A Bayesian calculator for estimating the incidence of errors in large populations

Bhojnarine R. Rambharat*
Office of the Comptroller of the Currency
Washington, DC 20024
e-mail: ricky.rambharat@occ.treas.gov.

[DRAFT VERSION]

February 3, 2014

Abstract

The inspection of *errors* (or *exceptions*) in a given population arises as a problem in various applied fields. An important question that has been exhaustively researched in the literature is what is the proportion of errors in a population. This question is relevant to quality assurance (or quality control) settings where it is necessary to understand if an underlying process functions effectively. For instance, the proportion of defective parts in a manufacturing plant may be of interest to a quality assurance / control team, and the underlying construction process may be deemed sub-standard if the proportion of errors statistically exceeds a given threshold. Additionally, errors can have multiple dimensions such that there may be graduated degrees of errors. Thus, the population can no longer be modeled using a binary framework (error / no error), but, rather, must utilize more complex, multivariate paradigms. Further, any information about the distribution of errors in the population is usually unknown. We estimate the *number* of errors in large population using a Bayesian Markov chain Monte Carlo (MCMC) approach. Our calculator provides a framework to reconcile prior assumptions about the distribution of errors in a given population with what is observed once a sample is evaluated. The present paper provides a calculator when a population can be regarded as binary (error or no error) as well as when a population contains errors with multiple levels. In the latter case, we use the NORTA ("Normal To Anything") algorithm in Chen (2001) along with a multivariate Poisson simulator described in Yahav and Shmueli (2012) in an MCMC framework to produce posterior distributions of the number of errors in each dimension within the population. Our analysis provides a method to reconcile "prior" and "posterior" beliefs about how errors are modeled in a population and, indeed, our results could be used for more informed, applied studies about errors.

Key words: Bayesian, Calculator, Sample design, Population, Total errors, Prior belief, Posterior outcome, Evaluate, Reconcile.

*The views expressed in this paper are solely those of the author. This paper neither represents the opinions of the U.S. Department of the Treasury nor the Office of the Comptroller of the Currency (OCC). All errors are the author's responsibility. As this paper is in **draft status**, please do not circulate or quote without first conferring with the author.

1 Introduction

The incidence of aberrations in a given population presents opportunities for the application of advanced statistical methods that help to uncover significant insights. Typically, the populations are large and inspection of each item is practically infeasible. Hence, statistical modeling, primarily through sample design and analysis, become indispensable for understanding the nuances of the population, particularly where errors are concerned. There are various applied settings where the analysis of errors is important, particularly those under the purview of quality assurance / quality control (QA/QC) stakeholders. Hence, the question of how to properly inspect a population for errors naturally arises among such stakeholders.

The steps to understanding how errors are modeled in a population usually involve i) prior beliefs about errors, ii) collecting a sample of data, and iii) evaluating sample results to form post-sample (or posterior) beliefs about the errors. These steps present a ripe application for the use of Bayesian methodologies. The use of a Bayesian framework allows for the adaptation of a viable scientific rubric when analyzing the incidence of errors in a given population. The agreement / disagreement between posterior outcomes and prior beliefs provides a means to reconcile preconceived notions about errors, and this could be informative for more in-depth studies. The study of errors in a population is intimately linked to the notion of sample size determination since inference about the incidence of errors in a given population will typically rely on a sample.

One classification of errors in a population is to think of the presence or absence of errors. The analysis of errors in this binary framework is well-studied in the literature. The seminal work in Cochran (1977) provides a thorough treatment of how to analyze errors in a given population and keen attention is given to sample size determination results. The treatment in Thompson (2012) also provides important results on how to estimate the prevalence of errors in a given population. A few examples of Bayesian research work addressing the analysis of sample size determination when analyzing errors in a population include M'Lan et al. (2008), Pham-Gia and Turkkan (1992) and Adcock (1987). The aforementioned list is by no means exhaustive, but it provides background about the conceptual framework surrounding the modeling of errors in a binary population (error vs. no error).

A related, albeit more complex problem is the study of errors in a population where the classification is no longer binary. Specifically, an "error" can have multiple levels of severity (e.g., High, Medium, or Low). This problem requires the use of a multinomial modeling paradigm, and a non-exhaustive reference list includes Thompson (1987), Alam and Thompson (1972), and Mavridis and Aitken (2009).[1] As in the case of the binary (binomial) problem described in the preceding paragraph, the references listed here contemplate the modeling of errors along with a sampling design (or sample size calculator) in order to determine how to sample from the population in order to reliably estimate the incidence of errors.

The problem of determining how many errors exist in a population is not tied to a specific "statistical ideology" *per se* – i.e., Bayesian or Classical. Moreover, as the aforementioned, partial bibliography on analyzing error incidence shows, the sample size problem is intricately connected to understanding errors

[1] The references in Mavridis and Aitken (2009) provide some additional background to the estimation of error incidence.

in a population. The present analysis will be primarily concerned with how to reconcile prior and posterior beliefs about the incidence of errors. We do not prescribe a specific sample size determination methodology but, rather, align the initial sample size determination problem to prior beliefs about the incidence of errors in a given population. Upon observing a sample of items from the population, the posterior outcome is reconciled with the prior beliefs, and if subsequent samples need to be drawn, sample sizes can be derived using sampling models that reflect the updated posterior outcomes.

