STATISTICAL ESTIMATION OF UNDETECTED ERRORS

PATRICK BARDSLEY

Abstract. We summarize and rederive a Bayesian framework which quantifies uncertainty in detecting errors in a given data source (i.e., detecting number of faulty items within a population), originally derived in [3]. The number of errors within the data source are assumed to occur, *a priori*, as a Poisson random variable. A set of inspectors then proofreads and tabulates errors within the data source, *independently* and *in parallel*. Given these prior assumptions and error tabulations, we form Bayesian-posterior estimates for the error occurrence. We then extend these posterior estimates to the case where the underlying data source itself is only a representative sample of a larger population.

Contents

1. Introduction

A recurring problem in data science is that of estimating the number of errors or faults in a particular data source. For example, one may be concerned with data labeling errors and estimating the number of files that have been labeled or marked-up incorrectly. In this document, we summarize and adapt (and in some cases correct) a Bayesian framework developed by Jewell [3]. This Bayesian methodology allows us to quantify not only the number of errors in a data source, but also our uncertainty in forming such estimates. The general idea, is the target data source is proofread by several inspectors, in parallel. Upon collecting each inspector's error tabulations, and imposing prior beliefs (i.e, prior distributions) on the error occurrence rate in the data source, posterior distributions can then be computed.

The remainder of this document is organized as follows. In §2 we set up the model we utilize for proofreading, error occurrence, and error tabulation. Then in §3 we provide various estimates for the errors within the data source, from maximum likelihood estimates to posterior distribution estimates derived using particular assumptions. In §4 we extend the Bayesian estimates to the situation where only a subsample of the full underlying data source is proofread. We give a simple numerical example in §5 and finally provide concluding remarks in §6.

2. Modeling errors as Poisson random variables

The notations and models used hereafter are essentially a refactored form of those used by Jewell [3], though we do make some updates for correctness and clarity. We defer all interested readers to this original work, of which we are merely rederiving and summarizing key results.

2.1. **Error tabulations.** Assume the existence of an underlying data source that is to be proofread for errors. These data can be anything from a single document (i.e., collection of words/formulas), to a large collection of physical or virtual files. The data are to be proofread by I inspectors using a parallel search strategy, i.e., each inspector $i = 1, \ldots, I$, will proofread the same data *simultaneous to and independent of* the other inspectors. After each inspector is finished with their proofreading, they will have detected and tabulated errors so that we have the following summary statistics:

$$\widetilde{\mathcal{S}} := \{s_1, s_{12}, s_{123}, \ldots, s_i, s_{ij}, s_{ijk}, \ldots, s_{12\cdots I}\},$$

where s_i denotes the number of errors found only by inspector i, s_{ij} denotes the number of common errors found by inspectors i and j, s_{ijk} denotes the number of common errors found by inspectors i, j and k, etc. Note, as defined above, these statistics carry redundancies in that $s_{i_1 i_2 \cdots i_n} = s_{\pi(i_1, i_2, \cdots, i_n)}$ where π is any permutation mapping. Thus, we further define $\mathcal{S}$ as the set of *unique* summary statistics and identify permutation elements with one representative element (e.g., $s_{132} \equiv s_{123}$):

$$\mathcal{S} = \widetilde{\mathcal{S}}/\!\sim.$$

The total number of errors found by inspector i can be expressed as

$$s(i) = \sum_{s_\alpha \in \mathcal{S}:\ i \in \alpha} s_\alpha,$$

where $\alpha = (j_1, j_2, \cdots, j_n)$ denotes a multi-index and in a slight abuse of notation we assume $i \in \alpha \iff j_k = i$ for some $k = 1, \ldots, n$. Similarly, the total number of *unique* errors found by all inspectors can be found by summing all of the elements in $\mathcal{S}$:

$$s_T = \sum_{s \in \mathcal{S}} s.$$

Given a total number of errors N within the data source, we can also define the number of *undetected* errors as

$$s_0 = N - s_T.$$

2.2. **Error detection probability.** We now impose assumptions on the probability of an inspector detecting an error. Namely, we assume p_i is the probability that inspector i will detect a given error, and further assume this probability is independent of previously detected errors by any inspector (including inspector i). With this assumption in place, it is straightforward to express the probability of error detections, jointly for *detected* and *undetected* errors, as a multinomial distribution (see, e.g., [2, 4]):

$$\begin{aligned} p(s_0, \mathcal{S}|N, \mathbf{p}) = \delta(s_0 + s_T - N)\binom{N}{s_0\ \ \mathcal{S}} &((1-p_1)(1-p_2)\cdots(1-p_I))^{s_0} \times \\ &(p_1(1-p_2)\cdots(1-p_I))^{s_1} \times \\ &(p_1p_2(1-p_3)\cdots(1-p_I))^{s_{12}} \times \\ &\cdots \\ &(p_1p_2\cdots p_I)^{s_{12\cdots I}} \\ = \delta(s_0 + s_T - N)\binom{N}{s_0\ \ \mathcal{S}} &Q^{s_0+s_T}\prod_{i=1}^{I}\left(\frac{p_i}{q_i}\right)^{s(i)}, \end{aligned}$$

where we have implicitly defined the miss-probabilities as

$$q_i := 1 - p_i$$

and

$$Q := \prod_{i=1}^{I} q_i.$$

Note, the appearance of the term $\delta(s_0 + s_T - N)$ is to explicitly show $p(s_0, \mathcal{S}|N, \mathbf{p})$ is supported (i.e., has non-zero mass) only for consistent values of s_0, s_T and N.

