A comparison of statistical methods for evaluating matching performance of a biometric identification device- a preliminary report

Michael E. Schuckers
Department of Mathematics, Computer Science and Statistics
St Lawrence University, Canton, NY 13617, and
Center for Identification Technology Research,
West Virginia University
schuckers@stlawu.edu

Anne Hawley
Department of Statistics
North Carolina State University, Raleigh, NC 27695

Katie Livingstone, Nona Mramba
Department of Mathematics, Computer Science and Statistics
St. Lawrence University, Canton, NY 13617

Abstract

Confidence intervals are an important way to assess and estimate a parameter. In the case of biometric identification devices, several approaches to confidence intervals for an error rate have been proposed. Here we evaluate six of these methods. To complete this evaluation, we simulate data from a wide variety of parameter values. This data are simulated via a correlated binary distribution. We then determine how well these methods do at what they say they do: capturing the parameter inside the confidence interval. In addition, the average widths of the various confidence intervals are recorded for each set of parameters. The complete results of this simulation are presented graphically for easy comparison. We conclude by making a recommendation regarding which method performs best.

KEYWORDS: *False accept rate, False reject rate, Logit transformation, Beta-binomial, Intra-individual correlation, Bootstrap, Method of moments*

1 Background

For any system, assessing how well it performs is an important function to understanding the system. This is especially true for biometric identification devices. In order to assess their matching performance several confidence interval methodologies have been proposed. In this paper we compare these methodologies using statistical methodology to determine which methods work appropriately and under what circumstances they do so. In this preliminary paper we report on methods proposed by Doddington et al. (2000), Mansfield

and Wayman (2002), Bolle et al. (2000), Schuckers (2003b), Schuckers (2003a) and Michaels and Boult (2001). Specifically we are interested how well these methods do what they say they do, which is capture the overall error rate inside the limits of their intervals.

To appropriately address the performance of a confidence interval, it is necessary to do one of two things. Either one must prove, in a mathematical sense, that there are theoretical reasons for the interval to have the desired properties of a confidence interval or one must show empirically through simulations the proposed interval have appropriate coverage. Here we use coverage in the statistical sense meaning the empirical probability that a proposed confidence interval captured the desired parameter over a large number of samples from the same population. Note that calling an interval a 95% confidence interval does not guarantee that it has 95% coverage. Here we only consider 95% confidence intervals for those methodologies that where the confidence level can be chosen. Performance for other confidence levels should be similar.

To decide which of the methods performs best we will use two criteria. Our first criterion for determining which method or methods perform best is how often it achieves the desired coverage, typically 95%. The second criterion will be the average widths of the intervals. That a 95% confidence interval should give coverage of 95% is a clear and intuitive criterion. The average width of the interval is important because the narrower the width the more information is known about where the parameter you are estimating falls. Smaller here is better. Thus if two intervals give the same coverage we will prefer the one with smaller average width. It is also worth noting that the width of an interval is a function of the square of the information contained in the data. So if making an interval half as big requires four times as many observations.

The rest of this paper is organized in the following way. Section 2 describes the different methodologies that are tested here. This includes some modifications to methods that were necessary to fit the format of testing used here. That is followed by Section 3 which describes how the methods were tested. The approach we use here is a Monte Carlo approach to estimating coverage by generating data from a correlated binary distribution. Section 4 outlines the results from these methods and gives graphical summaries for each method. That is followed by Section 5 which discusses these results and makes a recommendation concerning these methods.

2 Confidence Interval Methods

In this paper we look at several different methods for making a confidence interval for the error rate. These methods span the typical statistical inference methodology though a Bayesian approach is absent. (Schuckers (2002) offers a Bayesian approach when zero errors are observed.) We begin this section by introducing notation that we will use throughout this paper. This notation is vital to understanding the methodologies that we consider here. With the exception of Doddington's Rule - see below for details - we focus on the creation of 95% confidence intervals.