Typically, investigators seek to estimate a *rate* of errors in a given population. The error rate is a convenient statistic as asymptotic theory can be harnessed to facilitate meaningful, applied analysis. Our concern in this paper is estimating the *total* number of errors as we believe this can help key stakeholders make informed decisions about the underlying population, especially where specific costs are concerned. While the error rate has obvious benefits, the quality of statistical inference about it may be limited when the rate is very low or very high. We demonstrate, using numerical experiments, that inference about the *total* number of errors is more reliable near these boundaries. Additionally, inference about the total number of errors could help to better inform analysts about potential deep-dive studies for a given population.

Our analysis begins in section 2 with a discussion of the incidence of total errors, treating the binary and multiple error cases separately. Additionally, we discuss the special "boundary case" where errors are considered rare – at least in an *a priori* sense. Next, we discuss in section 3 the common sampling models utilized when creating sample designs and computing sample sizes under various assumptions. Section 4 presents two Bayesian estimation routines that provide MCMC estimates of the posterior distribution of the *total* number of errors in a given population. One algorithm is devoted to the binary case and another treats the multiple error case. Section 5 presents some basic simulation results under different experimental conditions. Finally, section 6 concludes with a discussion and opportunities for future work.

2 Incidence of total errors

The modeling of the incidence of total errors requires a discrete distribution as the underlying statistical instrument. We discuss the case of modeling a binary population, and then modeling a population where there are errors that have multiple levels. We also discuss the special case where errors are treated as "rare" in a given population, and what that entails for statistical modeling strategies.

An important point to consider when modeling errors is the reconciliation between prior beliefs about errors and the actual evaluation of sample outcomes. For example, if errors are presumed to be non-rare *a priori*, the question of how might one evaluate a sample where the incidence of errors is rare presents non-trivial statistical issues in a practical setting. Additionally, the question of how to model errors if they are believed to be rare requires specific modeling strategies, and the reconciliation between *a priori* beliefs and evaluated outcomes should use a robust statistical rubric. Indeed, if errors are believed to be be rare and they are not, then operational consequences could be significant. We address these issues in both the univariate and multivariate cases in the ensuing sections.

2.1 Univariate errors

The most basic model of errors in a given population is one where the population is binary – i.e., either a given population element is an error or not. Figure 2.1 provides a schematic of how to conceptualize univariate errors. The "error" vs. "non-error" is depicted by red dots vs. blue dots in part A of the schematic.

A common approach to modeling errors in this situation is to adapt the *Binomial* model and estimate the *rate* of errors in the population. This model is applicable in large populations where the sampling *with* replacement assumption is plausible. Additionally, one could also use the *Hyper-Geometric* model if there is a need to carefully account for the sampling *without* replacement assumption. In the case of the Hyper-Geometric model, however, it would most likely be the total number of errors that will be modeled.

Furthermore, large-sample theory could be used where the Normal model applies since it is well-known that the sample proportion of errors has an asymptotic Normal distribution (cf. DeGroot and Schervish (2002), for example). This approach is very common in the survey sampling literature, and the use of the Normal model is described in Cochran (1977). In fact, the Normal modeling paradigm is typically invoked in sample size calculations, although, other modeling paradigms (to be described in section 3) could also be used for sample size calculations. One limitation of the Normal model is that it is not reliable when the error rate is close to boundary points (i.e., 0 or 1).

It is straightforward to model the rate of errors in a given population as various statistical modeling tools could be harnessed to this purpose. Moreover, the rate of errors provides key stakeholders, including QA/QC teams, a sense of "risk" in terms of the incidence of errors in a population. Our approach will be to model the total number of errors. We believe that an estimate of the distribution of the total number of errors will be useful, particularly if one is interested in more precise cost-related calculations. Arguably, the error rate could be translated to a total, however, the result is not necessarily accurate, especially when the rate is close to 0 or 1. The sampling model of choice for our problem will be the Hyper-Geometric model as we intend to work in the more general "sampling without replacement" framework.

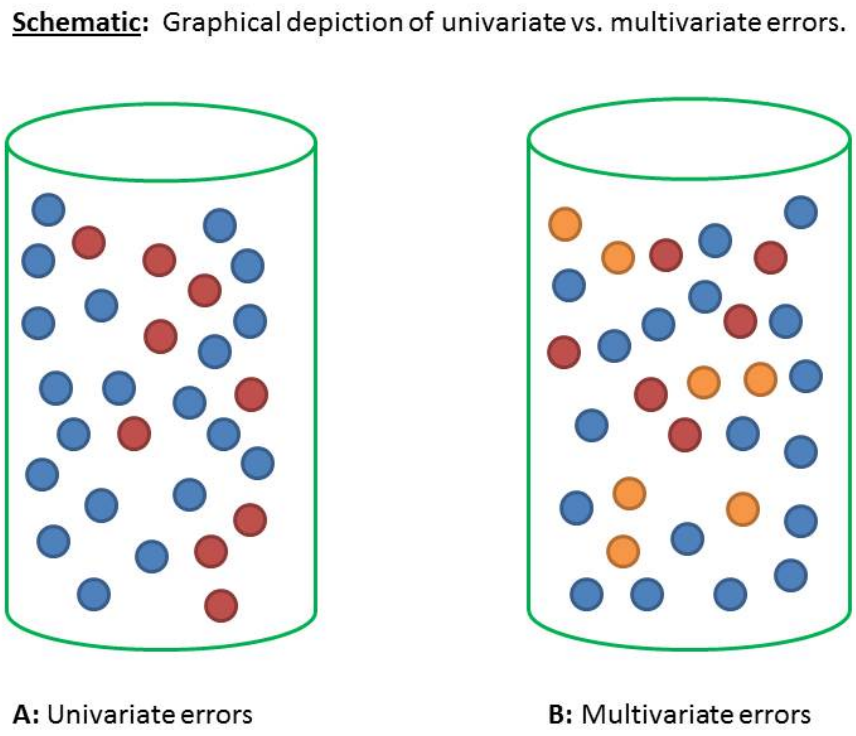
Apart from the case where an error has only a "single shade," there are problems of applied interest where errors have multiple dimensions. For example, an error may be classified as "moderate" or "serious." Thus, one needs to use the appropriate statistical tools in this situation, which we now describe in the following sub-section.