Since the number of errors in the data source, N, is an unknown and essentially random quantity for different data sources, we model it using a Poisson distribution with average number of errors λ:

$$p(N|\lambda) = \frac{\lambda^N e^{-\lambda}}{N!}, \quad N \in \mathbb{N}. \tag{1}$$

It follows then, that the probability of error detections given the parameter λ is

$$\begin{aligned} p(s_0, \mathcal{S}|\lambda, \mathbf{p}) &= \sum_N p(s_0, \mathcal{S}, N|\lambda, \mathbf{p}) \\ &= \sum_N p(s_0, \mathcal{S}|N, \lambda, \mathbf{p})p(N|\lambda, \mathbf{p}) \\ &= \sum_N \delta(s_0 + s_T - N)\binom{N}{s_0\ \ \mathcal{S}} Q^{s_0+s_T}\prod_{i=1}^{I}\left(\frac{p_i}{q_i}\right)^{s(i)} \frac{\lambda^N e^{-\lambda}}{N!} \qquad (2) \\ &= \binom{s_0 + s_T}{s_0\ \ \mathcal{S}} Q^{s_0+s_T}\prod_{i=1}^{I}\left(\frac{p_i}{q_i}\right)^{s(i)} \frac{\lambda^{(s_0+s_T)} e^{-\lambda}}{(s_0+s_T)!} \\ &= \binom{1}{\mathcal{S}} \frac{(\lambda Q)^{s_0} e^{-\lambda}\lambda^{s_T}}{s_0!}\prod_{i=1}^{I} p_i^{s(i)} q_i^{s_T - s(i)}. \end{aligned}$$

3. Estimating the number of undetected errors

By leveraging expression (2), we can derive various formulas for estimating

- the data source error parameter λ,
- the set of inspector error-detection probabilities $\mathbf{p} = [p_1, p_2, \ldots, p_I]$,
- the number of undetected errors s_0.

In §3.1, we derive maximum likelihood estimators for these quantities, which can be immediately related to the so-called Petersen-Chapman-Darroch estimators [7, 5]. Then, in §3.2, following [3], we derive Bayesian estimators for the distributions of these quantities. These estimates follow a conjugate-prior derivation with natural prior distributions assumed for the various quantities.

3.1. **Maximum likelihood estimators.** Since s_0 is an unknown and unmeasurable quantity (whereas we can explicitly measure the set of statistics $\mathcal{S}$), we integrate this quantity out of (2):

$$\begin{aligned} p(\mathcal{S}|\lambda, \mathbf{p}) &= \sum_{s_0 \in \mathbb{N}} p(s_0, \mathcal{S}|\lambda, \mathbf{p}) \\ &= \sum_{s_0 \in \mathbb{N}} \binom{1}{\mathcal{S}} \frac{(\lambda Q)^{s_0} e^{-\lambda} \lambda^{s_T}}{s_0!} \prod_{i=1}^{I} p_i^{s(i)} q_i^{s_T - s(i)} \\ &= \binom{1}{\mathcal{S}} e^{-\lambda(1-Q)} \lambda^{s_T} \prod_{i=1}^{I} p_i^{s(i)} q_i^{s_T - s(i)}. \end{aligned} \tag{3}$$

It follows then, given a set of *detected* error statistics $\mathcal{S}$, maximum likelihood estimators (MLEs) for λ and $\mathbf{p}$ satisfy the nonlinear relationships

$$\begin{aligned} \widehat{\lambda} &= s_T + \widehat{\lambda} \prod_{i=1}^{I} \left(1 - \frac{s(i)}{\widehat{\lambda}}\right) \quad \text{and} \\ \widehat{p_i} &= \frac{s(i)}{\widehat{\lambda}} \quad \text{for } i = 1, \ldots, I. \end{aligned} \tag{4}$$

Aside from the relatively simple expression one can obtain for $I = 2$ (see [5]), iterative procedures exist to estimate these values (see, e.g., [7]).

We note that upon estimating $\widehat{\lambda}$ using these MLEs and a set of sample statistics $\mathcal{S}$, one has obtained an estimate for the average number of errors occurring in the underlying data source. More precisely, since we have assumed errors occur according to a Poisson distribution with parameter λ, we can use either the mean or mode of the Poisson distribution to estimate the number of errors in the data source:

- Mean: $\widehat{N} = \widehat{\lambda}$
- Mode: $\widehat{N} = \lfloor \widehat{\lambda} \rfloor$

Estimating the number of undetected errors, s_0, is then a trivial computation of $\widehat{s_0} = \widehat{N} - s_T$.

3.2. **Bayesian estimators.** Rather than relying only on the likelihood function and point estimates for λ and $\mathbf{p}$, we now extend our estimators to a Bayesian framework [6] by assuming a prior distribution on both λ and $\mathbf{p}$. Note, the MLEs derived in (4) can also be placed in this framework by simply assuming a uniform prior distribution on λ and $\mathbf{p}$. Extending to the Bayesian framework, will give us

more flexibility in how we estimate the number of undetected errors. In particular, obtaining a distributional estimate of s_0 will allow us to use credibility regions to give a confident upper bound estimate, along with other measures to quantify our uncertainty in measuring s_0.

