Approximate Confidence Intervals for Estimation of Matching Error Rates of Biometric Identification Devices

Travis J. Atkinson[1] and Michael E. Schuckers[1,2]

[1] Department of Mathematics, Computer Science and Statistics
St. Lawrence University, Canton NY 13617, USA
[2] Center for Identification Technology Research (CITeR)
schuckers@stlawu.edu

Abstract. Assessing the matching error rates of a biometric identification devices is integral to understanding its performance. Here we propose and evaluate several methods for creating approximate confidence intervals for matching error rates. Testing of biometric identification devices is recognized as inducing intra-individual correlation. In order to estimate error rates associated with these devices, it is necessary to deal with these correlations. In this paper, we consider extensions of recent work on adjustments to confidence intervals for binomial proportions to correlated binary proportions. In particular we propose a Agresti-Coull type adjustment for estimation of a proportion. Here that proportion represents an error rate. We use an overdispersion model to account for intra-individual correlation. To evaluate this approach we simulate data from a Beta-binomial distribution and assess the coverage for nominally 95% confidence intervals.

1 Background

Errors often result from the matching process of a biometric identification device. Depending on the individual attempting to gain access, these errors are often classified as being either false accepts or false rejects. It is important that users of these devices have an understanding of the magnitude of these error rates. Several authors have noted the importance of this knowledge to testers of biometric devices and for users, see e.g. [1] and [2]. More recently a National Science Foundation Workshop found that "there is a need to develop statistical understanding of biometric systems sufficient to produce models useful for performance evaluation and prediction", [3]. Because of the binary nature of the outcome of the matching decision from a biometric device, it is often supposed that a binomial distribution is appropriate to analyze such data. Several authors including [1] have noted that a binomial distribution is not appropriate for biometric data under most circumstances. The problem with using the binomial distribution to describe the data is that it does not allow for intra-individual correlation in testing. [4] proposed that the Beta-binomial distribution is appropriate for describing data taken from a biometric identification device because

it permits correlation within individuals. In that same paper, [4] proposed a methodology for making confidence intervals using this approach. An updated version is given in [5].

The purpose of this paper is to determine whether or not the methodology of interval estimation proposed by Agresti and Coull in [6] can be applied to a Beta-binomial distribution, and to establish an appropriate manner in which to do so. The approach of Agresti and Coull was to add four additional observations (two successes and two failures). They applied their approach to the binomial distribution, specifically for estimation of proportions. We are motivated in applying their methodology to Beta-binomial proportions because data taken from a biometric device often follows a Beta-binomial distribution. Again, our application is to overall error rates. To extend Agresti and Coull's approach to the Beta-binomial, we consider several methods for applying their approach and we will evaluate these by simulating data from a variety of parameter combinations. We will then compare these methods to the unaugmented approach of [5].

2 Background and Notation

We use a parametric approach for modelling the data from a biometric device. We follow the approach of [5] and use the notation found therein. Assume that we are interested in estimating an overall error rate – either the false accept rate (FAR) or false reject rate (FRR)– of π. Further, let n be the number of individuals tested and m_i be the number of times that the individual is tested, $i = 1, \ldots, n$. Agresti and Coull [6] suggested adding two "successes" and two "failures" to binomial data as a way to improve estimation. In the case of the binomial since all trials are independent, augmenting the data is straightforward. For correlated binary data any augmentation is not straightforward because of the correlated structure induced by having multiple individuals tested. We shall consider several ways to distribute these additional "observations". With these changes, some individuals will have up to four additional tests added. As will become apparent, this is not the case for Beta-binomial data. We call the occurrence of an error a "success", S; a "failure", F, would be the event that an error does not occur. Let

$$X_i = \sum_{j=1}^{m_i} \omega_{ij}.$$

And let

$$\omega_{ij} = \left\{ \begin{matrix} 0 & \text{if } S \\ 1 & \text{if } F \end{matrix} \right\} \tag{1}$$