2.2 Multivariate errors

The incidence of errors where there are multiple levels presents additional nuances that ought to be taken into account when using statistical modeling tools. Figure 2.1 illustrates the multivariate error situation in part B of the schematic. An "error" could be an error along multiple dimensions. For example, administrative errors may be less egregious (orange dots) relative to more systemic errors (red dots) that could have dire consequences for an entity.

The extension of the binomial modeling framework in this situation would be the multinomial model. In the multinomial setting, sampling is also assumed to be executed with replacement, but there are multiple

Figure 1: A graphical illustration of the univariate error problem vs. the multivariate error problem. The blue dots are non-errors in both cases. In the case of univariate errors (A), we aim to estimate the total number of red dots in the population, whereas in the case of multivariate errors (B), we aim to estimate the total number of orange dots (less egregious errors) and the total number of red dots (more egregious errors).



rates to estimate. For example, one would seek to estimate the proportion of orange dots and proportion of red dots, and from these two one could recover the proportion of blue dots in Figure 2.1 (part B). The multinomial problem is also well-studied as is demonstrated by the works cited in section 1. Both Classical and Bayesian approaches have been used to study the multinomial problem. Estimation is not as straightforward as in the binomial setting, but established algorithms are available. Additionally, more elaborate and complex statistical modeling in the categorical case can be uncovered in Fienberg (1980).

Our approach is to estimate the *total* number of errors in each category. In this instance, we use the multivariate extension of the Hyper-Geometric model to accommodate the general situation of "sampling without replacement." We make the argument for estimating total errors as opposed to error rates as the former could be more useful, particularly in terms of cost-related constraints, relative to the latter.

2.3 Rare vs. non-rare errors

One key issue to consider is the notion of rare errors. This would be the case where, say, the red dots in Figure 2.1 are significantly outnumbered by the blue dots. Additionally, in the case where there are multiple types of errors, it may be that one or more type is dominated by the non-errors. A common model for rare occurrences is the Poisson model, which is described in more detail in section 3.

The work in Cochran (1977) or, more recently, in Valliant et al. (2013) provides some insights about how to estimate rare items (errors) in a population with due attention paid to sample size calculations. The use of an appropriate statistical rubric, typically in the form of a hypothesis test, should be employed so that an analyst can carefully reconcile prior beliefs about errors (e.g., if they are rare or not) with what is actually observed in a sample. While judgment is of value when evaluating a sample, a robust statistical approach should be used when reconciling rare vs. non-rare errors as part of sample evaluation activities.

3 Sampling models

The two fundamental probabilistic paradigms for sampling designs are 1) *sampling with replacement*, and 2) *sampling without replacement*. The latter paradigm is likely the most appropriate for practical settings because review of sampled items typically entail costs, hence items are usually not replaced once sampled. The probabilistic model relevant for the sampling without replacement case is the *Hyper-Geometric* (HG) model whose probability mass function (pmf) is given by:

$$\Pr(X = k|N, M, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \tag{1}$$

where X is the random variable (r.v.) taking values between 0 and N (the population size), n is the sample size, M is the number of exceptions in the population, and k is the observed number of exceptions in the sample. While this model is the exact model for the sampling without replacement case, it is computationally burdensome to manipulate for sample size calculations, especially when the population size, N, is large. As noted in authoritative references on sampling like Cochran (1977) and Thompson (2012), reliable approximations for large populations exist, with commonly applied approximations based on the Binomial and Normal distributions.

The Binomial distribution is a sampling *with replacement* model. The probability of finding an exception in this sampling paradigm is given by the Binomial pmf:

$$\Pr(X = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{2}$$

where the r.v. X takes values between 0 and n (the sample size), k is the observed number of errors is the sample, which is drawn from a population without replacement, and p is the probability of finding an error in the population. Since it is assumed that sampling is done with replacement, the parameter p is assumed constant. Given an assumption about p, it is possible to compute the requisite sample size such that the

probability of observing exactly k errors (or at least k errors) in a given sample is equal to a stipulated value τ.[2] Typically, this requires a numerical procedure in order to solve for the required sample size.

