is the high cost associated with verifying parts. The context of the study implies this high cost because if the gold standard is not burdensome in some way, the gold standard would be used in place of the BMS being assessed.

In medical studies the set of patients verified by the gold standard is often not based on a well documented sampling protocol. Sometimes patients cannot be verified with the gold standard because the patient is unable or unwilling to continue in the study or because it may be dangerous to apply the gold standard to some patients. Sometimes the decision to verify is based on the medical opinion of a doctor. There is oftentimes a good medical reason for these decisions but this biased selection of patients for verification can invalidate the results of the statistical study. There is extensive research studying this problem, which is referred to as verification bias. This literature can be best understood by using the framework of missing data analysis. For an overview of missing data analysis see Little & Rubin (2002). The first article studying verification bias was Begg & Greenes (1983) which critiques removing units with missing x_i values from estimation and suggests an alternative using the missing-at-random (MAR) assumption. Other papers have suggested that even a MAR assumption is not justified, see Baker (1995). Intuitively one would not expect the MAR assumption to hold unless the basis for verifying and not-verifying is fully documented and understood.

In the targeted verification plans proposed in this thesis, parts are selected using a known sampling procedure. Thus the missing mechanism is explicitly known and can be modelled appropriately. There is only one paper that examines partial verification in this context; see Albert & Dodd (2008). This paper considers selecting parts for verification completely at random and shows the improvement in precision and robustness compared to latent class models. It also briefly considers something like targeted verification which it refers to as over-sampling. Please see Chapter 2 for a discussion of the results of this paper.

1.3 Conditional Sampling and Baseline Information

Conditional sampling and baseline information are useful design elements in assessment plans for BMS. Chapter 4 discusses the value of these design elements when used in tandem with targeted verification.

Baseline information refers to a large number of single y_i measurements. As previously stated this pass-rate information alone is not sufficient to assess the misclassification probabilities of a BMS. However when used alongside other information it can significantly improve estimation. The concept of using pass-rate information to improve estimation was first considered in Danila et al. (2008) and made explicit in Danila et al. (2010). Conditional sampling refers to sampling from parts that either passed or failed a single inspection with the BMS being assessed. This technique is useful because in many industrial applications the conforming probability, $\pi_C = P\left(X_i = 1\right)$, is close to one. Therefore a study using parts from the general population has a lot of information about conforming parts, which allows for precise estimation of $\mu_B = P\left(Y_i = 0 \mid \mathrm{X}_i = 1\right)$, but not much information of about non-conforming parts, making estimation of $\mu_A = P\left(Y_i = 1 \mid \mathrm{X}_i = 0\right)$ difficult. Alternately in medicine when assessing a screening test, the probability of a person being diseased can be very close to zero, which causes the same problem. The concept of conditional sampling was first proposed in Haitovsky and Rapp (1992) which was conceived as an extension to work of Tenenbein (1972). However both of these papers focused on assessing π_C, with the BMS being used primarily as a tool to improve the estimation thereof. Danila et al. (2008, 2010) applied this methodology with the purpose of assessing the properties of a fallible BMS, specifically μ_A and μ_B.

1.4 Outline

This thesis will examine the effectiveness of targeted verification in assessing a BMS. It examines different assessment plans and different sampling protocols for determining which parts are verified. Additionally different estimation procedures will be compared for the proposed targeted verification plans.

Chapter 2 examines targeted verification in a simple framework where n parts are each measured r times. The chapter examines the best sampling protocol for verifying parts. It finds that verifying parts that have roughly equal numbers of passed and failed inspections $\left(\frac{s_i}{r} \simeq \frac{1}{2}\right)$ improves estimate precision much more than verifying parts that either passed most inspections $s_i \simeq r$ or failed most inspections $s_i \simeq 0$. It also finds that verifying a small number of parts with each s_i value helps the robustness of and numerical stability of the estimation. Based on these principles a two-phase targeted verification plan is suggested. This chapter uses maximum likelihood estimation (MLE), formalized by Fisher (1922). It also uses the asymptotic theory for MLE standard error calculation, developed by Fisher (1925).

Chapter 3 derives a set of closed form estimates that can be used for the plans considered in Chapter 2, where n parts are each measured r times. It also derives a closed form approximation for the variance of those estimates. It compares the underlying assumptions of the parametric model used in Chapter 2 to the implicit assumptions of the closed form estimates. It derives the optimal sampling protocol for verifying parts under the assumptions of the closed form estimates and compares that to the findings in Chapter 2. Finally, it gives a performance comparison between the ML estimates used in Chapter 2 and the closed form estimates.

Chapter 4 considers the use of targeted verification in plans that make use of baseline information and conditional sampling. It demonstrates the efficiency gains of baseline information and conditional sampling found in full verification plans are also possible when using a targeted verification plan. This chapter will also make adjustments to the plan suggested in Chapter 2 to account for these design elements. Chapter 4 uses the ML estimates assuming a beta-binomial model.

Chapter 5 considers an alternate method of estimation for the plans that use baseline information and conditional sampling. Unfortunately, there is no apparent way to incorporate baseline data into the closed form estimates seen in Chapter 3. However, Chapter 5 continues in the spirit of Chapter 3, developing estimates that make few assumptions and have relevant theoretical properties. The chapter then compares the effectiveness of the new estimates with the beta-binomial ML estimates used in Chapter 4.

Chapter 6 summarizes and discusses the contributions of the thesis. It also provides directions for future work that make use of the Targeted Verification concept.

Chapter 2 Targeted Verification Plan

2.1 Foreword

An article based upon the findings of this chapter has been published in the Journal of Quality Technology (Vol. 48, No. 2, April 2016, p. 128-138).

2.2 Introduction

Binary measurement systems (BMS) are an important part of quality improvement in manufacturing. They arise whenever a specification is qualitative and pass/fail is the only way to quantify a measurement, such as the presence or absence of ghosting (a surface defect) on a painted fascia. They also occur when many measurements are combined to assess overall performance, for example, a system that checks conformance of numerous continuous characteristics on a camshaft, as in the example of Section 2.5. One major objective of a BMS is to prevent customers from receiving out-of-specification product or parts.

Many quality plans require the routine assessment of critical-to-the-customer continuous measurement systems, often with a gauge R&R study. Binary measurement systems used for 100% inspection do not receive the same attention. I suspect that the reason for this neglect is that assessment studies require very large sample sizes (100's of parts) in order to estimate the characteristics of the BMS with useful precision. This is especially true for high quality processes and binary measurement systems that make few errors.

The traditional assessment plan for a BMS requires all parts in the study to be measured with the gold standard. See, for example, Danila et al. (2008). The situation implies that using the gold standard is burdensome; were it not, the gold standard would simply be used in place of the BMS being assessed. Our goal is to lower the cost of assessing binary measurement systems by reducing the number of parts verified with the gold standard.

Many BMS assessment plans repeatedly measure a random sample of parts. Using repeated measurements improves efficiency and reduces the use of the gold standard. However there is a complication associated with repeated measurements. In practice it is found that individual parts have varying misclassification rates; that is, some parts are harder to classify than others. Different modeling approaches have been proposed in the statistics literature to deal with this heterogeneity. A model from the medical literature was introduced by Qu et al. (1996). Another model was proposed by De Mast et al. (2011). This chapter uses the model introduced by Danila et al. (2012, 2013). Each of these models use random effects to account for varying misclassification rates.

One approach to mitigate the burden of the gold standard is to introduce latent variables which represent the conforming status of each part. See, for example, Boyles (2001). This allows for assessment of a BMS without the use of the gold standard, provided there are a sufficient number of repeated measurements. While the latent class approach has merit, previous literature on medical diagnostic tests demonstrates significant flaws. Albert and Dodd (2004) show estimates of the characteristics of the BMS using a latent class approach have significant bias when the model is misspecified. Given that there is little information to determine which of the many models available should be used, this problem cannot be ignored. In contrast, Albert and Dodd showed that when a gold standard is used, the estimates are robust to model misspecification. It seems the use of a gold standard cannot be completely eliminated; however I show the gold standard need not be used on every part.

I propose a two-phase plan. In the first phase, a random sample of parts is measured repeatedly by the BMS, hereafter called the repeated measurement phase. Then in the second phase, parts are selected to be verified based on the outcomes in the repeated measurement phase; this is referred to as the verification phase. The information gained by verifying a part differs dramatically depending on the number of times that part passed inspection in the repeated measurement phase. Verifying parts that either passed or failed inspection all or almost all of the time provides effectively no benefit, while verifying parts that had roughly equal number of passes and failures is of tremendous benefit. If verification is done selectively, performance closely matching that of a full verification plan can be obtained while verifying only a small fraction of the sampled parts.

2.3 Beta-Binomial Model

The chapter uses the model developed in Danila et al. (2012) which has five parameters. The first three are the quantities of interest introduced in Chapter 1: π_C the probability a part randomly selected from the population is conforming, μ_A the probability of misclassifying a non-conforming part, μ_B the probability of misclassifying a conforming part. The other two, γ_A and γ_B, are nuisance parameters related to the correlation between repeated measurements of non-conforming and conforming parts respectively.

The model assumes independence between parts. That is, the conforming status and measurements of one part do not depend on the conforming status or measurements of any other part. Therefore it is important to make sure the sample of parts used in the study are selected at random over a long enough time frame to be representative of the current manufacturing and measurement process.

The model introduced in Danila et al. (2012) proposes that each part has its own misclassification rate. The misclassification rates of non-conforming parts and conforming parts are distributed according to different beta distributions, each with its own parameter values. For non-conforming parts, the mean of the beta distribution is μ_A, the overall average misclassification rate for non-conforming parts or, equivalently, the probability a randomly selected non-conforming part passes a single inspection with the BMS. The variance of the beta distribution is $\frac{\gamma_A}{(1+\gamma_A)}\mu_A(1-\mu_A)$, which can be adjusted using γ_A. When γ_A is equal to zero, the variance is zero and the misclassification rate is constant. In this case, the r repeated measurements will be distributed according to a binomial distribution with probability μ_A; this also implies no correlation between repeated measurements. When γ_A approaches infinity, the distribution of the misclassification rates converges to a Bernoulli distribution with probability μ_A. In this case, the r repeated measurements will either be all failures or all passes with probabilities $1-\mu_A$ and μ_A respectively; this implies perfect positive correlation between repeated measurements. The distribution of misclassification rates for conforming parts mirrors that of non-conforming parts with parameters μ_B and γ_B in place of μ_A and γ_A respectively.

When conducting an assessment study, the part specific misclassification rates are not observed. Rather they are represented by random variables that are integrated out of the proposed probability mass function for S_i, the number of times part i passes inspection.

Let α represent the misclassification rate for a given non-conforming part; then the probability of observing s passes in r measurements can be expressed as,

$$P(S_i = s \mid X_i = 0) = \int_0^1 P(S_i = s \mid \alpha)\mathrm{f}_A(\alpha)d\alpha$$

$$= \int_0^1 \left(\binom{r}{s}(\alpha)^s(1-\alpha)^{r-s}\right)\left(\frac{(\alpha)^{\frac{\mu_A}{\gamma_A}-1}(1-\alpha)^{\frac{1-\mu_A}{\gamma_A}-1}}{\mathrm{Beta}\left(\frac{\mu_A}{\gamma_A},\frac{1-\mu_A}{\gamma_A}\right)}\right)d\alpha$$

$$= \binom{r}{s}\frac{\mathrm{Beta}\left(s+\frac{\mu_A}{\gamma_A}, r-s+\frac{1-\mu_A}{\gamma_A}\right)}{\mathrm{Beta}\left(\frac{\mu_A}{\gamma_A},\frac{1-\mu_A}{\gamma_A}\right)}. \qquad (2.1)$$

Similarly, for conforming parts,

$$P(S_i = s \mid X_i = 1) = \binom{r}{s} \frac{\mathrm{Beta}\left(r - s + \dfrac{\mu_B}{\gamma_B}, s + \dfrac{1-\mu_B}{\gamma_B}\right)}{\mathrm{Beta}\left(\dfrac{\mu_B}{\gamma_B}, \dfrac{1-\mu_B}{\gamma_B}\right)} . \tag{2.2}$$

In Equations (2.1) and (2.2), $\mathrm{Beta}(a,b)$ is the beta function. The following terms are defined for convenience to be used later: $q_s = P(S_i = s \mid X_i = 0)P(X_i = 0)$ and $p_s = P(S_i = s \mid X_i = 1)P(X_i = 1)$. Recall that $P(X_i = 1) = \pi_C$.

2.4 Two-Phase Plan

This section details how to assess a BMS with a two-phased plan that allows for targeted verification. In the repeated measurement phase, n parts are measured r times with the BMS being assessed. Parts are then separated into bins based on the number of times a part passed inspection. The bins are indexed by $s \in \{0,1,2,...,\mathrm{r}\}$ which represents the number of times parts in said bin passed inspection. Let n_s denote the number of parts that end up in bin s.

In the verification phase, the experimenter will decide how many parts to verify from each bin. Let v_s denote the number of parts verified from bin s where $0 \le v_s \le n_s$. Setting v_s equal to zero indicates no parts are verified from bin s, whereas setting v_s equal to n_s indicates all parts from that bin are verified. Any other choice for v_s indicates a subset of parts is verified; this subset is to be selected using simple random sampling. The recommended choice for v_s will be detailed in Section 2.6. After determining which parts will be verified, measure those parts with the gold standard recording u_s the number from each bin that conform to specification. If no parts are verified in bin s, then u_s is automatically equal to zero. The resulting data can be summarized as in Table 2.1.

Treatment of v_s

Notice that v_s is bounded above by n_s, which is a result of the repeated measurement phase. This can make the treatment of v_s problematic. It is possible to define a set way to choose v_s as a function of the repeated measurement phase data, making it a random variable. However it is also possible to leave the choice of v_s open to the experimenter. Thus there are two ways to treat v_s, as a design parameter or as a random variable. In this thesis I treat each v_s as a design parameter and thus all inference will be done conditional upon the values of v_s, see the conditionality principle in Cox & Hinkley (1974).

Table 2.1 - Data Summary

Number of Passes (Bin #)	0	1	$\ldots$	r
Number of Parts (Repeated Measurement Phase)	n_0	n_1	$\ldots$	n_r
Number Verified	v_0	v_1	$\ldots$	v_r
Number Conforming among Verified (Verification Phase)	u_0	u_1	$\ldots$	u_r

The repeated measurement phase data (n_s) has a multinomial distribution while the verification phase data (u_s) is distributed according to a sequence of independent binomial distributions. The log-likelihood for the two-phase model is derived as follows,

$$\begin{aligned}
\mathcal{L}(\theta) &= \left(\binom{n}{n_0\ n_1 \ldots n_r} \prod_{s=0}^{r} P(S_i = s)^{n_s} \right) \\
&\qquad * \left(\prod_{s=0}^{r} \binom{v_s}{u_s} P(X_i = 1 \mid S_i = s)^{u_s} P(X_i = 0 \mid S_i = s)^{v_s - u_s} \right) \\
&= \left(\binom{n}{n_0\ n_1 \ldots n_r} \prod_{s=0}^{r} (p_s + q_s)^{n_s} \right) \left(\prod_{s=0}^{r} \binom{v_s}{u_s} \left(\frac{p_s}{p_s + q_s} \right)^{u_s} \left(\frac{q_s}{p_s + q_s} \right)^{v_s - u_s} \right) \text{ (Bayes' Rule)} \\
&= \binom{n}{n_0\ n_1 \ldots n_r} \prod_{s=0}^{r} (p_s + q_s)^{n_s} \binom{v_s}{u_s} \left(\frac{p_s}{p_s + q_s} \right)^{u_s} \left(\frac{q_s}{p_s + q_s} \right)^{v_s - u_s} \\
&= \binom{n}{n_0\ n_1 \ldots n_r} \prod_{s=0}^{r} (p_s + q_s)^{n_s - v_s} \binom{v_s}{u_s} p_s{}^{u_s} q_s{}^{v_s - u_s},
\end{aligned}$$

then taking the logarithm, the log-likelihood is

$$\ell(\theta) = k + \sum_{s=0}^{r} (n_s - v_s) \log(p_s + q_s) + u_s \log p_s + (v_s - u_s) \log q_s, \tag{2.3}$$

where $\theta = (\mu_A, \mu_B, \pi_C, \gamma_A, \gamma_B)$ and k is a constant which does not depend on θ.

Equation (2.3) is used with data recorded as in Table 2.1 to calculate ML estimates. In order to maximize this expression, numerical optimization must be used. Equation (2.3) constitutes a proper likelihood function; thus under certain regularity conditions the estimates calculated using Equation (2.3) are asymptotically unbiased. Additional code is available for the calculation of standard errors based on the asymptotic variance theory developed by Fisher (1925).

This general two-phase plan includes as a special case, the full verification plan where all parts are verified, i.e. set v_s equal to n_s for all s. The general plan also includes, as a special case, the no verification plan presented in Danila et al. (2012), where v_s is set equal to zero for all s.

2.5 Camshaft Example

The context is real, the data are realistic. An automated gauge determines whether or not the lobes on a camshaft are within specification with respect to their geometry. Each of the twelve lobes is checked for six critical characteristics. If one or more of these characteristics are out of specification for any lobe, the camshaft is rejected for scrap or rework. Individual gauge R&R studies on specific continuous characteristics are conducted by lobe on a regular basis – it was known that these characteristics are correlated. To assess the overall performance of the gauge, 500 camshafts were measured five times each and the number of times that each camshaft passed was recorded. The geometry of 40 problematic camshafts was measured using a high precision coordinate measuring machine, here taken to be the gold standard. Five of the seven camshafts that passed twice in the first part of the study were found to be defective. None of the 33 camshafts with three initial passes were out-of-specification for any characteristic. Table 2.2 summarizes the data.

Table 2.2 - Camshaft Data

Number of Passes (s)	0	1	2	3	4	5
Number of Camshafts (n_s)	29	9	7	33	132	290
Number Verified (v_s)	0	0	7	33	0	0
Number Conforming among Verified (u_s)	0	0	2	33	0	0

Maximum likelihood estimates and their associated asymptotic standard errors are calculated with and without the data from the verification phase. See Section 2.12 for justification of the asymptotic approximation. The results are given in Table 2.3.

Table 2.3 – Camshaft Example Estimation Summary

Parameter	μ_A	μ_B	π_C	γ_A	γ_B
Without Verification					
Estimate	0.0661	0.0935	0.9208	0.0483	0.0301
Std. Error	0.0690	0.0093	0.0181	0.3032	0.0336
With Verification					
Estimate	0.0902	0.0896	0.9141	0.0886	0.0103
Std. Error	0.0239	0.0061	0.0126	0.1081	0.0177
Reduction of Std. Err.	65.4%	33.8%	30.4%	64.3%	47.2%

Using targeted verification results in a large reduction in the standard error of all the parameter estimates, particularly those relating to non-conforming parts i.e. $\hat{\mu}_A$ and $\hat{\gamma}_A$. Of the three primary quantities of interest, $\hat{\mu}_A$ is estimated with the least precision. This is problematic because μ_A affects the customer, making it perhaps the most important quantity. Fortunately, verification improves most the precision of the estimates related to non-conforming parts, thus mitigating this inconvenient disparity. The two nuisance parameters, γ_A and γ_B, are poorly estimated, particularly when no verification is used.

2.6 Proposed Plan

The method for selecting which parts to verify in the camshaft example was effective, providing large improvement with relatively little work since only 8% of the parts were verified with the gold standard. The plan for the camshaft example is similar to the plan proposed in this chapter. The recommended plan is defined as follows:

Repeated Measurement Phase:	•Measure n parts five times each with the BMS. •Separate parts into six bins based upon the number times they passed inspection.
Verification Phase:	•Verify all parts in the bins representing two or three out of five passes. •Verify five randomly selected parts from each of the other bins (where possible)

The recommendation needs justification. Section 2.7 discusses the decision to verify all parts in the middle two bins. Section 2.8 discusses repeated measurement phase planning including the number of repeated measurements. Section 2.9 shows the performance of the plan with comparisons to both the full and no verification plans. Section 2.10 demonstrates the robustness of the recommended plan to model misspecification with comparison to the no verification plan, and gives a justification for verifying parts in the non-central bins.

Many of these sections make arguments using the results of a full factorial experiment with factors defined by the five parameters using a common set of levels. The levels are chosen to correspond to values thought to represent the most likely ranges for a BMS in practice. The levels are outlined in Table 2.4. Each experiment has 32 runs and the results are summarized by box plots constructed over these runs.

Table 2.4 – Factorial Experiment Levels

Factor	μ_A	μ_B	π_C	γ_A	γ_B
Levels	0.05 0.10	0.05 0.10	0.90 0.95	0.05 0.20	0.05 0.20

2.7 Verification Strategy

This section discusses which parts to verify in order to minimize the standard error of the estimates of μ_A, μ_B and π_C. One might argue that the three standard errors cannot be simultaneously minimized. While technically true, the best way to verify is approximately the same for all three.

After the repeated measurement phase, parts are selected for verification based on the number of times they passed inspection. Verifying parts that always passed or always failed is futile because the conforming status of those parts is already known with reasonable certainty; this is because the BMS is assumed to be reasonably good. Rather, it is best to verify parts where the conforming status is the most uncertain; that is, verify parts with an approximately equal number of passes and fails, or parts “in the middle”.

To demonstrate that selecting from the middle is the most effective strategy, I carried out the following study. Suppose you can verify all parts in one bin and one bin only: Figure 2.1 shows the resulting reduction in standard error for verifying each bin separately. The parameter values, sample size and number of repeated measurements are based on the plan and estimates in the camshaft example. The repeated measurement phase data are based on the expected values of n_s for this set of parameter values and are displayed in the lower right subplot. The remaining three plots show the standard errors of $\hat{\mu}_A, \hat{\mu}_B$ and $\hat{\pi}_C$ when you verify all the parts of one bin and no other parts. Each of these plots has two dashed lines for reference; the higher line represents the standard error of the no verification plan while the lower line represents the standard error for the full verification plan. All calculations for this experiment are based on asymptotic results.

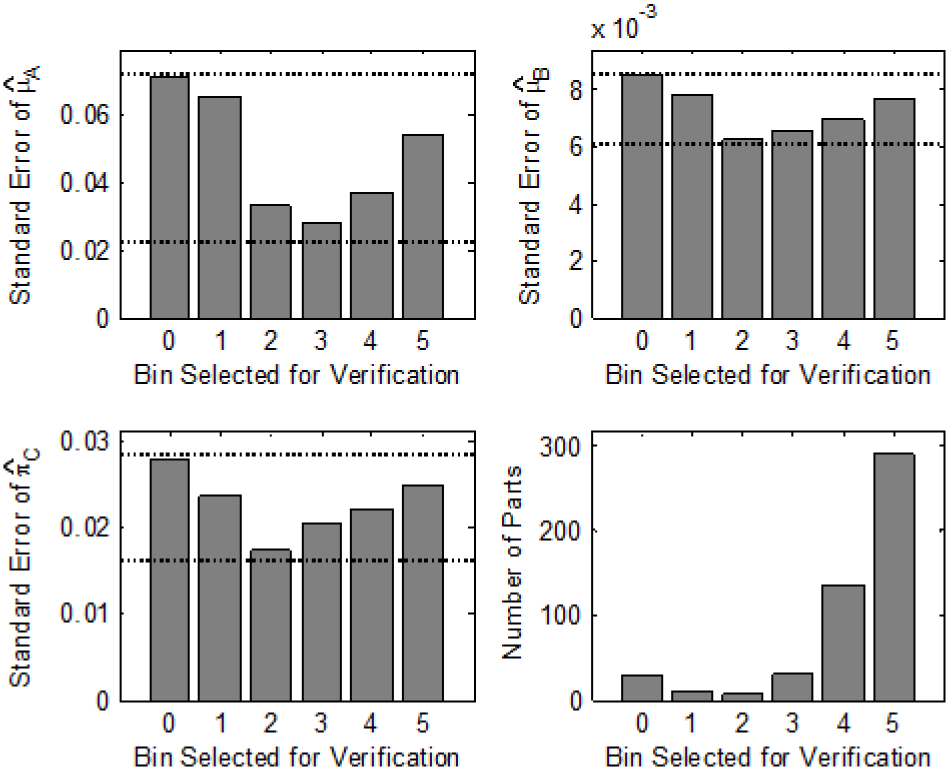

Figure 2.1 – One Bin Verification Example

$n = 500, r = 5, \mu_A = 0.0902, \mu_B = 0.0896, \pi_C = 0.9141, \gamma_A = 0.0886, \gamma_B = 0.0103$

Dashed lines represent the standard errors of full verification (lower) and no verification (higher) plans.

Notice that in Figure 2.1, verifying the few parts with two passes improves estimation more than verifying the approximately 300 parts that had always passed inspection. This result shows how wasteful the full verification plan can be.

Figure 2.1 shows results for only one set of parameter values. To obtain more general conclusions, I conducted a factorial experiment for each combination of parameter values in Table 2.4 and recorded the optimal bin for each of the three model parameters of primary interest. For example, in the previous set of parameters used in Figure 2.1, bin three was optimal for μ_A because verifying bin three resulted in the lowest standard error for $\hat{\mu}_A$. Similarly bin two was optimal for μ_B and π_C. A summary of the results is displayed in Table 2.5. As with Figure 2.1, standard errors are calculated using asymptotic results.

Table 2.5 – Optimal Bin Factorial Experiment

Percentage of time each bin is optimal for reducing standard error of various parameters

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Bin #	μ_A	μ_B	π_C
0	0	0	0
1	0	0	0
2	43.8	100	87.5
3	56.2	0	12.5
4	0	0	0
5	0	0	0

Table 2.5 shows that selecting to verify either bin two or bin three is best in all cases tested. Bin two seems the best overall when considering all three parameters. It is clear that selecting from the middle is the best strategy for verification over these sets of parameter values which were chosen to represent typical values for BMS in industry.

Verifying from the middle is optimal for many extreme scenarios as well; the only exception I found is when the parameters γ_A and γ_B have values greater than one. This case is unrealistic because it implies the distribution of misclassification rates is U-shaped and is clustered around 0% and 100% as opposed to the average misclassification rate.

Changing underlying model parameters typically does not change the optimal verification strategy. However, model parameters and the plan of the repeated measurement phase do affect the potential benefit of verification. Increasing the number of repeated measurements reduces the potential improvement from verification. As the conforming rate increases the optimal bin to verify from shifts slightly towards more failures; however this typically does not change the best bin from which to verify when the number of repeated measurements is less than ten.

Notice in Figure 2.1 that bins two and three are equally close to the middle and yet verifying parts in bin two gives more improvement *per part*. I speculate the reason is that verifying parts with fewer successes provide more information about the parameters μ_A and γ_A, which are the most poorly estimated parameters in the no verification plan. This leads to a simple yet effective order for selecting parts to verify:

> "Select the next part from the bin closest to the middle that has not yet been exhausted. If two bins are equally close, choose from the bin with fewer passes."

The rule is optimal for μ_A, in most cases, so long as γ_A and γ_B are not very large. This might seem to be in conflict with the results in Table 2.5 where bin three was optimal 56.2% of the time. However, the experiment associated with Table 2.5 considered verifying entire bins. Recall that while selecting bin three for verification was optimal for the scenario shown in Figure 2.1, it is obvious that bin two provided more improvement *per part*.

Having established that it is best to verify starting with parts in the middle, I now examine how many parts should be verified. I consider how the standard errors of the three parameters of interest decrease as the proportion of verification increases according to the order proposed earlier in this chapter. Figure 2.2 summarizes the results for the parameter estimates found in the camshaft example. The standard errors in Figure 2.2 are calculated using asymptotic expressions.