2.1 Notation

Assume that n individuals are tested each of m_i, $i = 1, \ldots, n$ times. We are interested in the number of errors from these test. Therefore, let x_{ij} be 1 for an error and 0 for no error for the j^{th} attempt by the i^{th} individual. Assume that there is an overall error rate, π which remains constant throughout testing. This is the estimand. Further let $X_i = \sum_{j=1}^{m_i} x_{ij}$. Thus X_i represents the number of errors for the i^{th} individual. Further, let $\mathbf{X} = (X_1, \ldots, X_n)^T$. We can represent the observed proportion of errors for each individual as $p_i = \frac{X_i}{n}$. Each of the methods below will utilize the information found in the X_i's in different ways. Often the average number of tests per individual will be of interest, $\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}$. One common point estimator for the overall error rate and that is used by several methods is the following:

$$\hat{\pi} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} m_i}. \tag{1}$$

Each of the methods below also attempts to estimate the sampling variability in the estimate. Statistically, the estimated standard deviation of the sampling distribution is known as the standard error. These methods differ in how they proceed toward this estimation. Another common value that is worth mentioning is the 97.5^{th} percentile of a Normal distribution which is often written as $z_{0.025} = 1.96$. We mention it here since it will appear multiple times below without comment.

2.2 Doddington's Rule(DR)

Based on Doddington et al. (2000), this approach to inference is based on a binomial distribution. This methodology also appears in Mansfield and Wayman (2002). Although this methodology is intended to only be utilized when $S = \sum_{i=1}^{n} X_i \geq 30$ we considers it's performance for all values of S. The interval here is

$$\hat{\pi} \pm 0.30\hat{\pi}. \tag{2}$$

Note that this methodology is meant to create a 90% confidence interval, rather than the 95% nominal interval that we desire here. We will take this into account when we compare the results below.

2.3 Best Practices(BP)

We refer to the methodology in this section as the "Best Practices" approach since it is taken from Mansfield and Wayman (2002). They specify this methodology as pertaining to the false match rate. We use it here more generally. The BP approach is a method-of-moments one. This method takes $\hat{\pi}$ as the point estimate of π. The variance of the $p_i's$ is

then used to estimate the standard error or $\hat{\pi}$. The confidence interval is then

$$\hat{\pi} \pm 1.96\sqrt{\frac{\sum_{i=1}^{n}(m_i(p_i - \hat{\pi}))^2}{\bar{m}^2 n(n-1)}}. \tag{3}$$

2.4 Subset Bootstrap(SB)

The "subset bootstrap" approach of Bolle et al. (2000) is an inherently non-parametric approach to interval estimation. It is based on the bootstrap resampling approach to estimating the standard error. The approach is this.

1. For the i^{th} individual take a sample of size m_i with replacement from the observations $x_{i1}, \ldots, x_{im_i}$. Note that this is equivalent to generating m_i Bernoulli variates each having probability p_i or sampling from a Binomial distribution with m_i trials and probability of success p_i. Call this sample $X_i^* = (x_{i1}^*, \ldots, x_{im_i}^*)^T$.
2. Repeat the previous step for all n individuals. Having done that we then have a replicate data set, call it $\mathcal{X} = (X_1^{*T}, \ldots, X_n^{*T})^T$.
3. Find $\hat{\pi}^*$ from $\mathcal{X}$ using equation (1).
4. Repeat the previous steps T times storing $\hat{\pi}^*$ each time.
5. Sort the $\hat{\pi}^*$'s and find the 2.5^{th} and 97.5^{th} percentiles from these T replicates of $\hat{\pi}$.

These percentiles then form a 95% confidence interval for π. Note that implicit in the structure of the replicate generation is the assumption that the X_i's are conditionally independent. We will use $T = 1000$.

2.5 Beta-binomial(BB)

The Beta-binomial approach that we use here is based on Schuckers (2003b). Schuckers (2003b) took a maximum likelihood approach to interval estimation, whereas Schuckers (2003a) used analysis of variance or ANOVA-type estimation due to Lui et al. (1996). In both of these papers the approach is parametric in that Schuckers specifies a model for the mean and variance and estimates the parameters of that model. This model is

$$\begin{aligned} E[X_i] &= m_i\pi \\ Var[X_i] &= m_i\pi(1-\pi)(1+(m_i-1)\rho) \end{aligned} \tag{4}$$

where ρ represents the intra-individual correlation. In addition, this model assumes that each of the X_i's are conditionally independent. This type of model is often referred to as an extra-variation model since it allows for more variation than would be expected under the typical model - in this case the binomial. The following quantities are used for estimation by this method.