The Normal distribution presents another approach to estimate sample sizes, and, indeed, it is perhaps one of the most widely used methods to compute sample sizes in audit contexts. If $\hat{p}$ represents the proportion of exceptions in a sample of size n, then the asymptotic distribution of $\hat{p}$ is Normal with mean p and variance $p(1-p)/n$, where p is the true population proportion of exceptions. If we invoke standard facts about the Normal distribution, we can arrive at a formula for a sample size. As stated in Cochran (1977), a formula for the sample size n is:

$$n = \frac{z_\alpha^2 p(1-p)/d^2}{1 + \frac{1}{N}\left(\frac{z_\alpha^2 p(1-p)}{d^2} - 1\right)}, \tag{3}$$

where z_α is the quantile of the standard Normal distribution that reflects the level of confidence (e.g., 90%, 95%, etc.), p is the probability of finding an error in the population of size N, and d is the degree of *precision* that the analyst desires to have around p. The Normal distribution is typically used for moderate values of p as it may not be optimal if either p or $1-p$ is very small.

As noted, in a sampling without replacement context, the HG model is the correct sampling model to adopt. The two approximations we noted, the Binomial and the Normal, are not at all exhaustive. For example, the Poisson distribution with parameter λ is also commonly adopted to model errors, especially when errors are considered rare. The pmf for the Poisson distribution is:

$$\Pr(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{4}$$

where X is a random variable taking values on the non-negative integers, and $\lambda > 0$ is a parameter that represents both the mean and variance of X. As an approximation to the Binomial distribution, the parameter $\lambda = np$, where n is the sample size and p is the Binomial probability of success.

In the case of investigating multiple levels of errors, the Multinomial model generalizes the Binomial model. The pmf of the Multinomial model is

$$\Pr(X_1 = x_1, \ldots, x_k|n, p_1, \ldots, p_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}, \tag{5}$$

where, p_i ($i = 1, \ldots, k$) is the probability of selecting an item in category i, x_i is the realized number of items from category i, which is modeled by the random variable X_k, and n is the sample size. In the above specification, once $X_1, \ldots, X_{k-1}$ (equivalently, $p_1, \ldots, p_{k-1}$) are determined, then X_k (correspondingly, p_k) is also determined based on the usual multinomial constraints. As in the case of the Binomial model, the context for the Multinomial model is a *sampling with replacement* framework.

For completeness, we provide the multivariate extensions to the Hyper-Geometric and Poisson models.

[2]For example, τ could take on values such as 0.90, 0.95, 0.99, etc.

The multivariate HG pmf is given by:

$$\Pr(Y_1 = j_1, Y_2 = j_2, \ldots, Y_k = j_k | m_1, \ldots, m_k, n) = \frac{\binom{m_1}{j_1}\binom{m_2}{j_2}\cdots\binom{m_k}{j_k}}{\binom{m}{n}}, \tag{6}$$

where each $(j_1, \ldots, j_k) \in \mathbb{N}^k \sum_i j_i = n$. This can be used to model the incidence of the red, orange, and yellow dots in part B of Figure 2.1 assuming that sampling is done *without* replacement.

Finally, the multivariate Poisson pmf is given by:

$$\Pr(X = x, Y = y | \lambda_0, \lambda_1, \lambda_2) = e^{-(\lambda_0+\lambda_1+\lambda_2)} \frac{\lambda_1^x}{x!} \frac{\lambda_2^y}{y!} \sum_{i=0}^{\min(x,y)} \binom{x}{i}\binom{y}{i} i! \left(\frac{\lambda_0}{\lambda_1\lambda_2}\right)^i. \tag{7}$$

We will employ this model as part of our MCMC algorithm in the multivariate case where it will be used as the *proposal* distribution. Hence, it will be integral to the stochastic search that will facilitate the MCMC estimation routine. Additional details about the particulars of MCMC estimation algorithms can be found in Gelman et al. (1995).

The algorithms that we present in our analysis will mainly rely on the exact HG model combined with an MCMC routine to estimate the total number of errors in a given population. We construct an MCMC algorithm that uses the log-probabilities of the HG model in order to estimate the posterior distribution of the total number of errors in a given population. The next section outlines our estimation strategy where we focus attention on the cases where errors have both univariate and multivariate characteristics.

4 Bayesian estimation of total errors

We use a Bayesian approach to estimate the posterior distribution of the total number of errors in a given population. We discuss two separate algorithms, one for the univariate error case and the other for the multivariate error case.

4.1 Univariate errors

In the case of univariate errors, we model the total number of errors using a Hyper-Geometric (HG) likelihood where the probability mass function (pmf) of the HG model is specified in equation (1). The proposal distribution is a Poisson distribution whose pmf is specified in equation (4).

ALGORITHM I: Bayesian calculator for univariate errors

1. Initialize a prior distribution on the total number of errors. (We use a Poisson distribution as the prior in our simulation examples.)

2. Choose a proposal distribution to execute the MCMC algorithm. (We use a Poisson[3] distribution as

[3]It ought to be demonstrated that the Poisson distribution satisfies the *detailed balance condition* (cf. Gelman et al.

the proposal in our simulation examples.)

3. Establish the likelihood for the data: we use the Hyper-Geometric model as we use the "sampling without replacement" framework.

4. For $i = 1, \ldots, M$,[4]

   (a) Evaluate the log-likelihood (and log-posterior) at the current point in the MCMC chain. Denote this by LP_{cur}.

   (b) Evaluate the log-likelihood (and log-posterior) at the proposed point in the MCMC chain. Denote this by LP_{pro}.