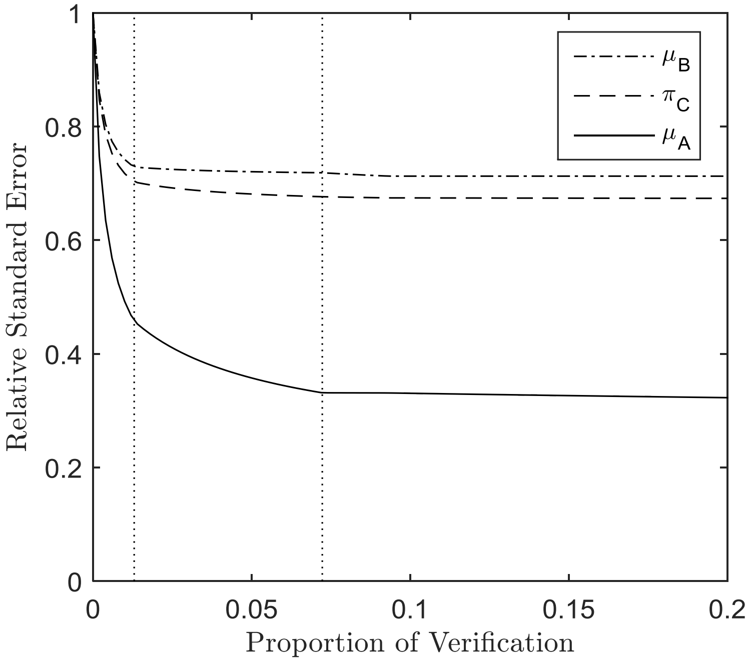

Figure 2.2 – Verification Proportion Plot

$n = 500, r = 5, \mu_A = 0.0902, \mu_B = 0.0896, \pi_C = 0.9141, \gamma_A = 0.0886, \gamma_B = 0.0103$

Figure 2.2 shows that the majority of the benefit of verification occurs very quickly. This is because there are few parts in the middle bins. Once the parts in the middle have been verified, the improvement in standard errors is negligible.

Notice also that in the line corresponding to μ_A there are two discontinuities in the derivatives. The first discontinuity occurs at around two percent verification and represents when the first bin that was verified, the bin representing two out of five passes, was exhausted and verification began on the next bin, the bin representing three out of five passes. The second discontinuity occurs when the second bin is exhausted. After this, there is negligible reduction in the standard errors of $\hat{\mu}_A$, $\hat{\mu}_B$ or $\hat{\pi}_C$.

2.8 Repeated Measurement Phase Planning

This section discusses how to conduct the first phase in order to maximize estimate precision. As expected, increasing the repeated measurement phase sample size, n , as well as the number of repeated measurements used, r , both improve estimate precision. For a fixed budget one must trade-off between n and r . It is natural to leave n as a design parameter which the experimenter will determine based on the precision requirements of the study. For r however, some guidance is needed.

To make a fair comparison between different r values on an equal cost basis, I conducted an experiment where $n*r$, i.e. the total number of measurements by the BMS is fixed. The number of repeated measurements is varied from three through nine with n being adjusted to keep the total number of measurements fixed. The range of r starts at three because the model parameters cannot be identified with fewer repeated measurements. The number of verifications is kept the same for all values of r and is equal to the number of verifications that would be required if the proposed plan would be used with five repeated measurements. The verifications are allocated using the rule described in the verification strategy section after five parts have been allocated to each bin. For each level of r , the asymptotic standard error of each parameter estimate is calculated. The experiment is first conducted with the parameter values taken from the camshaft example. The results are found in Table 2.6.

Table 2.6 – Optimal r Experiment

Asymptotic Standard Errors as n and r vary with $n*r=2500$ fixed

$\mu_A=0.0902, \mu_B=0.0896, \pi_C=0.9141, \gamma_A=0.0886, \gamma_B=0.0103$

	r = 3 n = 833	r = 4 n = 625	r = 5 n = 500	r = 6 n = 416	r = 7 n = 357	r = 8 n=312	r = 9 n = 277
μ_A	0.0596	0.0265	**0.0239**	0.0241	0.0244	0.0248	0.253
μ_B	0.0063	**0.0061**	0.0061	0.0061	0.0062	0.0062	0.0062
π_C	0.0122	**0.0114**	0.0126	0.0138	0.0148	0.0159	0.0169

For the set of parameter values used in Table 2.6, using five repeated measurements is best for estimating μ_A while using four repeated measurements is best for μ_B and π_C. Typically more importance is placed on μ_A because it is the most poorly estimated quantity and the most important for the customer. The standard errors for five to six repeated measurements are very similar and close to optimal. Table 2.6 shows the results for μ_A , μ_B and π_C but not the nuisance parameters, γ_A and γ_B. The precision of the estimates of the nuisance parameters improves when a higher number of repeated measurements are used.

The experiment was also done for the grid of parameter values in Table 2.4. The results for μ_A are shown in Figure 2.3. The results are given in a relative to the optimal standard error basis. For example, if the results in Table 2.6 were included in Figure 2.3, the value for $r=5$ would be 1, whereas the value for $r=6$ would $0.0241/0.0293=$ 1.0086. The results for $r=3$ are not shown because they are so far from optimal that they would make differences between the other choices for r difficult to observe.

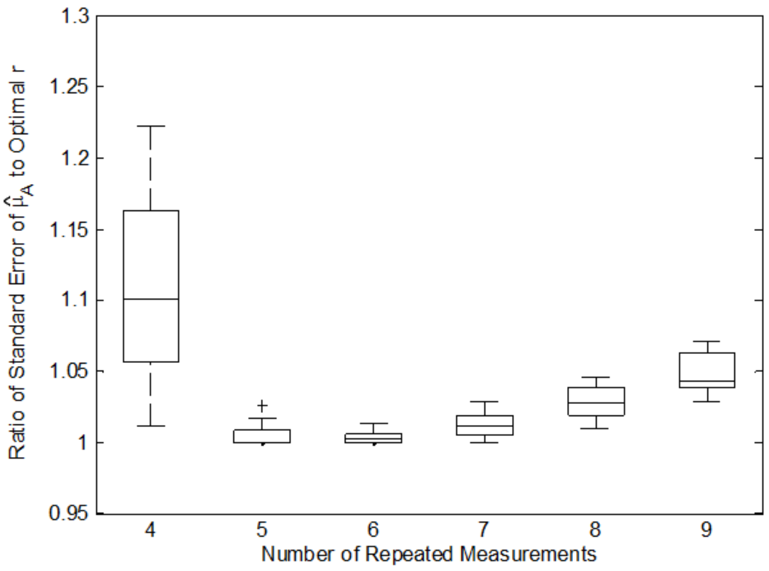

Figure 2.3 – Optimal r Experiment

Factorial Experiment run at all combinations $n*r \simeq 2500$

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

As can be seen in Figure 2.3, using five repeated measurements seems to be optimal or close to optimal in all cases. Using five repeated measurements is optimal in 53% of cases tested, within 2% of optimal in 97% of cases tested, is always within 3% of optimal for all cases tested. While the optimal choice of r depends on the parameter values of the BMS, using five repeated measurements is so close to optimal that it can be done indiscriminately.

Note that the number of verifications was chosen to suit the five repeated measurements plan and thus the results may slightly favor choosing five repeated measurements. However, I did this type of experiment changing the number of verifications in a variety of ways and five, six or seven repeated measurements always resulted in the greatest efficiency.

2.9 Performance

To test the performance of the proposed plan, another factorial experiment was conducted with the levels described in Table 2.4. For each combination of model parameter values, the standard errors were calculated for the full verification plan, the no verification plan and the proposed plan. Recall that the proposed plan verifies all parts in the two central bins and five parts in the non-central bins. From these three quantities, performance measures are calculated as described in Figure 2.4. The performance is shown in Figure 2.5 using box plots calculated over the 32 different combinations of parameter values. Asymptotic standard errors were used for the full verification plan and the proposed plan while standard errors for the no verification plan were estimated using simulation. For each combination of the parameter values, 1000 data sets were generated.

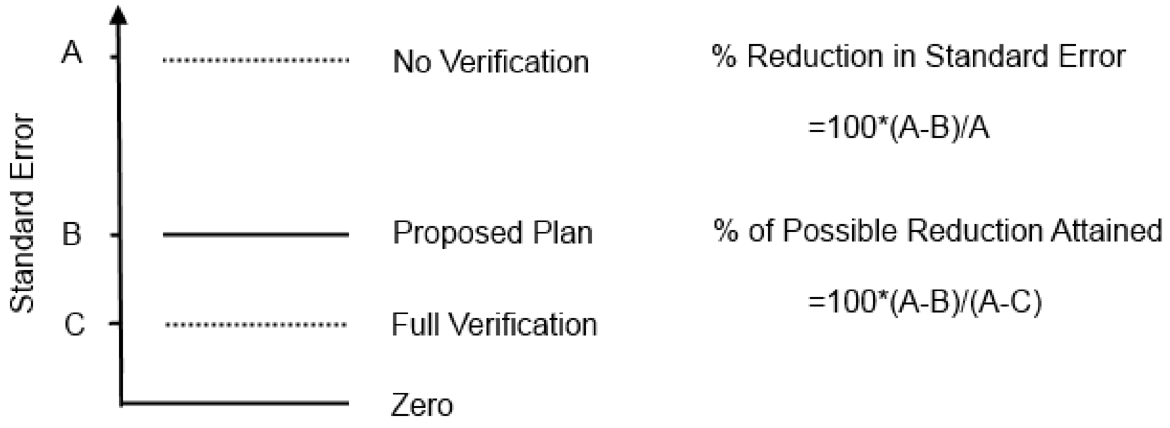

Figure 2.4 – Performance Measures Summary

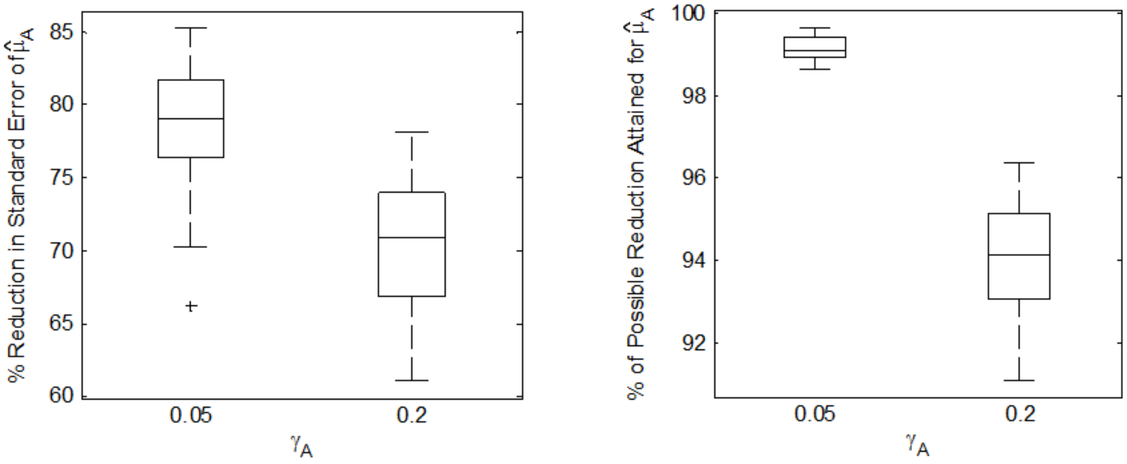

Figure 2.5 – Proposed Plan Performance
Factorial experiment run at all combinations with $n = 500$ using proposed plan
$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

We see from Figure 2.5 that the proposed plan offers a huge reduction in standard error compared to the no verification plan and that it attains the majority of the potential improvement to be had by verifying all parts with the gold standard. The improvement in standard error is dramatic, with most cases seeing even more improvement than in the camshaft example. The standard error of $\hat{\mu}_A$ under the proposed plan is typically reduced to less than a third of that same standard error under the no verification plan. Furthermore, almost all of the potential gain is realized by verifying the few parts in the two middle bins, as well as the five from each of the other bins. On average, 97% of the possible reduction available through verification was attained using the proposed plan.

The number of parts that must be verified under the proposed plan varies and is affected by all five model parameters. Since most parts are conforming, μ_B plays the dominant role in determining how many parts will fall in two middle bins and thus need verification. For the grid of parameters values specified in Table 2.4, the percentage of parts verified averaged 9% when μ_B was 0.05 and 15% when μ_B was 0.1.

There are two ways to summarize the benefits of the proposed plan. First, it gives large gains in precision over the no verification plan for little additional cost. And second, it attains comparable performance to the full verification plan while eliminating the majority of the cost associated with using the gold standard.

2.10 Robustness

One of the problems with the no verification plan is that it is not robust to model misspecification with respect to bias. See Albert and Dodd (2008).

In the recommended plan five parts are verified from the four non-central bins in addition to all the parts in Bins 2 and 3. These extra verifications were added to account for some oddities in the likelihood surface. When no verifications were taken from the other groups sometimes ML estimates would describe a U-shaped beta distribution when this was not appropriate. Verifying a few observations from each of the non-central bins makes the likelihood surface better behaved and eliminates these undesirable estimates. This improves the robustness properties of the recommended plan. While I recommended five verifications in the non-central bins this is not set in stone. Generally I would recommend some verifications in the non-central bins for the benefits just described. However if the experimenters are willing to tolerate marginally more bias, when the model is misspecified, they could reduce the number or if the assessment study is very large and important the number of verifications could be increased.

I conducted another simulation to demonstrate that the proposed plan has robustness properties similar to that of the full verification plan. The data were generated using the Gaussian Random Effect (GRE) model developed by Qu et al. (1996) over all combinations of the parameters in Table 2.4. The parameter values for the GRE model were chosen to match the mean and variance of the beta distribution for each combination of parameter values. ML estimates that assumed the beta-binomial model calculated from simulated data for both the proposed plan and the no verification plan. This was done 1000 times to estimate any bias present in the estimation procedures. The results are summarized in Figure 2.6.

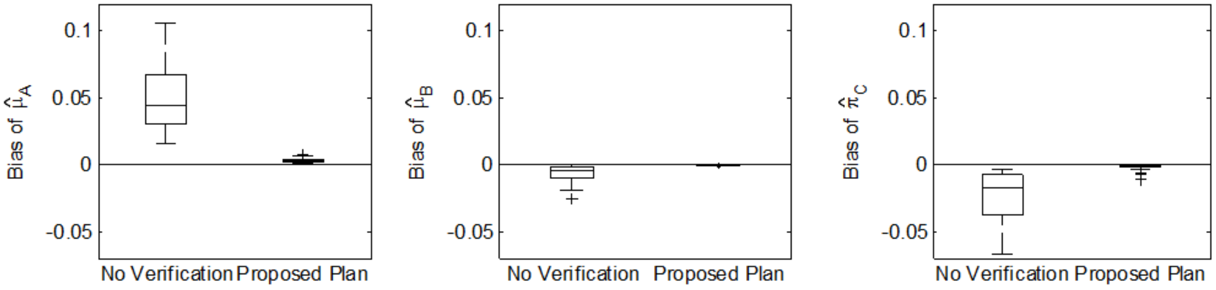

Figure 2.6 – Gaussian Random Effect Bias Experiment

Factorial experiment run at all combinations *with* $n = 500$

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 2.6 shows a high level of bias for the no verification plan and negligible bias for the proposed plan. This is a very positive result and shows that the proposed plan is robust. I must be careful not to overstate the findings here because robustness in general is impossible to prove. Typically if one maliciously designs the underlying model so that it cannot easily be matched by the assumed model one can break the robustness property. This is most likely the case here. What this experiment does demonstrate however is that the proposed plan is not overly sensitive to model misspecification like the no verification plan.

2.11 Discussion

The proposed plan has comparable performance and robustness to the full verification plan while eliminating the majority of the cost inherent in using the gold standard. This is possible because the amount of information gained in verifying parts is not the same. It is better to first repeatedly measure parts with the BMS and then verify only parts that have roughly equal number of passes and failures. Using this idea, I proposed a simple and effective plan for assessing binary measurement systems which I feel confident in recommending to practitioners. The recommended plan is effective for a wide range of different parameter values. This plan, as well as similar plans built on the targeted verification concept, can dramatically reduce the burden of using the gold standard and should encourage practitioners to assess the binary measurement systems with the same regularity that they assess other measurement systems.

While this plan stands on its own there is room for further study and possible extensions. One possible extension is to include baseline information and use conditional sampling as in Danila et al. (2012), which is considered in Chapter 4. This can be a useful alteration to the plan when the conforming rate π_c is high. Note also that the plan can be used with any underlying model such as GRE or even one that assumes conditional independence.

2.12 Asymptotic Variance Justification

Determining an analytic expression for the maximum likelihood estimates or their variance is not feasible, thus an approximation for this variance is needed. This paper uses asymptotic variance results due to Fisher (1925). The purpose of this section is to assess the reliability of these asymptotic results at different sets of parameter values. I conducted a factorial experiment with six factors: sample size, n , and all five model parameters. For each treatment, one thousand datasets were simulated from the beta-binomial model discussed in Section 2.3. For each data set the parameters were estimated using MLE. Parts were selected and verified according the proposed plan. The bias and standard error of these estimates were calculated and recorded for comparison to the asymptotic standard error approximation. Figure 2.7 shows the ratio of the simulated standard errors and the asymptotic standard errors for each combination of parameter values. The results are separated by sample size, n . Thus each box represents 32 combinations of parameter values as described in Table 2.4.

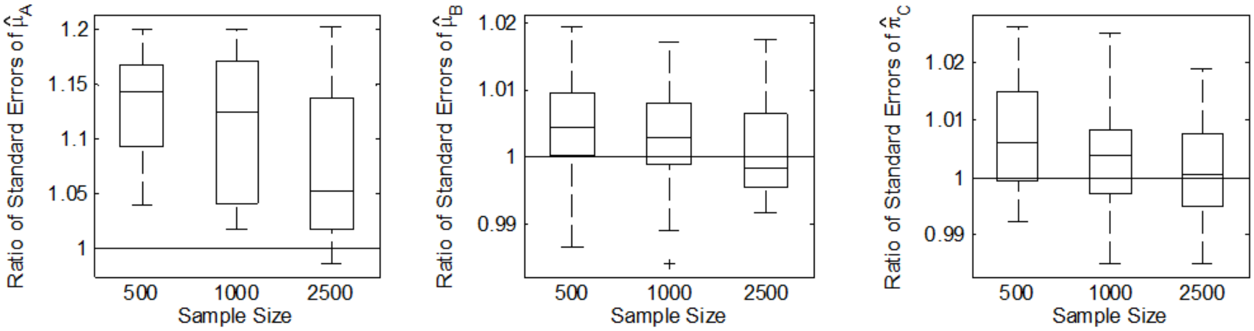

Figure 2.7 – Asymptotic Vindication Experiment

Ratio of Simulated and Asymptotic Standard Errors

$\mu_A, \mu_B = 0.05, 0.1;\ \pi_C = 0.9, 0.95;\ \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

The ratios are typically close to one indicating that the asymptotic variance is a reasonable approximation, and thus sufficiently accurate for the manner in which it is used.

Chapter 3 Closed Form Estimates for Repeated Measurement Study

3.1 Introduction

This chapter will consider an alternate form of estimation for the plan developed in Chapter 2. The alternative estimates are closed form and have no implicit or explicit assumptions that cannot be justified in an absolute mathematical sense. As will be shown the estimates have both advantages and disadvantages when compared to the beta-binomial ML estimates used in Chapter 2. The beta-binomial model makes some subjective assumptions and while these assumptions are reasonable approximations they are not true in a mathematical sense. Unfortunately this means that the theoretical properties of MLE may not be applicable for the beta-binomial ML estimates in practice. In contrast, because the closed form estimates make no subjective modeling assumptions, the theoretical properties derived in this chapter can be relied upon in practice; this is the primary advantage of the closed form estimates. Another attractive feature of the estimates is the simplicity of their form. The closed form estimates for μ_A, μ_B, and π_C are defined as,

$$\hat{\pi}_C = \sum_{s=0}^{r} \frac{n_s}{n} \frac{u_s}{v_s} \;,\; \hat{\mu}_A = \frac{1}{1-\hat{\pi}_C} \sum_{s=0}^{r} \frac{s}{r} \frac{n_s}{n} \frac{v_s - u_s}{v_s},\; \hat{\mu}_B = \frac{1}{\hat{\pi}_C} \sum_{s=0}^{r} \frac{r-s}{r} \frac{n_s}{n} \frac{u_s}{v_s}, \tag{3.1}$$

where n_s, v_s, and u_s are data obtained from a targeted verification plan described in Chapter 2; see Table 2.1. Note: Define $\hat{\mu}_B$ as 0 if $u_s = 0$ for all s and define $\hat{\mu}_A$ as 0 if $u_s = v_s$ for all s.

In this chapter I will derive these estimates and their theoretical properties. I will then compare the performance of these closed form estimates with the MLE used in Chapter 2 using simulation studies.

3.2 Basic Quantities

Notice that the estimates $\hat{\mu}_A$ and $\hat{\mu}_B$ in Equation (3.1) have random variables in the denominator while $\hat{\pi}_C$ does not. This makes the properties of $\hat{\mu}_A$ and $\hat{\mu}_B$ different and more complicated than those of $\hat{\pi}_C$. For this reason I will initially work with an alternative set of basic quantities, which do not have random variables in the denominator. The advantage of these quantities is that I can derive unbiased estimates and unbiased variance estimates thereof. These quantities are defined in Table 3.1.

Table 3.1 – Basic Quantity Definitions

	$Y_i = 0$	$Y_i = 1$	
$X_i = 0$	$\pi_{00} = P(Y_i = 0, X_i = 0)$	$\pi_{10} = P(Y_i = 1, X_i = 0)$	$1 - \pi_C = P(X_i = 0)$
$X_i = 1$	$\pi_{01} = P(Y_i = 0, X_i = 1)$	$\pi_{11} = P(Y_i = 1, X_i = 1)$	$\pi_C = P(X_i = 1)$
	$1 - \pi_P = P(Y_i = 0)$	$\pi_P = P(Y_i = 1)$	1

π_P is called the pass-rate in manufacturing. The misclassification probabilities can be expressed as a ratio of these basic quantities,

$$\mu_A = \frac{\pi_{10}}{\pi_{00} + \pi_{10}} = \frac{\pi_{10}}{1 - \pi_C} \quad , \mu_B = \frac{\pi_{01}}{\pi_{01} + \pi_{11}} = \frac{\pi_{01}}{\pi_C} \quad .$$

I will now derive an estimate for π_{10}. In the following expression I will use both Y_i and S_i. This is best understood by thinking of S_i as a result of r measurements/inspections that have already been recorded and Y_i as the next single measurement. In this way I can pose the question: given that a part passed inspection s out of r times what is the probability it will pass the next inspection?

First, I use the law of total probability conditioning on S_i,

$$\pi_{10} = P(Y_i = 1, X_i = 0) = \sum_{s=0}^{r} P(Y_i = 1,\ X_i = 0,\ S_i = s).$$

Then using the definition of conditional probability twice,

$$\pi_{10} = \sum_{s=0}^{r} P(Y_i = 1 \mid X_i = 0,\ S_i = s) P(X_i = 0 \mid S_i = s) P(S_i = s).$$

This gives an expression with terms that can be estimated using the data from a targeted verification plan, so long as $v_s > 0$ for all s. To move from the theoretical quantity to the estimate I simply replace each of the three parts of each summand with a simple estimate based on the observed quantities,

$$\hat{P}(Y_i = 1 \mid S_i = s,\ X_i = 0) = \tfrac{s}{r},\ \hat{P}(X_i = 0 \mid S_i = s) = \tfrac{v_s - u_s}{v_s},\ \hat{P}(S_i = s) = \tfrac{n_s}{n},$$

which yields,

$$\hat{\pi}_{10} = \sum_{s=0}^{r} \frac{s}{r} \frac{n_s}{n} \frac{v_s - u_s}{v_s} .$$

Note that the estimate for $P(Y_i = 1 | S_i = s, X_i = 0)$ ignores the information about X_i and simply uses the observed pass rate for each bin. It is natural to ask whether the information about X_i would affect the probability and thus whether it should be incorporated in the estimate. However this simple estimate makes the basic quantity estimators unbiased as will be shown in Section 3.8. This same process can be repeated for any of the basic quantities. Please see Table 3.2.

Table 3.2 – Closed Form Basic Quantity Estimates - Standard Form

	$Y_i = 0$	$Y_i = 1$	
$X_i = 0$	$\hat{\pi}_{00} = \sum_{s=0}^{r} \frac{r-s}{r} \frac{n_s}{n} \frac{v_s - u_s}{v_s}$	$\hat{\pi}_{10} = \sum_{s=0}^{r} \frac{s}{r} \frac{n_s}{n} \frac{v_s - u_s}{v_s}$	$1 - \hat{\pi}_C = \sum_{s=0}^{r} \frac{n_s}{n} \frac{v_s - u_s}{v_s}$
$X_i = 1$	$\hat{\pi}_{01} = \sum_{s=0}^{r} \frac{r-s}{r} \frac{n_s}{n} \frac{u_s}{v_s}$	$\hat{\pi}_{11} = \sum_{s=0}^{r} \frac{s}{r} \frac{n_s}{n} \frac{u_s}{v_s}$	$\hat{\pi}_C = \sum_{s=0}^{r} \frac{n_s}{n} \frac{u_s}{v_s}$
	$1 - \hat{\pi}_P = \sum_{s=0}^{r} \frac{r-s}{r} \frac{n_s}{n}$	$\hat{\pi}_P = \sum_{s=0}^{r} \frac{s}{r} \frac{n_s}{n}$	1

3.3 Part-by-Part Alternative Form

To show that the estimates in Table 3.2 are unbiased, the summation needs to be rewritten in terms of $i \in \{1, 2, ..., n\}$ as opposed to $s \in \{1, 2, ..., r\}$. To do this I will introduce some new notation. Let $z_i = 1, 0$ represent whether part i was verified or not verified respectively. Let Z_i be the associated random variable, with $P(Z_i = 1 | S_i = s) = \frac{v_s}{n_s}$. These properties of Z_i are known because the sampling protocol used to verify parts is well described in Chapter 2. To help understand this rewriting I present a data set represented part-by-part and sorted by s_i then by z_i, then by $z_i x_i$, in Table 3.3.