$$\hat{\pi} = \frac{\sum X_i}{mn}$$

$$
\begin{aligned}
\hat{\rho} &= \frac{BMS - WMS}{BMS + (m_0 - 1)WMS} \\
\text{where} & \\
BMS &= \frac{\sum m_i (p_i - \hat{\pi})^2}{n-1}, \\
WMS &= \frac{\sum m_i p_i (1 - p_i)}{n(\bar{m} - 1)} \\
\text{and} & \\
m_0 &= \bar{m} - \frac{\sum_{i=1}^{n} (m_i - \bar{m})^2}{n\bar{m}}.
\end{aligned} \tag{5}
$$

The confidence interval for π in then

$$
\hat{\pi} \pm 1.96 \left[\frac{\hat{\pi}(1-\hat{\pi})(1 + (m_o - 1)\hat{\rho})}{\bar{m}n} \right]^{1/2}. \tag{6}
$$

We note here that the BP method can be thought of as another method for estimating the parameters of the model in (4). Both methods should converge to the same estimates of π and ρ when n is large. (The methods should also agree when m is large; however this is rarely the case in testing of biometric identification devices.) Consequently, both methods give nearly identical confidence intervals for the cases we consider below. Thus, we will treat these two methods - developed independently - as a single approach because the results we give below for both methods are identical. We'll refer to this approach as BP-BB.

2.6 Logit Beta-binomial(LBB)

The logit Beta-binomial confidence interval is derived from the Beta-binomial (Schuckers, 2003a). The logit function is the log-odds

$$
logit(y) = log(\frac{y}{1-y}). \tag{7}
$$

The idea behind this interval is to create a confidence interval for $logit(\pi)$ and then transform that interval back to the interval $(0, 1)$ via the inverse of the logit function

$$
ilogit(x) \equiv logit^{-1}(x) = \frac{e^x}{1 + e^x}. \tag{8}
$$

The motivation behind this approach is that it often improves coverage.

The transformed confidence interval for the $logit(\pi)$ is

$$
logit(\hat{\pi}) \pm 1.96 \sqrt{\frac{1 + (\bar{m} - 1)\hat{\rho}}{\hat{\pi}(1-\hat{\pi})\bar{m}n}}. \tag{9}
$$

We calculate $\hat{\pi}$ and $\hat{\rho}$ as in the BB method above. We then take the inverse logit of the upper and lower endpoints for this interval, U and L, respectively. To get the interval $(ilogit(L), ilogit(U))$. Note that this interval is guaranteed to fall inside $(0, 1)$.

2.7 Balance Repeated Replicates(BRR)

The balanced repeated replication method of Michaels and Boult (2001) is an approach similar to the SB above since both using resampling methods. Unlike the SB, the BRR does not create a complete replicate data set. Rather, it samples a single value is selected from each individual. The approach here is to treat individuals as if each represents a stratum and use methods appropriate for finite population sampling (Michaels and Boult, 2001). This methodology requires that the $m_i = m$ for all i and that m_i is prime. Since these conditions are not always met, we use a Monte Carlo approach to estimating the standard error based on this approach. In our testing, the Monte Carlo approximation gave results which were nearly identical to the exact method when m_i was prime. Therefore, we use the Monte Carlo approach outlined below. We use the notation from the SB approach above.

1. For the i^{th} individual we sample a single observation ω_i from the observations $x_{i1}, x_{i2}, \ldots, x_{im_i}$. This is equivalent to generating a single Bernoulli trial with probability of success p_i.
2. Repeat the previous step for all n individual. Having done that we have $(\omega_1, \ldots, \omega_n)^T$.
3. Calculate $\tilde{\pi} = \frac{\sum_{i=1}^{n} \omega_i}{n}$.
4. Repeat the previous steps T times storing $\tilde{\pi}$ each time.
5. Sort the $\tilde{\pi}$'s and find the 2.5^{th} and 97.5^{th} percentiles from these T replicates of $\tilde{\pi}$.

These values then form the endpoints for a 95% confidence interval. As with SB, we use T=1000.

