Interval Estimates When No Failures Are Observed

Michael E. Schuckers
Department of Statistics and Center for Identification Technology Research
410 Hodges Hall
West Virginia University
Morgantown, WV 26506-6330
mschucke@stat.wvu.edu

Abstract

In this paper we discuss ways to use Bayesian methodology to estimate the matching performance of a biometric identification device when no errors are detected. One of the drawbacks to the classical or frequentist statistical estimation methods is that it is not possible to create a confidence interval for the error rate when no errors are observed. In this paper we begin by discussing the relevant work in statistical estimation. We then introduce the Bayesian approach to estimation and present examples.

1. Introduction

Biometric Identification Devices (BID) compare a physiological measure of a subject to a database of stored templates. The goal of any BID is to correctly match those two quantities. As acknowledged in a variety of papers, e.g. [9], [8], there is a pressing need for assessing the uncertainty in estimating the performance of a BID. The primary summaries for estimating performance are the false non-match rates (FNMR) and false match rates (FMR).

At present there is no widely accepted methodology in usage for creating interval estimates of FNMR and FMR. At present there is no universally accecpted methodology for assessing the performance of a BID. This is especially true when multiple individuals are tested multiple times. In particular, as [10] and others have noted the binomial distribution underestimates the variability when more than one individual attempts to match. [6] introduced a parametric frequentist methodology for assessing BID performance using a Beta-binomial distribution. Other non-parametric attempts have been suggested by, for example, [2], [10]. Both of these papers use resampling techniques to derive appropriate standard errors. One of the drawbacks of both the parametric frequentist approach of [6] and the nonparametric methods of [2] and [10] is that they cannot produce interval estimates when there are no failures. The goal of this paper is to present a methodology that will remedy that gap in the Biometric Identification literature.

2. Bayesian Methodology

Bayesian methodology derives its name from the Rev. Thomas Bayes who developed Bayes' Theorem which is at the heart of the methodology. The basic premise of Bayesian techniques is that all unknowns are treated as random variables and knowledge of these quantities is summarized via a probability distribution. Thus both the data, $\mathbf{Y}$ and the parameters, $\boldsymbol{\theta}$, in the frequentist framework are treated as random quantities. Bayesian methodology then takes what is known about the unknown parameters and updates that with the observed information found in the data.

Assume that we have knowledge about a, possibly vector valued, parameter, $\boldsymbol{\theta}$. Let that information be contained in a probability distribution $f(\boldsymbol{\theta})$. This distribution is often referred to as the prior distribution of $\boldsymbol{\theta}$. Assume that we have a vector of n observations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. We then suppose that we have a sampling distribution that is conditional upon $\boldsymbol{\theta}$, $f(\mathbf{Y} \mid \boldsymbol{\theta})$. This is often called the data distribution or the likelihood function. Note that We will use $f(\cdot)$ to denote any distribution. The arguments of the function will determine the precise distribution that is meant. We then follow the laws of probability to obtain the conditional distribution of $\boldsymbol{\theta}$ given the observed values, $f(\boldsymbol{\theta} \mid \mathbf{Y})$. Bayes Theorem gives us:

$$f(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{f(\mathbf{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\mathbf{Y})}, \tag{1}$$

where $f(\mathbf{Y})$ is the marginal distribution of $\mathbf{Y}$. Then $f(\boldsymbol{\theta} \mid \mathbf{Y})$ is called the posterior distribution of $\boldsymbol{\theta}$ and it describes that updated information that we possess about $\boldsymbol{\theta}$ having observed the data, $\mathbf{Y}$.

The posterior distribution, $f(\boldsymbol{\theta} \mid \mathbf{Y})$ is the basis for all Bayesian inference. Thus, all summaries about the quantity

$\boldsymbol{\theta}$ – point estimates or interval estimates – are based on the posterior distribution. For this paper we are particularly interested in interval estimates. Under the Bayesian paradigm such intervals are called credible sets or credible intervals. The typical (1-α)$\times$ 100% interval is created by finding the $100\times\alpha/2^{th}$ percentile and the $100\times(1-\alpha/2)^{th}$ percentile. An added benefit of this approach is that interpretation of this interval is a probabilistic one. That is, if U and L are the upper and lower limits for a 95% credible interval for θ, then there is a 95% probaility that θ is between U and L, i. e. $Pr(L < \theta < U) = 0.95$. This interpretation is a more intuitive one then the usual frequentist interpretation.

An excellent introduction to Bayesian methodology can be found in [7]. A more thorough description of these techniques can be found in [3].