From Bayes theorem [6, 8], the posterior distribution of the parameters is given by

$$p(\lambda, \mathbf{p}|\mathcal{S}) = \frac{p(\mathcal{S}|\lambda, \mathbf{p})p(\lambda, \mathbf{p})}{p(\mathcal{S})}.$$

The probability mass function $p(\mathcal{S}|\lambda, \mathbf{p})$ is given by (3). Thus it remains to specify the prior distribution $p(\lambda, \mathbf{p})$ and compute the total probability $p(\mathcal{S})$ via

$$p(\mathcal{S}) = \int_{\mathbb{R}^+ \times [0,1]^I} p(\mathcal{S}|\lambda, \mathbf{p})p(\lambda, \mathbf{p})d\lambda d\mathbf{p}.$$

Fortunately, for some choices of prior distributions $p(\lambda, \mathbf{p})$, the posterior distribution form is known exactly. Thus, with particular selections of this prior, we can simply use a conjugate-prior lookup table (see, e.g., [1]) to derive the posterior distribution, saving us tedious computation. Finally, we can use the posterior distribution for the parameters, to determine the posterior-predictive distribution for the number of undetected errors:

$$p(s_0|\mathcal{S}) = \int_{\mathbb{R}^+ \times [0,1]^I} p(s_0|\mathcal{S}, \lambda, \mathbf{p})p(\lambda, \mathbf{p}|\mathcal{S})d\lambda d\mathbf{p},$$

where

$$p(s_0|\mathcal{S}, \lambda, \mathbf{p}) = \frac{p(s_0, \mathcal{S}|\lambda, \mathbf{p})}{p(\mathcal{S}|\lambda, \mathbf{p})} = \frac{(\lambda Q)^{s_0} e^{-\lambda Q}}{s_0!},$$

i.e., $s_0|\mathcal{S}, \lambda, \mathbf{p} \sim \text{Poisson}(\lambda Q)$.

In the following subsections, we specify various prior distributions and their resulting posterior distributions. We proceed in the same manner as Jewell [3]. In §3.2.1, we assume detection probabilities $\mathbf{p}$ are known exactly while the data-error parameter λ is unknown. Then in §3.2.2 we reverse this by assuming the parameter λ is known exactly while detection probabilities $\mathbf{p}$ must be estimated. Finally, in §3.2.3 we give the general result when both λ and $\mathbf{p}$ must be estimated.

3.2.1. *Known detection probabilities, unknown error rate.* If the error detection probabilities $\mathbf{p} \equiv \mathbf{p}_*$ are known exactly, we have $p(\lambda) \equiv p(\lambda, \mathbf{p}) = p(\lambda)\delta(\mathbf{p} - \mathbf{p}_*)$.

- **Prior:**
$$\lambda \sim \text{Gamma}(a, b)$$
$$\iff p(\lambda|a, b) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}$$
- **Likelihood:**
$$p(s_T|\lambda, \mathbf{p}) = \sum_{\mathcal{S}' : \left(\sum_{s \in \mathcal{S}'} s\right) = s_T} p(s_0, \mathcal{S}'|\lambda, \mathbf{p}) = C(s_T)e^{-\lambda(1-Q)}\lambda^{s_T}$$
- **Posterior:**
$$\lambda|s_T, \mathbf{p} \sim \text{Gamma}(a + s_T, b + 1 - Q)$$
$$\iff p(\lambda|s_T, p) = \frac{(b + 1 - Q)^{a+s_T} \lambda^{a+s_T-1} e^{-\lambda(b+1-Q)}}{\Gamma(a + s_T)}$$

- **Posterior-predictive:**

$$s_0|\mathcal{S},\mathbf{p} \sim \text{NegBinom}(a+s_T, Q/(b+1)))$$
$$\iff p(s_0|\mathcal{S},\mathbf{p}) = \binom{s_0+a+s_T-1}{a+s_T-1}(1-Q/(b+1))^{s_0}(Q/(b+1))^{a+s_T}$$

Here the posterior and posterior-predictive for s_0 are found by direct computation. Quantities which are specified in terms of total detected errors, s_T, are computed by marginalizing over error detection partitions, $\mathcal{S}$, with fixed total sum s_T. The Gamma hyperparameters a and b are selected to give a desired shape and scale for the λ prior distribution (e.g., a fixed mean and variance).

3.2.2. *Unknown detection probabilities, known error rate.* If the error rate $\lambda \equiv \lambda_*$ is known exactly, we have $p(\mathbf{p}) \equiv p(\lambda,\mathbf{p}) = p(\mathbf{p})\delta(\lambda-\lambda_*)$.

- **Prior:**

$$p_i \text{ independently distributed with } p_i \sim \text{Beta}(\alpha_i,\beta_i)$$
$$\iff p(p_i|\alpha_i,\beta_i) = \frac{\Gamma(\gamma_i)p_i^{\alpha_i-1}q_i^{\beta_i-1}}{\Gamma(\alpha_i)\Gamma(\beta_i)}$$

  where $q_i = 1-p_i, \gamma_i = \alpha_i+\beta_i$

- **Likelihood:**

$$p(\mathcal{S}|\lambda,\mathbf{p}) = \binom{1}{\mathcal{S}}e^{-\lambda(1-Q)}\lambda^{s_T}\prod_{i=1}^{I}p_i^{s(i)}q_i^{s_T-s(i)}$$

- **Posterior:**

$$p(\mathbf{p}|\mathcal{S},\lambda) = \left(\sum_k c_k\right)^{-1}\sum_j c_j\prod_{i=1}^{I}p(p_i|\alpha_i+s(i),\beta_i+s_T-s(i)+j)$$

  where

$$c_j = \frac{\lambda^j}{j!}\prod_{i=1}^{I}\frac{\Gamma(\beta_i+s_T-s(i)+j)\Gamma(\gamma_i+s_T)}{\Gamma(\beta_i+s_T-s(i))\Gamma(\gamma_i+s_T+j)}$$

- **Posterior-predictive:**

$$p(s_0|\mathcal{S},\lambda) = p(0|\mathcal{S},\lambda)\left(\frac{\lambda^{s_0}}{s_0!}\right)\prod_{i=1}^{I}\frac{\Gamma(\beta_i+s_T-s(i)+s_0)\Gamma(\gamma_i+s_T)}{\Gamma(\beta_i+s_T-s(i))\Gamma(\gamma_i+s_T+s_0)}$$