   (c) Simulate $u \sim \text{Unif}[0, 1]$.

   (d) If $\log(u) < \quad LP_{\text{pro}} - LP_{\text{cur}}$, then update the current point with the proposed point. Else, reject the proposed point and keep the current point.

5. Output the posterior distribution of the *total* number of errors in the population.

Algorithm I is a standard MCMC algorithm that provides a method to estimate the total number of errors in a population (recall part A of Figure 2.1). The output from Algorithm I would be the posterior distribution of the total number of errors in the population, and this could be of value to key decision-makers. Key summary measures, or basic descriptive statistics, might also help to make more informed decisions about the prevalence of errors in a population. Moreover, an estimate of the total number of errors might also provide insights about the consequences of the errors. We now extend this estimation strategy to the multivariate case, which we describe in the following sub-section.

4.2 Multivariate errors

The case of multivariate errors presents additional estimation challenges beyond the univariate case. As noted above, one approach would be to adopt a multinomial model where sampling is assumed to be implemented without replacement. Consequently, one estimates an error rate for each type of error in the population. On the other hand, we estimate the posterior distribution for the total amount of each type of error in the population. We believe that where the rate has limitations, the total value could provide added value. The details of the multivariate estimation routine is provided in Algorithm II below.

ALGORITHM II: Bayesian calculator for multivariate errors

(1995)).

[4]The quantity M in the MCMC for-loop refers to the number of MCMC iterations.

1. Initialize a prior distribution on the total number of errors of each type. (We use independent Poisson priors in our simulation examples.[5])

2. Choose a proposal distribution to execute the MCMC algorithm. (We use a multivariate Poisson[6] that relies on the NORTA ("Normal To Anything") algorithm, which is described in Chen (2001) and Yahav and Shmueli (2012). We also add a small amount of negative correlation between the types of errors in the population.)

3. Establish the likelihood for the data: we use the multivariate Hyper-Geometric model as we use the "sampling without replacement" framework.

4. For $i = 1, \ldots, M$,[7]

   (a) Evaluate the log-likelihood (and log-posterior) at the current point in the MCMC chain. Denote this by LP_{cur}.

   (b) Evaluate the log-likelihood (and log-posterior) at the proposed point in the MCMC chain. Denote this by LP_{pro}.

   (c) Simulate $u \sim \text{Unif}[0, 1]$.

   (d) If $\log(u) < \ LP_{\text{pro}} - LP_{\text{cur}}$, then update the current point with the proposed point. Else, reject the proposed point and keep the current point.

5. Output the posterior distribution of the *total* number of errors of each type in the population.[8]

As noted in Algorithm II, we use a multivariate Hyper-Geometric model (cf. equation (6) for the likelihood of errors in this framework). Analogous to the univariate case, we extend the Poisson distribution to the multivariate version as use it as our core stochastic search mechanism. The output of Algorithm II is the joint posterior distribution of the multiple levels of errors in the population. For example, in part B of Figure 2.1, Algorithm II ought to output the posterior distribution of red dots and orange dots in the population.[9]

We now present the results of some preliminary numerical experiments. We also discuss the ramifications of these results, especially to policy-makers. We discuss the univariate and multivariate error cases separately, and we point out some important research objectives, particularly for the multivariate case.

[5]This assumption could be enhanced to accommodate an *a priori* dependence structure between the different types of errors.

[6]It ought to be demonstrated that the multivariate Poisson distribution satisfies the detailed balance condition (cf. Gelman et al. (1995)).

[7]The quantity M is the number of MCMC iterations.

[8]Note that a block-Metropolis version of Algorithm II is also possible to implement.

[9]Although it may not be of immediate interest, the posterior distribution of the non-errors is also available.

5 Preliminary results

First, we illustrate the univariate results and then we present the results of the multivariate case. We did more extensive testing in the univariate case. However, the testing that we did in the multivariate case pointed out some opportunities for on-going research work.

5.1 Univariate results

We report the results of 8 numerical experiments in the univariate case where the goal is to estimate the proportion of errors in a given population. Specifically, this would be in the context of estimating the proportion of red dots in Figure 2.1. The parameters of our planned simulation study are reported in Table 5.1 below.

Table 1: Setup of numerical experiments for univariate errors.

Experiment	Pop. size	Sample size	Prior	Obs. errors
set 1	10,000	500	5,000	3
set 2	10,000	500	50	3
set 3	10,000	100	5,000	3
set 4	10,000	100	50	3
set 5	500,000	5,000	50	2
set 6	500,000	5,000	250,000	2
set 7	500,000	100	50	2
set 8	500,000	100	250,000	2

As an illustration, consider experiment 1 where we assume a population size of 10,000 and a sample size of 500. (At this point, we take the sample size calculation for granted, but we return to this issue in our discussion below.) Further, we assume an expected amount of errors of 5,000 (or 50% of the population). Finally, in the sample of size 500, we observe 3 errors. The application of Algorithm I produces the posterior distribution in Figure 5.1 corresponding to set 1. Additionally, Table 5.1 reports key posterior summary measures. Based on the results from set 1, we find that the posterior outcomes shift toward what is observed in the sample, thus updating the prior beliefs. Essentially, the analyst starts with an expectation of 50%

errors in the population and find less than 1% errors in the sample. Thus, prior beliefs are updated and reconciled through this posterior analysis.