for the j^{th} attempt by the i^{th} individual. Hence, X_i represents the number of errors by the i^{th} individual and let $p_i = X_i/m_i$ represent the percentage of errors in m_i attempts made by the i^{th} individual. In this paper we follow [4] in assuming an extravariation or overdispersion model for the data. Formally we

assume that

$$\begin{aligned} E[X_i] &= m_i\pi_i \\ &\text{and} \\ Var[X_i] &= m_i\pi_i(1-\pi_i)(1+(m_i+1)\rho) \end{aligned} \tag{2}$$

where ρ is the intra-individual correlation of the observed data [4]. The intra-individual correlation, measures the similarity of outcomes within a single individual [7] relative to the overall variability of the X_i's. It is worth mentioning here that if $\rho = 0$ then this model simplifies to the binomial model. Using the notation described above, we can now estimate π and ρ. Estimates of parameters will be denoted with a "hat". Once these values are estimated we can derive confidence intervals for the error rate, π. To begin, suppose that we have tested a biometric identification device and have observed the X_i's. For our purposes, instead of taking data from an actual biometric identification device, we conducted simulations using the statistical software R. This allowed us to test performance of the proposed confidence intervals for a variety of different values for the parameters π and ρ. For estimation we then use,

$$\begin{aligned} \hat{\pi} &= \frac{\sum X_i}{mn} \\ \hat{\rho} &= \frac{BMS - WMS}{BMS + (m_0 - 1)WMS} \\ &\text{where} \\ BMS &= \frac{\sum m_i(p_i - \hat{\pi})^2}{n-1}, \\ WMS &= \frac{\sum m_i p_i(1-p_i)}{n(\bar{m}-1)} \\ &\text{and} \\ m_0 &= \bar{m} - \frac{\sum_{i=1}^{n}(m_i - \bar{m})^2}{n\bar{m}}. \end{aligned} \tag{3}$$

The confidence interval for π in then

$$\hat{\pi} \pm 1.96\left[\frac{\hat{\pi}(1-\hat{\pi})(1+(m_o-1)\hat{\rho})}{\bar{m}n}\right]^{1/2}. \tag{4}$$

The above methodology for estimating ρ is based on an ANOVA-type estimator given by [8]. In Equation (3) BMS represents between mean squares and WMS represents within mean squares.

3 Proposed Methods

The augmentation of the binomial proposed by [6] involves adding two "successes" and two "failures" to the observed successes and failures. For our purposes this is equivalent to adding two additional errors on four additional attempts. As mentioned above, we assume data is collected on n individuals each

of which is tested m_i times $i = 1, \ldots, n$. Hence, we have n individuals with m_i binary observations each. The correlated structure of this data implies that adding two "successes" and two "failures" is not simple. We need to consider how to distribute these "successes" and "failures" among the n individuals. We have developed four specific approaches for doing this. We refer to these methods as A1, A2, A3, and A4. These methods will be compared to the unaugmented approach of [9] which we label A0. For all of these methods, we implicitly assume that the individuals are ordered randomly. If that is not the case then the individuals who receive additional "observations" as part of the augmentation need to be randomly chosen.

3.1 Approach 1 (A1)

The first approach involves adding two "successes" and two "failures" to a single individual, in this case the first individual. For simplicity with this and other approaches we assume $m_i = m$ for all i initially. Extensions to the more general case are straightforward. Thus,

$$m_i = \begin{cases} m+4 & \text{for } i = 1 \\ m & \text{for } i = 2, \ldots, n \end{cases} \tag{5}$$

and

$$X_i = \begin{cases} \sum_{j=1}^{m_i} \omega_{ij} + 2 & \text{for } i = 1 \\ \sum_{j=1}^{m_i} \omega_{ij} & \text{for } i = 2, \ldots, n. \end{cases} \tag{6}$$

We then use Equation (4) to develop a confidence interval based on the updated values for the X_i's and the m_i's. Hence, the first individual tested effectively undergoes four additional tests, two of which result in a "success" and two of which result in a "failure". For example, if $n = 1000$ and each individual underwent 5 tests, then we let $m_1 = 5 + 4 = 9$ and $X_1 = \sum_{j=1}^{5} \omega_{ij} + 2$, while $m_i = 5$ and $X_i = \sum_{j=1}^{5} \omega_{ij}$ for all $i = 2, 3, \ldots, 1000$.