Clearly, the posterior and posterior-predictive distributions are quite complicated, with the former only expressed as an infinite series. This arises from a coupling term $e^{\lambda Q}$ present in the likelihood expression, followed by a series expansion of this exponential. However, relatively simple methods exist to estimate these distributions. For the posterior distribution $p(\mathbf{p}|\mathcal{S},\lambda)$, one can simply compute the values of c_j until they become negligible. The distribution can then be estimated using these terms and the corresponding truncated series. For the posterior-predictive distribution, we first appeal to a recursion formula it satisfies:

$$p(s_0+1|\mathcal{S},\lambda) = p(s_0|\mathcal{S},\lambda)\prod_{i=1}^{I}f_{q_i}\left(1-\frac{s(i)}{s_T+s_0};s_T+s_0,\gamma_i\right),$$

where

$$f_y\left(\hat{y};\mu,\nu\right) = \frac{\mu}{\mu+\nu}\hat{y}+\frac{\nu}{\mu+\nu}E(y)$$

is a weighted *credibility* formula for the estimator $\hat{y}$ given its prior $E(y)$ and credible weights μ and ν. By setting $p(0|\mathcal{S},\lambda) \equiv 1$, one can compute probabilities for other values of s_0 until they become negligible, relative to $p(0|\mathcal{S},\lambda)$, and finally renormalize.

3.2.3. *Unknown detection probabilities, unknown error rate.* Here we provide the general formulas for the case when both $\mathbf{p}$ and λ must be estimated. These results follow from analogous arguments as those given in §3.2.1 and §3.2.2.

- **Prior:**

$$\lambda, \mathbf{p} \sim \text{Gamma}(a,b)\prod_{i=1}^{I}\text{Beta}(\alpha_i,\beta_i)$$

$$\iff p(\lambda,\mathbf{p}|a,b,\{\alpha_i,\beta_i\}) = \frac{b^a\lambda^{a-1}e^{-b\lambda}}{\Gamma(a)}\prod_{i=1}^{I}\frac{\Gamma(\alpha_i+\beta_i)p_i^{\alpha_i-1}q_i^{\beta_i-1}}{\Gamma(\alpha_i)\Gamma(\beta_i)}$$

- **Likelihood:**

$$p(\mathcal{S}|\lambda,\mathbf{p}) = \binom{1}{\mathcal{S}}e^{-\lambda(1-Q)}\lambda^{s_T}\prod_{i=1}^{I}p_i^{s(i)}q_i^{s_T-s(i)}$$

- **Posterior:**

$$p(\lambda,\mathbf{p}|\mathcal{S}) = \left(\sum_k d_k\right)^{-1}\sum_j d_j p(\lambda|a+s_T+j,b+1)\times \prod_{i=1}^{I}p(\mathbf{p}|\{\alpha_i+s(i),\beta_i+s_T-s(i)+j\}) \tag{5}$$

where

$$d_j(\mathcal{S}) = \frac{(b+1)^{-j}\Gamma(a+s_T+j)}{j!\Gamma(a+s_T)}\prod_{i=1}^{I}\frac{\Gamma(\beta_i+s_T-s(i)+j)\Gamma(\gamma_i+s_T)}{\Gamma(\beta_i+s_T-s(i))\Gamma(\alpha_i+\beta_i+s_T+j)}$$

- **Posterior-predictive:**

$$p(s_0|\mathcal{S}) = \frac{d_{s_0}(\mathcal{S})}{\sum_k d_k(\mathcal{S})} \tag{6}$$

Again, the posterior distributions are complicated and expressed via infinite series. However, similar to the methods discussed in §3.2.2, we can estimate these distributions by truncating the series expressions. That is, we compute the coefficients d_j until they become negligible, after which $p(\lambda,\mathbf{p}|\mathcal{S})$ or $p(s_0|\mathcal{S})$ can be estimated through appropriate normalization (i.e., dividing by the truncated series).

Remark 1 (Gamma quotients)**.** *Numerically estimating the posterior distributions* (5) *and* (6) *involves computing the terms d_j for sufficiently many terms j until they become negligible. Naively computing d_j can result in numerical overflow issues for even moderate values of $s_T, s(i)$ and j, due to the form of the d_j coefficients and more precisely their quotients of gamma functions (recall the gamma function is an analytic generalization of the factorial function satisfying $\Gamma(n) = (n-1)!$ for*

$n \in \mathbb{N}$). We can overcome this issue by recursively computing the d_j coefficients as

$$d_0 = 1,$$
$$d_j = d_{j-1} \frac{(b+1)^{-1}(a+s_T+j-1)}{j} \prod_{i=1}^{I} \frac{\beta_i + s_T - s(i) + j - 1}{\alpha_i + \beta_i + s_T + j - 1} \quad \textit{for } j \geq 1, \tag{7}$$

which follows immediately from the gamma function property $\Gamma(z+1) = z\Gamma(z)$.