3. Beta-binomial Sampling Distribution

In this section we introduce the Beta-binomial distribution. We will use this parametric distribution as the sampling distribution for our Bayesian inference. The Beta-binomial is derived in the following manner. Suppose that we have m individuals and each of those individuals is tested n_i times where $i = 1, \ldots, m$. Assume that $X_i \mid n_i, p_i \sim Bin(n_i, p_i)$ where X_i is the number of failures, and that

$$\mathrm{P}(X = x) = \binom{n_i}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}. \qquad (2)$$

We then further model each of the p_i's as conditionally independent draws from a Beta distribution. The Beta distribution is a continuous distribution on the interval [0,1] and it is parametrized with two quantities, α and β. Letting p_i have a Beta distribution, the probability density function is then

$$f(p_i \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1}. \qquad (3)$$

The mean and the variance for a Beta random variable are given by $\pi = \frac{\alpha}{\alpha+\beta}$ and $\pi(1-\pi)\frac{1}{\alpha+\beta+1}$, respectively. If we let $\pi = \frac{\alpha}{\alpha+\beta}$, then $E[p_i \mid \alpha, \beta] = \pi$ and $Var[p_i \mid \alpha, \beta] = \pi(1-\pi)(\alpha+\beta+1)^{-1}$.

The joint distribution is then,

$$\begin{aligned} f(\mathbf{x}, \mathbf{p} \mid \alpha, \beta, \mathbf{n}) &= f(\mathbf{x} \mid \mathbf{p}, \mathbf{n}) f(\mathbf{p} \mid \alpha, \beta) \\ &= \prod_{i=1}^{m} \binom{n_i}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i} \\ &\times \prod_{i=1}^{m} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1} \end{aligned}$$

where $\mathbf{p} = (p_1, \ldots, p_m)^T, \mathbf{x} = (x_1, \ldots, x_m)^T$, and $\mathbf{n} = (n_1, \ldots, n_m)^T$.

Now inference for this hierarchical model should be focused on α and β, since they define the overall probability of success. Consequently, we can integrate out the p_i's since they are now nuisance parameters. Thus,

$$\begin{aligned} f(\mathbf{x} \mid \alpha, \beta, \mathbf{n}) &= \int f(\mathbf{x}, \mathbf{p} \mid \alpha, \beta, \mathbf{n}) d\mathbf{p} \\ &= \int f(\mathbf{x} \mid \mathbf{p}, \mathbf{n}) f(\mathbf{p} \mid \alpha, \beta) d\mathbf{p} \\ &= \prod_{i=1}^{m} \binom{n_i}{x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\ &\times \frac{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)}{\Gamma(\alpha+\beta+n_i)} \qquad (4) \end{aligned}$$

Equation 4 is referred to as a joint Beta-binomial distribution or the product of Beta-binomial distributions. Under the Beta-binomial distribution,

$$\begin{aligned} E[X_i] &= n_i \frac{\alpha}{\alpha+\beta} = n_i \pi \\ Var[X_i] &= n_i \pi(1-\pi) C \end{aligned}$$

where $C = (\alpha+\beta+n_i)(\alpha+\beta+1)^{-1}$.

A few comments about the Beta-binomial distribution are warranted. The Beta-binomial distribution is often called an extravariation model. The reason for this is that it allows for greater variability among the variates, x_i, than the binomial distribution. The additional term, C, allows for additional variability beyond the $\pi(1-\pi)$ that is found under the binomial model. Letting $\gamma = \alpha+\beta$, γ determines the amount of variability in the Beta-binomial random variable beyond that found in the binomial. If the Beta-binomial model is correct and the binomial model is used then the variance will be understated by a factor of C and, hence, confidence intervals will be understated by a factor of $C^{\frac{1}{2}}$. $1 \le C \le n_i$. If γ is large, then $C \approx 1$ and the variability approaches $n_i\pi(1-\pi)$. When γ is large, it indicates that individual probabilities, the $p_i's$ are very similar and hence the variability resembles that of the binomial distribution. If γ is small, then $C \approx n_i$ and the variability approaches $n_i^2\pi(1-\pi)$. The flexibility that the Beta-binomial exhibits is a result of the two-parameter nature of the Beta-binomial distribution. Thus, it can exhibit more plasticity than the one-parameter binomial.

The appropriateness of the Beta-binomial distribution is dependent upon how well the Beta distribution can represent the population of p_i's. Several authors, e.g. [1], have noted the flexibility of the Beta distribution. Taking values on the interval $[0,1]$, the distribution is unimodal if $\alpha > 1$ and $\beta > 1$. If both α and β are 1, then the Beta distribution is equivalent to the continuous Uniform distribution on that interval. If either of these parameters are less than 1, then the distribution is J-shaped or reverse J-shaped. If both are less than 1, the distribution is U-shaped.