3.2 Approach 2 (A2)

Another approach would be to distribute the additional attempts and errors equally among two individuals. This is our second approach. Thus,

$$m_i = \begin{cases} m+2 & \text{for } i = 1, 2 \\ m & \text{for } i = 3, \ldots, n. \end{cases} \tag{7}$$

and

$$X_i = \begin{cases} \sum_{j=1}^{m_i} \omega_{ij} + 1 & \text{for } i = 1, 2 \\ \sum_{j=1}^{m_i} \omega_{ij} & \text{for } i = 3, \ldots, n. \end{cases} \tag{8}$$

Hence, there were two additional tests added to both the first and second individuals. Each of these two individuals receives one "failure" and one "success". Again we assume that individuals are numbered randomly.

3.3 Approach 3 (A3)

The third approach adds an additional test to four separate individuals. Two of those individuals receive a "success"; the other two receive a "failure". Thus we augment the data in the following manner

$$m_i = \begin{cases} m+1 & \text{for } i = 1,2,3,4 \\ m & \text{for } i = 5,\ldots,n \end{cases} \tag{9}$$

and

$$X_i = \begin{cases} \sum_{j=1}^{m_i} \omega_{ij} + 1 & \text{for } i = 1,2 \\ \sum_{j=1}^{m_i} \omega_{ij} & \text{for } i = 3,\ldots,n. \end{cases} \tag{10}$$

By this method the four individuals effectively undergo one additional test each. For individuals 1 and 2 this test results in a "success", while for individuals 3 and 4 the additional test results in a "failure".

3.4 Approach 4 (A4)

The final augmentation approach involves adding an $n+1^{st}$ individual who receives two "successes" and two "failures". Note that this methodology is different from the previous approaches in that we are also altering n. Thus,

$$m_i = \begin{cases} m & \text{for } i = 1,\ldots,n \\ m+4 & \text{for } i = n+1 \end{cases} \tag{11}$$

and

$$X_i = \begin{cases} \sum_{j=1}^{m_i} \omega_{ij} + 1 & \text{for } i = 1,\ldots,n \\ 2 & \text{for } i = n+1. \end{cases} \tag{12}$$

4 Evaluation

Our approach to evaluating these methods is to use simulated data. We choose this approach over application to observed test data because it allows us to test many different parameter values and parameter combinations. To evaluate each of these methods, we simulate data from a variety of parameter values. To test the estimated proportion we use Equation (4) which provides a nominally 95% confidence interval. As has been stated, the goal is to test the performance of

Table 1. Coverage for A0

	$n = 1000$ $m = 5$				$n = 1000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.931	0.928	0.926	0.866	0.951	0.953	0.904	0.854
0.004	0.947	0.951	0.942	0.883	0.941	0.939	0.932	0.907
0.008	0.955	0.940	0.926	0.931	0.944	0.945	0.945	0.933
0.010	0.937	0.951	0.952	0.926	0.952	0.954	0.956	0.938

	$n = 2000$ $m = 5$				$n = 2000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.947	0.950	0.930	0.908	0.934	0.942	0.937	0.907
0.004	0.944	0.953	0.948	0.922	0.948	0.945	0.931	0.935
0.008	0.947	0.955	0.940	0.948	0.941	0.963	0.945	0.944
0.010	0.937	0.957	0.950	0.932	0.935	0.951	0.939	0.932

Table 2. Coverage for A1

	$n = 1000$ $m = 5$				$n = 1000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.969	0.962	0.963	0.936	0.961	0.955	0.950	0.907
0.004	0.953	0.960	0.952	0.942	0.959	0.956	0.942	0.932
0.008	0.955	0.956	0.951	0.933	0.941	0.949	0.942	0.927
0.010	0.946	0.954	0.948	0.949	0.958	0.950	0.941	0.949