4. Extension to subsample proofreadings

Assume we are given a data set with set of M objects. We then let each one of I inspectors proofread this data source, independent of one another, looking for the number of objects which contain errors. If an inspector finds an error, they simply record which object contained the error, append this to a list, and move on to the next object. After this proofreading process, we will have generated the error statistics:

$$\{s(1), s(2), \ldots, s(I), s_T\}.$$

From this empirical sample of statistics, we can generate the joint posterior distribution $p(\lambda, \mathbf{p}|\mathcal{S})$, given by (5). The posterior-predictive mass function is then computed via

$$p(s_0|\mathcal{S}) = \int_{\mathbb{R}^+ \times [0,1]^I} p(s_0|\mathcal{S}, \lambda, \mathbf{p}) p(\lambda, \mathbf{p}|\mathcal{S}) d\lambda d\mathbf{p}, \tag{8}$$

and can be shown by direct computation (and crucially utilizing the Taylor expansion of $e^{\lambda Q}$) to simplify to (6). It follows then that

$$p(N|\mathcal{S}) = p(s_0 = N - s_T|\mathcal{S}) = \begin{cases} 0, & N < s_T, \\ \frac{d_{N-s_T}(\mathcal{S})}{\sum_k d_k(\mathcal{S})}, & N \geq s_T. \end{cases}$$

From this mass function, it is then possible to form different estimates of N (e.g., a 95% credible interval for N).

Now, assume we are only given a *representative* subsample of $M_s < M$ from the full set of objects. This subsample should be representative in the sense that important metadata compositions and proportions are preserved in the sampling process. We can perform the same proofreading process as before, but upon forming our posterior estimate of λ, we note this value now relates to average number of errors occurring in a data source of M_s objects, not the original M objects as required.

Ideally, we would adjust for the smaller sample size M_s and estimate the number of errors occurring in the full set of M objects, by computing a dilated predictive posterior via

$$p_\mu(s_0|\mathcal{S}) := \frac{1}{\mu} \int_{\mathbb{R}^+ \times [0,1]^I} p(s_0|\mathcal{S}, \lambda, \mathbf{p}) p\left(\frac{\lambda}{\mu}, \mathbf{p}|\mathcal{S}\right) d\lambda d\mathbf{p}. \tag{9}$$

However, there is no closed form expression for $p_\mu(s_0|\mathcal{S})$, except for relatively small values of μ which is not our intended use-case. This follows from the dilation in the posterior density ensuring there is no longer a cancellation of $e^{\lambda Q}$ terms, which keeps the $\mathbf{p}$ and λ integrals intertwined (see also Remark 2).

As a workaround, we suggest a dilated and linearly interpolated probability mass function for the posterior predictive:

$$p_\mu(s_0|\mathcal{S}) \propto p(\lfloor \mu^{-1} s_0 \rfloor|\mathcal{S})(1 - z(s_0)) + p(\lceil \mu^{-1} s_0 \rceil|\mathcal{S}) z(s_0), \tag{10}$$

where

$$z(s_0) = \frac{s_0 - \lfloor \mu^{-1} s_0 \rfloor}{\lceil \mu^{-1} s_0 \rceil - \lfloor \mu^{-1} s_0 \rfloor}$$

satisfies $z(\lfloor \mu^{-1} s_0 \rfloor) = 0, z(\lceil \mu^{-1} s_0 \rceil) = 1$, and $z(s) \in [0, 1]$. Here, the symbol $\propto$ is to indicate the mass function must be normalized to account for both the dilation parameter μ, as well as the linear interpolation. Finally, we obtain an estimate for the number of errors in the full data source of M objects using the dilated mass function:

$$p_\mu(N|\mathcal{S}) := \begin{cases} 0, & N < \mu^{-1} s_T, \\ p_\mu(s_0 = N - \mu^{-1} s_T|\mathcal{S}) & N \geq \mu^{-1} s_T. \end{cases} \tag{11}$$

Note, while this expression looks complicated, it is nothing more than a rightward shift of $p_\mu(s_0|\mathcal{S})$ by a value of $\mu^{-1} s_T$.

Remark 2 (Sub-sample dilation)**.** *Note, the ideal dilation we propose in* (9) *essentially assumes the same Poisson prior for N (and thus $s_0|\mathcal{S}$ by virtue of $N = s_T + s_0$), while assuming our prior for λ must first be dilated to the appropriate scale to account for data size mismatches. Indeed a closed form solution can be found for small values of μ using a series expansion (see Appendix A). However, we anticipate practical values of μ will be significantly larger, for which the series expansion no longer converges.*

5. Numerical Example

Here we present a simple numerical example of this Bayesian estimation procedure. For this example, we assume only $I = 2$ inspectors and work with a data source of $M = 500$ objects. We generate synthetic error tabulations from each inspector in the following manner. Letting the underlying error rate $\lambda = 0.2$, we randomly select $M\lambda = 100$ of the integers in $[1, 500]$ as the labels of objects which contain errors. Then, for each inspector, we randomly assign a detection label ($0 =$ undetected, $1 =$ detected) to each of these objects with probability of detection 80%. Finally, the set of error labels that are detected by inspector $i = 1, 2$ are used to compute $\mathcal{S} = \{s_1, s_2, s_{12}\}$.

From the tabulated errors, we make a very crude estimate (recall only two inspectors are used) of the mean (μ) and variance (σ^2) of the number of errors in the 500 object data source. These estimates serve only to set the hyperparameters a and b for our Gamma distribution prior for λ:

$$a = \mu^2/\sigma^2, \quad b = \mu/\sigma^2.$$

The resulting gamma density is shown in Figure 1. Next, for simplicity we assume, *a priori*, each inspector detects errors (independently) with probability p distributed as a beta random variable, with shape parameters

$$\alpha = 5, \quad \beta = 5/9.$$

These shape parameters are admittedly somewhat arbitrary, but chosen to produce a particular density shape and a mean detection probability of 90%. This beta distribution density is depicted in Figure 2.