In set 2, we repeat the experiment but we start with prior beliefs that reflect a rare presence of errors in the population. Again, the posterior outcomes are updated according to what is observed in the sample, which is also indicative of rare error incidence. The point to note about the cases like set 2 is that with a preconceived belief that errors are rare in a population, commonly applied sampling models, such as the Normal model, may not be applicable. Note that our modeling framework is robust to assumptions about rare or non-rare errors.

Set 3 and Set 4 repeat the previous two experiments but reduces the sample size from 500 to 100. In this case, we see that the posterior is not as responsive to the sample results, although it is updated in the direction of what is observed in the sample (cf. Figure 5.1 and Table 5.1). Thus, the sample size calculation ought not to be taken for granted, and large enough samples should be drawn where possible.

Additionally, the effect of the population size is investigated in set 5 through set 8 where we increase the population size from 10,000 to 500,000. These experiments also point to effects due to sample size, and its appears that the larger sample size (5,000 in this case) naturally provides a richer information set about the distribution of errors than the smaller sample size of 100. For example, an analyst is clearly more informed in set 5 compared to set 7, which uses a smaller sample. Generally, however, posterior outcomes can be utilized to update prior beliefs and help one better reconcile sample results.

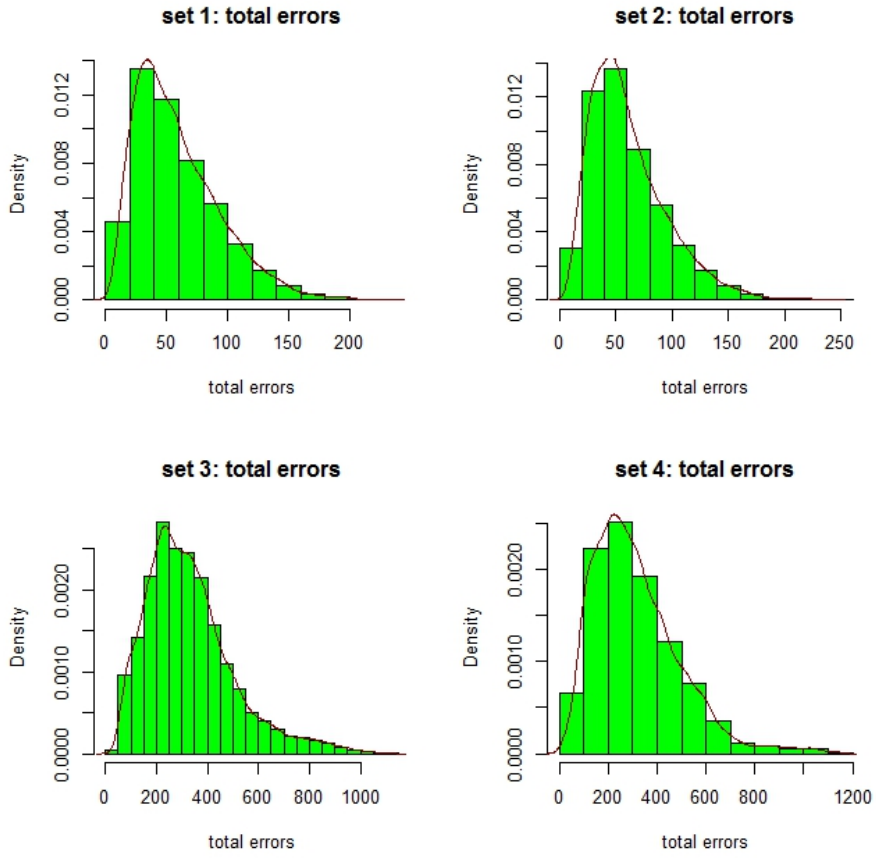
5.2 Multivariate results

The multivariate analysis of errors extends the methodology from the univariate case, but the underlying sampling models are more intricate as is described in Algorithm II. Table 5.2 provides the setup for the numerical experiment for multiple error incidence. Additionally, Figure 5.2 and Table 5.2 illustrate the corresponding results for the multivariate error case. We consider 2 types of errors: A and B. One can think of the type A errors as the orange dots in part B of Figure 2.1 and the type B errors as the red dots in part B of Figure 2.1.

The numerical testing that we have executed for the multivariate case only investigates the effects of sample size. These experiments start with a rare view of errors (both of the A and B type), and then the sample outcomes indicate that the errors may not be as rare as initially believed. It appears that the larger sample size does provide more information, but the effect seems marginal, especially for the case of error B. However, this may be due to the mixing quality of the MCMC algorithm described in Algorithm II. While the multivariate Poisson appears to function as a reasonable proposal distribution, we may need to adjust its specification to better search the parameter space. Currently, it appears that the weight of the prior information is significant in the multivariate results. Alternatively, it may well be the case that multivariate errors require larger sample sizes to better understand the underlying distribution of errors in the population.

The sample size issue also needs to be resolved in the multivariate case. Clearly, there is an effect due to sample size. At this point, the sample size calculation is taken for granted but this can be formalized through the multivariate Hyper-Geometric specification. Additionally, we may need to enhance either the

Figure 2: MCMC posterior output from numerical experiments 1 through 4 (univariate cases).