	$n = 2000$ $m = 5$				$n = 2000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.964	0.963	0.958	0.946	0.946	0.949	0.945	0.926
0.004	0.952	0.944	0.959	0.947	0.947	0.951	0.950	0.924
0.008	0.947	0.941	0.947	0.948	0.949	0.943	0.955	0.944
0.010	0.957	0.952	0.948	0.944	0.957	0.959	0.954	0.951

the confidence interval given above. We use simulations of data given n, m, π, ρ. Running the simulations under a given set of parameters, we create a confidence interval for each data set and then compare this interval to our actual error rate. We then determine whether or not the interval contains, or "captures", π. For repeated generation of data, we expect that the percent of times that π falls in the interval created is 95%. (Recall that the definition of a 95% confidence interval is that $P(\hat{\pi}_L < \pi < \hat{\pi}_U) = 0.95$ where $\hat{\pi}_L$ and $\hat{\pi}_U$ are the lower and upper endpoints of that interval respectively.) We will refer to the proportion of times that π is within the range of the 95% confidence interval as the coverage. We will use this value to assess the overall performance of an approach. Since we use a 95% confidence interval of the estimated error rate, we expect the coverage to be 95% percent. We desire coverage that is close to 95%. The closer the coverage is to this value, the better the performance of the approach. Coverage that is too low is deficient in that it does not meet the criteria for a 95% confidence interval. Coverage that is too high is too conservative in the statistical sense that it yields intervals that are too wide.

Table 3. Coverage for A2

	$n = 1000$ $m = 5$				$n = 1000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.967	0.961	0.958	0.922	0.949	0.947	0.941	0.891
0.004	0.949	0.952	0.952	0.938	0.946	0.947	0.938	0.927
0.008	0.954	0.956	0.941	0.936	0.955	0.964	0.958	0.934
0.010	0.944	0.952	0.964	0.951	0.964	0.950	0.952	0.939

	$n = 2000$ $m = 5$				$n = 2000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.942	0.956	0.956	0.942	0.943	0.958	0.954	0.939
0.004	0.949	0.957	0.951	0.956	0.949	0.952	0.952	0.939
0.008	0.960	0.949	0.945	0.943	0.947	0.953	0.947	0.944
0.010	0.950	0.951	0.951	0.943	0.943	0.952	0.955	0.953

The parameters used in the simulation of each approach are: $n = (1000, 2000)$, $m = (5, 10)$, $\pi = (0.002, 0.004, 0.008, 0.01)$, and $\rho = (0.001, 0.01, 0.1, 0.4)$. 1000 data sets were generated under each scenario. Data were generated from a Beta-binomial distribution. The scenarios – combinations of parameters – are investigated to see how the changes in the parameters make a difference in the coverage.

The values of π were chosen to be representative of error rates produced by a BID [5]. The results of these simulations are given in Tables 1-5. Again we use simulated data because it allows us to test the performance of various parameter combinations.

Table 4. Coverage for A3

	$n = 1000$ $m = 5$				$n = 1000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.965	0.944	0.952	0.929	0.950	0.946	0.949	0.896
0.004	0.954	0.950	0.944	0.925	0.953	0.957	0.946	0.926
0.008	0.961	0.949	0.940	0.950	0.948	0.947	0.938	0.928
0.010	0.956	0.957	0.942	0.939	0.944	0.950	0.953	0.939

	$n = 2000$ $m = 5$				$n = 2000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.940	0.942	0.955	0.939	0.957	0.958	0.951	0.913
0.004	0.948	0.945	0.952	0.941	0.949	0.941	0.942	0.935
0.008	0.960	0.956	0.957	0.934	0.943	0.952	0.950	0.943
0.010	0.958	0.951	0.957	0.953	0.954	0.944	0.947	0.941

5 Results

As we examine the data in the tables, we begin with the trends that occur in all data sets regardless of the approach used. First, as π, n and m increase, the coverage increases. Coverage also increases as ρ decreases. These trends hold for all approaches. It is also apparent that as ρ increases above 0.1, the coverage decreases significantly. Again, ρ governs the degree of independence in the data. High values of this parameter imply less independent information in the data. Another way to think of this is that the effective sample size is $\frac{n\bar{m}}{1+(\bar{m}-1)\rho}$, which is equivalent to the number of independent observations in the data. From this formula, it is clear that the effective sample size decreases as ρ increases.