Table 3.3 – Part-by-Part Data Representation for Basic Plan

s_i	0	0	…	0	0	0	…	0	0	0	…	0	1	1	…	1	1	1	…	1	1	1	…	1	…	…	…	…	r	r	…	r	r	r	…	r	r	r	…	r
	n_0												n_1																n_r											
z_i	1	1	…	1	1	1	…	1	0	0	…	0	1	1	…	1	1	1	…	1	0	0	…	0	…	…	…	…	1	1	…	1	1	1	…	1	0	0	…	0
	v_0												v_1																v_r											
$z_i x_i$	1	1	…	1	0	0	…	0	0	0	…	0	1	1	…	1	0	0	…	0	0	0	…	0	…	…	…	…	1	1	…	1	0	0	…	0	0	0	…	0
	u_0												u_1																u_r											

I start with an alternate form for the estimate $\hat{\pi}_{10}$ written in terms of $i \in \{1, 2, ..., n\}$,

$$\hat{\pi}_{10}{}^{ALT} = \frac{1}{n}\sum_{i=1}^{n}\frac{s_i}{r}\frac{z_i - z_i x_i}{\left(v_{s_i}/n_{s_i}\right)},$$

and prove that it is equal to the standard form found in Table 3.2. I rewrite the summands as a sum over $s \in \{1, 2, ..., r\}$ using indicator functions as follows,

$$\hat{\pi}_{10}{}^{ALT} = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=0}^{r} I\left(s_i = s\right)\frac{s_i}{r}\frac{z_i - z_i x_i}{\left(v_{s_i}/n_{s_i}\right)}.$$

Then I can replace the $z_i - z_i x_i$ part of the summand with the appropriate conditions in the indicator function. Then because the summand is non-zero only when $s_i = s$, I can substitute s for s_i in the summand, which yields

$$\hat{\pi}_{10}{}^{ALT} = \frac{1}{n}\sum_{i=1}^{n}\sum_{s=0}^{r} I\left(s_i = s, z_i = 1, z_i\,\mathrm{x}_i = 0\right)\frac{s}{r} n_s \frac{1}{v_s}.$$

Changing the order of summation gives,

$$\hat{\pi}_{10}{}^{ALT} = \sum_{s=0}^{r}\sum_{i=1}^{n} I\left(s_i = s, z_i = 1, z_i\,\mathrm{x}_i = 0\right)\frac{s}{r}\frac{n_s}{n}\frac{1}{v_s}.$$

Next I move parts of the inner summand to the outer summation because they do not depend on i, giving

$$\hat{\pi}_{10}{}^{ALT} = \sum_{s=0}^{r}\frac{s}{r}\frac{n_s}{n}\frac{1}{v_s}\sum_{i=1}^{n} I\left(s_i = s, z_i = 1, z_i\,\mathrm{x}_i = 0\right).$$

And finally I evaluate the inner summation, giving

$$\hat{\pi}_{10}{}^{ALT} = \sum_{s=0}^{r} \frac{s}{r} \frac{n_s}{n} \frac{v_s - u_s}{v_s} ;$$

please refer to Table 3.3 to understand this step. Therefore $\hat{\pi}_{10}{}^{ALT} = \hat{\pi}_{10}$. This process can be repeated for all the basic quantity estimates; please see Table 3.4.

Table 3.4 - Closed Form Basic Quantity Estimates – Part-by-Part Representation

	$Y_i = 0$	$Y_i = 1$	
$X_i = 0$	$\hat{\pi}_{00} = \frac{1}{n} \sum_{i=1}^{n} \frac{r - s_i}{r} \frac{z_i - z_i x_i}{\left(v_{s_i} / n_{s_i}\right)}$	$\hat{\pi}_{10} = \frac{1}{n} \sum_{i=1}^{n} \frac{s_i}{r} \frac{z_i - z_i x_i}{\left(v_{s_i} / n_{s_i}\right)}$	$1 - \hat{\pi}_C = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i - z_i x_i}{\left(v_{s_i} / n_{s_i}\right)}$
$X_i = 1$	$\hat{\pi}_{01} = \frac{1}{n} \sum_{i=1}^{n} \frac{r - s_i}{r} \frac{z_i x_i}{\left(v_{s_i} / n_{s_i}\right)}$	$\hat{\pi}_{11} = \frac{1}{n} \sum_{i=1}^{n} \frac{s_i}{r} \frac{z_i x_i}{\left(v_{s_i} / n_{s_i}\right)}$	$\hat{\pi}_C = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i x_i}{\left(v_{s_i} / n_{s_i}\right)}$
	$1 - \hat{\pi}_P = \frac{1}{n} \sum_{i=1}^{n} \frac{r - s_i}{r}$	$\hat{\pi}_P = \frac{1}{n} \sum_{i=1}^{n} \frac{s_i}{r}$	1

3.4 Expectation of Basic Quantity Estimators

Now given this alternative form for the estimates of the basic quantities it is possible to prove they are unbiased. I will show the proof for the estimator $\tilde{\pi}_{10}$ only, as the other proofs are similar. Starting with the definition of the estimator, I have

$$E\left[\tilde{\pi}_{10}\right] = E\left[\frac{1}{n} \sum_{i=1}^{n} \frac{S_i}{r} \frac{Z_i - Z_i X_i}{\left(v_{S_i} / n_{S_i}\right)} \right].$$

First I move the expectation inside the summation. Then because each summand is identical I can simplify the expression to

$$E\left[\tilde{\pi}_{10}\right] = \frac{1}{n} \sum_{i=1}^{n} E\left[\frac{S_i}{r} \frac{Z_i - Z_i X_i}{\left(v_{S_i} / n_{S_i}\right)} \right] = E\left[\frac{S_i}{r} \frac{Z_i - Z_i X_i}{\left(v_{S_i} / n_{S_i}\right)} \right].$$

Next I use the law of total expectation conditioning on S_i and move the appropriate terms to the outer expectation, yielding

$$E\left[\tilde{\pi}_{10}\right]=E\left[E\left[\frac{S_i}{r}\frac{Z_i-Z_iX_i}{\left(v_{S_i}/n_{S_i}\right)}\middle|S_i\right]\right]=E\left[\frac{S_i}{r}\frac{E\left[Z_i-Z_iX_i|S_i\right]}{\left(v_{S_i}/n_{S_i}\right)}\right].$$

Next, I can split up the term $E\left[Z_i-Z_iX_i|S_i\right]$ because Z_i and X_i are conditionally independent given S_i , this is because the sampling protocol by which parts are verified is explicitly defined and only depends on S_i. Doing do gives,

$$E\left[\tilde{\pi}_{10}\right]=E\left[\frac{S_i}{r}\frac{E\left[Z_i|S_i\right]}{\left(v_{S_i}/n_{S_i}\right)}E\left[1-X_i|S_i\right]\right].$$

But $P\left(Z_i=1|S_i=s\right)=\frac{v_s}{n_s}$ therefore $E\left[Z_i|S_i\right]=\frac{v_{S_i}}{n_{S_i}}$. So I substitute this into the equation and then again use the law of total expectation on the simplified expression, yielding

$$E\left[\tilde{\pi}_{10}\right]=E\left[\frac{S_i}{r}\frac{\left(v_{S_i}/n_{S_i}\right)}{\left(v_{S_i}/n_{S_i}\right)}E\left[1-X_i|S_i\right]\right]=\frac{1}{r}E\left[E\left[S_i\left(1-X_i\right)\middle|S_i\right]\right]=\frac{1}{r}E\left[S_i\left(1-X_i\right)\right].$$

But S_i is a sum of r repeated measurements, so I can express it as $S_i=\sum_{j=1}^{r}Y_{ij}$ and substituting this into the equation. I then move the expectation inside the summation which gives,

$$E\left[\tilde{\pi}_{10}\right]=\frac{1}{r}E\left[\sum_{j=1}^{r}Y_{ij}\left(1-X_i\right)\right]=\frac{1}{r}\sum_{j=1}^{r}E\left[Y_{ij}\left(1-X_i\right)\right].$$

Because each summand is identical the summation and division by the number of summands r can be removed and replaced by any one of the summands, say $E\left[Y_{i1}\left(1-X_i\right)\right]$. Thus I have that,

$$E\left[\tilde{\pi}_{10}\right]=E\left[Y_{i1}\left(1-X_i\right)\right]=P\left(Y_{i1}=1,X_i=0\right)=\pi_{10}.$$

The proof for the other basic quantities is very similar. It is also possible to prove $\tilde{\pi}_C$ is an unbiased estimate without using the alternative part-by-part expression.

3.5 Variance and Covariance of Basic Quantity Estimators

It is also possible to derive unbiased variance estimators for each basic quantity. In Section 3.15, I derive the variance estimate for $\hat{\pi}_{10}$. I start with the variance for the corresponding estimator $\tilde{\pi}_{10}$, which replaces n_s and u_s in the estimate $\hat{\pi}_{10}$ with their random variable counterparts N_s and U_s,

$$Var\left(\tilde{\pi}_{10}\right)=Var\left(\sum_{s=0}^{r}\frac{s}{r}\frac{N_s}{n}\frac{v_s-U_s}{v_s}\right).$$

I then decompose the variance of the summation over $s\in\{1,2,\ldots,r\}$ into a sum of the individual variances and covariances, yielding

$$Var\left(\tilde{\pi}_{10}\right)=\sum_{s=0}^{r}\left(\frac{s}{r}\right)^2 Var\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s}\right)+2\sum_{s<t}\frac{s}{r}\frac{t}{r}Cov\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s},\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right). \tag{3.2}$$

I then show the distribution of $\left(N_0,N_1,\ldots,N_r\right)$ is multinomial and thus the marginal distribution of N_s is binomial. I also show that given v_s the distribution of U_s is binomial. I prove that given v_s, N_s and U_s are independent and use the properties to derive unbiased estimators for

$$Var\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s}\right),\ Cov\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s},\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right).$$

I substitute those unbiased estimators into Equation (3.2) to obtain an unbiased estimator for the variance of $\tilde{\pi}_{10}$. I then define the associated unbiased variance estimate. The derivations of the variance for the other basic quantity estimates are very similar to that of $\tilde{\pi}_{10}$. Below I give the variance estimates for all the basic quantities,

$$\begin{aligned}Var\left(\hat{\pi}_{00}\right)=&\sum_{s=0}^{r}\frac{(r-s)^2}{r^2}\left(\frac{n_s^{\,2}}{n^2}\frac{\left(v_s-u_s\right)^2}{v_s^{\,2}}-\frac{n_s^{\,2}-n_s}{n(n-1)}\frac{\left(v_s-u_s\right)^2-\left(v_s-u_s\right)}{v_s\left(v_s-1\right)}\right),\\&-\frac{2}{n-1}\sum_{s<t}\frac{r-s}{r}\frac{r-t}{r}\frac{n_s}{n}\frac{n_t}{n}\frac{v_s-u_s}{v_s}\frac{v_t-u_t}{v_t}\end{aligned}$$

$$\begin{aligned}Var\left(\hat{\pi}_{10}\right)=&\sum_{s=0}^{r}\frac{s^2}{r^2}\left(\frac{n_s^{\,2}}{n^2}\frac{\left(v_s-u_s\right)^2}{v_s^{\,2}}-\frac{n_s^{\,2}-n_s}{n(n-1)}\frac{\left(v_s-u_s\right)^2-\left(v_s-u_s\right)}{v_s\left(v_s-1\right)}\right),\\&-\frac{2}{n-1}\sum_{s<t}\frac{s}{r}\frac{t}{r}\frac{n_s}{n}\frac{n_t}{n}\frac{v_s-u_s}{v_s}\frac{v_t-u_t}{v_t}\end{aligned}$$

$$Var\left(\hat{\pi}_{01}\right)=\sum_{s=0}^{r}\frac{\left(r-s\right)^2}{r^2}\left(\frac{n_s{}^2}{n^2}\frac{u_s{}^2}{v_s{}^2}-\frac{n_s{}^2-n_s}{n\left(n-1\right)}\frac{u_s{}^2-u_s}{v_s\left(v_s-1\right)}\right)-\frac{2}{n-1}\sum_{s<t}\frac{r-s}{r}\frac{r-t}{r}\frac{n_s}{n}\frac{n_t}{n}\frac{u_s}{v_s}\frac{u_s}{v_t},$$

$$Var\left(\hat{\pi}_{11}\right)=\sum_{s=0}^{r}\frac{s^2}{r^2}\left(\frac{n_s{}^2}{n^2}\frac{u_s{}^2}{v_s{}^2}-\frac{n_s{}^2-n_s}{n\left(n-1\right)}\frac{u_s{}^2-u_s}{v_s\left(v_s-1\right)}\right)-\frac{2}{n-1}\sum_{s<t}\frac{s}{r}\frac{t}{r}\frac{n_s}{n}\frac{n_t}{n}\frac{u_s}{v_s}\frac{u_s}{v_t},$$

$$Var\left(\hat{\pi}_P\right)=\sum_{s=0}^{r}\frac{s^2}{r^2}\left(\frac{n_s{}^2}{n^2}-\frac{n_s{}^2-n_s}{n\left(n-1\right)}\right)-\frac{2}{n-1}\sum_{s<t}\frac{s}{r}\frac{t}{r}\frac{n_s}{n}\frac{n_t}{n},$$

$$Var\left(\hat{\pi}_C\right)=\sum_{s=0}^{r}\left(\frac{n_s{}^2}{n^2}\frac{u_s{}^2}{v_s{}^2}-\frac{n_s{}^2-n_s}{n\left(n-1\right)}\frac{u_s{}^2-u_s}{v_s\left(v_s-1\right)}\right)-\frac{2}{n-1}\sum_{s<t}\frac{n_s}{n}\frac{n_t}{n}\frac{u_s}{v_s}\frac{u_s}{v_t}.$$

Notice that the variance estimates, except for $\hat{\pi}_P$, only go to zero when both $n \to \infty$ and $v_s \to \infty$ for all s.

The work involved in deriving these variance expressions can also be used to derive the covariance of the basic quantity estimators. To illustrate, I will derive the covariance of $\tilde{\pi}_{10}$ with $1-\tilde{\pi}_C$. I start with,

$$Cov\left(\tilde{\pi}_{10},1-\tilde{\pi}_C\right)=Cov\left(\sum_{s=0}^{r}\frac{N_s}{n}\frac{v_s-U_s}{v_s},\sum_{t=0}^{r}\frac{t}{r}\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right).$$

I decompose the above equation into summations of the variance and covariance of the summands, yielding

$$Cov\left(\tilde{\pi}_{10},1-\tilde{\pi}_C\right)=\sum_{s=0}^{r}\frac{s}{r}Var\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s}\right)+\sum_{s<t}\frac{s+t}{r}Cov\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s},\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right).$$

Then using the unbiased estimators derived for the variance and covariance terms in Section 3.15, I define an unbiased estimator and the associated estimate. I will not list estimates for all covariance terms to save space. Instead I will give the two unbiased estimates needed for Section 3.8, which are

$$Cov\left(\tilde{\pi}_{10},1-\tilde{\pi}_C\right)=\sum_{s=0}^{r}\frac{s}{r}\left(\frac{n_s{}^2}{n^2}\frac{\left(v_s-u_s\right)^2}{v_s{}^2}-\frac{n_s{}^2-n_s}{n\left(n-1\right)}\frac{\left(v_s-u_s\right)^2-\left(v_s-u_s\right)}{v_s\left(v_s-1\right)}\right)$$
$$-\frac{1}{n-1}\sum_{s<t}\frac{s+t}{r}\frac{n_s}{n}\frac{n_t}{n}\frac{v_s-u_s}{v_s}\frac{v_t-u_t}{v_t}$$

,

$$Cov\left(\tilde{\pi}_{01},\tilde{\pi}_{C}\right)=\sum_{s=0}^{r}\frac{r-s}{r}\left(\frac{n_s^{\,2}}{n^2}\frac{u_s^{\,2}}{v_s^{\,2}}-\frac{n_s^{\,2}-n_s}{n\left(n-1\right)}\frac{u_s^{\,2}-u_s}{v_s\left(v_s-1\right)}\right)$$
$$-\frac{1}{n-1}\sum_{s<t}\frac{r-s+r-t}{r}\frac{n_s}{n}\frac{n_t}{n}\frac{u_s}{v_s}\frac{u_t}{v_t}.$$

3.6 Uniformly Minimum Variance Unbiased Estimators

In Section 3.4, I showed that the basic quantities estimators are unbiased. Furthermore, in Sections 3.5 and 3.15 I showed that variance and covariance estimators for the basic quantity estimators are unbiased. Additionally both of these quantities are functions of the statistic $T=\left(n_0,n_1,\ldots,n_r,u_0,u_1,\ldots,u_r\right)$. I define the most general probability model for the Two-phase Plan, letting the repeated measurement phase be modeled using a multinomial distribution, and the verification phase be modeled conditionally upon v_s using a series of binomial distributions. I give the likelihood for this model below.

$$\mathcal{L}\left(\theta\right)=\left(\binom{n}{n_0\;n_1\;\ldots\;n_r}\prod_{s=0}^{r}P(S_i=s)^{n_s}\right)\left(\prod_{s=0}^{r}\binom{v_s}{u_s}P(X_i=1\,|\,S_i=s)^{u_s}P(X_i=0\,|\,S_i=s)^{v_s-u_s}\right)$$

It is well known that $\left(n_0,n_1,\ldots,n_r\right)$ is the complete sufficient statistic for the multinomial model used in the first phase, and that u_s is the complete sufficient statistic for the binomial model used in modeling the verification phase for bin s. Furthermore given each v_s , the results of the first phase and the second phase are independent and the binomial distributions are independent of one another, see Section 3.15. Given these facts it should be clear that T is a complete sufficient statistic for this model. The proof of this is tedious and not enlightening and is thus omitted. Thus the unbiased property and the complete sufficient properties allows for use of the Lehmann & Scheffé theorem (1950, 1955) to prove that both the basic quantity estimators and their variance estimators are uniformly minimum variance unbiased estimators for the model described in the above likelihood expression.

3.7 Consistency of Basic Quantity Estimators

Section 3.4 showed that the estimates for the basic quantities found in Table 3.2 are unbiased. Furthermore it is clear the variance estimates for the basic quantities in Section 3.5 approach zero as the appropriate design parameters n, v_s go to infinity. However if you consider Equation (3.2), and Equations (3.9) and (3.14) it is clear the theoretical variance for $\hat{\pi}_{10}$ approach zero as well. Showing the equivalent results for the other basic quantities is similar. Therefore the basic quantity estimates are consistent, that is, $\hat{\pi}_* \to \pi_*$ when $n \to \infty$ and $v_s \to \infty$ for all s, where π_* can be any of the basic quantities.

3.8 Estimates of Ratios of Basic Quantities

Recall that two of the quantities of interest μ_A and μ_B are not basic quantities but rather are ratios of basic quantities. I define an estimate for ratios of basic quantities as the ratio of the basic quantity estimates, that is,

$$\hat{\mu}_A = \frac{\hat{\pi}_{10}}{\hat{\pi}_{00} + \hat{\pi}_{10}} = \frac{\hat{\pi}_{10}}{1 - \hat{\pi}_C}, \ \hat{\mu}_B = \frac{\hat{\pi}_{01}}{\hat{\pi}_{01} + \hat{\pi}_{11}} = \frac{\hat{\pi}_{01}}{\hat{\pi}_C} \tag{3.3}$$

Unfortunately these ratio estimates are biased, and there is no apparent way to derive an unbiased variance estimates for these ratio estimates. However the amount of bias is small when compared to the bias in other estimation techniques. I show the results of an experiment to quantify this bias in Section 3.10.

Since the basic quantity estimators are consistent, see Section 3.7 above, it is a clear consequence of Slutsky's theorem (1925), that the ratio estimators are consistent as well. That is, $\hat{\pi}_* / \hat{\pi}_{**} \to \pi_* / \pi_{**}$ when $n \to \infty$ and $v_s \to \infty$ for all s, where π_* and π_{**} can be any of the basic quantities, provided $\pi_{**} > 0$.

To estimate the variance of the estimates in Equation (3.3), I propose using the Taylor series approximation for a ratio of random variables in Equation (3.4) below,

$$Var\left(\tfrac{N}{D}\right) \simeq \frac{E^2[N]}{E^2[D]}\left(\frac{Var(N)}{E^2[N]} - \frac{Cov(N,D)}{E[N]E[D]} + \frac{Var(D)}{E^2[D]}\right). \tag{3.4}$$

To find a specific variance estimate I use the results of Section 3.5 for the variance and co-variance terms and the estimates themselves for the expectation terms. I will not state the complete variance estimate for $\hat{\mu}_A$ and $\hat{\mu}_B$ because the expression is so long that it can only be understood by parsing it back into the various components. The accuracy of the variance estimates will be evaluated in Section 3.10.

3.9 Camshaft Example

To give a tangible example of how the estimates are used, I calculated the closed form estimates and their associated standard error estimates for the Camshaft example used previously in Section 2.5. The data set is slightly altered from that found in Section 2.5. Five verifications were added in the non-central bins, making the verifications conform to the recommended plan. The addition was necessary because the closed form estimates cannot be calculated without some verification from all bins in which parts fell.

Table 3.5 – Camshaft Example Data Set

Number of Passes (s)	0	1	2	3	4	5
Number of Camshafts (n_s)	29	9	7	33	132	290
Number Verified (v_s)	5	5	7	33	5	5
Number Conforming among Verified (u_s)	0	0	2	33	5	5

Table 3.6 – Camshaft Example Closed Form Estimates vs. Beta-Binomial MLE

Parameter	μ_A	μ_B	π_C
Closed Form			
Estimate	0.0884	0.0893	0.9140
Std. Error	0.0248	0.0062	0.0126
Maximum Likelihood			
Estimate	0.0903	0.0894	0.9139
Std. Error	0.0236	0.0061	0.0126

The closed form estimates are quite similar to the ML estimates in Camshaft example but with slightly higher standard errors. The largest difference occurs in the estimate of μ_A.

3.10 Simulation Study of Closed Form Estimates

A simulation with one million replications was conducted for each of the 32 sets of parameter combinations described in Table 2.4. For each replication the estimates and their associated standard error estimates were calculated. The mean and variance of the one million replications was then calculated for each quantity. The properties of the closed form estimates are compared to the simulation results from Chapter 2. The accuracy of the standard error estimates is also examined.

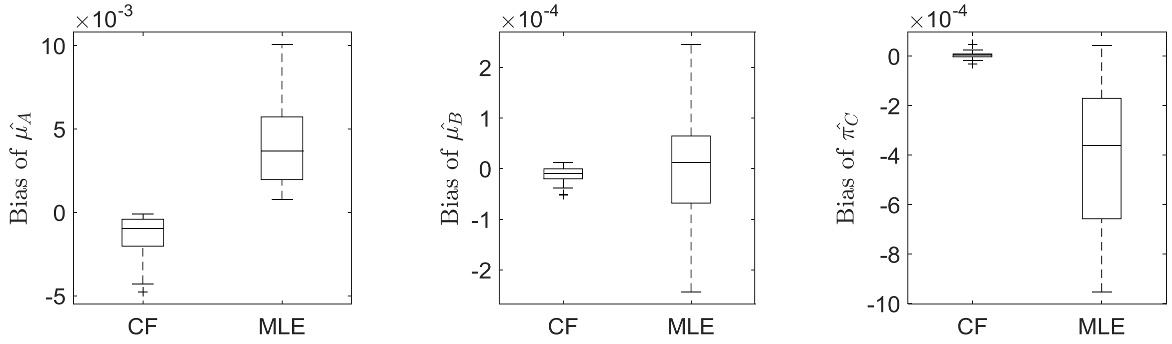

Figure 3.1 – Bias Comparison - Closed Form(CF) vs. MLE – Beta-Binomial data

Factorial experiment run at all combinations with $n = 500$

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 3.1 shows that the bias for the closed form (CF) estimates is less than that of the ML estimates even though the underlying beta-binomial assumption is met. Oddly the direction of the bias of $\hat{\mu}_A$ is positive for the ML estimates and negative for the CF estimates. Recall the closed form basic quantity estimate $\hat{\pi}_C$ is unbiased and thus the bias present can be used to gage the size of simulation error.

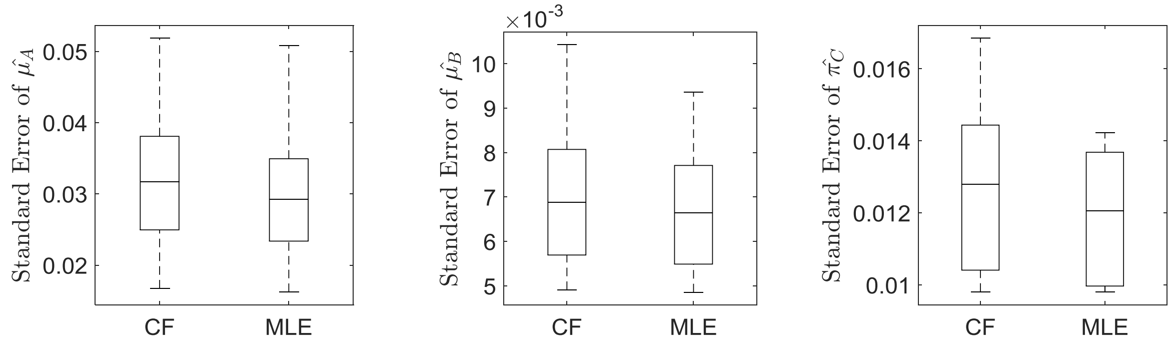

Figure 3.2 –Standard Error Comparison - Closed Form(CF) vs. MLE – Beta-Binomial Data

Factorial experiment run at all combinations with $n = 500$

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 3.2 shows a comparison of the standard errors of the CF and ML estimates. The standard errors of the ML estimates are slightly lower than the CF estimates for all three quantities of interest. This is expected because assumptions like those of the beta-binomial model are often made to reduce the number of parameters and thus the reduce variance. The beta-binomial model has five free parameters while the closed form estimates implicitly assume multinomial and binomial models with a total of $2r+1$ free parameters.

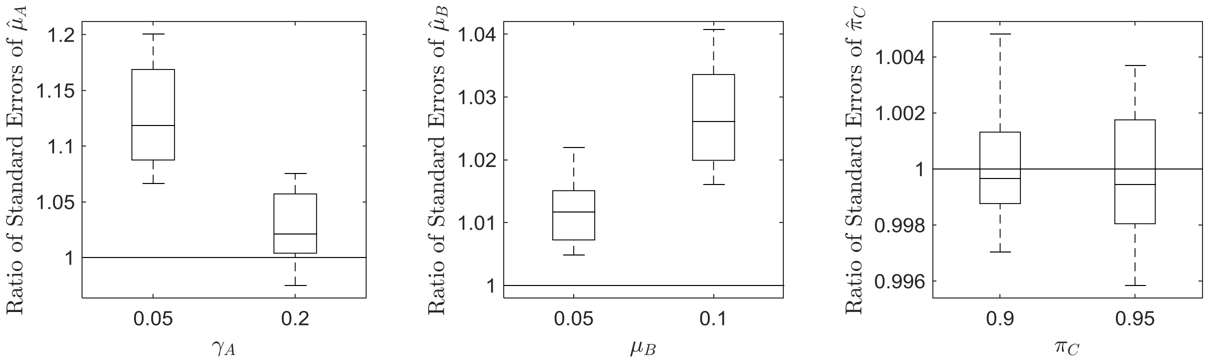

Figure 3.3 –CF Standard Error – Mean Estimated over Simulated – Beta-Binomial Data

Factorial experiment run at all combinations with $n = 500$

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 3.3 shows the ratio of mean of the standard error estimates in the simulation over the standard deviation of the parameter estimates in the simulation. The closer the mean of the standard error estimates is to the observed standard deviation in the estimates, the better I deem the standard error estimates to be. Thus the closer the ratio is to one the better the standard error estimate is. The ratio for μ_A is a little above one, indicating the standard error estimates slightly overestimate the standard deviation observed in the simulation. The ratio for μ_B is closer to one indicating that the standard error estimate for $\hat{\mu}_B$ is more accurate than that of $\hat{\mu}_A$. The standard error estimate for the basic quantities are unbiased, thus the plot for π_C can be used to gage the simulation error present. Please see Section 2.12 for the simulation results for the asymptotic standard error estimates for the beta-binomial ML estimates. The ratios are close enough to one to indicate that the standard error estimates described in Sections 3.5 and 3.8 are sufficiently accurate.

Rare Events

In the industrial setting a part that passes inspection five out of five times in the repeated measurement phase is almost assuredly a conforming part. However the probability that a part in Bin 5 is a non-conforming part is non-zero. In simulations I observe rare events where not all parts verified in Bin 5 are conforming. Unfortunately these rare events have a dramatic impact on the estimation of the quantities of interest, μ_A especially.