4. Posterior Estimation for the Beta-binomial

Having introduced the Beta-binomial distribution, we now present methods for determining posterior inference for the For Bayesian inference, we need two pieces. We need the sampling distribution, $f(\mathbf{Y} \mid \boldsymbol{\theta})$ and a prior distribution. The prior distribution is a description of the state of knowledge about the parameter, $\boldsymbol{\theta}$, before any data is observed. Here it is worth noting that the marginal distribution, $f(\mathbf{Y})$, can be found by integrating the joint distribution over $\boldsymbol{\theta}$.

$$f(\mathbf{Y}) = \int f(\mathbf{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{5}$$

Since $f(\mathbf{Y})$ is a constant relative to the posterior distribution, the posterior is often written as proportion to the product of the sampling distribution and the prior.

$$f(\boldsymbol{\theta} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}). \tag{6}$$

Assuming that we have m individuals and each of those individual is tested n times, we will use a Beta-binomial distribution for our sampling distribution. Recall that x_i represents the number of failure for the i^{th} individual. Thus our sampling distribution is the following:

$$\prod_{i=1}^{m} \binom{n_i}{x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)}{\Gamma(\alpha+\beta+n_i)}. \tag{7}$$

For the prior distribution, we have chosen a distribution that is flat on the proportion, π, and which is fairly uninformative about the sum of the parameters α and β. The prior mean of the sum, $\alpha + \beta$, which is intergral to inference about the quantity C is 10 while the prior variance is 100.

$$f(\alpha, \beta) \propto e^{-\frac{(\alpha+\beta)}{10}}. \tag{8}$$

Thus our posterior distribution is

$$\begin{aligned} f(\alpha, \beta \mid \mathbf{X}) \quad &\propto \quad e^{-\frac{(\alpha+\beta)}{10}} \\ &\times \quad \left\{ \prod_{i=1}^{m} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right. \\ &\times \quad \left. \frac{\Gamma(\alpha+x_i)\Gamma(\beta+n-x_i)}{\Gamma(\alpha+\beta+n)} \right\}, \end{aligned} \tag{9}$$

where $\mathbf{X} = (x_1, x_2, ..., x_n)^T$ and the combinatoric terms have been removed since it does not depend on either of the parameters. Evaluating equation 9 is not a straightforward task. The integral is not of a known form and, consequently, we cannot arrive at an analytic solution. We can, however, use simulation methods to obtain the quantities of interest about this distribution. Specifically we can use Markov chain Monte Carlo (MCMC) techniques to obtain the quantities of interest.

Table 1. Posterior summaries

		Percentiles		
m	n	2.5^{th}	50^{th}	97.5^{th}
10	2	0.00191	0.03446	0.16664
10	5	0.00079	0.01666	0.08156
10	10	0.00037	0.00900	0.04921
10	20	0.00039	0.00664	0.03945
20	2	0.00083	0.01750	0.08779
20	5	0.00035	0.00826	0.04131
20	10	0.00016	0.00467	0.02581
20	20	0.00011	0.00325	0.01999
50	2	0.00036	0.00736	0.03855
50	5	0.00013	0.00303	0.01573
50	10	0.00012	0.00205	0.01052
50	20	0.00006	0.00152	0.01132
100	2	0.00015	0.00371	0.01951
100	5	0.00008	0.00154	0.00838
100	10	0.00004	0.00095	0.00527
100	20	0.00003	0.00062	0.00401

Developed over the last decade, MCMC methods are a way to draw samples from a target distribution. They create a Markov Chain that converges to the target distribution to draw realizations from that distribution. Once convergence has been determined, the draws are assumed to come from the target distribution. We will follow the methodology of [4] for assessing convergence. More information on MCMC methods can be found in [5].

5. Application

In this section, we provide examples of the results using the posterior distribution given by equation 9. As before m will be the number of individuals tested and n_i will represent the number of times each individual is tested. For simplicity we will assume that each person is tested the same number of times, i. e. that $n_i = n$ for all i. The goal of this paper is to provide a methodology for creating interval estimates when no failures are observed. Consequently we will assume that the observed data $\mathbf{X} = (0, 0, ..., 0)^T$ is all zeros. Under this scenario we consider different values for m and n. Specifically, we present posterior intervals for π, the mean failure rate for the population, for all possible pairs of values where $m = 10, 20, 50, 100$ and $n = 2, 5, 10, 20$.