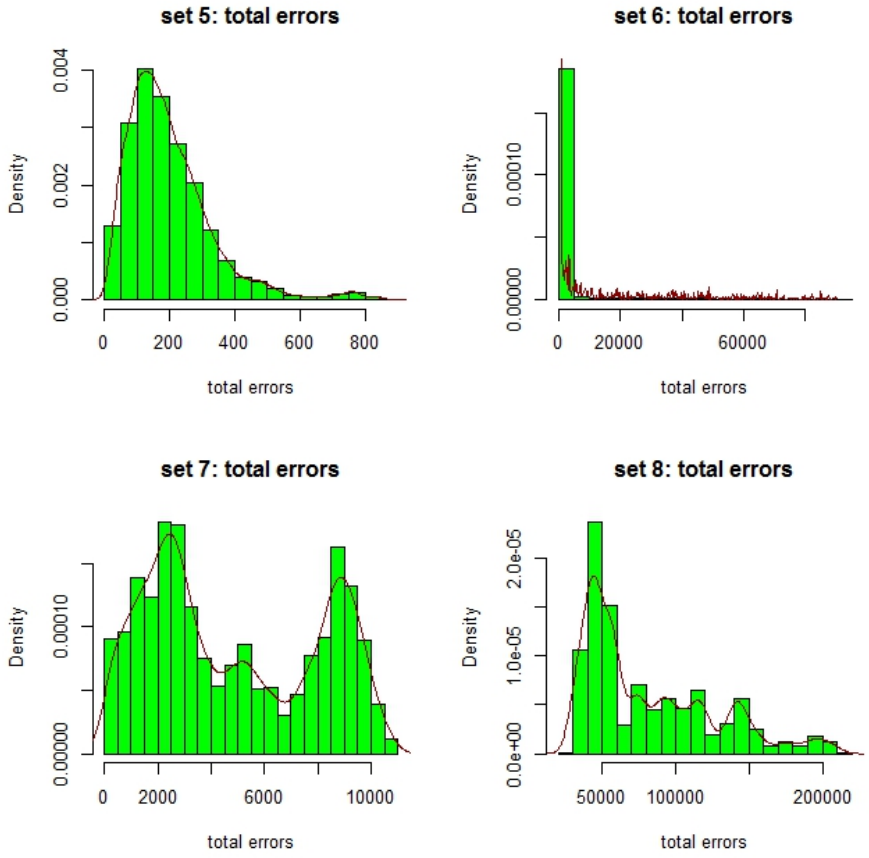


proposal (or prior specification) to be reflective of the empirical outcome. Thus far, we have tested the length of the MCMC chain and found that it has negligible effects on the results. Nonetheless, the raw results of Algorithm II are in the correct direction in the sense that the posterior results better inform an evaluation of the sample outcome, and hence prior beliefs can be updated accordingly. Additional work continues on the multivariate case, particularly for the sample size determination problem. Moreover, a benchmark against the multinomial (sampling with replacement) paradigm would serve as a meaningful comparison.

6 Discussion

We provide a Bayesian algorithmic approach to reconcile prior beliefs and sample outcomes when seeking to estimate the rate of errors in a given population. We illustrate this problem in the univariate case where a population element is either an error or not. Moreover, we have studied the multivariate problem where an error can have multiple shades – i.e., some errors are more egregious than others. Our approach provides

Figure 3: MCMC posterior output from numerical experiments 5 through 8 (univariate cases).



estimates of the *total* number of errors in a given population as opposed to the *rate* of errors that is commonly studied. Furthermore, we utilize the exact sampling paradigm under the (multivariate) Hyper-Geometric model, which assumes that sampling is done without replacement.

We believe that estimation of the total number of errors can be useful to policy-makers as it may help to prepare for certain operational contingencies. Furthermore, an understanding of the total number of errors will enhance potential future deep-dive studies about a given population, perhaps informing how to stratify such a population. Additionally, inference about the the total number of errors in a population, while computationally intensive, overcomes any limitations for the case where inference is made about the rate of errors in a population. Our analysis does not depend on large-sample theory or any assumption about the sampling mechanism, and hence it is robust as an analytical tool.

The current state of our results do point to the need for future research in some important areas, especially for the case of multiple shades of errors in a given population. For example, the effects of sample size needs to be addressed, and the stochastic search mechanism in Algorithm II ought to be made more efficient. Ultimately, we seek to provide a computational rubric that can help analysts draw a sample of a given

Table 2: Key posterior summaries from univariate analysis of errors. Note that there are $M = 20,000$ MCMC runs with a 5% burn-in.

Experiment	Min	1st quart.	Med.	Mean	Std. dev.	Mode	3rd quart.	Max.
set 1	4	33	52	58.85	34.31	35	79	231
set 2	4	36	54	61.06	33.50	46	79	241
set 3	26	209	303	330.90	174.27	234	415	1,119
set 4	14	185	282	315.60	179.96	222	411	1,148
set 5	2	108	172	198.50	131.85	130	257	882
set 6	5	112	198	2,873	11,390.13	107	381	90,190
set 7	60	2,131	3,978	4,823	3,157.11	2,449	8,216	10,930
set 8	29,280	44,730	60,510	81,200	44,410.12	44,242	112,500	211,700

size, and using that sample, obtain a reliable estimate of the total number of errors in a population, thereby updating any prior notions about the incidence of errors in the population.