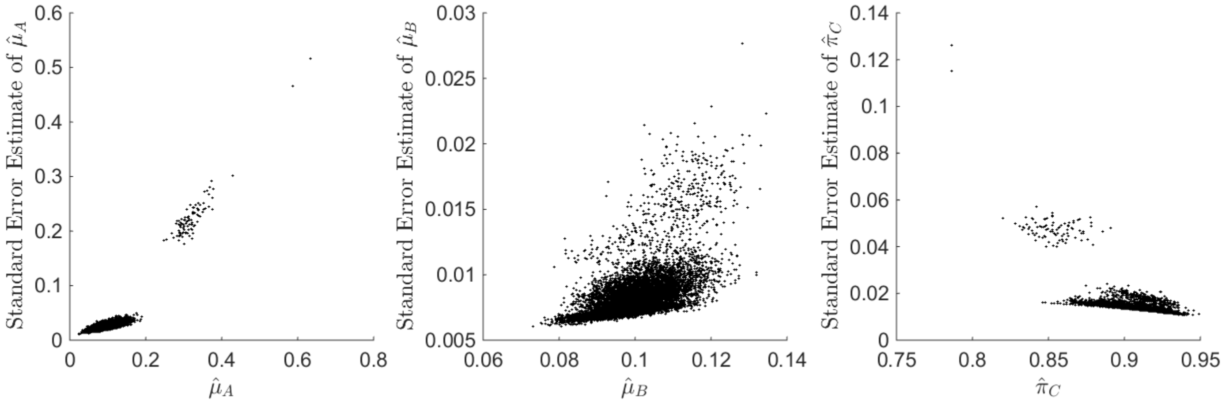

Figure 3.4 – Rare Events Scatter Plot

$n = 500 \;\; \mu_A, \mu_B = 0.1; \; \pi_C = 0.9; \; \gamma_A, \gamma_B = 0.1;$

Figure 3.4 shows a scatter plot of ten thousand simulated estimates and their corresponding standard error estimates. The data were generated using a beta-binomial model with μ_A and μ_B equal to 0.1, π_C equal to 0.9, and γ_A and γ_B equal to 0.1. First notice that there are two outliers in the plots for μ_A and π_C. These correspond to instances where only four of the five parts verified from Bin 5 were found to be conforming. Thus the CF estimate implicitly estimates that 20% of the parts in Bin 5 are non-conforming and thus misclassified; this dramatically inflates $\hat{\mu}_A$. Fortunately the standard error estimate in this scenario is also inflated and thus gives an appropriate warning that the estimate is not accurate. Additionally this anomaly should be visible when viewing the data table, thus giving yet another warning. While an equivalent problem does exist in estimating μ_B the results are not as dramatic. There is also a second group of outliers, this represents the less rare event when only four out of five parts verified from Bin 4 are found to be conforming.

Note that these rare events greatly influence the standard deviation of the CF estimates in Figure 3.2, and that removing these rare events would improve the performance of the CF estimates. Practically speaking there is no completely satisfying way to deal with these rare events without doing further verifications. One possible way to avoid these scenarios would be to verify no parts from Bins 4 and 5 and assume the probability of a part from either of those bins being non-conforming is zero. This would eliminate the rare events and reduce the variance. However this would go against one of the primary advantages and motivations for the CF estimates, that is, they do not have unnecessary assumptions. Ultimately these events are very rare and when someone comes across such an event, the standard error estimates give an appropriate warning. Note that this also happens when using the ML estimates. While the ML estimates are less affected in these rare scenarios both estimates are so bad that in practice remedial action should be taken regardless of the estimation technique used. I would recommend taking 20 more verifications from the bin where the non-conforming part was found.

3.11 Optimal Allocation

Using the Lagrange Multiplier method I can determine the optimal way to allocate a fixed number of verifications V across the different bins for each of the basic quantities. Unfortunately this cannot be used in practice because it will require exact knowledge of the properties of the measurement system and process before the study has been conducted. However it can give us insight into how verifications should be allocated.

McNamee (2002) considered optimal allocation for a similar two-phased plan for assessing measurement systems. However, McNamee considered a categorical measurement system that could be mapped to a BMS instead of a BMS being applied repeatedly. She also allowed different costs for verifying subjects from different categories. Some of content of the paper is similar to this section however the allocation is based on different estimates and incorporates different verification costs into the allocations. The application of the results is very different because McNamee presents the optimal allocation solution as though it could be implemented in practice.

I will derive the optimal allocation for π_{10} and then give the results for all basic quantities. Starting with the Lagrangian, I expand the variance term to get,

$$L = Var\left(\sum_{s=0}^{r}\frac{s}{r}\frac{n_s}{n}\frac{v_s - u_s}{v_s}\right) + \lambda\left(\sum_{s=0}^{r} v_s - V\right)$$

$$L = \sum_{s=0}^{r}\left(\frac{s}{r}\right)^2 Var\left(\frac{N_s}{n}\frac{v_s - U_s}{v_s}\right) + 2\sum_{s<t}\frac{s}{r}\frac{t}{r}Cov\left(\frac{N_s}{n}\frac{v_s - U_s}{v_s}, \frac{N_t}{n}\frac{v_t - U_t}{v_t}\right) + \lambda\left(\sum_{s=0}^{r} v_s - V\right).$$

But in Section 3.15 I show that,

$$Cov\left(\frac{N_s}{n}\frac{v_s - U_s}{v_s}, \frac{N_t}{n}\frac{v_t - U_t}{v_t}\right) = E\left[\frac{v_s - U_s}{v_s}\right]E\left[\frac{v_t - U_t}{v_t}\right]Cov\left(\frac{N_s}{n}, \frac{N_t}{n}\right).$$

And since $E\left[\frac{v_s - U_s}{v_s}\right] = P\left(X_i = 1 | S_i = s\right)$ does not depend on v_s, I have that

$$\frac{\partial Cov}{\partial v_s}\left(\frac{N_s}{n}\frac{v_s - U_s}{v_s}, \frac{N_t}{n}\frac{v_t - U_t}{v_t}\right) = \frac{\partial Cov}{\partial v_t}\left(\frac{N_s}{n}\frac{v_s - U_s}{v_s}, \frac{N_t}{n}\frac{v_t - U_t}{v_t}\right) = 0.$$

Therefore the derivative of L with respect to v_s is equal to

$$\frac{\partial L}{\partial v_s} = \frac{\partial}{\partial v_s}\left(\sum_{k=0}^{r} Var\left(\frac{k}{r}\frac{N_k}{n}\frac{v_k - U_k}{v_k} \right) + \lambda\left(\sum_{k=0}^{r} v_k - V \right) \right) = \frac{\partial Var}{\partial v_s}\left(\frac{s}{r}\frac{N_s}{n}\frac{v_s - U_s}{v_s} \right) + \lambda .$$

Then using results from Section 3.15 I expand this to

$$\frac{\partial L}{\partial v_s} = \frac{s^2}{r^2}\frac{\partial}{\partial v_s}\left(E\left[\frac{{N_s}^2}{n^2} \right] E\left[\frac{(v_s - U_s)^2}{{v_s}^2} \right] - E^2\left[\frac{N_s}{n} \right] E^2\left[\frac{v_s - U_s}{v_s} \right] \right) + \lambda .$$

But again $E\left[\frac{v_s - U_s}{v_s} \right] = P(X_i = 1 | S_i = s)$ does not depend on v_s. I exploit this multiple times to simplify the expression to

$$\frac{\partial L}{\partial v_s} = -\frac{s^2}{r^2} E\left[\frac{{N_s}^2}{n^2} \right] \frac{\partial}{\partial v_s}\left(E\left[\frac{(v_s - U_s)^2}{{v_s}^2} \right] \right) + \lambda$$

$$= -\frac{s^2}{r^2} E\left[\frac{{N_s}^2}{n^2} \right] \frac{\partial}{\partial v_s}\left(Var\left[\frac{v_s - U_s}{v_s} \right] + E^2\left[\frac{v_s - U_s}{v_s} \right] \right) + \lambda$$

$$= -\frac{s^2}{r^2} E\left[\frac{{N_s}^2}{n^2} \right] \frac{\partial}{\partial v_s}\left(Var\left(\frac{v_s - U_s}{v_s} \right) \right) + \lambda .$$

I then substitute in the expression derived in Section 3.15,

$$\frac{\partial L}{\partial v_s} = -\frac{s^2}{r^2}\left(\frac{1}{n} P(S_i = s) + \frac{(n-1)}{n} P^2(S_i = s) \right) \frac{\partial}{\partial v_s}\left(\frac{1}{v_s} P(X_i = 0 | S_i = s) P(X_i = 1 | S_i = s) \right) + \lambda .$$

Differentiating I have that

$$\frac{\partial L}{\partial v_s} = -\frac{1}{{v_s}^2}\frac{s^2}{r^2} f_s + \lambda ,$$

where

$$f_s = P(X_i = 0 | S_i = s) P(X_i = 1 | S_i = s) \left(\frac{1}{n} P(S_i = s) + \frac{(n-1)}{n} P^2(S_i = s) \right).$$

Now doing the typical Lagrange's multiplier method steps,

$$\frac{\partial L}{\partial v_s} = 0 \Rightarrow \lambda = \frac{1}{{v_s}^2}\frac{s^2}{r^2} f_s \Rightarrow v_s = \frac{s}{r}\sqrt{\frac{f_s}{\lambda}} \Rightarrow V = \sum_{s=0}^{r} \frac{s}{r}\sqrt{\frac{f_s}{\lambda}} \Rightarrow \lambda = \left(\frac{\sum_{s=0}^{r} \frac{s}{r}\sqrt{f_s}}{V} \right)^2 ,$$

I have that,

$$v_s = V\frac{\frac{s}{r}\sqrt{f_s}}{\sum_{m=0}^{r}\frac{s}{r}\sqrt{f_m}} = V\frac{\frac{s}{r}\sqrt{P(X_i=1|S_i=s)P(X_i=0|S_i=s)\left(\frac{1}{n}P(S_i=s)+\frac{n-1}{n}P^2(S_i=s)\right)}}{\sum_{m=0}^{r}\frac{m}{r}\sqrt{P(X_i=1|S_i=m)P(X_i=0|S_i=m)\left(\frac{1}{n}P(S_i=m)+\frac{n-1}{n}P^2(S_i=m)\right)}}.$$

So as, $n\to\infty$,

$$v_s \to V\frac{\frac{s}{r}P(S_i=s)\sqrt{P(X_i=1|S_i=s)P(X_i=0|S_i=s)}}{\sum_{m=0}^{r}\frac{m}{r}P(S_i=m)\sqrt{P(X_i=1|S_i=m)P(X_i=0,S_i=m)}}\ .$$

I used a similar derivation for each of the other basic quantities; please see the results in Table 3.7.

Table 3.7 – Optimal Allocation for Basic Quantities

π_{00} and π_{01}	$v_s \propto \frac{s}{r}P(S_i=s)\sqrt{P(X_i=1|S_i=s)P(X_i=0|S_i=s)}$
π_{10} and π_{11}	$v_s \propto \frac{r-s}{r}P(S_i=s)\sqrt{P(X_i=1|S_i=s)P(X_i=0|S_i=s)}$
π_C	$v_s \propto P(S_i=s)\sqrt{P(X_i=1|S_i=s)P(X_i=0|S_i=s)}$
π_P	Variance does not depend on v_s

These expressions give some insight into why verifying parts in the middle tends to be most effective in targeted verification. A basic understanding of stratified sampling explains why the weights are proportional to the various constants and the probability of a part falling in that bin, $P(S_i=s)$, which is analogous to the size of the stratum, and the product term, $\sqrt{P(X_i=1|S_i=s)P(X_i=0|S_i=s)}$, which is analogous to the standard deviation of the response for the stratum. It is this last factor that explains why so many verifications should be done in the middle, despite the few parts found there. In order for the product term to be greater than zero by some practically significant amount it must be plausible that parts from a particular bin can be either conforming or non-conforming.

3.12 Camshaft Example - Optimal Allocation

Unfortunately, the analytical solution in Section 3.11 does not incorporate the limitations that $v_s \le n_s$ for all s. Incorporating these constraints into the problem leaves no analytical solution. However the solution may be found using numerical optimization. The optimization procedure reallocates one verification at a time. Specifically it reduces the number of verifications in one bin and increases the number of verification in another bin. It calculates the decrease in the estimated variance for either μ_A, μ_B or π_C, see Sections 3.5 and 3.8, for all possible reallocations and selects the one that reduces the specified estimated variance the most. In order for the variance estimate to be defined I impose the constraint that, $v_s \ge 2$ except when $v_s = n_s < 2$. The optimization procedure will never consider reallocations that violate these constraints. To calculate the optimal allocation, n_s and $P\left(X_i = 1 \middle| S_i = s\right)$ must be known for all s.

For a concrete example I calculated the optimal allocation for the Camshaft example. The observed repeated measurement phase data from the camshaft example is used for n_s while $P\left(X_i = 1 \middle| S_i = s\right)$ is calculated based on the parameter values estimated in Table 2.3. The algorithm for finding the optimal allocation minimizes the expected standard error for a specified quantity of interest. Table 3.8 shows the calculated optimal allocations for each of the three quantities. Table 3.9 shows the expected standard errors for each of the allocations as well as the observed standard errors for the complete data set shown in Table 3.5.

Table 3.8 – Camshaft Example Optimal Allocation

Number of Passes (s)	0	1	2	3	4	5
Number of Camshafts (n_s)	29	9	7	33	132	290
Number Verified (v_s)						
-Recommended Plan	5	5	7	33	5	5
-Optimal Allocation μ_A	2	2	7	18	19	12
-Optimal Allocation μ_B	6	9	7	26	10	2
-Optimal Allocation π_C	2	5	7	21	17	8

The differences in the optimal allocations are easily explained by analytical results found in Table 3.7. The optimal allocation for μ_A is skewed towards the bins representing more passes because of the constant $\frac{s}{r}$, while the optimal allocation for μ_B is skewed towards those bins representing fewer passes because of the constant $\frac{r-s}{r}$, while the optimal allocation for π_C is somewhere in between. Notice that Bin 2 was completely verified in all three optimal allocations. Furthermore Bin 3 was allocated the greatest number of verifications in two of the three optimal allocations and that it was very close to the most in the third. This confirms that overall Bins 2 & 3 are the two most important bins to verify as was found in Section 2.7. The main difference between the recommended plan and the optimal allocation plans seems to be that not all parts in Bin 3 were verified, and that those verifications were allocated to other bins, primarily Bin 4.

Table 3.9 – Camshaft Example Optimal Allocation Standard Errors

Parameter	μ_A	μ_B	π_C
Recommended Plan	0.0469	0.0064	0.0131
Optimal Allocation			
For μ_A	0.0355	0.0066	0.0128
For μ_B	0.0477	0.0064	0.0130
For π_C	0.0365	0.0064	0.0128
% Reduction of Standard Error Optimal vs. Recommended			
For μ_A	24.3%	-2.0%	2.1%
For μ_B	-1.7%	0.6%	0.1%
For π_C	22.2%	-0.2%	2.6%

Table 3.9 shows the expected standard errors for each of the allocations shown in Table 3.8. It also shows the observed standard error data in Table 3.5. It gives the percentage the optimal plan reduced the expected standard error over the recommended plan for each of the optimal allocations. The optimal allocations for μ_A and π_C both give a meaningful reduction in the expected standard error of $\hat{\mu}_A$ when compared to the recommended plan. The other changes in expected standard errors are not significant. Overall the optimal allocation for π_C seems to be the best all around, but as discussed previously this cannot be implemented in practice because the true parameters values are unknown.

Optimal allocations were also calculated for 1000 simulated data sets for each of the 32 parameter combinations possible with Table 2.4. In order to properly interpret the allocation data it must be compared to the number of camshafts in each bin. And given that there are multiple bins the data are in some sense three-dimensional. There is no immediately obvious way to visually represent all three dimensions and attempts to reduce the dimension of the data have been misleading. Exploring the data in a variety of ways, I found that the results are very much the same as the camshaft example with skew towards parts that passed more often in the optimal allocation for μ_A and skew towards parts that passed less often in the optimal allocation for μ_B. Bin 2 was almost always completely verified, and Bin 3 most often had the largest share of verifications. The optimal allocations for μ_A and μ_B traded off gains in the one being optimized for with losses in the other. The optimal allocation scheme for π_C was again the best overall, with a reduction in standard error of 13.6%, 5.2%, and 4.8% respectively of average for $\hat{\mu}_A$, $\hat{\mu}_B$, and $\hat{\pi}_C$ respectively when compared the recommended plan. These gains are not trivial but reasonably small, indicating that the recommended plan is reasonably close to that optimal performance which is only attainable with pre-knowledge of the quantities of interest.

3.13 Discussion

The closed form estimates have small sample performance similar to the beta-binomial ML estimates. Additionally, the closed form estimates have theoretical properties that are relevant in practice. Most importantly, provided that the assessment study is conducted properly, as described in Chapter 2, the closed form estimates are consistent. Additionally the estimates for the basic quantities have very attractive theoretical properties, including being unbiased and more particularly the UMVUE for the most general model. I would recommend using the closed form estimates particularly in a larger study where bias is a concern.

3.14 Future Work

Chapter 5 develops an alternative estimation technique for the probabilities related to the verification phase which could be applied to the closed form estimates. Unfortunately, using this alternative would mean the estimates would no longer be closed form, at least in any practical sense. However in Chapter 5, this restriction was seen to reduce the variation of the estimates, and is reasonable to think that the restriction would have the same impact on the closed form estimation of this chapter.

Additionally, the plan used in this chapter was developed with the beta-binomial ML estimates in mind. It may be worth investigating whether choosing some of the design parameters, like the number of repeated measurements, differently would improve efficiency when using the closed form estimates.

3.15 Variance Derivation

In this section I give a full derivation of the variance estimator of $\hat{\pi}_{10}$; this section was briefly summarized in Section 3.5.

Distribution Properties

Many of the distribution properties derived herein are obvious when considered in isolation. However, when one considers these properties in the context of the sequential plan, doubts may arise. This section shows these properties in the context of the sequential plan.

To evaluate the summands in Equation (3.2), I first derive distributions of U_s and $\left(N_0, N_1, \ldots, N_r\right)$. To do this I introduce some new notation. Let $\Omega_s = \{i \mid S_i = s, Z_i = 1\}$ represent the set of indices of the parts verified for each number of passes, s. U_s is distributed according to a binomial distribution with number of trials v_s and probability $P\left(X_i = 1 \mid S_i = s\right)$. To show this I will use moment generating functions. First I use the definition of a moment generating function then substitute an expression for U_s written in terms of Ω_s, yielding

$$M_{U_s}(t) = E\left[e^{tU_s}\right] = E\left[\exp\left(t\sum_{i\in\Omega_s} X_i\right)\right].$$

I then use the law of total expectation conditioning on Ω_s to get

$$M_{U_s}(t) = E\left[E\left[\exp\left(t\sum_{i\in\Omega_s} X_i\right)\middle|\Omega_s\right]\right] = E\left[E\left[\prod_{i\in\Omega_s} e^{tX_i}\middle|\Omega_s\right]\right].$$

But the random variables X_i are independent given Ω_s, so the expectation of the product can be written as the product of the expectations, therefore

$$M_{U_s}(t) = E\left[\prod_{i\in\Omega_s} E\left[e^{tX_i}\middle|\Omega_s\right]\right]. \tag{3.5}$$

But X_i is independent of the properties of all other parts. Additionally X_i is independent of Z_i given S_i. Therefore $E\left[e^{tX_i}\middle|\Omega_s\right] = E\left[e^{tX_i}\middle|S_i = s, Z_i = 1\right] = E\left[e^{tX_i}\middle|S_i = s\right]$. Then using the definition of expectation it is clear that $E\left[e^{tX_i}\middle|\Omega_s\right] = P\left(X_i = 0 \mid S_i = s\right) + P\left(X_i = 1 \mid S_i = s\right)e^t$, which can be substituted into Equation (3.5), giving

$$M_{U_s}(t) = E\left[\prod_{i\in\Omega_s} P(X_i = 0 | S_i = s) + P(X_i = 1 | S_i = s)e^t\right].$$

However all parts have identical distributions. Therefore each multiplicand is identical; therefore I can replace the product with an exponent; thus

$$M_{U_s}(t) = E\left[\left(P(X_i = 0 | S_i = s) + P(X_i = 1 | S_i = s)e^t\right)^{|\Omega_s|}\right].$$

But the sampling protocol discussed in Chapter 2 implies that $|\Omega_s| = |\{i | S_i = s, Z_i = 1\}| = v_s$; thus I have,

$$M_{U_s}(t) = E\left[\left(P(X_i = 0 | S_i = s) + P(X_i = 1 | S_i = s)e^t\right)^{v_s}\right].$$

Now inside the expectation there are no longer any random variables. Therefore,

$$M_{U_s}(t) = \left(P(X_i = 0 | S_i = s) + P(X_i = 1 | S_i = s)e^t\right)^{v_s}.$$

Therefore by characterization of its moment generating function, U_s is distributed according to a binomial distribution with number of trials v_s and probability $P(X_i = 1 | S_i = s)$.

The distribution of $(N_0, N_1, ..., N_r)$ is multinomial with number of trials n and probabilities $(P(S_i = 0), P(S_i = 1), ..., P(S_i = r))$. The proof of this is very simple. S_i follows a categorical distribution with support $\{0, 1, ..., r\}$ and probabilities $(P(S_i = 0), P(S_i = 1), ..., P(S_i = r))$. Each S_i is independent of all S_j for all $j \neq i$. Thus $(N_0, N_1, ..., N_r)$ represents a count of the outcome of n iid categorical random variables with probabilities $(P(S_i = 0), P(S_i = 1), ..., P(S_i = r))$. Therefore $(N_0, N_1, ..., N_r)$ has multinomial distribution with number of trials n and probabilities $(P(S_i = 0), P(S_i = 1), ..., P(S_i = r))$. This implies that marginally N_s is distributed according to a binomial distribution with number of trials n and probability $P(S_i = s)$.

I will now show that given v_s, N_s and U_s are independent. To do this I must introduce new notation. Let $\Upsilon_s = \{i | S_i = s\}$ represent the set of indices of the parts which passed inspection s times. I will show independence using moment generating functions. First I start with the definition of a joint moment generating function and then apply the law of total expectation conditioning on Υ_s and Ω_s; this yields

$$M_{N_s,U_s}(t_1, t_2) = E\left[e^{t_1 N_s + t_2 U_s}\right] = E\left[E\left[e^{t_1 N_s + t_2 U_s} \middle| \Upsilon_s, \Omega_s\right]\right].$$

But N_s is simply a constant when Υ_s is given, therefore it can be moved to the outer expectation, giving

$$M_{N_s,U_s}(t_1,t_2)=E\left[e^{t_1N_s}E\left[e^{t_2U_s}\middle|\Upsilon_s,\Omega_s\right]\right]. \tag{3.6}$$

But U_s is independent of Υ_s given Ω_s. Therefore $E\left[e^{t_2U_s}\middle|\Upsilon_s,\Omega_s\right]=E\left[e^{t_2U_s}\middle|\Omega_s\right]$. And in the proof of the binomial distribution of U_s I showed that $E\left[e^{t_2U_s}\middle|\Omega_s\right]=M_{U_s}(t_2)$. Therefore substituting this result into Equation (3.6) I get,

$$M_{N_s,U_s}(t_1,t_2)=E\left[e^{t_1N_s}M_{U_s}(t_2)\right]=E\left[e^{t_1N_s}\right]M_{U_s}(t_2)=M_{N_s}(t_1)M_{U_s}(t_2).$$

This proves that N_s and U_s are independent given v_s.

Covariance term derivation

Now I will derive an estimate for the covariance terms seen in Equation (3.2). First I apply the definition of covariance, giving

$$Cov\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s},\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right)$$

$$=E\left[\frac{N_s}{n}\frac{v_s-U_s}{v_s}\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right]-E\left[\frac{N_s}{n}\frac{v_s-U_s}{v_s}\right]E\left[\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right].$$

Then I use the fact that U_s and N_s are independent for any given v_s to get,

$$Cov(\ldots)=E\left[\frac{v_i-U_s}{v_s}\right]E\left[\frac{v_t-U_t}{v_t}\right]E\left[\frac{N_s}{n}\frac{N_t}{n}\right]$$

$$-E\left[\frac{N_s}{n}\right]E\left[\frac{N_t}{n}\right]E\left[\frac{v_s-U_s}{v_s}\right]E\left[\frac{v_t-U_t}{v_t}\right].$$

I then simplify the expression using the definition of covariance, yielding

$$Cov(\ldots)=E\left[\frac{v_s-U_s}{v_s}\right]E\left[\frac{v_t-U_t}{v_t}\right]Cov\left(\frac{N_s}{n},\frac{N_t}{n}\right). \tag{3.7}$$

The terms in Equation (3.7) can be evaluated using basic properties of the distributions of $(N_0,N_1,\ldots,N_r)$ and U_s. First note that,

$$E\left[\tfrac{N_s}{n}\right]=\tfrac{E[N_s]}{n}=\tfrac{nP(S_i=s)}{n}=P(S_i=s),$$

$$Cov\left(\tfrac{N_s}{n},\tfrac{N_t}{n}\right)=-\tfrac{1}{n}P(S_i=s)P(S_i=t),$$

$$E\left[\tfrac{N_s}{n}\tfrac{N_t}{n}\right]=Cov\left(\tfrac{N_s}{n},\tfrac{N_t}{n}\right)+E\left[\tfrac{N_s}{n}\right]E\left[\tfrac{N_t}{n}\right]=\tfrac{n-1}{n}P(S_i=s)P(S_i=t)\ ,$$

$$E\left[\tfrac{v_s-U_s}{v_s}\right]=\tfrac{v_s-v_sP(X_i=1|S_i=s)}{v_s}=1-P(X_i=1|S_i=s)=P(X_i=0|S_i=s).$$

I use these four properties to show that $\frac{-n_s n_t}{(n-1)n^2}$ is an unbiased estimate for $Cov\left(\frac{N_s}{n},\frac{N_t}{n}\right)$,

$$E\left[\frac{-N_s N_t}{(n-1)n^2}\right]=\frac{-1}{n-1}E\left[\frac{N_s}{n}\frac{N_t}{n}\right]=\frac{-1}{n-1}\left(\frac{n-1}{n}P(S_i=s)P(S_i=t)\right)$$
$$=-\frac{1}{n}P(S_i=s)P(S_i=t)=Cov\left(\frac{N_s}{n},\frac{N_t}{n}\right).$$

Furthermore, it is clear that $\frac{v_s-U_s}{v_s}$ is an unbiased estimate for $E\left[\frac{v_s-U_s}{v_s}\right]$. Substituting these unbiased estimates into Equation (3.7) I have that an unbiased estimate for the covariance terms,

$$Cov\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s},\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right)=-\frac{1}{n-1}\frac{v_s-u_s}{v_s}\frac{v_t-u_t}{v_t}\frac{n_s}{n}\frac{n_t}{n}. \tag{3.8}$$

Additionally substituting in the theoretical quantities into Equation (3.7) gives the theoretical covariance, i.e. the quantity being estimated in terms of the underlying model properties,

$$Cov\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s},\frac{N_t}{n}\frac{v_t-U_t}{v_t}\right)$$
$$=-\frac{1}{n}P(\mathrm{X}_i=1|\mathrm{S}_i=s)P(\mathrm{X}_j=1|\mathrm{S}_j=t)P(S_i=s)P(S_j=t) \tag{3.9}$$

Variance term derivation

Next I derive an estimate for the variance terms seen in Equation (3.2). I start with the definition and use the independence of U_s and N_s to get,

$$Var\left(\frac{N_s}{n}\frac{v_s-U_s}{v_s}\right)=E\left[\frac{{N_s}^2}{n^2}\frac{(v_s-U_s)^2}{{v_s}^2}\right]-E^2\left[\frac{N_s}{n}\frac{v_s-U_s}{v_s}\right]$$
$$=E\left[\frac{{N_s}^2}{n^2}\right]E\left[\frac{(v_s-U_s)^2}{{v_s}^2}\right]-E^2\left[\frac{N_s}{n}\right]E^2\left[\frac{v_s-U_s}{v_s}\right]. \tag{3.10}$$

I will now evaluate the quantities seen in Equation (3.10), give estimates for those quantities and prove the corresponding estimators are unbiased. I use the previously shown result that N_s has binomial distribution with number of trials n and probability $P(S_i=s)$ to show that,

$$E\left[\frac{N_s}{n}\right]=\frac{E[N_s]}{n}=\frac{nP(S_i=s)}{n}=P(S_i=s), \tag{3.11}$$

$$E\left[\frac{{N_s}^2}{n^2}\right]=\frac{E\left[{N_s}^2\right]}{n^2}=\frac{nP(S_i=s)(1-P(S_i=s)+nP(S_i=s))}{n^2}=\frac{1}{n}P(S_i=s)+\frac{(n-1)}{n}P^2(S_i=s). \tag{3.12}$$

Defining an estimate for $E\left[\frac{{N_s}^2}{n^2}\right]$ is very easy because if I define the estimate as $\frac{{n_s}^2}{n^2}$, then the corresponding estimator is $\frac{{N_s}^2}{n^2}$ has $E\left[\frac{{N_s}^2}{n^2}\right]$ as its expectation. The estimate for $E^2\left[\frac{N_s}{n}\right]$ is not as obvious. However defining the estimate as $\frac{n_s^2 - n_s}{n(n-1)}$, makes the corresponding estimator $\frac{N_s^2 - N_s}{n(n-1)}$ an unbiased estimator of $E^2\left[\frac{N_s}{n}\right]$. To show this I use the results found in Equations (3.11) and (3.12) , which yields

$$E\left[\frac{N_s^2 - N_s}{n(n-1)}\right] = E\left[\frac{N_s^2}{n(n-1)}\right] - E\left[\frac{N_s}{n(n-1)}\right] = \frac{n}{n-1}\left(E\left[\frac{N_s^2}{n^2}\right] - \frac{1}{n}E\left[\frac{N_s}{n}\right]\right)$$
$$= \frac{n}{n-1}\left(\frac{1}{n}P(S_i = s) + \frac{(n-1)}{n}P^2(S_i = s) - \frac{1}{n}P(S_i = s)\right)$$
$$= \frac{n}{n-1}\left(\frac{(n-1)}{n}P^2(S_i = s)\right) = P^2(S_i = s).$$

Then using the result in Equation (3.11) again, I have that $E\left[\frac{N_s^2 - N_s}{n(n-1)}\right] = E^2\left[\frac{N_s}{n}\right]$ which shows $\frac{n_s^2 - n_s}{n(n-1)}$ is an unbiased estimate of $E^2\left[\frac{N_s}{n}\right]$.