A phenomenon that can be seen from the data tables below is what has been called the "Hook Paradox" and occurred similarly in [5]. The paradox is that when ρ is large as the number of tests per individual, m_i increases, the coverage decreases. This is an unexpected result. It is especially apparent in

the data when $\rho = 0.4$ and the error rate π is small. For example, in Table 2 when $n = 1000, m = 5, \pi = 0.002, \rho = 0.4$, the coverage is 0.936, but when m is changed to 10, the coverage decreases to 0.907. Intuitively, the coverage should increase as m_i increases since there are a larger number of tests taken. This pattern seems to occur for several confidence interval methods [9].

Turning to individual approaches, it is evident that all four approaches (A1-A4) that are modified using the Agresti and Coull adjustment (Tables 2-5) perform better than the original approach (A0) (Table 1). One method to examine coverage is to determine if it is within two standard deviations of the nominal confidence level. Here for 95% confidence intervals, the standard deviation is calculated to be $\sqrt{\frac{0.95(1-0.95)}{1000}}$. (Note that coverage probabilities should follow a binomial sampling distribution assuming the nominal coverage probability of 95% is correct.) Thus, a reasonable range of values for coverage is $0.950 \pm 2\sigma$ or $(0.936, 0.964)$. Below we report the minimum coverage, the maximum coverage and the mean coverage.

A0, proposed by [4] and seen in Table 1, had the worse coverage values of the five approaches. The coverage values for the A0 range from 0.866 to 0.963 and the mean coverage was 0.937. The data for A1-A3, seen in Tables 2-4, indicates that they all have approximately the same performance. A1 has a minimum coverage value, mean coverage value, and maximum coverage value of 0.907, 0.949, and 0.969, respectively. The same values for A2, given in the same order are 0.891, 0.948, and 0.967. These values given for A3 are 0.896, 0.946, and 0.965. The data for A4 shows that the coverage values were higher than the other four approaches. The range of values went from 0.907 for a minimum up to 0.985 for a max, which is well above the max coverage values of the other approaches. The mean of 0.957 is also high, but still better in comparison with the low mean of 0.937 that A0 generated.

Additionally, we summarize these results in Table 6. The percentages of coverage values falling below $0.95 - 2\sigma$ are given in column 2. Using this manner of evaluation, we see that A0 has the worst performance with 34.38% of coverages falling below the acceptable range. A2 and A4 performed the best each having 4.69% of values below the acceptable range. It is noteworthy that approximately 2.50% should fall below this range by chance alone.

Next, we compare the minimum coverage value, the mean coverage, and the maximum coverage value for each approach (columns 4-6 of Table 6). Looking at minimum coverage values, A0 has the lowest value with 0.866. The highest minimum coverage value belongs to A1 and A4 with 0.907. The mean coverage tells us a lot about overall performance of an estimation method. A1-A4 all perform very well, much better than the A0. They all have a mean close to 0.950. As can be seen from a max coverage of 0.985, A4 gives coverage values that are relatively high in comparison with the other approaches, but every augmented approach gives higher coverage values than the original unaugmented Beta-binomial approach.