Applying the methodology of the previous sections we get the results found in Table 1. Along with m and n, the $2.5^{th}, 50^{th}$ and 97.5^{th} percentiles from each posterior distribution are given there. For all of the analyses presented in this section, four MCMC chains of length 10000 were run and the last half of each chain was used to calculate the percentiles in Table 1. Several patterns are apparent from this table. We will focus on the median or 50^{th} percentile for this discussion since this is the center of the posterior distribution. First, if we consider m fixed and focus on the changes due to n, we note that there are diminishing returns for increasing n. For each of the different values of m, as the value of n was increased from 2 to 5, there was a nearly 50% reduction in the median. However, to get an additional 50% decrease in the median it would be necessary to increase n to well over 10. Intuitively this is quite reasonable especially in the case of no failures. Our estimated failure rate for each person is not changing as we move from an n of 5 to an n of 10; thus, we need more and more evidence to decrease the overall estimate. Note that this assumes that additional attempts per person do not result in a single failure.

The second pattern that is apparent is a linear trend due to changes in m. Focusing on the estimated median, as we increase the value of m the estimated value of π seems to be decrease in a linear fashion. Again, this is only true of the increase number of individuals being tested does not produce a single failure. Consider the case when $n = 5$. As m increases from 10 to 20, the drop in the estimated median is almost exactly in half from 0.03446 to 0.01750. Similarly from $m = 20$ to $m = 50$ and from $m = 50$ to $m = 100$, the drop in the 50^{th} percentile of the posterior distribution is proportional to the increase in the value of m. Likewise this pattern is generally followed for the endpoints of the posterior 95% credible interval, the 2.5^{th} and 97.5^{th} percentiles. Thus, we can conclude that as a general rule increasing the number of individuals is more effective in terms of reducing variability in our estimation than increasing the number of attempts per individual. We reiterate that there is no guarantee that these relationships will continue to hold if failures are observed.

To determine the impact of changes in m on the upper limit of the posterior credible set, we undertook a second study. Recall that the posterior credible set gives a range of possible values for π based on the posterior distribution, in this case $f(\pi \mid \mathbf{X})$. The upper bound plays an important role in inference when no failures are observed since it gives the highest probable value of π. Thus it is this value which will gives an upper found on the performance of the system based upon the data. For this study we fixed the value of n to be 5 and expanded previous values of m to include 500, 1000 and 2000. Table 2 give the value of the 97.5^{th} percentile for each of these scenarios. We can interpret the table in the following manner: if we test 500 people 5 times each and get no failures, then the maximum error rate we could infer would be 16 errors per 10,000 attempts. Similarly testing 2000 individuals 5 times each results in an upper bound on the error rate of 48 errors per 100,000. The primary pattern here is a linear one. As we increase m the decrease in the upper limit is linear, assuming that no failures are observed.

Table 2. Upper limits for n=5

m	n	97.5^{th} percentile
10	5	0.08156
20	5	0.04131
50	5	0.01573
100	5	0.00838
500	5	0.00161
1000	5	0.00081
2000	5	0.00048

6. Summary

In this paper we have presented a methodology for creating an inferential interval when no failures occur in testing of a biometric identification device. This methodology is general; however, we have presented examples and explicit details for a Beta-binomial sampling distribution. We have chosen a proper prior distribution to ensure that our posterior is proper. Additional work is underway to determine the effect of prior choice on the posterior inference. Evidence from other studies and analytic work suggests that the impact of the prior is small when there is a large amount of data. In addition to our discussion of the methodology we have provided examples and discuss what impact these may have.

References

[1] D. A. Berry. *Statistics: A Bayesian Perspective*. Duxbury Press, New York, 1996.

[2] R. M. Bolle, N. Ratha, and S. Pankanti. Evaluation techniques for biometrics-based authentication systems. In *Proceedings of the 15^{th} International Conference on Pattern Recognition*, pages 835–841, 2000.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, 1995.

[4] A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.

[5] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, pages 1–20. Chapman & Hall, New York, 1996.

[6] M. E. Schuckers. Using the Beta-binomial distribution to assess performance of a biometric device. submitted to International Journal of Image and Graphics, 2001.

[7] B. Stephenson and H. S. Stern. A primer on the Bayesian approach to statistical inference. *Stats*, (23):3–9, 1998.

[8] U.K. Biometrics Working Group. Best practices in testing and reporting performance of biometric devices. on the web at www.cesg.gov.uk/biometrics, 2000.

[9] J. L. Wayman. *Biometrics:Personal Identification in Networked Society*, chapter 17. Kluwer Academic Publishers, Boston, 1999.

[10] J. L. Wayman. Confidence interval and test size estimation for biometric data. In *Proceedings of IEEE AutoID '99*, pages 177–184, 1999.