Now $\left(\frac{v_s - u_s}{v_s}\right)^2$ is an unbiased estimate of $E\left[\left(\frac{v_s - U_s}{v_s}\right)^2\right]$ just as $\frac{n_s^2}{n^2}$ was an unbiased estimate of $E^2\left[\frac{N_s}{n}\right]$. Additionally $\frac{(v_s - U_s)^2 - (v_s - U_s)}{v_s(v_s - 1)}$ is an unbiased estimator of $E^2\left[\frac{v_s - U_s}{v_s}\right]$ in the same way $\frac{n_s^2 - n_s}{n(n-1)}$ was an unbiased estimate of $E^2\left[\frac{N_s}{n}\right]$. These derivations are very similar because both N_s and U_s have binomial distributions.

Now substituting the estimates into Equation (3.10) I have that,

$$Var\left(\frac{N_s}{n}\frac{v_s - U_s}{v_s}\right) = \frac{n_s^2}{n^2}\frac{(v_s - u_s)^2}{v_s^2} - \frac{n_s^2 - n_s}{n(n-1)}\frac{(v_s - u_s)^2 - (v_s - u_s)}{v_s(v_s - 1)} \tag{3.13}$$

Additionally substituting the theoretical quantities found in Equation (3.11) and (3.12) into Equation (3.10) I have that

$$\begin{aligned} Var\left(\frac{N_s}{n}\frac{v_s - U_s}{v_s}\right) = &\left(\frac{1}{n}P(S_i = s) + \frac{(n-1)}{n}P^2(S_i = s)\right) \\ &* \left(\frac{1}{v_s}P(X_i = 1 | S_i = s) + \frac{(v_s - 1)}{v_s}P^2(X_i = 1 | S_i = s)\right) \\ &- P^2(S_i = s)P^2(X_i = 1 | S_i = s) \end{aligned} \tag{3.14}$$

Finally substituting the unbiased estimates found in Equations (3.8) and (3.13) into Equation (3.2) yields an unbiased estimate for $Var\left(\hat{\pi}_{10}\right)$.

$$\begin{aligned}Var\left(\hat{\pi}_{10}\right)=\sum_{s=0}^{r}\frac{s^2}{r^2}\left(\frac{{n_s}^2}{n^2}\frac{\left(v_s-u_s\right)^2}{{v_s}^2}-\frac{{n_s}^2-n_s}{n\left(n-1\right)}\frac{\left(v_s-u_s\right)^2-\left(v_s-u_s\right)}{v_s\left(v_s-1\right)}\right)\\ -\frac{2}{n-1}\sum_{s<t}\frac{s}{r}\frac{t}{r}\frac{n_s}{n}\frac{n_t}{n}\frac{v_s-u_s}{v_s}\frac{v_t-u_t}{v_t}\end{aligned} \qquad (3.15)$$

Chapter 4 Targeted Verification with Conditional Sampling and Baseline Information

4.1 Introduction

Assessing a BMS becomes more difficult as the rarity of non-conforming parts increases. There exist many situations where any reasonably sized random sample from the manufacturing process will not contain a sufficient number of non-conforming parts to accurately estimate μ_A. An effective and practical solution to this was presented by Danila et al. (2010) using baseline information and conditional sampling. These design elements were shown to improve the precision of $\hat{\mu}_A$, $\hat{\mu}_B$, and $\hat{\pi}_C$ and are easy to implement for a BMS that is currently in use. Baseline information is a set of single y_i measurements done on a large sample of parts from the manufacturing process. This data is considered 'free' because it is most often collected independently from the assessment plan as part of ongoing manufacturing operations. Conditional sampling means re-sampling from parts that either passed or failed inspection once and only once by the BMS. A sample from parts that already failed inspection will yield many more non-conforming parts than a sample from the manufacturing process.

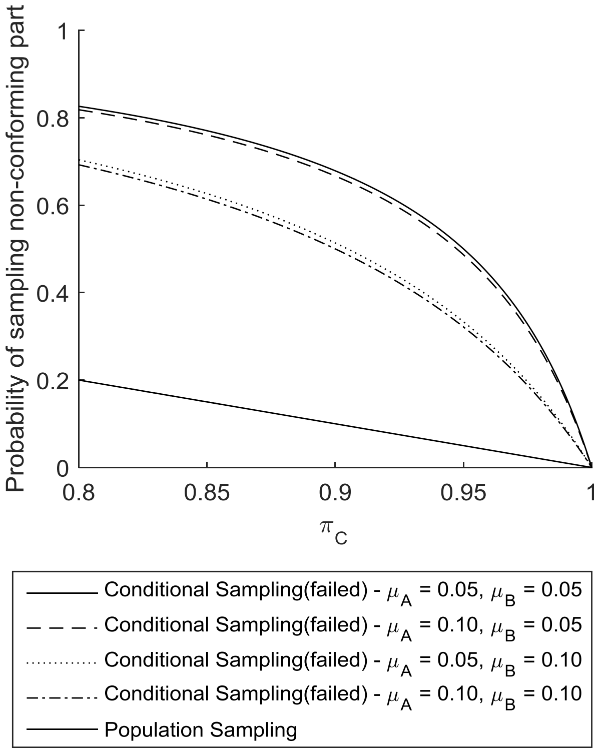

Figure 4.1 – Conditional Sampling Justification

Figure 4.1 shows how using conditional sampling will increase the probability of sampling non-conforming parts. This property is very useful because, when π_C is very close to one, the relative scarcity of non-conforming parts compared to conforming parts makes the estimate of μ_A far less precise than the estimate of μ_B. This makes conditional sampling preferable to population sampling when roughly equal estimate precision for $\hat{\mu}_A$ and $\hat{\mu}_B$ is desirable, which is true in most assessment studies. Thus by using conditional sampling, one can improve the efficacy of a BMS assessment study by bringing the number of conforming and non-conforming parts closer to balance. Actually when possible it may be advantageous to have more non-conforming parts than conforming parts because baseline information provides more information about μ_B than μ_A.

The structure of this chapter will be like that of Chapter 2. I will first propose a general three-stage plan, and derive a likelihood expression for said plan. I will argue that verifying from the middle bins is also effective in conditional sampling plans, and provide a recommended plan. I will then demonstrate the value of targeted verification within a conditional sampling plan. Next, I will compare the effectiveness of said plan, including bias, precision, and robustness to the recommended plan of Chapter 2. Finally, I discuss the impact of the baseline size on the plan and give a justification for the asymptotic variance approximations used.

4.2 Three-Phase Plan

This section details how to assess a BMS with a three-phased plan that allows for targeted verification. In the baseline phase, n_B parts are measured once with the BMS being assessed. Parts are then separated into groups of passing and failing parts. Let y_B denote the number of parts that passed inspection in the baseline phase. In the repeated measurement phase, n_P parts are selected randomly from the y_B passing parts and are measured r times. Similarly n_F parts are selected randomly from the $n_B - y_B$ failing parts and measured r times. The parts are then separated into bins based on the number of times a part passed inspection, including the initial baseline inspection. The bins are indexed by $s \in \{0,1,2,...,r+1\}$ which represents the number of times parts in said bin passed inspection. As earlier, let n_s denote the number of parts in bin s. Note that it is not necessary to retain information about how many parts in each bin passed or failed the initial baseline measurement; this will be addressed in Section 4.3.

In the verification phase, the experimenter decides how many parts to verify from each bin. Let v_s denote the number of parts verified from Bin s where $0 \le v_s \le n_s$. Setting v_s equal to zero indicates no parts are verified from bin s, whereas setting v_s equal to n_s indicates all parts from that bin are verified. Any other choice for v_s indicates a subset of parts is verified; this subset is to be selected using simple random sampling. A recommendation for choosing each v_s will be given in Section 4.5. After determining which parts will be verified, measure those parts with the gold standard recording u_s, the number from each bin that conformed to specification. If no parts are verified in bin s, then u_s is automatically equal to zero. The resulting data can be summarized as in Table 4.1.

Table 4.1 –Conditional Data Summary

Number of Parts Measured in Baseline	n_B	Number of Parts Sampled From Baseline Rejects	n_F
Number of Parts Passing Inspection in Baseline	y_B	Number of Parts Sampled From Baseline Passes	n_P

Number of Passes (Bin #)	0	1	$\cdots$	r	$r+1$
Number of Parts (Repeated Measurement Phase)	n_0	n_1	$\cdots$	n_r	n_{r+1}
Number Verified (Verification Phase)	v_0	v_1	$\cdots$	v_r	v_{r+1}
Number Conforming among Verified (Verification Phase)	u_0	u_1	$\cdots$	u_r	u_{r+1}

4.3 Likelihood Derivation

It is not obvious why the multinomial results of repeatedly measuring passing parts, $\left(n_0{}^P, n_1{}^P, \ldots, n_r{}^P\right)$, can be combined with the results of repeatedly measuring failing parts, $\left(n_0{}^F, n_1{}^F, \ldots, n_r{}^F\right)$ into one data set representing the total number of passes, including the baseline measurement, as

$$\left(n_0 = n_0{}^F, n_1 = n_0{}^P + n_1{}^F, n_2 = n_1{}^P + n_2{}^F, \ldots, n_r = n_{r-1}{}^P + n_r{}^F, n_{r+1} = n_r{}^P\right).$$

However this section, as part of the derivation of the likelihood, will show that $\left(n_0, n_1, n_2, \ldots, n_r, n_{r+1}\right)$ is a sufficient statistic for the probability model being used.

Some new notation is required. Let Y_i represent the outcome of the single binary measurement from the baseline phase for part i. Let S_i be the sum of the r binary measurements from the repeated measurement phase for part i , and let $T_i = Y_i + S_i$ represent the sum of the outcome of all binary inspections for part i . Note that not all parts are selected for the repeated measurement phase. Also note that T_i can be modeled in the same way as S_i, as seen in Chapter 2, but with $r+1$ repeated measurements instead of r repeated measurements.

Below I present the likelihoods for each of the three phases. In the baseline phase, the results are modeled using a binomial distribution. In the repeated measurement phase, the data is modelled using two multinomial distributions one for those sampled from failing parts and one those sampled from passing parts. Finally the verification phase is modeled using a series of binomial distributions, one for each of the $r+1$ bins. For each of the phases the probabilities for the binomial and multinomial distributions are based upon the beta-binomial model described in Section 2.3. The likelihoods for each phase are,

$$\mathcal{L}_B(\theta) = \binom{n_B}{y_B} P(Y_i = 1)^{y_B} P(Y_i = 0)^{n_B - y_B},$$

$$\mathcal{L}_{RM}(\theta) = \binom{n_P}{n_1{}^P\ n_2{}^P \ldots n_{r+1}{}^P} \left(\prod_{s=0}^{r} P(S_i = s \mid Y_i = 1)^{n_s{}^P} \right) \binom{n_F}{n_0{}^F\ n_1{}^F \ldots n_r{}^F} \left(\prod_{s=0}^{r} P(S_i = s \mid Y_i = 0)^{n_s{}^F} \right),$$

$$\mathcal{L}_V(\theta) = \prod_{s=0}^{r+1} \binom{v_s}{u_s} P(X_i = 1 \mid T_i = s)^{u_s} P(X_i = 0 \mid T_i = s)^{v_s - u_s}.$$

I will now derive an expression for $\mathcal{L}_{RM}(\theta)$ written in terms of T_i as opposed to $S_i \mid Y_i$. To do this I will start by using the definition of conditional probability, and replacing the factorial pieces with k_{RM} to get,

$$\mathcal{L}_{RM}(\theta) = k_{RM} \left(\prod_{s=0}^{r} \left(\frac{P(S_i = s, Y_i = 1)}{P(Y_i = 1)} \right)^{n_s{}^P} \right) \left(\prod_{s=0}^{r} \left(\frac{P(S_i = s, Y_i = 0)}{P(Y_i = 0)} \right)^{n_s{}^F} \right).$$

Then collecting the terms $P(Y_i = 1)$ and $P(Y_i = 0)$ noting that $n_1{}^P + n_2{}^P + \ldots + n_{r+1}{}^P = n_P$ and $n_0{}^F + n_1{}^F + \ldots + n_r{}^F = n_F$, I get,

$$\mathcal{L}_{RM}(\theta) = k_{RM} P(Y_i = 1)^{-n_P} P(Y_i = 0)^{-n_F} \left(\prod_{s=0}^{r} P(S_i = s, Y_i = 1)^{n_s{}^P} \right) \left(\prod_{s=0}^{r} P(S_i = s, Y_i = 0)^{n_s{}^F} \right). \quad (4.1)$$

Now I will find an expression for $P(S_i = s, Y_i = 1)$ in terms of T_i. First I recombine the conditions in a linear way such that both are preserved and recognize that $T_i = Y_i + S_i$, to obtain,

$$P(S_i = s, Y_i = 1) = P(Y_i + S_i = s+1, Y_i = 1) = P(T_i = s+1, Y_i = 1) .$$

Then I use the definition of conditional probability and then symmetry among the individual measurements to evaluate the conditional probability statement, yielding.

$$P(T_i = s+1, Y_i = 1) = P(Y_i = 1 \mid T_i = s+1) P(T_i = s+1) = \tfrac{s+1}{r+1} P(T_i = s+1) .$$

Using similar argumentation one can show that $P(S_i = s, Y_i = 0) = \tfrac{r+1-s}{r+1} P\left(T_i = s\right)$. Substituting these expressions into Equation (4.1) I have that

$$\mathcal{L}_{RM}\left(\theta\right) = k_{RM} P\left(Y_i = 1\right)^{-n_P} P\left(Y_i = 0\right)^{-n_F} \left(\prod_{s=0}^{r} \left(\tfrac{s+1}{r+1} P(T_i = s+1)\right)^{n_s{}^P} \right) \left(\prod_{s=0}^{r} \left(\tfrac{r+1-s}{r+1} P(T_i = s)\right)^{n_s{}^F} \right).$$

Then combining terms for the repeated measurement phases and using the combined notation, i.e. $\left(n_0 = n_0{}^F, n_1 = n_0{}^P + n_1{}^F, n_2 = n_1{}^P + n_2{}^F, ..., n_r = n_{r-1}{}^P + n_r{}^F, n_{r+1} = n_r{}^P\right)$, $\mathcal{L}_{RM}\left(\theta\right)$ can be written as,

$$\mathcal{L}_{RM}\left(\theta\right) = k_{RM} P\left(Y_i = 1\right)^{-n_P} P\left(Y_i = 0\right)^{-n_F} \left(\prod_{s=0}^{r+1} P(T_i = s)^{n_s} \right) \left(\prod_{s=1}^{r+1} \left(\tfrac{s+1}{r+1}\right)^{n_s{}^P} \right) \left(\prod_{s=0}^{r} \left(\tfrac{r+1-s}{r+1}\right)^{n_s{}^F} \right).$$

Notice that the only place where $\left(n_0{}^P, n_1{}^P, \ldots, n_r{}^P\right)$ and $\left(n_0{}^F, n_1{}^F, \ldots, n_r{}^F\right)$ appear is at the end of the expression that does not involve any model parameters. Thus by the factorization criteria for sufficiency $\left(n_0, n_1, n_2, ..., n_r, n_{r+1}\right)$ is a sufficient statistic for the model used in the repeated measurement phase. Combining the likelihood from the three phases, I have

$$\begin{aligned} \mathcal{L}\left(\theta\right) = k * P\left(Y_i = 1\right)^{y_B - n_P} P\left(Y_i = 0\right)^{n_B - y_B - n_F} \left(\prod_{s=0}^{r+1} P(T_i = s)^{n_s} \right) \\ * \left(\prod_{s=0}^{r+1} P(X_i = 1 \mid T_i = s)^{u_s} P(X_i = 0 \mid T_i = s)^{v_s - u_s} \right), \end{aligned} \qquad (4.2)$$

where

$$k = \binom{n_B}{y_B} \binom{n_P}{n_1{}^P \; n_2{}^P \ldots n_{r+1}{}^P} \binom{n_F}{n_0{}^F \; n_1{}^F \ldots n_r{}^F} \left(\prod_{s=1}^{r+1} \left(\tfrac{s+1}{r+1}\right)^{n_s{}^P} \right) \left(\prod_{s=0}^{r} \left(\tfrac{r+1-s}{r+1}\right)^{n_s{}^F} \right) \left(\prod_{s=0}^{r+1} \binom{v_s}{u_s} \right).$$

The likelihood derived in Equation (4.2) does not depend on the beta-binomial model assumptions and can be used with other models. Equation (4.3) gives the equivalent log-likelihood expression substituting in the expressions from the beta-binomial model, see Section 2.3,

$$\begin{aligned}\ell(\theta) = \log(\mathrm{k}) &+ (n_B - y_B - n_F)\log\left(\pi_C\mu_B + (1-\pi_C)(1-\mu_A)\right) \\ &+ (y_B - n_P)\log\left(\pi_C(1-\mu_B) + (1-\pi_C)\mu_A\right) \\ &+ \sum_{s=0}^{r+1}(n_s - \mathrm{v}_s)\log(p_s + q_s) + u_s \log p_s + (v_s - u_s)\log q_s \ .\end{aligned} \tag{4.3}$$

The likelihood expression in Equation (4.3) is used with data recorded as in Table 4.1 to calculate ML estimates.

4.4 Example

In order to give a tangible example, I will calculate estimates for the dataset used in Danila et al. (2013). The data is reproduced in Table 4.2 and the corresponding estimates are given in Table 4.3. The example has full verification data, and to get the targeted verification data, some of the verification information is discarded. This could have introduced some randomness into the outcome of the verification phase because in some bins a random sample of five parts was selected to be verified. However, all bins where a random sample was needed had either all conforming parts or all non-conforming parts, so there is only one possible outcome for the verification phase data under this proposed targeted verification plan. The fact that the parts in non-central bins were either all conforming or all non-conforming shows the wastefulness of full verification and the primary advantage of targeted verification.

Table 4.3 shows that the estimates for the robust targeted verification scheme, see Table 4.2, were very close to the full verification scheme estimates and the standard errors of the three quantities of interest are only marginally different. However this robust targeted verification scheme required only 34 gold standard measurements compared to 100 for the complete verification plan. Thus almost identical estimates were obtained with roughly one third of the effort. The standard targeted verification scheme estimates also performed well. There is some deviation in estimates for μ_A while the deviations of the other parameters is very low. The standard error estimates for $\hat{\mu}_A$ and $\hat{\pi}_C$ increased by 38% and 19% respectively. The estimates are performing slightly worse than the full verification scheme, however this verification scheme only used 14 gold standard measurements as compared to the 100 for the full verification plan. The targeted verification schemes are far more efficient in the use of gold standard measurements.

Table 4.2 – Example Data

Number of Parts Measured in Baseline	1243	Number of Parts Sampled From Baseline Rejects	100
Number of Parts Passing Inspection in Baseline	960	Number of Parts Sampled From Baseline Passes	0

Repeated Measurement Phase							
Number of Passes (Bin #)	0	1	2	3	4	5	6
Number of Parts	41	18	5	9	5	22	0
Verification Phase – Full Verification Plan							
Number Verified	41	18	5	9	5	22	0
Number Conforming among Verified	0	0	0	5	5	5	0
Verification Phase – Targeted Verification – Robust Scheme							
Number Verified	5	5	5	9	5	5	0
Number Conforming among Verified	0	0	0	5	5	5	0
Verification Phase – Targeted Verification – Standard Scheme							
Number Verified	0	0	5	9	0	0	0
Number Conforming among Verified	0	0	0	5	0	0	0

Table 4.3 – Example Estimates

	Parameter	μ_A	μ_B	π_C	γ_A	γ_B
Full Verification Plan						
	Estimate	0.134	0.086	0.820	0.141	0.020
	Standard Error	0.029	0.013	0.016	0.098	0.030
Robust Targeted Verification Plan						
	Estimate	0.136	0.086	0.819	0.151	0.021
	Standard Error	0.031	0.012	0.017	0.109	0.029
Standard Targeted Verification Plan						
	Estimate	0.146	0.085	0.816	0.187	0.022
	Standard Error	0.040	0.012	0.019	0.145	0.030

4.5 Recommended Conditional Plan

Here I present a recommended conditional sampling plan. The plan is consistent with principles developed in Chapter 2. A justification of the plan will be given in Section 4.6. The recommended plan is presented with a robustness option that can be used at the experimenter's discretion. One situation where the robustness add-on is suggested is if the standard plan is performed and the gamma values are estimated as greater than one half. In this case, the additional data needed for the robustness option should be collected and the estimates recalculated.

Baseline Phase

- Measure n_B parts once and record the number of parts that passed inspection as y_B .

Repeated Measurement Phase

- Randomly select n_F parts that failed inspection in the baseline phase and measure them seven additional times. Separate them into nine bins based upon the total number of times each part passed inspection
- *Robustness Option*: In addition, randomly select $n_P = 10$ parts that passed inspection in the baseline phase and measure them seven additional times. Separate them into nine bins based upon the total number of times each part passed inspection including the baseline measurement.
- Record the total number of parts in each bin as n_s

Verification Phase

- Measure all parts in the bins representing three or four out of seven total passes with the gold standard
- *Robustness Option*: In addition, randomly selected five parts from each of the other bins (where possible) and measure them with the gold standard
- For each bin record the number of parts that were measured to be conforming as u_s

4.6 Justification of Recommended Conditional Plan

Verification Phase

In order to justify the recommended conditional plan, first I present some experiments to justify the Verification Phase; specifically the choice to verify all of Bin 3 and 4. Recall in Chapter 2, I found that verifying parts in the central bins provided tremendous benefit while verifying non-central bins provided almost no benefit. Recall also the following verification order that gave close to optimal results: that is, "Select the next part from the bin closest to the middle that has not yet been exhausted. If two bins are equally close, choose from the bin with fewer passes." Figure 4.2 shows that verifying in this order is very effective when dealing with conditional sampling data as well. The dashed lines represent when each of the first 5 bins is exhausted. The near linear reduction of the standard errors between the dashed lines indicates that the decision to verify parts can essentially be done on a bin-by-bin basis, which dramatically simplifies the problem of determining a verification strategy.

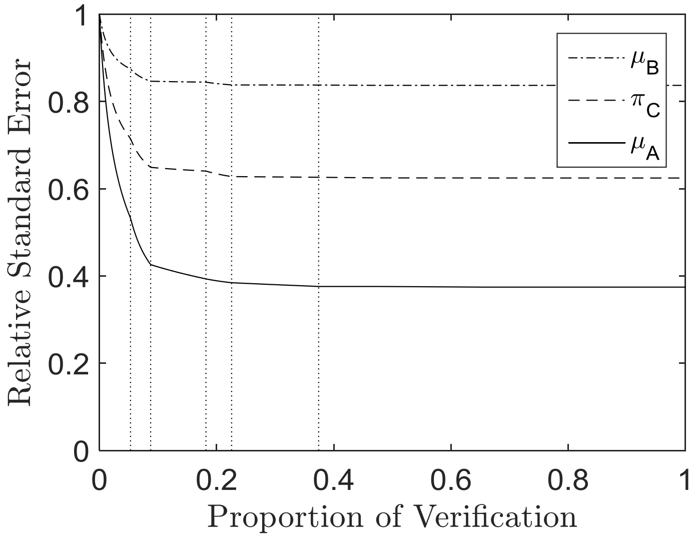

Figure 4.2 – Verification Proportion Plot – Conditional Sampling

$n_F = 500, n_P = 0, r = 7, n_B = 10000,\ \mu_A = 0.075, \mu_B = 0.075, \pi_C = 0.925, \gamma_A = 0.125, \gamma_B = 0.125$

Vertical dashed lines represent when each of the first 5 bins are exhausted

The effectiveness of verifying in this order is not isolated to this one set of parameters but holds for all the 32 sets of parameter values laid out in Table 2.4. To demonstrate this I show how much verifying different bins would reduce the standard error of the estimates for the quantities of interest. This is done in a successive fashion selecting the best bin to verify at each step.

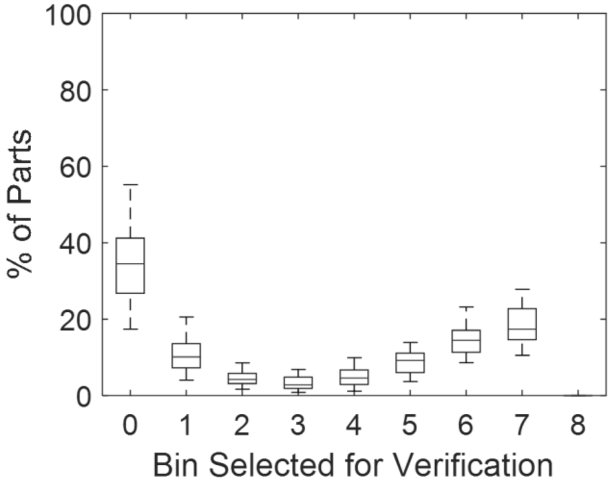

Figure 4.3 – Percentage of Parts in Each Bin

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $r = 7$,

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 4.3 shows what percentage of parts are in each bin, to give an impression of the relative amount of effort involved in verifying each bin. Each of the boxes represents the expected percentage for each of the 32 sets of parameter values described in Table 2.4. Notice that verifying Bins 3 & 4 requires verifying few parts relative to many of the other bins.