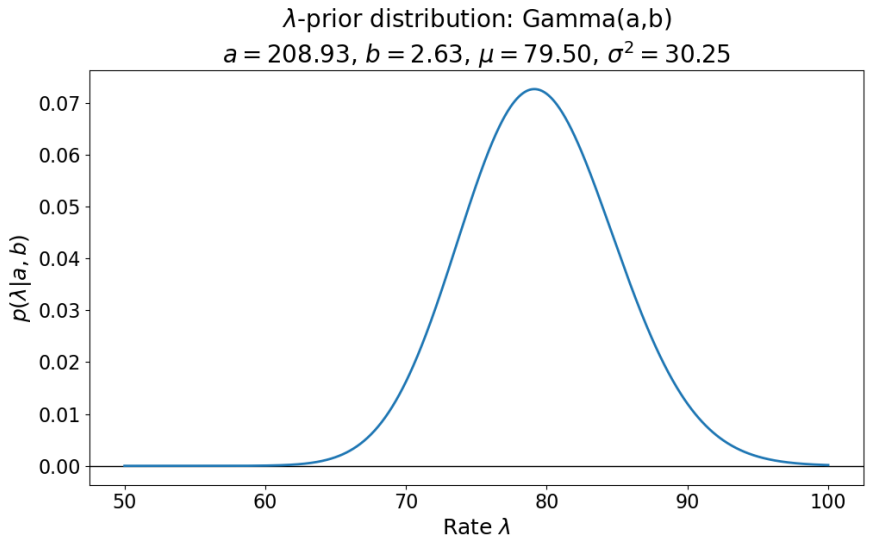

FIGURE 1. The probability density $p(\lambda|a,b)$ of the Gamma(a,b) distribution, used as the prior distribution for the mean error rate λ.

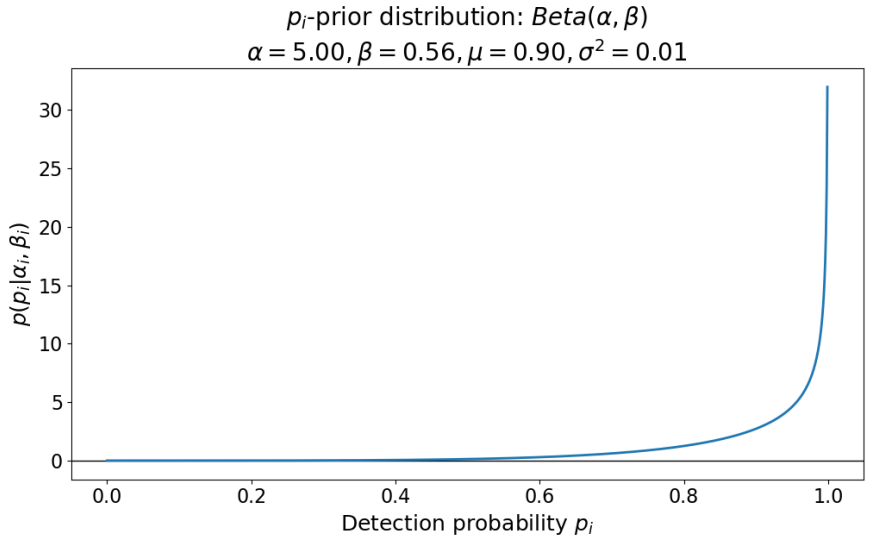

FIGURE 2. The probability density $p(p_i|\alpha_i,\beta_i)$ of the Beta(α_i,β_i) distribution, used as the prior distribution for the error detection rate p_i for each inspector $i = 1, 2$.

We compute the λ-posterior distribution by integrating (5) over all values of $\mathbf{p}$:

$$p(\lambda|\mathcal{S}) = \left(\sum_k d_k\right)^{-1} \sum_j d_j p(\lambda|a + s_T + j, b + 1).$$

Here we make use of Remark 1 to efficiently compute the coefficients d_j. The resulting posterior density is shown in Figure 3. Then in Figure 4, we show the posterior predictive probability mass function (6) for the number of undetected errors, i.e., $p(s_0|\mathcal{S})$.

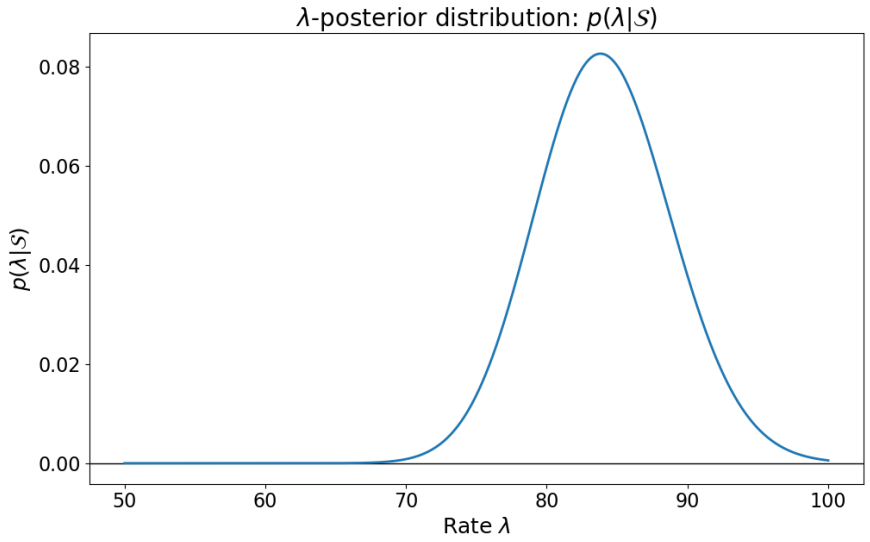

FIGURE 3. The posterior probability density $p(\lambda|\mathcal{S})$ obtained from preceding prior distribution assumptions, and the error tabulations $\mathcal{S}$ for inspectors $i = 1, 2$.