3 Data Simulation

We statistically evaluate the methodologies in the following way. For each value of m, n, π and ρ and for each method, we generate $\mathbf{X} = (X_1, \ldots, X_n)^T$1000 times. Note that for computational speed and simplicity we let $m = m_i$ for all i. Each X_i is assumed to be conditionally independent and generated from a correlated binary distribution (Oman and Zucker, 2001). This distribution assumes the first two moments and does not assume a particular form for the distribution. For each $\mathbf{X}$ we create a confidence interval and determine whether π is captured by that interval. We then calculate coverage as the percentage of times that π is "captured" inside the various intervals for that set of parameters. This process is then repeated for varying values of the parameters. In addition to the coverage, the average width at each set of parameter values was recorded. The values we used were $\pi = (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1)$, $\rho = (0.001, 0.01, 0.05, 0.1, 0.2, 0.4, 0.8)$, $m = (2, 5, 8, 12, 15, 20, 25, 30)$ and $n = (100, 500, 1000, 5000, 10000, 50000, 100000)$. These values focus on parameters that are plausible for biometric identification devices based on published results.

Table 1: Guide to colors in Figure 1 and 2

Color	Range of Coverages	
	Lower	Upper
Tan	0.000	0.010
Yellow	0.010	0.250
Yellow-Green	0.250	0.500
Green	0.500	0.750
Teal	0.750	0.900
Light Blue	0.900	0.925
Blue	0.925	0.950
Purple	0.950	1.000

4 Results

The results for the six methodologies considered here are given below. Overall all of the methods performed well when $n\pi > 10$ and poorly when $n\pi < 0.05$. That was expected. The logit Beta-binomial approach performed the best among these methods in the sense that it gives appropriate coverage for smaller sample sizes, m and n, or larger correlations ρ than the other methods in this evaluation. In addition we note that the Beta-binomial/Best Practices approach and the logit Beta-binomial method each produced the anomalous result that under certain circumstances increasing m results in a decrease in coverage. This certainly warrants further study because of it's counterintuitive nature. We refer to this as the 'hook paradox' because of the appearance of a hook in graphical representations of the data. This hook can be seen in Figure 2.

To illustrate how these methods performed, we include Figures 1 and 2. Each of the graphs is composed of blocks. Each block represents a set of parameter values. For each block the values of n, the x-axis and m the y-axis are fixed; the particular values are given on the axis for each block with values increasing from left to right and bottom to top respectively. Within each block the values for π are on the x-axis increasing from left to right and the values for ρ are on the y-axis increasing from bottom to top. Each square within a block represents the coverage at the particular combination of n, m, π, ρ. The color given to each square is dependent on the coverage. Table 1 gives a breakdown of the coverages that produce each color. Recall that our target coverage is 0.950 or 95%. However, some fluctuation around that coverage is to be expected.

4.1 DR results

Individually, the method of Doddington performed well when $n\pi > 10$ or when $n\pi > 5$, $\rho < 0.2$ and $m > 8$. It performed poorly when $n\pi < 1$ and when $n\pi < 5$, $m < 8$, $\rho > 0.2$. Under the former circumstances the methodology performed too well giving intervals that gave coverage of 100% suggesting that the interval was too wide.

Figure 1: Results for DR, BP-BB and SB for various parameters
Darker values indicate higher coverage

4.2 BP-BB results

The Best Practices/Beta-binomial approach performed well when $n\pi > 5$ or when $n\pi > 1$ and $\rho < 0.01$ and $m > 8$. It performed poorly when $n\pi < 0.1$ and when $n\pi < 5$ and $m < 8$ and $\rho > 0.2$. Again it should be noted that these two methods produced nearly identical coverages and average interval widths and so we treat them as a single method.

4.3 SB results

It was anticipated prior to running these simulations that SB would perform the best. Resampling approaches are well understood and widely used because of their strong performance record especially in situations where estimation is difficult. SB did well when $n > 5000$ and $\rho < 0.2$. This method performed well when $\pi \geq 0.005$ and $\rho < 0.1$. Coverage was poor when $n \leq 1000$ and $\pi \leq 0.001$, and when $m > 12$ and $\rho > 0.05$.

4.4 LBB results

The LBB approach does well when $n\pi > 5$ or when $n\pi > 0.5$ and $\rho < 0.1$ and $m > 5$. The approach does poorly when $n\pi < 0.05$ or when $n\pi < 1$ and $m < 8$ and $\rho > 0.1$.