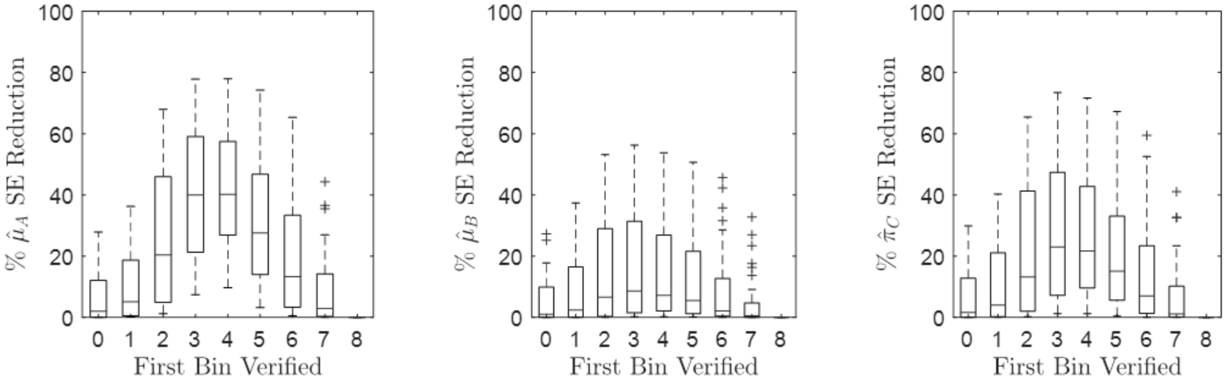

Figure 4.4 – Bin by Bin SE Reduction – Comparison to No Verification

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 4.4 shows how much the standard error is reduced when verifying each bin compared to the no verification plan. Each box-plot represents the 32 sets of parameter values described in Table 2.4 . All standard errors are based on the Fisher information asymptotic approximation. As was found in Chapter 2, the bins with the fewest parts, Bins 3 & 4, provide the greatest reduction in standard error. They provide on average approximately a 40% reduction in the standard error of $\hat{\mu}_A$, a 20% reduction in the standard error of $\hat{\pi}_C$, and a modest decrease in the standard error of $\hat{\mu}_B$. Now, suppose Bin 4 is verified first, I examine the value of verifying each of the other bins.

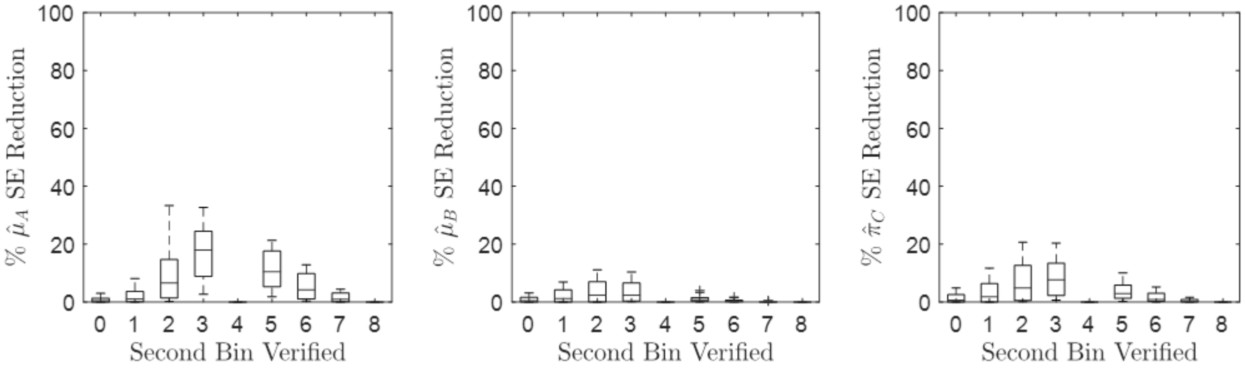

Figure 4.5 – Bin by Bin SE Reduction – Comparison to Bin 4 Only

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$,

$\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Figure 4.5 shows how much reduction in standard error verifying each bin provides after verifying only Bin 4. Each boxplot represents the 32 sets of parameter values described in Table 2.4 . All standard error figures used are based on the Fisher information asymptotic approximation. Given that Bin 3 and Bin 4 gave roughly equal reduction in standard errors in Figure 4.4, it is not surprising that after Bin 4 was verified Bin 3 becomes the best choice for further verification. Notice that the reduction in standard error is not as great as in Figure 4.4. But given that Bin 3 contains only 3.4% of the total number of parts on average and still gives a further 20% reduction in the standard error of $\hat{\mu}_A$, it is worth verifying the parts in Bin 3. Supposing then that Bins 3 & 4 have been verified, I will examine the merits of verifying each of the remaining bins, in a similar fashion.

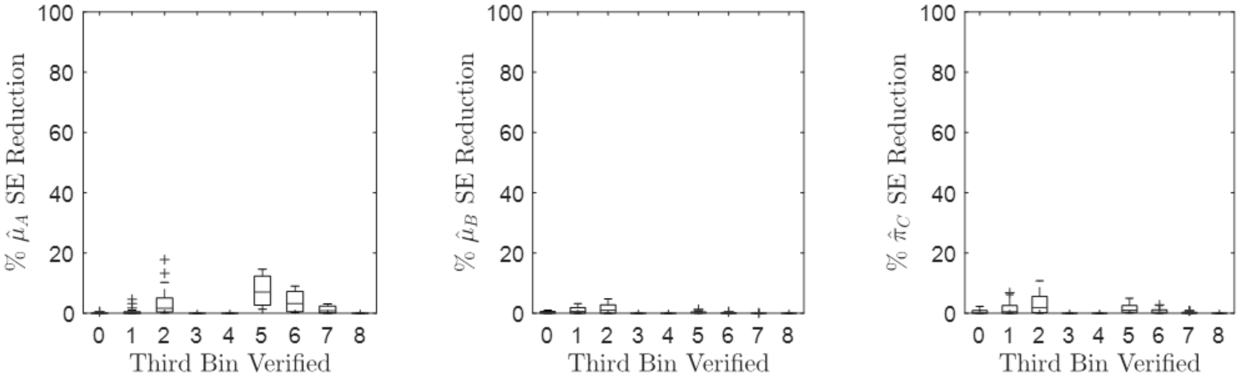

Figure 4.6 – Bin by Bin SE Reduction – Comparison to Bins 3 & 4 Only

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$,

$\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Figure 4.6 shows how much reduction in standard error verifying each bin provides compared to verifying only Bins 3 & 4. As was the case in Figure 4.2 the further reduction in standard error possible by verifying the other bins is very small, with Bin 5 giving the biggest possible reduction. However on average Bin 5 contains a greater percentage of parts than Bins 3 & 4 combined and provides a reduction of only about 5%.

Notice that in Figure 4.4, Figure 4.5, and Figure 4.6, there is considerable variation in the improvement provided by verifying the bins. In the cases where the improvement is smaller, there are very few parts in the central bins; this is typically the case when both the misclassification probabilities and gamma values are low. Thus small improvement typically goes along with small effort. Additionally in these cases the BMS is easier to assess, even with few verifications.

One interesting observation is that the bins with the fewest parts give the greatest reduction in standard error. Actually in cases where the verification order described above is not optimal this rule of verifying the bins with the fewest parts typically works well. The verification order described above works well for small r, but may be less reliable as r increases. For $r = 5$ it is optimal in all cases, for $r = 7$ it is optimal in 95% of cases and continues to be effective in the remaining 5%. However, for very large values of r perhaps a different order of verification could be used where bins with the fewest parts are verified first.

Repeated Measurement Phase

For this section, I take the order of verification that is optimal for smaller r as granted. This allows for a tractable discussion of how to conduct the repeated measurement phase. There are essentially three design parameters which can change: r and n_P , and n_F . Ideally there would be one design parameter left open for the user to set based on the standard error requirements. Having three open design parameters leaves the experimenter unsure of how to best design a plan. This section will try to give choices for r and n_P that are good for the BMS most likely encountered in industry, see Table 2.4, while leaving n_F for the experimenter to determine based on estimate precision requirements.

The number of parts re-sampled from passing parts, n_P , will be set at zero or some small number. In the industrial setting π_C is usually very close to one, representing a high quality process. This implies re-sampling from failed parts, n_F , will provide a better balance of conforming and non-conforming parts and thus greatly reduce the standard error of $\hat{\mu}_A$. Therefore it is best to keep n_P small. However in some cases the ML estimates can mistakenly fit gamma values that correspond to a “U-shape” beta distribution for the misclassifications rates. This is more likely when the beta-binomial model is not a good fit for the data. When this occurs verifying some parts from the non-central bins, that is the robustness option, is a good remedy. To implement this remedy it is necessary to have a small but non-zero n_P . However this issue is rare when dealing with conditional sampling and baseline information. Therefore I recommend not re-sampling from passed parts except when this “U-shape” problem arises. In such cases, sample ten parts from those that already passed inspection, i.e. $n_P = 10$, verifying five parts in non-central bins and calculating new estimates.

Increasing r will always improve estimate precision, thus to provide a relevant analysis of the best choice for r , I keep the total number of measurements in the repeated measurement phase constant. Figure 4.7 shows the standard error of $\hat{\mu}_A$ for different values of r , where $r\left(n_P + n_F\right) = 2500$, specifically $n_P = 0$, $n_F = \lfloor 2500/r \rfloor$. The number of verifications is increased from zero to ten percent of the total number of repeated measurements. Verifications are done in the order described in the verification phase justification section above.

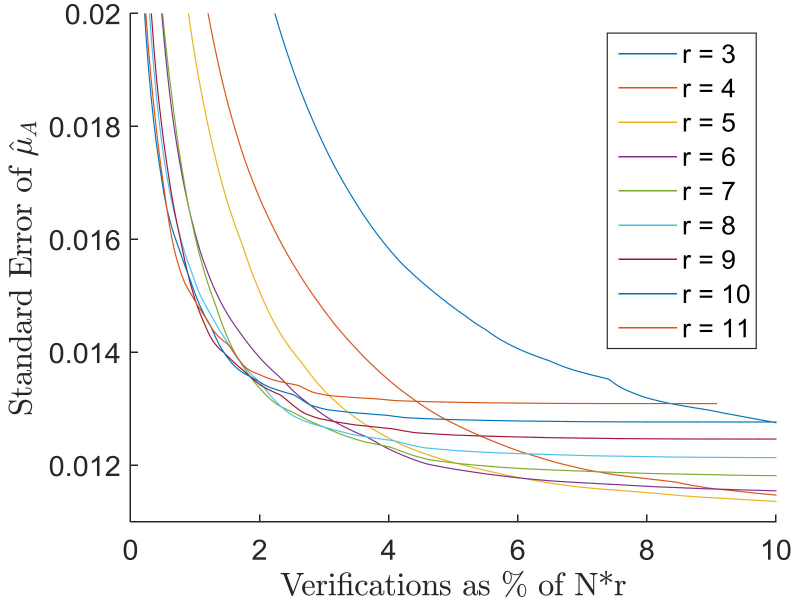

Figure 4.7 – Optimal r Experiment – Conditional Sampling

Factorial Experiment run at all combinations $n * r \simeq 2500$
$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 4.7 shows that at a low verification percentage a greater value of r results in the lowest standard error for the same number of total measurements while at a higher verification percentage, a lower value for r gives the lowest standard error. So the choice of r depends on the choice of verification percentage. Thus the choice of r and the choice of verification percentage must be made together. The best verification percentage depends on both the plan and the model parameters. Figure 4.7 shows that the reduction of standard error starts to slow down significantly around two percent verification, and that $r = 7$ is optimal or close to optimal around this verification percentage. While it is impossible to choose a value of r that is optimal in all cases, the choice of $r = 7$ gives standard errors close enough to optimal to recommend in general. Note: The whole range of the standard errors is not shown because the asymptotic standard errors for very low verification percentage may not be an accurate approximation for reasonable sample sizes, and furthermore there is substantial bias in estimates with very low verification percentage for reasonable sample sizes.

4.7 Conditional Sampling Plan Performance Summary

Having justified the recommended conditional sampling plan, I will now assess the performance of said plan with comparison to the full and no verification plans. First I show the reduction in standard errors that result from targeted verification. Figure 4.9 and Figure 4.10 are based on the results of ten thousand simulated data sets for each of the 32 parameter value combinations laid out in Table 2.4. The data sets were generated using the beta-binomial model. The baseline size was ten thousand, $n_B = 10000$, five hundred parts were re-sampled from failed parts, $n_F = 500$, zero parts were re-sampled from passed parts, $n_P = 0$, and the number of repeated measurements was seven, $r = 7$. Each data set was fit using the likelihood model outlined at the end of Section 4.3. The data sets were generated with full verification information. The full verification plan used the full data set, while the recommended conditional plan, see Section 4.5, and no verification plan used the appropriate "censored" data sets. The recommended conditional plan was compared to the full and no verification plans using the measures described in Figure 4.8.

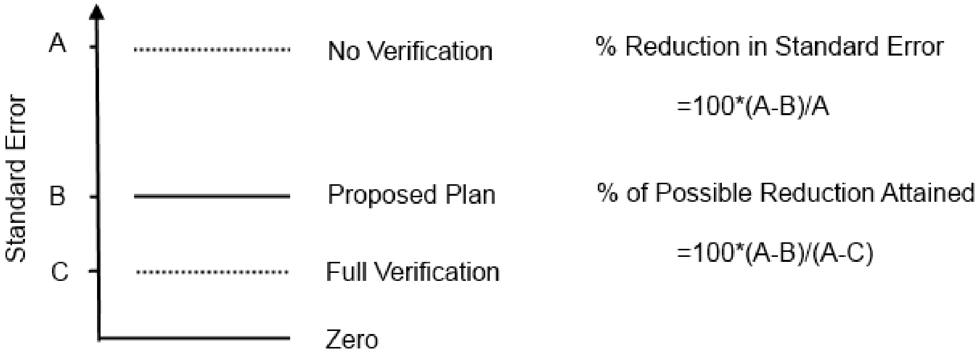

Figure 4.8 - Performance Measures Summary

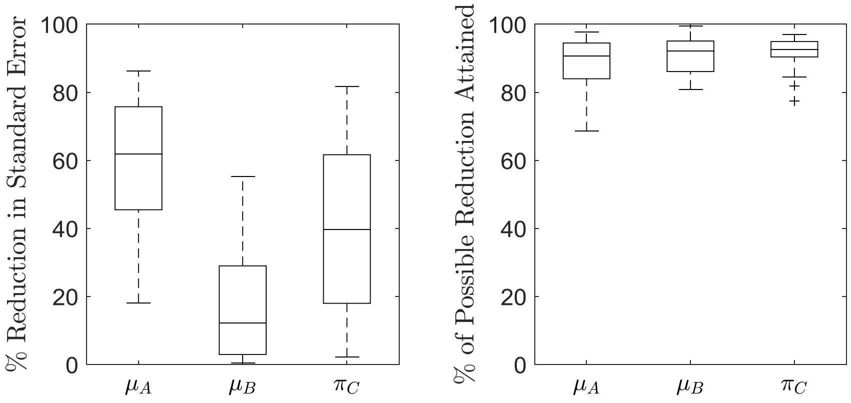

Figure 4.9 – Conditional Plan Performance – Impact of Targeted Verification on SE – Beta Binomial

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$,

$\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Figure 4.9 shows a great improvement by verifying Bins 3 & 4 which, on average, represents only 8.4% of the parts repeatedly measured. The reduction in the standard error of $\hat{\mu}_A$ is on average about 60% which represents 90% of the reduction possible through verification. This is a dramatic improvement for very little work. Targeted verification also reduces the standard errors of $\hat{\mu}_B$ and $\hat{\pi}_C$ by 18% and 41% respectively which in both cases represents on average 90% of the reduction possible.

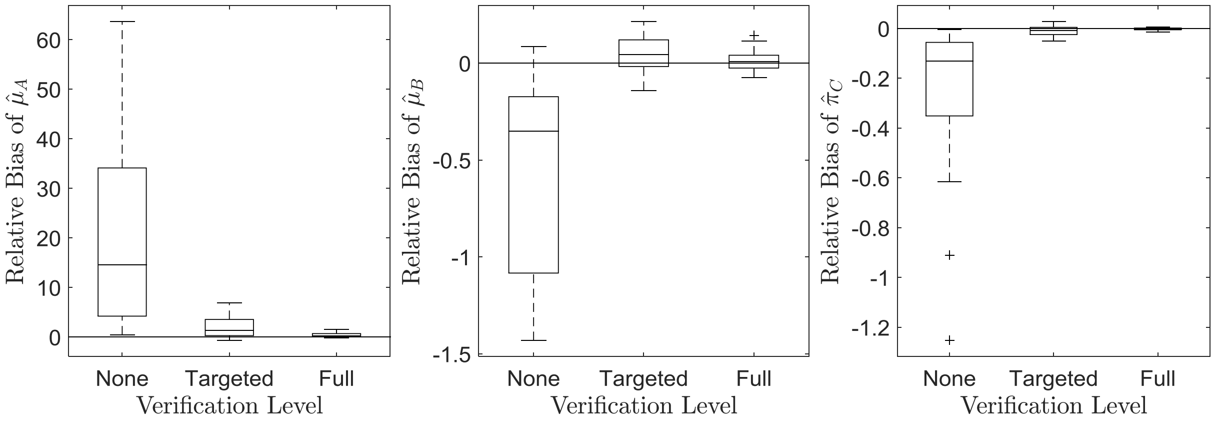

Figure 4.10 - Conditional Plan Performance – Relative Bias – Beta Binomial

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$,

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 4.10 shows a dramatic reduction in the bias for all three quantities of interest when compared to the no verification plan. The relative bias of quantities of interest in the recommended conditional plan is much closer to the relative bias in the full verification plan than the no verification plan. The reduction in biases compared to the no verification plan for $\hat{\mu}_A$, $\hat{\mu}_B$ and $\hat{\pi}_C$ are 84%, 70% and 83% on average respectively which represents around 90% of the possible reduction possible through verification. Recall this is all attained through verifying only 8.4%, on average, of the parts in the repeated measurement phase.

This section shows that the conclusions of Chapter 2, hold in the conditional sampling case as well. Specifically, I showed that targeted verification gives a great reduction in standard error and bias, and in fact has performance close to that of the full verification plan with relatively little work.

Robustness Consideration

This section considers the performance of the recommended conditional sampling plan when the model is misspecified. The comparison results are based on ten thousand simulated data sets generated with the GRE model but fit with the beta-binomial model. All elements of the plan including sample size and baseline size are identical to the simulations done with beta-binomial data used in the simulations summarized in Figure 4.9 and Figure 4.10. This was done to show that the recommended conditional sampling plan is robust to model misspecification. The recommended conditional plan was compared to the full and no verification plans using the measures described in Figure 4.8.

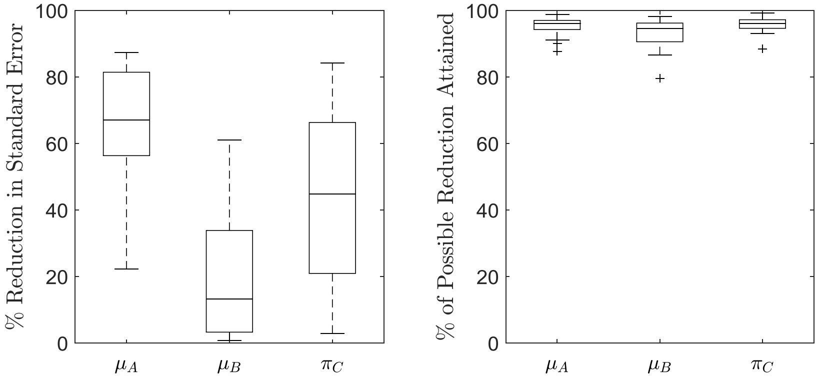

Figure 4.11 – Conditional Plan Performance – Standard Error Reduction - GRE

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$,

$\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Comparing Figure 4.11 to Figure 4.9 shows the improvements in standard error remain roughly the same for both GRE and beta-binomial data. One odd observation is that the percentage of possible reduction obtained is higher with GRE data. This is likely due to the shapes of the distribution for misclassification rates in the GRE model, which have density zero at the ends of the $[0,1]$ interval. This shape reduces the probability of the rare events that can result in pathological estimates.

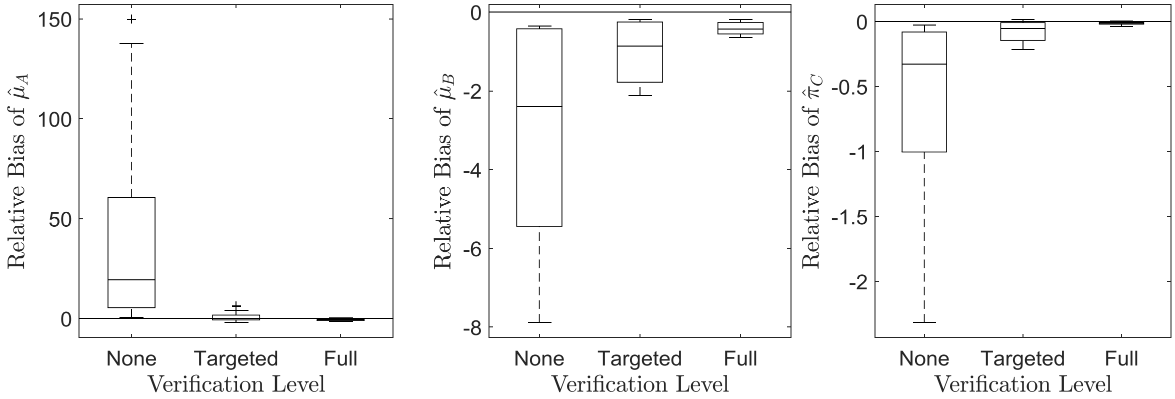

Figure 4.12 - Conditional Plan Performance – Relative Bias - GRE

Factorial Experiment run at all combinations $n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$,

$\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Comparing Figure 4.12 with Figure 4.10 shows that the no-verification plan has a significant increase in bias of $\hat{\mu}_A$ when the model is misspecified. But the targeted verification and full verification plans are not subject to any increase in bias of $\hat{\mu}_A$ due to the model misspecification. On the other hand, the bias of $\hat{\mu}_B$ and $\hat{\pi}_C$ increases for the no verification, targeted verification, and full verification plans when compared to the bias found in Figure 4.10. The bias remains negligible for the targeted and full verification plans, and this bias is transient and decreases as the size of the study is increased.

4.8 Conditional and Population Sampling Plan Performance Comparison

When assessing a BMS already in use, the baseline information is usually available with very large sample sizes, as businesses record the number of parts that pass and fail inspection as part of ongoing operations. Additionally finding parts that failed inspection is easy because parts failing inspection are often collected for scrap or rework. Sampling rejected parts is usually less intrusive to the manufacturing process because it does not interfere with production goals. If using conditional sampling and baseline information is not burdensome, then a direct comparison to population sampling approach is justified.

This comparison was done using ten thousand simulated data sets generated for each of the 32 sets of parameter values found in Table 2.4. This was done both for data for the population plan and separately for data for the conditional plan. The two sets of simulations keep the total effort the same between the population and conditional sampling plans. Specifically this means that the total number of measurements performed in each repeated measurement phase is equal and the number of verifications in each verification phase is equal. To accomplish the first, I used different sample sizes, $n = 500$ and $n_F = 357$ in order that the total number of measurements would be equal, $n * 5 = n_F * 7 \simeq 2500$. Second, in order to keep the number of verifications the same, I used a fixed number of verifications for each set of parameter values. That number changed for each set of parameter values and was approximately equal to the expected number of parts that would fall in the two central bins in the population sampling case, assuming the beta-binomial model. This slightly favours the population sampling case because the number of verifications is better adapted to that plan. The verifications were done in the order described in Section 2.7. Note that in either case no verifications were performed in non-central bins. Eliminating the verifications in the non-central bins was left out of the two-phase plan because doing so allowed a more direct comparison. Furthermore the additional verifications in non-central bins are not thought to have a dramatic influence on the estimates in either case when the assumed model, beta-binomial, is correctly specified. Figure 4.13 and Figure 4.14 are based on the results of these simulations. These figures use relative measures meaning, that they are expressed as a percentage of the parameter value being estimated.

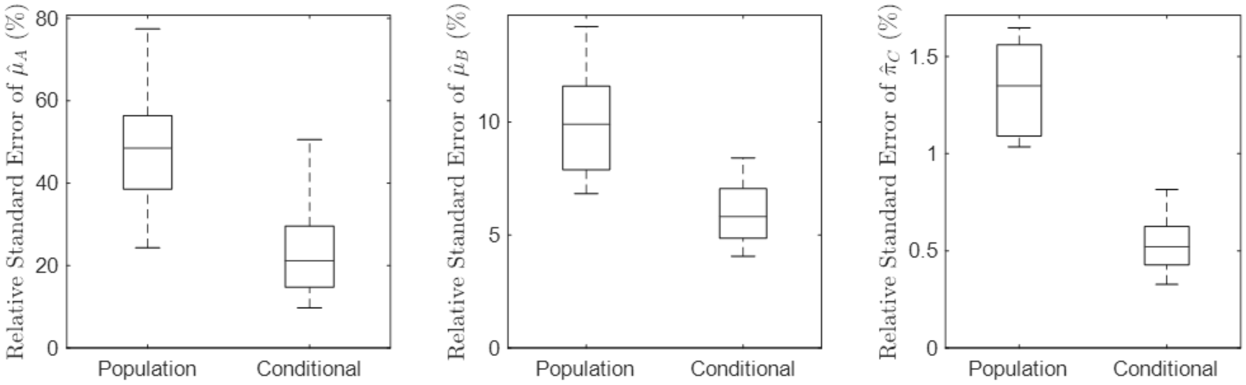

Figure 4.13 - Population vs. Conditional Sampling Performance Comparison – Standard Error
$n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$ vs. $n = 500$, $r = 5$

$\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Figure 4.13 shows dramatic improvements for all three quantities of interest when using the conditional plan rather than the population plan. Using the conditional sampling plan instead of the population plan yields reductions in the standard errors of $\hat{\mu}_A$, $\hat{\mu}_B$ and $\hat{\pi}_C$ of 53%, 39% and 60% on average respectively. Note that although the boxplots for population and conditional sampling overlap in Figure 4.13 the conditional sampling plan always provides an improvement over the population sampling plan in each of the 32 parameter value combinations.