Lastly, we demonstrate our subsampling and probability dilation procedure. Assuming our data source is a representative subsample of a larger data source, we compute the μ-dilated posterior predictive function for the *total* number of errors, namely, $p_\mu(N|\mathcal{S})$ defined in (11). Recall the dilation parameter μ essentially scales our error estimates by a specified ratio to account for total data set and sample data set size differences. To compute this function, we utilize the linear interpolated and dilated function $p_\mu(s_0|\mathcal{S})$ given in (10), which in turn leverages (6) and Remark 1. We show this dilated probability mass function for $\mu = 50$, along with a 95% highest posterior density interval (i.e., a 95% credibility interval) in Figure 5. Here the credible interval is $[4700, 4977]$.

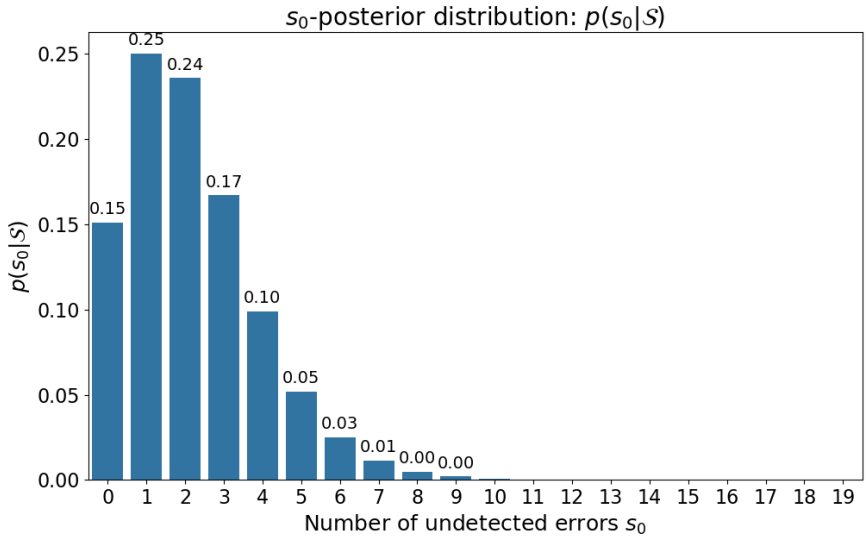

FIGURE 4. The posterior predictive probability mass function $p(s_0|\mathcal{S})$ for the number of unknown errors s_0, obtained using formula (6) where the coefficients d_j are computed as in Remark 1.

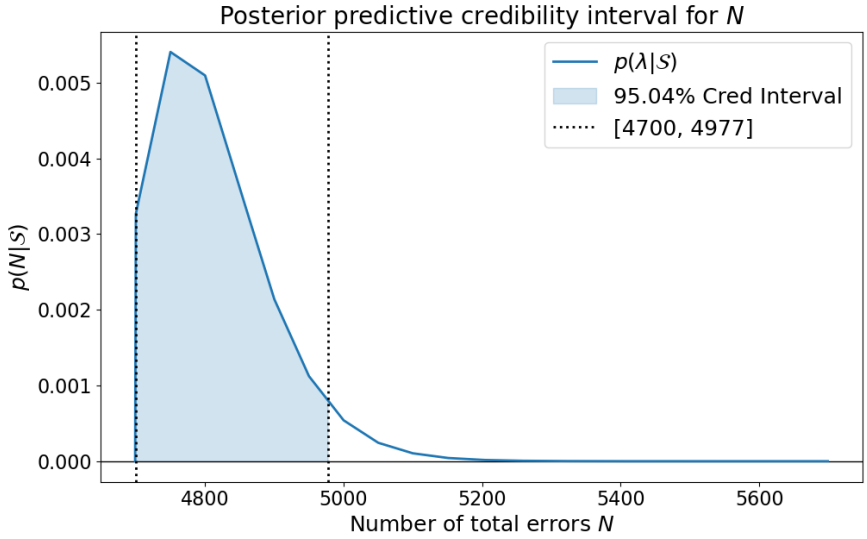

FIGURE 5. The dilated posterior predictive probability mass function $p_\mu(N|\mathcal{S})$ for the number of total errors N. Here we have utilized formula (11) for a dilation factor $\mu = 50$. The 95% highest posterior density interval $[4700, 4977]$ is depicted by the shaded region.

6. Concluding Remarks

We have rederived (and in some cases corrected) the original Bayesian framework developed in [3] for estimating the number of errors in a data source. These Bayesian estimates rely on

- an independent and parallel proofreading strategy,
- assuming a known prior distribution for error occurrence,
- assuming independent probability of error detection per inspector,
- inspectors proofread either an entire data source or a stratified and representative sample of a larger data source.

In future work, we would like to address the shortcomings of this framework that arise from the last two assumptions. That is, a more comprehensive estimation procedure should account for a case when only a subset of the entire data source can be feasibly proofread. This implies certain errors remain undetected with probability 1 as they are omitted from the proofreading process entirely. It is possible that allowing for such a form of error-occurrence prior and likelihood would still lead to closed form expressions for the posterior distributions, without the overly restrictive assumption of representative data source subsamples.