4.5 BRR results

The BRR methodology performed well. In general it performed well when $n\pi > 0.5$ or when $n > 1000$ and $\pi \geq 0.005$ and $\rho \leq 0.4$. When $\rho = 0.8$ or when $n\pi \leq 0.005$,

5 Conclusions

None of the methods do well when $n\pi$ is small. This is not surprising. All of the methods do well when $n\pi$ is large. This too is not surprising. (For the binomial distribution the usual suggestion is that for the usual large sample confidence interval that $n\pi > 10$ and $n(1-\pi) > 10$.) In addition most of the methods did poorly when ρ was large. This indicates the difficulty of trying to deal with data of this type. BRR and LBB seem to be the best choices overall in terms of coverage. This can be seen visually by looking at Figures 1 and 2 and determining which of these have the most blue. BP-BB lags slightly behind these three. DR is clearly not as effective as the others. Likewise SBB does surprisingly poorly. BRR seems to have difficulty estimating π when ρ is large (say greater than 0.4). LBB seems to do more poorly than the other two methods when ρ is large and π is small. To differentiate between BRR and LBB, we further consider the average widths of these two methods. We considered the ratio of average widths of the intervals created by these three methods. Looking at the ratio of BRR widths to LBB width, we found that the mean value for this ratio was 3.348, while the median ratio was 1.584. Therefore we recommend the following. As a general rule under the scenarios described above, LBB performs the best. If the data indicates that $\hat{\rho}$ is large (say greater than 0.4) and $n\hat{\pi}$ is small (say less than 1), then BRR may be preferable. Further, we recommend that the methodology found in Schuckers (2003a) be used for determining sample size since LBB clearly performed well.

Figure 2: Results for LBB and BRR for various parameters
Darker values indicate higher coverage

We recommend this methodology since no *a priori* sample size calculations are possible with SB.

There are several possible avenues for future work. Whether or not the data from a biometric identification device follows the correlated binary distribution proposed by Oman and Zucker (2001) is an important question that needs to be answered. We have also proposed to study and compare these methods under the Beta-binomial distribution. (The full report from this project will include this evaluation.) Preliminary analysis of data found in Ross and Jain (2003) indicates that the Beta-binomial distribution fits the much of the data found there. Assuming these methods are appropriate, the observed values for ρ will play an integral role in determining how well these methods perform. Additional research needs to be done on values of ρ between 0.2 and 1.0. Further reporting of these estimates is needed so that the biometric identification community can use these values for planning and inferential purposes. Additionally, more work is needed to characterize the "hook paradox" and to understand the exact conditions under which it occurs. Finally, there is a definite need in the biometric identification community to move beyond confidence intervals for π and begin to look at additional tools for inference including hypothesis tests.

Acknowlegements: This research was made possible by generous funding from the Center for Identification Technology Research (CITeR) at the West Virginia University. The authors would also like to thank Robin Lock, Stephanie Caswell Schuckers, Anil Jain, Larry Hornak, Jeff Dunn and Bojan Cukic for insightful discussions regarding this paper. Similarly we would like to thank Collen J. Knickerbocker for sharing his wisdom concerning computation matters.

References

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

Doddington, G. R., Przybocki, M. A., Martin, A. F., and Reynolds, D. A. (2000). The NIST speaker recognition evaluation: overview methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254.

Lui, K.-J., Cumberland, W. G., and Kuo, L. (1996). An interval estimate for the intraclass correlation in beta-binomial sampling. *Biometrika*, 52(2):412–425.

Mansfield, T. and Wayman, J. L. (2002). Best practices in testing and reporting performance of biometric devices. on the web at www.cesg.gov.uk/site/ast/biometrics/media/BestPractice.pdf.

Michaels, R. J. and Boult, T. E. (2001). Efficient evaluation of classification and recognition systems. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*.

Oman, S. D. and Zucker, D. M. (2001). Modelling and generating correlated binary variables. *Biometrika*, 88(1):287–290.

Ross, A. and Jain, A. K. (2003). Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125.

Schuckers, M. E. (2002). Interval estimates when no failures are observed. In Ratha, N. K. and Bolle, R. M., editors, *AutoID'02 Proceedings: Workshop on Automatic Identification Advanced Technologies*, pages 37–41, Tarrytown, NY. IEEE Robotics and Automation Society and The Association for Automatic Identification and Data Capture Technologies (AIM).

Schuckers, M. E. (2003a). Estimation and sample size calculations for matching performance of biometric identification devices. Center for Identification Technology Research technical report.

Schuckers, M. E. (2003b). Using the beta-binomial distribution to assess performance of a biometric identification device. *International Journal of Image and Graphics*, 3(3):523–529.