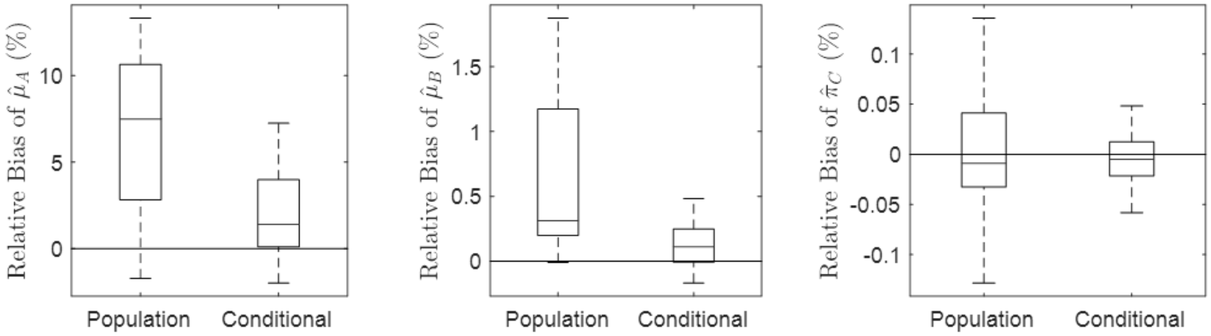

Figure 4.14 - Population vs. Conditional Sampling Performance Comparison – Relative Bias
$n_F = 500$, $n_P = 0$, $n_B = 10000$, $r = 7$ vs. $n = 500$, $r = 5$

$\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Figure 4.14 show that the bias is also reduced for each of the quantities of interest. The reduction in the average bias for $\hat{\mu}_A$, $\hat{\mu}_B$ and $\hat{\pi}_C$ is 65%, 75% and 59%, respectively. In conclusion, the conditional sampling plan greatly outperforms the population plan. So wherever possible I would recommend using the conditional sampling plan over the population sampling plan. This is particularly easy to do when a BMS is currently used in a manufacturing environment.

Considerations when π_C is very close to one

Simulations were done comparable to those used in Figure 4.13 and Figure 4.14, but with π_C set to 0.990 and 0.995 instead of 0.90 and 0.95. This represented a reduction in the order of magnitude of the number of non-conforming parts in the study. Most changes caused were expected, such as the standard error of $\hat{\mu}_A$ increasing by a factor of roughly $\sqrt{10}$, and the standard errors of $\hat{\mu}_B$ and $\hat{\pi}_C$ being reduced slightly. The only finding which was contrary to my expectations was that improvement of the conditional sampling plan over the population sampling plan did not represent a greater percentage improvement when π_C was closer to one, rather the percentage improvement stayed approximately the same. A potential solution for cases where π_C is very close to one is proposed in Section 4.11.

4.9 Impact of Baseline Size

This chapter did not treat the baseline size as a design parameter. This is because in a manufacturing environment with the BMS already in use, baseline information is typically already available with a large sample size. Thus the experimenter would not choose the baseline size as he chooses a design parameter, but rather would use all the baseline information available: provided it is considered representative of the current performance of the process and BMS. However when baseline information is not available some information about the impact of the size of the baseline sample is needed.

First, for a specified sample size for the repeated measurement phase, n_F , there is a recommended minimum sample size. Simulations revealed that when the baseline size, n_B, was not at least seven times the size of the repeated measurement phase n_F there were severe problems with the asymptotic standard error estimates. I recommend using this as an absolute lower bound for the baseline sample size; baseline information of this or greater size would typically be available for a BMS already in use. A recommended guideline for the size of the baseline is the expected number of parts produced by the process before n_F failing parts would be found: i.e. $n_F/(1-\pi_P)$. The baseline size of 10000 used in the previous simulations and calculations meets or exceeds this guideline for each set of parameter values described in Table 2.4.

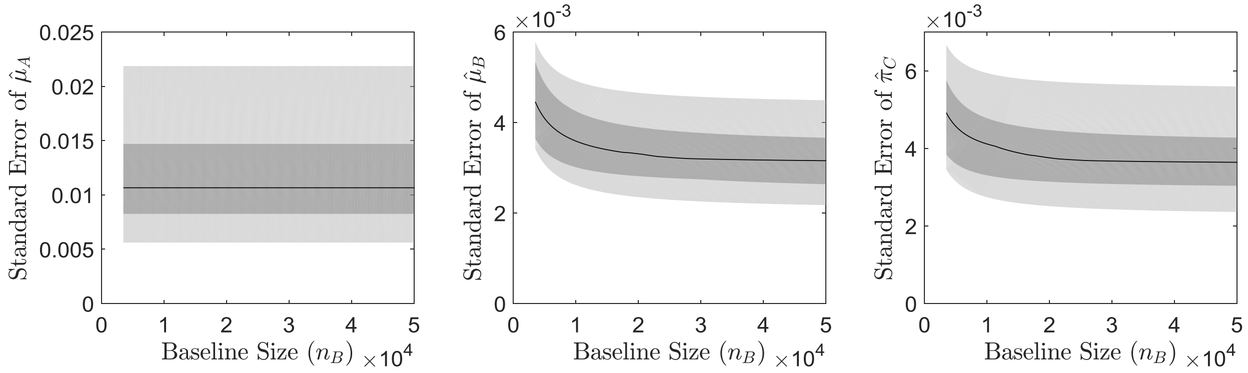

Figure 4.15 – Baseline Size Plot

$n_F = 500$, $n_P = 0$, $r = 7$, $\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Figure 4.15 displays the standard error of the 32 sets of parameter values described in Table 2.4, using an approach similar to a box plot. The dark line represents the median of the 32 parameter sets for each of the standard errors described in the vertical axis labels. The light shaded error represents the first and fourth quartiles, and the dark shaded areas represent the second and third quartiles. The plot could be described as a box plot with a continuous covariate.

Figure 4.15 show the impact of the baseline size on the standard error of the three quantities of interest. The plot leaves out the values for baselines sizes below 3500, seven times the repeated measurement phase sample size n_F. This was done because as discussed earlier, the asymptotic standard errors would not be reliable. The effect on μ_A beyond 3500 is negligible, though non-zero. The effect on μ_B and π_C is relatively small. The plots indicate that standard error of μ_B and π_C would increase a lot as the baseline size is reduced below 3500. Finally these plots show that for the parameter values used in Table 2.4 the baseline size of 10000 is sufficient to essentially attain all the reduction in standard error possible through baseline information. This gives some credibility to the recommendation of $n_B \geq n_F / (1 - \pi_P)$.

4.10 Discussion

The proposed conditional sampling plan has improved performance over the population sampling plan when compared over the sets of parameter values described in Table 2.4. This is because conditional sampling allows for a more balanced number of conforming and non-conforming parts, and because baseline information provides further information at no cost. The recommended plan is effective for a wide range of different parameter values. This plan would be very easy to implement for a BMS currently in use, and would in some cases be less intrusive to ongoing manufacturing operations than the population sampling plan, since failed parts are usually set aside as part of normal production.

4.11 Future Work

Higher Order Conditional Sampling

This section will discuss one possible sampling scheme that is a natural extension of conditional sampling that could be studied in the future. This sampling scheme will be useful for processes where non-conforming parts are extremely rare and even conditional sampling may not ensure a substantial number of non-conforming parts. The ability of conditional sampling to increase number of non-conforming parts is effected by μ_B. Consider the ratio of probabilities of obtaining non-conforming and conforming parts under conditional sampling from failed parts.

$$\frac{P\left(X_i=0|Y_i=0\right)}{P\left(X_i=1|Y_i=0\right)}=\frac{P\left(X_i=0,Y_i=0\right)}{P\left(X_i=1,Y_i=0\right)}=\frac{\left(1-\mu_A\right)\left(1-\pi_C\right)}{\mu_B\pi_C}\simeq\frac{1-\pi_C}{\mu_B}$$

When μ_B is less than $1-\pi_C$ or even of the same order of magnitude then conditional sampling is effective enough to obtain a sufficient number of non-conforming parts. However, when μ_B is much greater than $1-\pi_C$, conditional sampling may not obtain enough non-conforming parts. In this case it may be worth considering doing what I will call conditional sampling of order k. This scheme re-samples from parts that failed k times in k inspections, instead of from parts that failed inspection once. As in the case with conditional sampling, it may be possible to combine data with re-sampled parts that passed on some or all of those k inspections if needed. Unfortunately unlike in the conditional sampling order 1 case, there is unlikely to be parts available from ongoing operations that meet this criterion. Thus creating such a sampling pool must be part of the experiment. The most efficient way to obtain parts that failed inspection k out of k times would be in a set of sequential inspections where a part is only inspected again if it failed inspection, and parts that pass inspection at any point are set aside. To explore the effectiveness of these different order conditional sampling schemes, I prepared a plot similar to Figure 4.1 where the probability of sampling a non-conforming part is plotted against π_C.

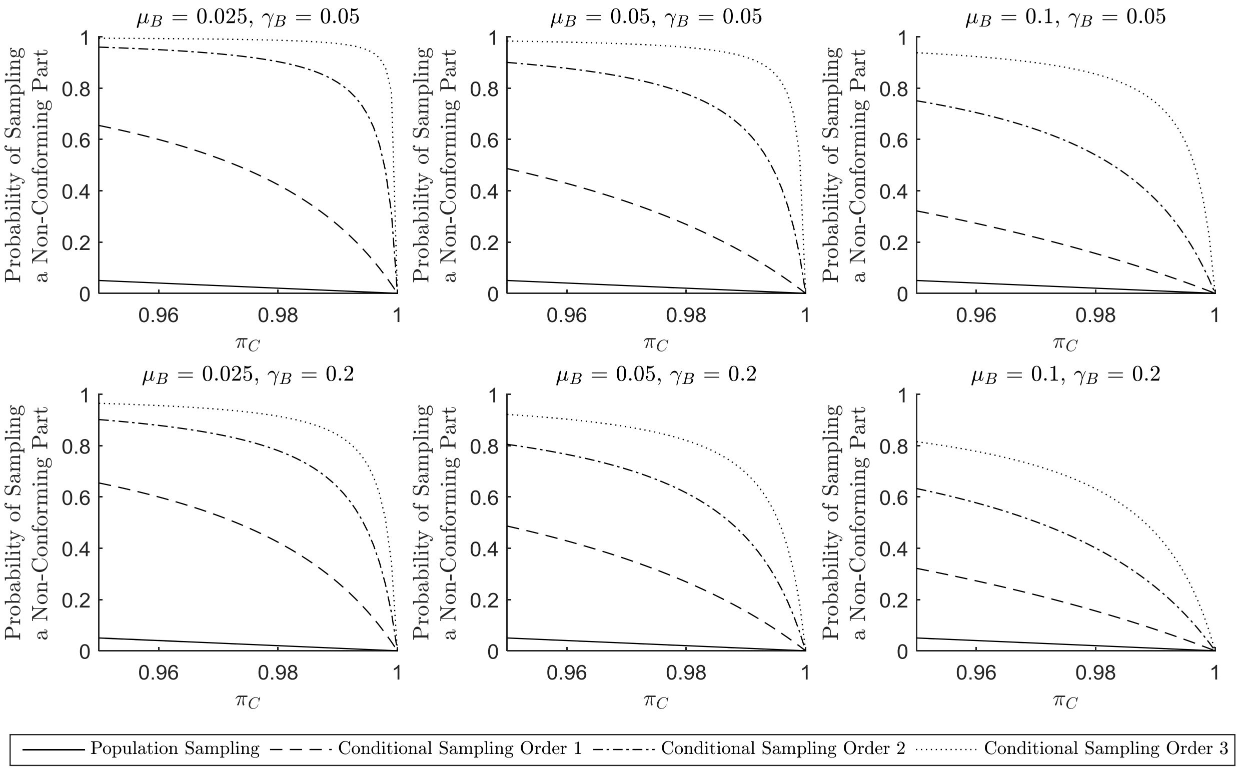

Figure 4.16 – Conditional Sampling Order K = 0,1,2,3 plot

Figure 4.16 shows that using a conditional sampling scheme of higher order can further increase the probability of sampling a non-conforming part. Figure 4.16 also demonstrates that a measurement system with higher μ_B or γ_B will make conditional sampling less effective at finding non-conforming parts. Note that for conditional sampling of order 1, γ_B has no effect.

Conditional sampling of different orders is left as future work. I feel the details of doing so are too tedious to include and that using the principles of this chapter a trained statistician could implement higher order conditional sampling method without further guidance.

Critique of Case Control Studies

One common question surrounding the practice of conditional sampling and is even more likely to be raised about using a conditional sampling scheme of higher order, is whether a case-control study would be more effective. One answer is that when pools of conforming and non-conforming parts do exist prior to the study, conditional sampling is much more efficient and cost effective than creating said pools and doing a case-control study. This is because to create these representative pools one must measure parts from the general population one by one with the gold standard. Another related answer would be that when those pools of conforming and non-conforming parts exist they are often not representative of the corresponding subpopulations and are often created using some unknown procedure. In particular, if a imperfect test is used to determine which parts are conforming or non-conforming, or even as part of a screening procedure to find parts that may be non-conforming then the resulting pools of parts may, and probably will not, be representative. This will create bias in the estimates from the case-control study, most notably underestimating μ_A. This bias could be demonstrated using simulation. However literature on this topic already exists, see Whiting, et al. (2004), and thus a proper literature review is necessary to determine the need of such work.

4.12 Asymptotic Variance Justification

It is not possible to find an analytic expression for the ML estimates or the associated standard errors, therefore the asymptotic variance results due to Fisher (1925) are used to estimate the standard errors. This section assesses the accuracy of these estimates for different sets of parameter values using a factorial experiment structure. There are 32 different treatments which are made up by varying the 5 model parameters, see Table 2.4. For each treatment, ten thousand datasets were simulated from the beta-binomial model, parts were selected and verified according the recommended conditional sampling plan and beta-binomial ML estimates were calculated. The bias and standard error of these estimates were calculated and recorded for comparison to the asymptotic standard error approximation. Figure 4.17 shows the ratio of the simulated standard errors and the expected asymptotic standard errors for each combination of parameter values. The design parameters were kept the same for all treatments with $n_P = 0$, $n_F = 500$, $n_B = 10000$ and $r = 7$.

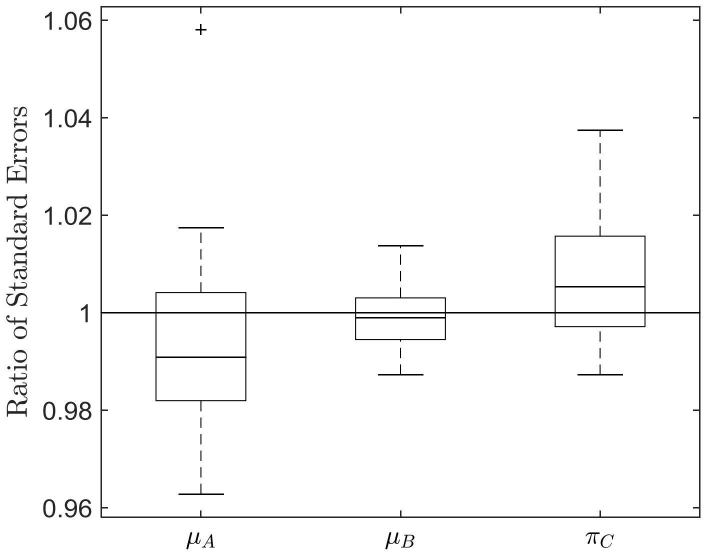

Figure 4.17 – Asymptotic Vindication Experiment

Ratio of Simulated and Asymptotic Standard Errors $n_F = 500$, $n_P = 0$, $n_B = 10000$ and $r = 7$

$\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2;$ (See Table 2.4)

Figure 4.17 shows that the ratios of simulated standard errors and asymptotic standard errors are very close to one, and thus sufficiently accurate.

Chapter 5 Multinomial-Based Estimates

5.1 Introduction

Chapters 2 and 4 use the beta-binomial model of Danila et al. (2012). The assumptions in this model are reasonable but unverifiable in the intended application. There are many competing models that could be used instead, such as the Gaussian Random Effects (GRE) model introduced in Qu (1996). In most cases fitting different models results in different estimates for the parameters of interest μ_A, μ_B and π_C. Additionally, estimates will almost surely not be consistent when the model is misspecified; this makes choosing a model a critical and perilous decision. While Section 2.10 shows that the Two-phase plan, from Section 2.4, was somewhat robust when using the beta-binomial model, it is worth considering a fully general model. Chapter 3 introduced closed form estimation, that had underlying it a very general model, see Section 3.6. The goal of this chapter is introduce a more general estimation procedure that can be used with data from the conditional sampling plan of Chapter 4.

5.2 Model Definition

Recall that the likelihood equation derived in Chapter 4 is

$$\mathcal{L}(\theta) = k * P(Y_i = 1)^{y_B - n_P} \, P(Y_i = 0)^{n_B - y_B - n_F} \left(\prod_{s=0}^{r+1} P(T_i = s)^{n_s} \right) \\ * \left(\prod_{s=0}^{r+1} P(X_i = 1 \mid T_i = s)^{u_s} \, P(X_i = 0 \mid T_i = s)^{v_s - u_s} \right).$$

Instead of using the beta-binomial model for each of the terms in this expression, I will instead define model parameters directly in terms of some of the quantities in this expression. The likelihood will be separated into two components. The likelihood for the baseline and repeated measurement phases is,

$$\mathcal{L}_{B+RM}(\Psi) = k_{B+RM} * \left(\sum_{s=0}^{r+1} \tfrac{s}{r+1} \psi_s \right)^{y_B - n_P} \left(1 - \sum_{s=0}^{r+1} \tfrac{s}{r+1} \psi_s \right)^{n_B - y_B - n_F} \left(\prod_{s=0}^{r+1} {\psi_s}^{n_s} \right), \tag{5.1}$$

where

$$\Psi = (\psi_0, \psi_1, \ldots, \psi_{r+1}),$$

$$\psi_s = P(T_i = s) \text{ , } 0 \le \psi_s \le 1 \text{ , and } \sum_{s=0}^{r+1} \psi_s = 1 \text{ for all } s \in \{0, 1, \ldots, r+1\}.$$

Note that using symmetry of the repeated measurements, I have that

$$P\left(Y_i=1\right)=\sum_{s=0}^{r+1}P\left(Y_i=1\,|\,T_i=s\right)P\left(T_i=s\right)=\sum_{s=0}^{r+1}\tfrac{s}{r+1}\psi_s\,.$$

The likelihood for the verification phase is

$$\mathcal{L}_V\left(\Phi\right)=k_V*\prod_{s=0}^{r+1}\phi_s^{u_s}\left(1-\phi_s\right)^{v_s-u_s}\,, \tag{5.2}$$

where

$$\Phi=\left(\phi_0,\phi_1,\ldots,\phi_{r+1}\right),$$

$$\phi_s=P\left(X_i=1\,|\,T_i=s\right)\ ,\ 0\le\phi_s\le1 \text{ for all } s\in\left\{0,1,\ldots,r+1\right\}.$$

The motivation for this model is the same as the motivation for estimates of Chapter 3, which is a more general alternative to the beta-binomial estimates used in Chapter 2. Unfortunately the plan used in Chapter 4 allows no closed form estimates akin to those in Chapter 3. Instead, the baseline and repeated measurement phase data along with the likelihood in Equation (5.1) are used to calculate $\hat{\Psi}$. Then the verification phase data and the likelihood in Equation (5.2) are used to calculate $\hat{\Phi}$. Defining the basic quantities of Chapter 3 in the terms of Ψ and Φ, it becomes obvious how to calculate the associated estimates for the basic quantities, please see Table 5.1. The estimates for μ_A and μ_B are then calculated as ratios of the basic quantity estimates as in Section 3.8.

Table 5.1 – Basic Quantities $\Psi\Phi$ Definition

	$Y_i=0$	$Y_i=1$	
$X_i=0$	$\pi_{00}=\sum_{s=0}^{r}\frac{r-s}{r}\psi_s\left(1-\phi_s\right)$	$\pi_{10}=\sum_{s=0}^{r}\frac{s}{r}\psi_s\left(1-\phi_s\right)$	$1-\pi_C=\sum_{s=0}^{r}\psi_s\left(1-\phi_s\right)$
$X_i=1$	$\pi_{01}=\sum_{s=0}^{r}\frac{r-s}{r}\psi_s\phi_s$	$\pi_{11}=\sum_{s=0}^{r}\frac{s}{r}\psi_s\phi_s$	$\pi_C=\sum_{s=0}^{r}\psi_s\phi_s$
	$1-\pi_P=\sum_{s=0}^{r}\frac{r-s}{r}\psi_s$	$\pi_P=\sum_{s=0}^{r}\frac{s}{r}\psi_s$	1

5.3 Verification Model Variant

I will consider two variations for the verification phase model, one as described above and another with the constraint $\phi_0 \leq \phi_1 \leq \ldots \leq \phi_{r+1}$ added. This constraint has the surface interpretation that a part that passed t out of $r+1$ inspections has a greater probability of being conforming than a part that passed fewer than t out of $r+1$ inspections. This interpretation has some intuitive meaning but the constraint may be better understood when expressed as a property of the measurement system. To get an equation that can be interpreted this way I use Bayes' rule and some basic algebraic manipulations. Note that the manipulation assumes some probabilities are non-zero, and though the quantities need not always be non-zero this manipulation helps with interpretation. The altered constraints are

$$\frac{P\left(T_i = t \middle| X_i = 0\right)}{P\left(T_i = s \middle| X_i = 0\right)} \leq \frac{P\left(T_i = t\right)}{P\left(T_i = s\right)} \leq \frac{P\left(T_i = t \middle| X_i = 1\right)}{P\left(T_i = s \middle| X_i = 1\right)} \text{ for all } 0 \leq s < t \leq r+1. \tag{5.3}$$

An interpretation of Equation (5.3) specific to the context of the study is that a conforming part is at least as likely to pass more inspections than a non-conforming part. That may be a bit abstract so let us consider the constraint in the case where there is only a single measurement. The constraint in that case is,

$$\frac{P\left(Y_i = 1 \middle| X_i = 0\right)}{P\left(Y_i = 0 \middle| X_i = 0\right)} \leq \frac{P\left(Y_i = 1\right)}{P\left(Y_i = 0\right)} \leq \frac{P\left(Y_i = 1 \middle| X_i = 1\right)}{P\left(Y_i = 0 \middle| X_i = 1\right)}. \tag{5.4}$$

Equation (5.4) can be interpreted that the BMS is at least as likely to pass a conforming part as a non-conforming part in inspection. When interpreted in this way the constraint is quite easy to understand. Furthermore it is clear to see that if this constraint is not true the measurement system is nonsensical because it would perform better if what was considered a pass and a failure was reversed. Note that when Equation (5.4) is written in terms of the quantities of interest the inequalities can be reduced to the constraint, $1-\mu_A-\mu_B \geq 0$.

Now this interpretation does not perfectly map back to the constraints found in Equation (5.3), because there are many constraints in Equation (5.3) and this allows for the possibility that some could hold and others not. However considering each constraint in isolation, it still makes sense that each one of them should hold, and it would take an exceptionally poor measurement system to break any of the constraints.

5.4 Treatment of Empty Bins

In cases where after the repeated measurement phase, one or more bins have zero parts, the maximum likelihood estimate for Φ will not be defined for the unrestricted model and may not be defined for the restricted model. When this problem arises and further sampling is not possible or practical, I recommend setting the undefined parameter, ϕ_s, to the parameter for the neighboring bin representing fewer passes, ϕ_{s-1}. When it is the bin representing zero passes that had zero parts set $\phi_0 = 0$.

This rule for special cases was not needed in Chapter 3, because under the plan in Chapter 2, a ϕ estimate could only be undefined when the corresponding ψ parameter had estimate zero. Thus no rule was necessary and the product of those ψ and ϕ terms was set to zero.

5.5 Standard Error Estimate Discussion

The procedure I use to calculate the standard error of the estimates is somewhat similar to that used in Chapter 3, see Section 3.5. However the estimates for the variance and covariance of the basic quantities estimates are different. In order to calculate these variance and covariance estimates I first need to estimate the variances and covariances of $\hat{\Psi}$ and $\hat{\Phi}$.

To estimate the variance and covariance of $\hat{\Psi}$ I use the maximum likelihood asymptotic theory of Fisher (1922). For the unrestricted estimate of Φ, the individual parameters can be estimated separately as binomial parameters, so the covariances are zero and the variances are estimated using its UMVUE.

For the estimate of the variance of $\hat{\Phi}$ a parametric bootstrapping method was used. The restricted estimates are treated as the true value Φ and a large number of data sets were generated using this assumption. Estimates were then calculated for each of the generated data sets and the variance and covariance of the estimates were calculated. The number of bootstrap samples was one thousand in simulations but this number could be increased in practice where only one set of bootstrap samples need to be generated.

Then using the estimates of the variances and covariances of $\hat{\Psi}$ and $\hat{\Phi}$ I calculate estimates of the variances and covariances of the basic quantities. As in Section 3.5 I give some details for one of the basic quantities $\tilde{\pi}_{10}$. I take Equation (3.2) and replace the estimators from Chapter 3 with the appropriate estimators from Chapter 5, yielding

$$Var\left(\tilde{\pi}_{10}\right)=\sum_{s=0}^{r+1}\left(\frac{s}{r}\right)^2 Var\left(\tilde{\psi}_s\left(1-\tilde{\phi}_s\right)\right)+2\sum_{s<t}\frac{s}{r}\frac{t}{r}Cov\left(\tilde{\psi}_s\left(1-\tilde{\phi}_s\right),\tilde{\psi}_t\left(1-\tilde{\phi}_t\right)\right).$$

Using the definitions of variance and covariance this becomes,

$$\begin{aligned}Var\left(\tilde{\pi}_{10}\right)=&\sum_{s=0}^{r+1}\left(\frac{s}{r}\right)^2\left(E\left[\tilde{\psi}_s^{\,2}\left(1-\tilde{\phi}_s\right)^2\right]-E^2\left[\tilde{\psi}_s\left(1-\tilde{\phi}_s\right)\right]\right)\\&+2\sum_{s<t}\frac{s}{r}\frac{t}{r}\left(E\left[\tilde{\psi}_s\tilde{\psi}_t\left(1-\tilde{\phi}_s\right)\left(1-\tilde{\phi}_t\right)\right]-E\left[\tilde{\psi}_s\left(1-\tilde{\phi}_s\right)\right]\mathrm{E}\left[\tilde{\psi}_t\left(1-\tilde{\phi}_t\right)\right]\right).\end{aligned}$$

I then use the independence of $\tilde{\Psi}$ and $\tilde{\Phi}$, conditional upon the number of verifications,

$$\begin{aligned}Var\left(\tilde{\pi}_{10}\right)=&\sum_{s=0}^{r+1}\left(\frac{s}{r}\right)^2\left(E\left[\tilde{\psi}_s^{\,2}\right]E\left[\left(1-\tilde{\phi}_s\right)^2\right]-E^2\left[\tilde{\psi}_s\right]E^2\left[\left(1-\tilde{\phi}_s\right)\right]\right)\\&+2\sum_{s<t}\frac{s}{r}\frac{t}{r}\left(E\left[\tilde{\psi}_s\tilde{\psi}_t\right]E\left[\left(1-\tilde{\phi}_s\right)\left(1-\tilde{\phi}_t\right)\right]-E\left[\tilde{\psi}_s\right]\mathrm{E}\left[\tilde{\psi}_t\right]\mathrm{E}\left[\left(1-\tilde{\phi}_s\right)\right]\mathrm{E}\left[\left(1-\tilde{\phi}_t\right)\right]\right).\end{aligned}$$

Estimates for each of the quantities used in the above expression can be calculated using the estimates for the first moments, variances and covariances of $\hat{\Psi}$ and $\hat{\Phi}$. Substituting those estimates will yield my estimate for the variance of $\hat{\pi}_{10}$. A similar procedure can be done for the variance of the other basic quantities as well as covariances between basic quantities. Then to estimate the variance of ratios of the basic quantities like μ_A and μ_B, a Taylor series approximation is used as described in Section 3.8.

5.6 Example

As was done in Section 4.4, I will fit the model to the dataset used in Danila et al. (2013); the data for this example is found in Table 4.2. The restricted and unrestricted estimates in Table 5.2 use the treatment for empty bins found in Section 5.4.