Appendix A. Computing dilated posterior for small μ

We denote

$$\begin{aligned} q_i &= 1 - p_i \quad \text{for } i = 1, \ldots, I, \\ Q &= \prod_{i=1}^{I} q_i, \\ a' &= a + s_T, \\ b' &= b + 1, \\ \alpha_i' &= \alpha_i + s(i), \\ \beta_i' &= \beta_i + s_T - s(i), \\ \gamma_i' &= \alpha_i + \beta_i + s_T, \end{aligned}$$

and compute the posterior predictive density (9) as

$$\begin{aligned}
p_\mu(s_0|\mathcal{S}) &= \int p(s_0|\mathcal{S},\lambda,\mathbf{p})p_\mu(\lambda,\mathbf{p}|\mathcal{S})d\lambda d\mathbf{p} \\
&= \mu^{-1}\int \sum_j \frac{(\lambda Q)^{s_0}e^{-\lambda Q}}{s_0!}\frac{(b')^{-j}\Gamma(a'+j)}{j!\Gamma(a')}\frac{(b')^{a'+j}\left(\mu^{-1}\lambda\right)^{a'+j-1}e^{-\lambda b'/\mu}}{\Gamma(a'+j)}\times \\
&\quad \prod_{i=1}^I \frac{\Gamma(\beta_i'+j)\Gamma(\gamma_i')}{\Gamma(\beta_i')\Gamma(\gamma_i'+j)}\frac{\Gamma(\gamma_i'+j)p_i^{\alpha_i'-1}q_i^{\beta_i'+j-1}}{\Gamma(\alpha_i')\Gamma(\beta_i'+j)}d\lambda d\mathbf{p} \\
&= \int \left(\frac{b'}{\mu}\right)^{a'}\frac{1}{s_0!\Gamma(a')}\left(\prod_{i=1}^I \frac{\Gamma(\gamma_i')p_i^{\alpha_i'-1}q_i^{\beta_i'+s_0-1}}{\Gamma(\alpha_i')\Gamma(\beta_i')}\right)\times \\
&\quad e^{-\lambda(b'/\mu+Q)}\lambda^{a'+s_0-1}\sum_j \frac{\lambda^j Q^j}{j!\mu^j}d\lambda d\mathbf{p} \\
&= \int \left(\frac{b'}{\mu}\right)^{a'}\frac{1}{s_0!\Gamma(a')}\left(\prod_{i=1}^I \frac{\Gamma(\gamma_i')p_i^{\alpha_i'-1}q_i^{\beta_i'+s_0-1}}{\Gamma(\alpha_i')\Gamma(\beta_i')}\right)d\mathbf{p}\times \\
&\quad \int \lambda^{a'+s_0-1}e^{-\lambda(b'/\mu+Q-Q/\mu)}d\lambda.
\end{aligned}$$

This last expression is a point of divergence in our derivation. The inner integral over λ can be evaluated in terms of the Gamma function $\Gamma(a'+s_0)$, but this leaves us with an unfortunate coupling of all of the $\mathbf{p}$ integrals from a term $(b'/\mu+Q(1-\mu^{-1}))^{-a'-s_0}$. Thus, for general μ, we are unable to find a closed form expression for $p_\mu(s_0|\mathcal{S})$.

However, for smaller values of μ, we can pursue a series expansion expression as follows:

$$\begin{aligned}
p_\mu(s_0|\mathcal{S}) &= \int \left(\frac{b'}{\mu}\right)^{a'}\frac{1}{s_0!\Gamma(a')}\left(\prod_{i=1}^I \frac{\Gamma(\gamma_i')p_i^{\alpha_i'-1}q_i^{\beta_i'+s_0-1}}{\Gamma(\alpha_i')\Gamma(\beta_i')}\right)d\mathbf{p}\times \\
&\quad \sum_j \frac{(-1)^j Q^j(1-\mu^{-1})^j}{j!}\left(\frac{\mu}{b'}\right)^{a'+s_0+j}\Gamma(a'+s_0+j) \\
&= \sum_j \frac{(-1)^j}{j!s_0!}\left(\frac{\mu}{b'}\right)^{s_0}\left(\frac{\mu-1}{b'}\right)^j\frac{\Gamma(a'+s_0+j)}{\Gamma(a')}\times \\
&\quad \prod_{i=1}^I \frac{\Gamma(\gamma_i')}{\Gamma(\alpha_i')\Gamma(\beta_i')}\int p_i^{\alpha_i'-1}(1-p_i)^{\beta_i'+s_0+j-1}dp_i \\
&= \sum_j c_j(s_0)\left(\frac{\mu}{b'}\right)^{s_0}\left(\frac{\mu-1}{b'}\right)^j,
\end{aligned}$$

where

$$c_j(s_0) = \frac{(-1)^j}{j!s_0!}\frac{\Gamma(a'+s_0+j)}{\Gamma(a')}\prod_{i=1}^I \frac{\Gamma(\gamma_i')\Gamma(\beta_i'+s_0+j)}{\Gamma(\beta_i')\Gamma(\gamma_i'+s_0+j)}.$$

Requiring convergence of this series expression and finite total mass (i.e, $\sum_{s_0} p_\mu(s_0|\mathcal{S}) = 1$) imposes a smallness restriction on μ of

$$\mu < b' = b + 1.$$

References

[1] Daniel Fink. A compendium of conjugate priors. 46, 1997. URL http://www.people.cornell.edu/pages/df36/CONJINTRnew%20TEX.pdf.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[3] William S Jewell. Bayesian estimation of undetected errors. Technical report, California University Berkeley Operations Research Center, 1983.

[4] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[5] George Polya. Probabilities in proofreading. *Am. Math. Monthly*, 83(1):42, 1976.

[6] H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Harvard Business School Publications. Division of Research, Graduate School of Business Adminitration, Harvard University, 1961. ISBN 9780875840178.

[7] G.A.F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Charles Griffin, 1982. ISBN 9780852642627.

[8] Alan Stuart. Kendall's advanced theory of statistics. *Distribution theory*, 1, 1994.