On the whole, every augmented approach performed very well. The performances of A1-A3 were approximately the same. There are not any factors that

Table 5. Coverage for A4

	$n = 1000$ $m = 5$				$n = 1000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.985	0.980	0.984	0.952	0.976	0.979	0.968	0.907
0.004	0.965	0.966	0.963	0.951	0.970	0.961	0.958	0.927
0.008	0.956	0.958	0.956	0.943	0.966	0.968	0.950	0.937
0.010	0.970	0.965	0.954	0.943	0.950	0.958	0.964	0.953

	$n = 2000$ $m = 5$				$n = 2000$ $m = 10$			
	ρ				ρ			
π	0.001	0.01	0.1	0.4	0.001	0.01	0.1	0.4
0.002	0.959	0.968	0.969	0.954	0.963	0.963	0.961	0.936
0.004	0.957	0.953	0.958	0.934	0.965	0.957	0.956	0.939
0.008	0.957	0.963	0.940	0.950	0.962	0.960	0.961	0.947
0.010	0.945	0.945	0.954	0.947	0.960	0.971	0.959	0.946

clearly show one of these approaches performing significantly better than the others. A4 performed very well, but quite often had coverage values much greater than 0.950. When $\rho \leq 0.1$, this approach gives values very near to 0.950. One alternative in the future would be to construct a confidence interval approach that utilizes one of the augmented approaches A1-A3 when ρ is small and A4 when ρ is large.

Table 6. Overall summary of results

Approach	% below $0.95 - 2\sigma$	minimum coverage	mean coverage	maximum coverage
A0	34.38	0.866	0.937	0.963
A1	7.81	0.907	0.949	0.969
A2	4.69	0.891	0.948	0.967
A3	12.50	0.896	0.946	0.965
A4	4.69	0.907	0.957	0.985

Several areas for future research are suggested by this work. First a study of the property of estimators of the intra-cluster correlation to determine the reasons for the hook paradox seems warranted. Each of the four augmented ap-

proaches performed better than the original approach, so we are able to conclude that the changes proposed by [6] improve inference for the mean error rate from a Beta-binomial distribution. However, since our estimators performed poorly as ρ increased, we may want to consider adding observations as a function of ρ. One such solution might be to add $2 * (1 + (\bar{m} - 1)\rho)$ "successes" out of $4 * (1 + (\bar{m} - 1)\rho)$ attempts. Note that $1 + (\bar{m} - 1)\rho$ is known as the design effect or *deff* and reduces to 1 for the binomial.

Our evaluation of each approach shows that with any one of the augmentation methods used, proportion estimation improved significantly. Hence, headway has been made in the estimation of error rates incurred by biometric identification devices. Of the four augmented approaches, A4 had the best performance and A1-A3 performed approximately the same. Although many of the coverage values were outside the desired range for A4, most of them were above 0.950. Hence the coverage of 95% was not only reached, but often also surpassed. Though we prefer our coverage values to be nearly 0.950, we are willing to trade a slightly wider interval for increased coverage in this case. In the end A4 had the best performance in error rate estimation, but all of the methods proposed here represents an improvement over the unaugmented approach, A0.

Acknowlegements: This research was made possible by generous funding from the Ronald E. McNair Postbaccalaureate Acheivement Program at St. Lawrence University.

References

1. James L. Wayman. *Biometrics:Personal Identification in Networked Society*, chapter 17. Kluwer Academic Publishers, Boston, 1999.
2. Tony Mansfield and James L. Wayman. Best practices in testing and reporting performance of biometric devices. on the web at www.cesg.gov.uk/site/ast/biometrics/media/BestPractice.pdf, 2002.
3. Edwin P. Rood and Anil K. Jain. Biometric research agenda. Report of the NSF Workshop, 2003.
4. Michael E. Schuckers. Using the beta-binomial distribution to assess performance of a biometric identification device. *International Journal of Image and Graphics*, 3(3):523–529, 2003.
5. Michael E. Schuckers. Estimation and sample size calculations for matching performance of biometric identification devices. Center for Identification Technology Research technical report, 2003.
6. Alan Agresti and Brent A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.
7. William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, third edition, 1977.
8. Kung-Jong Lui, William G. Cumberland, and Lynn Kuo. An interval estimate for the intraclass correlation in beta-binomial sampling. *Biometrika*, 52(2):412–425, 1996.
9. Michael E. Schuckers, Anne Hawley, Katie Livingstone, Nona Mramba, and Collen Knickerbocker. A comparison of statistical methods for evaluating matching performance of a biometric identification device- a preliminary report. Center for Identification Technology Research technical report, 2004.