Table 5.2 – Example Estimates

	Parameter	μ_A	μ_B	π_C
Full Verifications Data – Beta Binomial MLE Estimate				
	Estimate	0.134	0.086	0.820
	Standard Error	0.029	0.013	0.016
Robust Targeted Verification Data - Restricted Estimate				
	Estimate	0.129	0.089	0.822
	Standard Error	0.028	0.013	0.016
Robust Targeted Verification Data - Unrestricted Estimate				
	Estimate	0.129	0.089	0.822
	Standard Error	0.029	0.014	0.016

The restricted and unrestricted estimates with the targeted verification plan are almost identical to the original estimates with the full verification plan used in Danila et al. (2013). Additionally the standard error estimates are essentially the same as well. The restricted and unrestricted model estimates are exactly the same because the restriction is not relevant in this example dataset. However the estimated standard error for the estimates is slightly higher for the unrestricted model, which is expected. The estimates achieve equivalent performance in this example and yet did so with only 34 verifications compared to 100 for the full verification plan.

5.7 Simulation Study

To compare the multinomial-based estimates to the beta-binomial ML estimates ten thousand data sets were generated under beta-binomial assumptions and then the parameters of interest were estimated using MLE with the unrestricted and restricted parameter spaces as described in Section 5.2 and the beta-binomial approach detailed in Chapter 4. This was also done for data generated with GRE assumptions. The relative standard error of each set of estimates is calculated and presented in Figure 5.1 and the relative bias of each set of estimates is displayed in Figure 5.2.

Standard Error Comparison

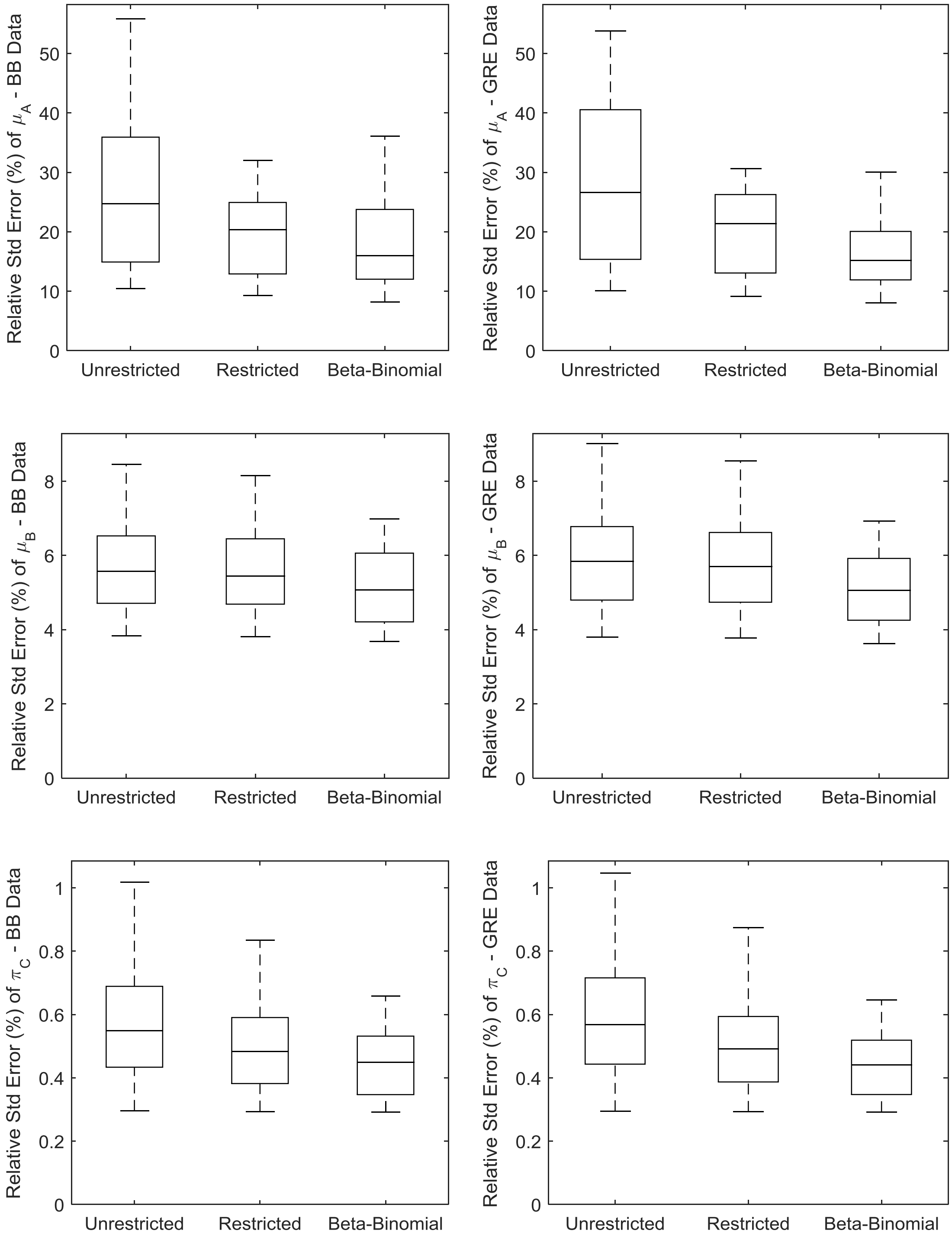

Figure 5.1 - Three-phase Simulation – Relative Standard Error

Factorial experiment run at all combinations with $n_F = 490$, $n_P = 10$, $n_B = 10000$, $r = 7$, $\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Bias Comparison

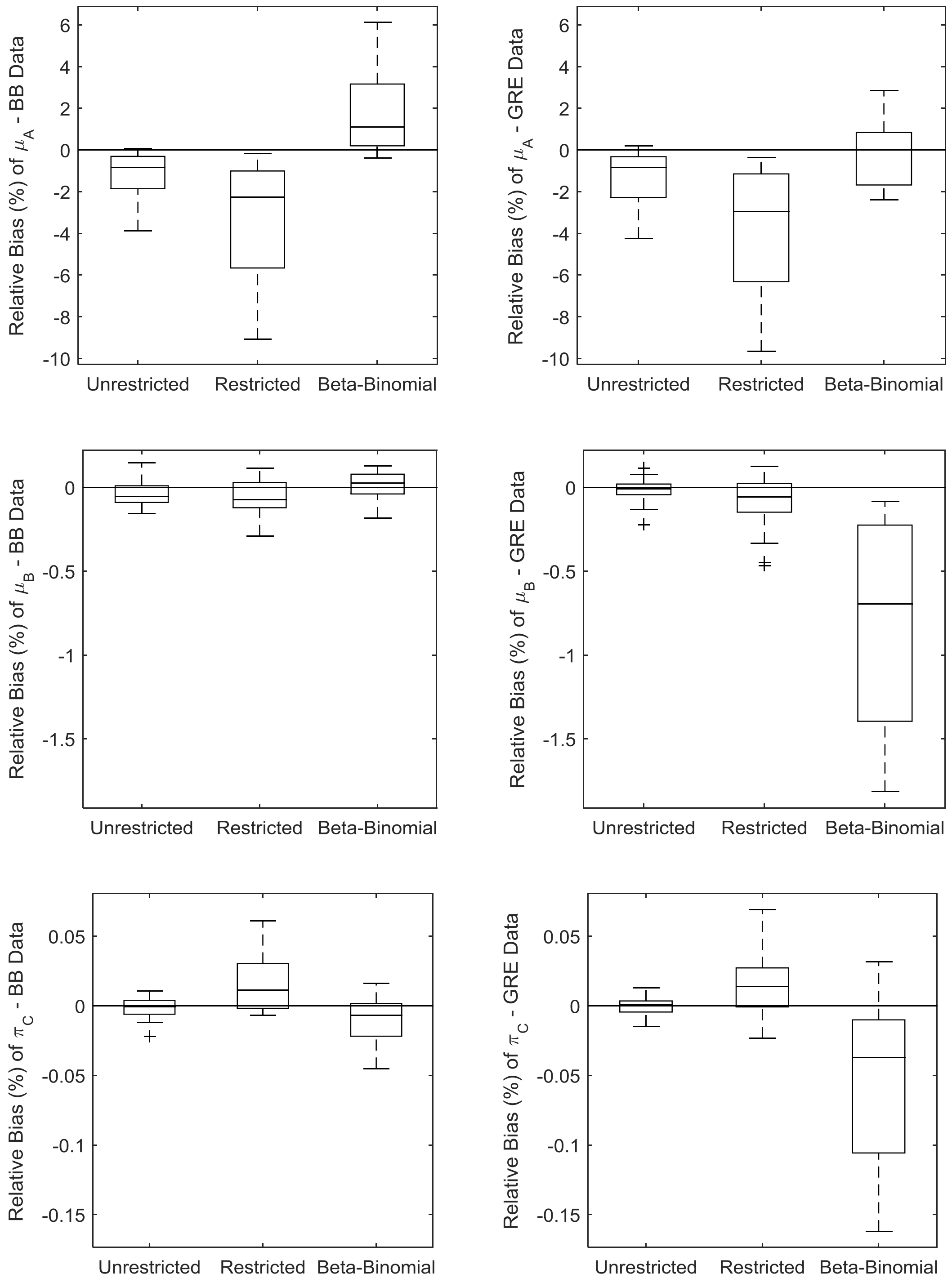

Figure 5.2 - Three-phase Simulation – Relative Bias

Factorial experiment run at all combinations with $n_F = 490$, $n_P = 10$, $n_B = 10000$, $r = 7$, $\mu_A, \mu_B = 0.05, 0.1; \pi_C = 0.9, 0.95; \gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Accuracy of Standard Error Estimates

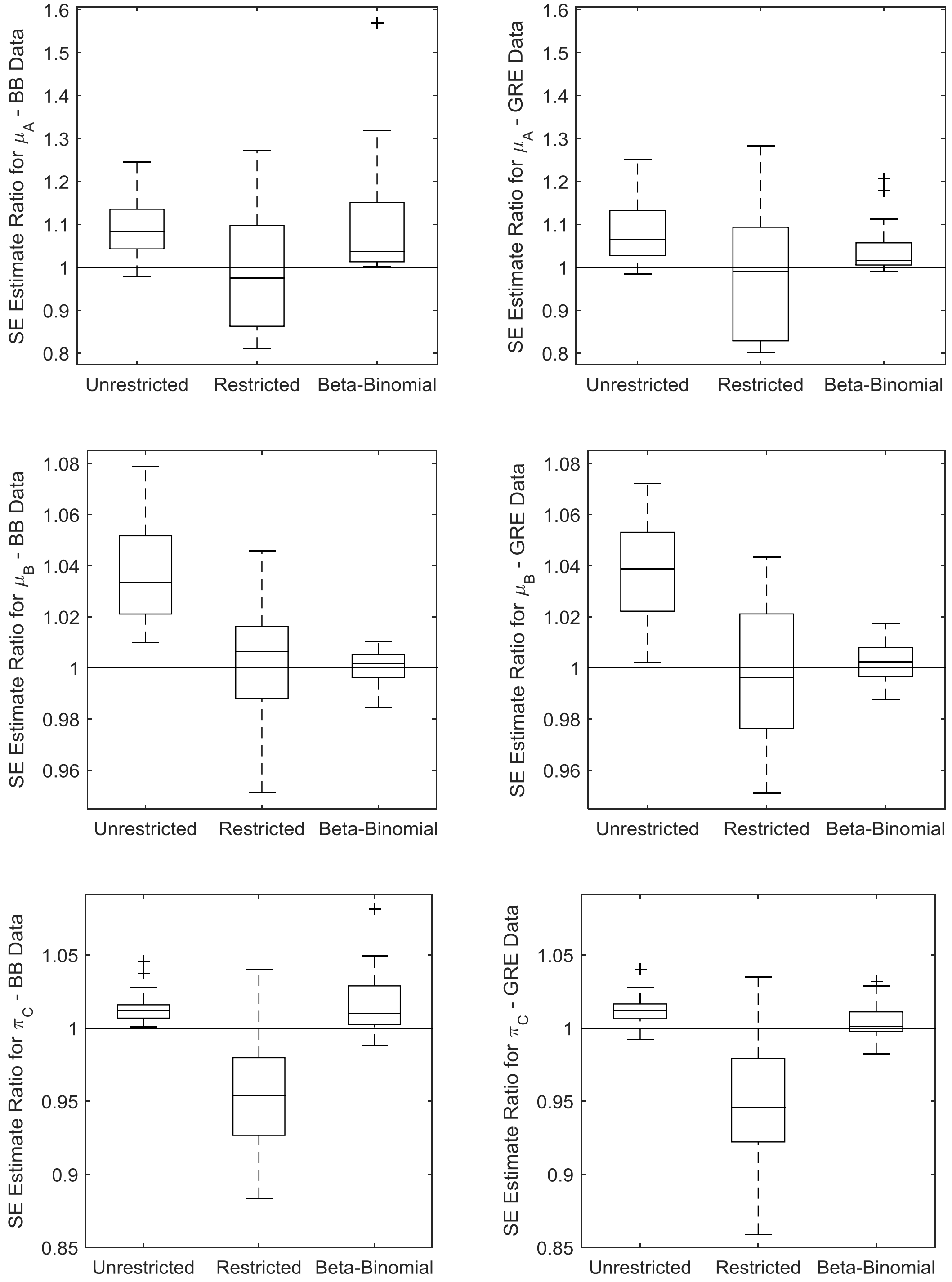

Figure 5.3 - Three-phase Simulation – Standard Error Estimate Accuracy

Factorial experiment run at all combinations with $n_F = 490$, $n_P = 10$, $n_B = 10000$, $r = 7$, $\mu_A, \mu_B = 0.05, 0.1$; $\pi_C = 0.9, 0.95$; $\gamma_A, \gamma_B = 0.05, 0.2$; (See Table 2.4)

Figure 5.1 shows that the beta-binomial ML estimates have slightly lower standard error than the restricted model estimates and significantly lower standard error than the unrestricted model estimates. The standard errors are the approximately the same for beta-binomial data and GRE data. The largest difference arising from the two data types is between the two beta-binomial ML estimates for μ_A. The restricted model seems to bridge the majority of the gap in precision between the unrestricted model estimates and the beta-binomial ML estimates.

Figure 5.2 shows that the unrestricted model estimates have the lowest bias for all three parameters of interest for both data types. The restricted model estimate for μ_A has larger bias than that of the beta-binomial model estimate. The restricted model estimates for μ_B and π_C have comparable bias to the beta-binomial ML estimates when beta-binomial data is used and less bias when GRE data is used.

Figure 5.3 shows that the standard error estimates are accurate enough for use; there is only one outlier in the beta-binomial standard error estimate for $\hat{\mu}_A$. The accuracy of the bootstrap standard error estimate for the restricted model varies the most, while the standard error estimate for the unrestricted model slightly over-estimates the standard error on average. The standard error estimates for the beta-binomial ML estimates seem to be mostly accurate; they are sometimes more accurate than that of the unrestricted estimate and sometimes less.

5.8 Discussion

Both restricted and unrestricted estimates make only assumptions that can be justified in practice and thus have relevant theoretical properties. They also have performance comparable to the beta-binomial model. The unrestricted estimators are consistent provided the experiment is conducted as prescribed in Chapter 4. The restricted estimators are consistent in the same circumstances provided the measurement system satisfies the non-decreasing ϕ constraint. In the finite sample cases explored, the unrestricted model is closer to being unbiased than either the restricted or beta-binomial models but has greater variance. Applying the restriction to the ϕ parameters reduces the variance in estimation so much that the restricted model estimates have variance closer to the beta-binomial ML estimates than the unrestricted model estimates. However applying the restriction increases the bias in finite samples. Considering all the findings of this chapter I would say when an assessment study with very limited resources is being conducted, and where bias is not a primary concern, the beta-binomial model estimates are preferable. However in an assessment study with greater resources, where asymptotic bias becomes more relevant I would recommend the restricted model. And finally, when no possibility of asymptotic bias can be tolerated, I would recommend the unrestricted model estimates.

5.9 Future Work

The plan used in Chapter 4 was made to suit the beta-binomial model estimates. Some of the details of the plan may not be optimal for these new multinomial-based estimates. In particular, the optimal value of r , the number of repeated measurements, may be lower for these estimates, since a larger value of r implies a larger number of parameters. This possibility will be left as future work.

Chapter 6 Discussion

6.1 Contributions

The primary contribution of this thesis is the concept of targeted verification. Prior to this work the advantages of this concept have not been explored in the assessment of a BMS.

Chapter 2, and the associated referencing publication in the Journal of Quality Technology, develops the concept of targeted verification in the simplest scenario where it is applicable: a plan with repeated measurements and with sampling from the population of interest. It develops a recommended plan, first for the verification phase, i.e. how targeted verification should be conducted, then for the repeated measurement phase. It then assesses the performance of the recommended plan, including its bias and standard error properties with comparison to the full and no verification plans. Upon the completion of the contents of Chapter 2, I found the results were so promising that developing further plans that integrated targeted verification was worthwhile.

Chapter 2 uses a beta-binomial model previously developed by Danila et al. (2012) to calculate ML estimates for the recommended targeted verification plan. Chapter 3 creates a new estimation procedure that combines traditional estimation techniques in a way that satisfies contemporary aversion to unnecessary assumptions. Therefore, the new estimates have theoretical properties that are relevant in practice, because there are no inherent assumptions that can’t be fully justified. The new estimates have comparable standard errors to the beta-binomial model estimates under the recommended targeted verification plan. The comparable finite sample size performance and preferable theoretical properties and robustness make them an advantageous alternative to the beta-binomial model estimates.

Chapter 4 takes the targeted verification concept and applies it in a more complicated scenario where conditional sampling is used and baseline information is available. It establishes a new recommended targeted verification plan and gives a justification for the various inherent design decisions. It compares the performance of the proposed targeted verification plan to the associated full and no verification plans, and finds that targeted verification attains performance similar to that of full verification with effort similar to that of no verification. It also compares the conditional sampling targeted verification plan with the population sampling targeted verification plan of Chapter 2. It shows that the improvements in efficiency from using conditional sampling and baseline information in targeted verification plans is very similar to the improvements these design elements make in full verification plans.

Chapter 5, continuing in the spirit of Chapter 3, develops a new estimation procedure for the plan developed in Chapter 4 that eliminates the assumptions of the beta-binomial model. By eliminating assumptions that cannot be verified, the new estimation procedure possesses reliable theoretical properties. It develops two variants of the new estimation procedure, one with a restriction on the parameter space that is intuitive and reasonable in practice. It compares the performance of the new estimation procedures to the beta-binomial model estimates of Chapter 4. The estimates have comparable finite sample performance and more relevant theoretical properties.

Ultimately I think the plans, and associated estimation techniques developed herein, are novel and are at the forefront of efficient and robust assessment of a BMS. I hope the contribution of this thesis will inspire innovations in other related areas but also lead to further improvements in the assessment of a BMS.

6.2 Future Work

Adaptation to Medical Studies

While the medical literature and application has been considered in this thesis, it is not the focus, and the plans are designed primarily for quality improvement and monitoring. Therefore there may be possibilities for future work in better adapting the methodology for medical studies. In particular considering cases where the number of repeated measurements may be restricted due to ethical considerations.

Sources of Variation in Binary Measurement Systems: (Multiple Operators)

This plan assumes repeated measurements come from a BMS that has varying results but does not consider or model any sources of variation. It may be of interested to record and model data that may affect the measurement system. The most obvious source of variation in a measurement system may be the operator. Being able to study and decompose the variation of a BMS will give a more complete understanding but also may provide direction for improvement of the BMS.

Targeted Verification with Gold Standard of Continuous Measurement Systems

Continuous measurement systems in industry are typically assessed with plans that make no gold standard measurements. The measurement system is assessed purely on the consistency of the repeated measurements. That is, if the repeated measurements on the same part have little variance then the continuous measurement system is deemed to be good. However this is not a full assessment of the accuracy of a continuous measurement. Even a measurement system that has perfectly consistent measurements, may still be very inaccurate, particularly measurements could be consistently wrong. To give a more complete assessment of a continuous measurement system a gold standard measurement is needed. Assessment studies with gold standard measurement systems are not thoroughly studied in the quality improvement literature for assessing continuous measurement systems. Additionally, the situation implies that the gold standard, which is not being considered for ongoing use, is expensive or otherwise burdensome to use. Thus many of the techniques used to assess a BMS which improve the efficiency of a plan with respect to gold standard usage can be applied to a continuous measurement system as well. That certainly includes the concept of targeted verification.

Extension to 2+ ordered categories

While this thesis considers how to use targeted verification to assess a BMS, the methods developed herein may be extended to the assessment of categorical or ordinal measurement systems. However, the number of bins in the repeated measurement phase would increase very quickly with the number of categories. Specifically the number of bins is $(r+1)^{c-1}$, where r is the number of repeated measurements and c is the number of categories. Either some model assumptions would have to be made to reduce the number of parameters or r would have to be kept small or both.

Comparison of two BMS

As opposed to assessing one measurement system, a study may instead compare two binary measurement systems to determine which is more suited for future use or to isolate important differences. Similarly the concept of targeted verification can be used here, but the number of bins would be $(r_1+1)(r_2+1)$ where r_i is the number of repeated measurements applied with test i. The number of bins is manageable for smaller values of r_i, but model assumptions may be needed to reduce the number of parameters when larger values for r_i are used.

Targeted Verification with an Accurate but Fallible Measurement System

The most obvious extension of targeted verification with a gold standard is targeted verification with a fallible but accurate reference measurement system. One way this could be implemented would be to use a reference measurement system that has known properties. The properties needed depend on the manner in which the reference measurement system is used, single measurement or repeated measurements, and the assumptions which the experimenter is willing to make in modeling. This type of assessment study can provide consistent estimates without the use of a gold standard, although a gold standard would have been required previously to assess the reference measurement system.

Another way this could be implemented is with a reference system that is thought to be reasonably accurate but has partially or even fully unknown statistical properties. This reference system could be used as a cheap screening mechanism to increase the number of non-conforming parts, or diseased subjects when struggling with rarity. The screening test could be used to stratify the population in a way that will allow for more efficient estimation. This approach would require the use of a gold standard measurement system.

Incorporating Other Types of Baseline Information

Some manufacturing processes implement the BMS being assessed in abnormal ways. For instance if a part fails inspection it may be retested due to lack of trust in the measurement system. If this procedure is well documented, this type of information can be incorporated into the estimation procedure and can provide improvements to estimate precision similar to those attained by incorporating standard baseline information.

Bibliography

- Akkerhuis, T. S. (2016). Measurement system analysis for binary tests Amsterdam: IBIS UvA
- Albert, P. S., & Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2), 427-435.
- Albert, P. S., & Dodd, L. E. (2008). On estimating diagnostic accuracy from studies with multiple raters and partial gold standard evaluation. *Journal of the American Statistical Association*, 103(481), 61-73.
- Automotive Industry Action Group (2010). *Measurement Systems Analysis*, 4th edition. Southfield, MI: AIAG. Reference Manual
- Baker, S. G. (1995). Evaluating multiple diagnostic tests with partial verification. *Biometrics 51:1,* 330-337.
- Begg, C. B. and R. A. Greenes (1983). Assessment of diagnostic test when disease verification is subject to selection bias. *Biometrics 39,* 207-215.
- Boyles R. A. (2001). Gauge Capability for Pass—Fail Inspection. *Technometrics* 43:2, 223-229.
- Bross, I. (1954). Misclassification in 2x2 tables. *Biometrics 10,* 474-486
- Chen, J., & Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in comp
- lex surveys. *Statistica Sinica*, 9(2), 385-406.
- Cochran, W. G. (1968). Errors of Measurement in Statistics. *Technometrics 10:4,* 637-666.
- Collins, J., M. Huynh (2014). Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine 33:24,* 4141-4169.
- Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.
- Danila, O., Steiner, S. H., & Mackay, R. J. (2008). Assessing a binary measurement system. *Journal of Quality Technology 40*(3), 310-318.
- Danila, O., Steiner, S. H., & MacKay, R. J. (2010). Assessment of a binary measurement system in current use. *Journal of Quality Technology 42*(2), 152.
- Danila, O., Steiner, S. H., & MacKay, R. J. (2012). Assessing a binary measurement system with varying misclassification rates using a latent class random effects model*. Journal of Quality Technology 44*(3), 179.
- Danila, O., Steiner, S. H., & MacKay, R. J. (2013). Assessing a binary measurement system with varying misclassification rates when a gold standard is available. *Technometrics*, 55(3), 335-345.
- Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, 20-28.
- De Mast, J., Erdmann, T.P. and Van Wieringen W.N. (2011). "Measurement system analysis for binary inspection: Continuous versus dichotomous measurands". *Journal of Quality Technology,* 43, 99-112.
- 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.
- Engel, B., Backer, J., & Buist, W. (2010). Evaluation of the accuracy of diagnostic tests from repeated measurements without a gold standard. *Journal of agricultural, biological, and environmental statistics*, 15(1), 83-100.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222, 309-368.

- Fisher, R.A. (1925). "Theory of Statistical Estimation". *Proceedings of the Cambridge philosophical Society*, 22, 700-725.
- Gart, J. J., A. A. Buck (1966). Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology 83:3,* 593-602.
- Haitovsky, Y. and J. Rapp (1992). Conditional resampling from misclassified multinomial data with application to sampling inspection. *Tecknometrics 34,* 473-483.
- Hui, S. L. and S. D. Walter (1980). Estimating the error rates of diagnostic test. *Biometrics 36,* 167-138.
- Lehmann, E. L., & Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation: Part I. Sankhyā: *The Indian Journal of Statistics*, (1933-1960), 10(4), 305-340.
- Lehmann, E. L., & Scheffé, H. (1955). Completeness, similar regions, and unbiased estimation: Part II. Sankhyā: *The Indian Journal of Statistics* (1933-1960), 15(3), 219-236.
- Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data. John Wiley & Sons.
- Mandel, J. (1959). The measuring process. *Technometrics*, 1(3), 251-267.
- McCaslin, J. A., Gruska, G. F. (1976). Analysis of attribute gage systems. *ASQC Technical Conference Transactions* (Vol. 30, pp. 392-399).
- McNamee, R. (2002). Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Statistics in medicine*, 21(23), 3609-3625.
- Neyman, J. (1947). Outline of Statistical Treatment of the Problem of Diagnosis. *Public Health Reports 62,* 1449-1456.
- 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 effect models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics 52,* 797-810.
- Quade D., P. A. Lachenburch, F. S. Whaley, D. K. McClish, and R. W. Haley (1980). Effect of misclassification of statistical inference in epidemiology. *American Journal of Epidemiology 111*(5), 503-515.
- Slutsky, E., 1925. Über stochastische Asymptoten und Grenzwerte. *Matematische Annalen* 5 (3) p. 3.
- Spiegelhalter, D. J. and P. G. I. Stovin (1983). An analysis of repeated biopsies following cardiac transplantation. *Statistics in Medicine 2,* 33-40.
- Tennenbein, A., (1969) Estimation From Data Subject to Measurement Error. (Ph.D. Dissertation, Statistics Department) Harvard University, Cambridge, Mass.
- Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of American Statistical Association 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. *Technometrics 14,* 187-202.
- Vacek, P. (1985). 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 75,* 211-218.

- Van Wieringen, W. N. and J. De Mast (2008). Measurement system analysis for binary data. *Technometrics 50,* 468-478.
- Whiting, P., Rutjes, A. W., Reitsma, J. B., Glas, A. S., Bossuyt, P. M., & Kleijnen, J. (2004). Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Annals of internal medicine*, 140(3), 189-202.
- Yerushalmy, J. (1947). Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Health Reports 63,* 1432-1449.
- Zhou, X. H., N. A. Obuchowski, and D. K. McClish (2011). *Statistical methods in diagnostic medicine.* New York: Wiley.