Acknowledgments

I would like to respectfully acknowledge comments from colleagues at the Office of the Comptroller of the Currency. Additionally, this paper benefited from comments provided by participants at the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference in Washington, DC.

References

Adcock, C. (1987). A Bayesian approach to calculating sample sizes for multinomial sampling. *The Statistician 36*(2/3), 155–159.

Alam, K. and J. Thompson (1972). On selecting the least probable multinomial event. *The Annals of Mathematical Statistics 43*(6), 1981–1990.

Table 3: Setup of numerical experiments for multivariate errors.

Experiment	Pop. size	Sample size	Prior A	Prior B	Obs. A errors	Obs. B errors
set 9	100,000	600	1,000	500	200	20
set 10	100,000	1,500	1,000	500	500	50

Table 4: Key posterior summaries from multivariate analysis of errors. Note that the number of MCMC runs, $M = 20,000$, and there is a 5% burn-in.

Experiment	Min	1st quart.	Med.	Mean	Std. dev.	Mode	3rd quart.	Max.
9 (error A)	1,096	1,175	1,196	1,196	31.38	1,193	1,217	1,313
9 (error B)	440	502	517	517.40	21.96	519	532	602
10 (error A)	1,362	1,468	1,488	1,489	31.58	1,486	1,510	1,618
10 (error B)	464	529	545	544.80	22.92	539	560	619

Chen, H. (2001). Initialization for NORTA: generation of random vectors with specified marginals and correlations. *INFORMS Journal on Computing 13*(4), 312–331.

Cochran, W. G. (1977). *Sampling Techniques* (3 ed.). New York: John Wiley & Sons.

DeGroot, M. and M. Schervish (2002). *Probability and Statistics* (3 ed.). Boston: Addison-Wesley.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data* (2 ed.). Cambridge: M.I.T. Press.

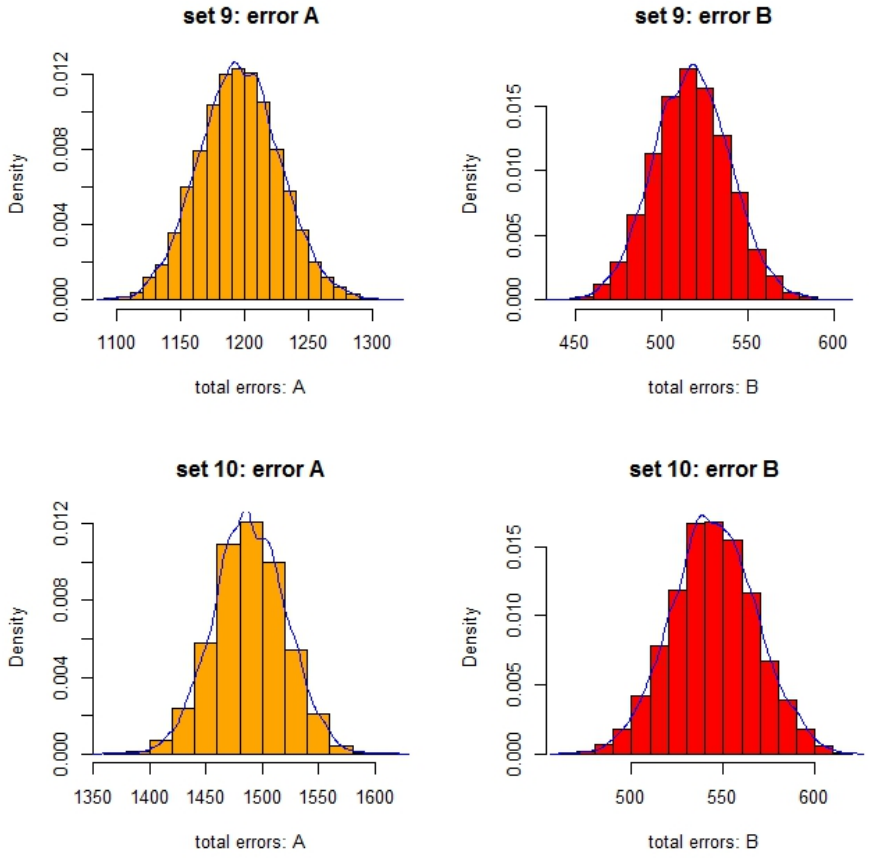
Gelman, A., J. Carlin, H. Stern, and D. Rubin (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Mavridis, D. and C. Aitken (2009). Sample size determination for categorical responses. *Journal of Forensic Science 54*(1), 135–151.

M'Lan, C., L. Joseph, and D. Wolfson (2008). Bayesian sample size determination for binomial proportions. *Bayesian Data Analysis 3*(2), 269–296.

Pham-Gia, T. and N. Turkkan (1992). Sample size determination in Bayesian analysis. *The Statistician 41*(4), 389–397.

Figure 4: MCMC posterior output from numerical experiments 9 through 10 (multivariate cases).



Thompson, S. (1987). Sample size for estimating multinomial proportions. *The American Statistician 44*(1), 42–46.

Thompson, S. K. (2012). *Sampling* (3 ed.). Hoboken: John Wiley & Sons.

Valliant, R., J. Dever, and F. Kreuter (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Yahav, I. and G. Shmueli (2012). On generating multivariate Poisson data in management science applications. *Applied Stochastic Models in Business and Industry 28*(1), 91–102.