Estimation and sample size calculations for correlated binary error rates of biometric identification devices

Michael E. Schuckers,211 Valentine Hall, Department of Mathematics
Saint Lawrence University, Canton, NY 13617 and
Center for Identification Technology Research, West Virginia University
schuckers@stlawu.edu

November 14, 2003

KEYWORDS: False Accept Rate, False Reject Rate, Logit Transformation, Beta-binomial, Intra-cluster Correlation

1 Background

Biometric Identification Devices (BID's) compare a physiological measurement of an individual to a database of stored templates. The goal of any BID is to correctly match those two quantities. When the comparison between the measurement and the template is performed the result is either an "accept" or a "reject." For a variety of reasons, errors occur in this process. Consequently, false rejects and false accepts are made. As acknowledged in a variety of papers, e.g. Wayman (1999), Mansfield and Wayman (2002), there is a need for assessing the uncertainty in these error rates for a BID.

Despite the binary nature of the outcome from the matching process, it is well known that a binomial model is not appropriate for estimating false reject rates (FRRs) and false accept rates (FARs). This is also implicitly noted by Wayman (1999). Thus, other methods that do not depend on the binomial distribution are needed. Some recent work has made headway on this topic. In Bolle et al. (2000) the authors propose using resampling methods to approximate the cumulative density function. This methodology, the so-called "subset bootstrap" or "block bootstrap", has received some acceptance. Schuckers (2003) has proposed use of the Beta-binomial distribution which is a generalization of the binomial distribution. Like the "subset bootstrap", this method assumes conditional independence of the responses for each individual. Utilizing the Beta-binomial, Schuckers outlines a procedure for estimating an error rate when multiple users attempt to match multiple times.

One of the primary motivations for assessing sampling variability in FARs and FRRs is the need to determine the sample size necessary to estimate a given error rate to within a specified margin of error,e. g. Snedecor and Cochran (1995). Sample size calculations exist for basic sampling distributions such as the Gaussian and the binomial. Calculations of this kind are generally simpler for parametric methods such as the Beta-binomial than they are for non-parametric methods such as the "subset bootstrap."

In this paper we present two methodologies for confidence interval estimation and sample size calculations. We follow a general model formulated by Moore (1987) for dealing with data that exhibits larger variation than would be appropriate under a binomial sampling distribution. In the next section we introduce this model and the notation we will use throughout this paper. Section 3 introduces confidence interval and sample size estimation based on this model. There we present a simulation study to assess the performance of this model. Section 4 proposes a second methodology for confidence interval and sample size estimation. This one is based on a log-odds, or logit, transformation of the error rate. We also present results from a simulation study for this model. Finally, Section 5 is a discussion of the results presented here and their implications.

2 Extravariation Model

Several approaches to modelling the error rates from a BID have been developed. Here we take a parametric approach. Schuckers (2003) presented an extravariation model for estimating FARs and FRRs. That approach was based on the Beta-binomial distribution which assumes that error rates for each individual follow a Beta distribution. See Schuckers (2003) for further details. Here we follow Moore (1987) in assuming the first two moments of that model but not in the form of the distribution; that is, we no longer use the assumptions of the Beta distribution. We give details of this model below.

In order to further understand this model, it is necessary to define some notation. We assume an underlying error rate, either FAR or FRR, of π. Following Mansfield and Wayman (2002), let n be the number of individuals tested and let m_i be the number of times that the i^{th} individual is tested, $i = 1, 2, \ldots, n$. For simplicity we will assume that $m_i = m$, for any i. All of the methods given below have generalizations that allow for different numbers of tests per individual. Then for the i^{th} individual, let X_i represent the observed number of errors from the m attempts and let p_i represent the percentage of errors from the m observed attempts, $p_i = X_i/m$.

We assume here that:

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

where ρ is a term representing the degree of extravariation in the model. ρ is often referred to as the intra-class correlation coefficient, e. g. (Snedecor and Cochran, 1995). In Appendix A, we give examples of data with identical values of $\hat{\pi}$ and two different values of ρ to illustrate how ρ influences the data that is observed. For the rest of this paper we will use traditional statistical notation for estimates of parameters by denoting them with a "hat." For example, $\hat{\pi}$ will represent the estimate of π, the overall error rate.

3 Traditional confidence intervals

In the previous section, we introduced an notation for the extravariation model. Here we will use this model for estimating π. Suppose that we have the observed X_i's from a test of a biometric identification device. We can then use that data to estimate the parameters of our model. Here we will focus on two aspects of inference: confidence intervals and sample size calculations. Specifically, we derive confidence intervals for the error rate π and the sample size needed to achieve a specified level of confidence. Let

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

BMS and WMS represent between mean squares and within mean squares respectively. This estimation procedure for ρ is given by Lui et al. (1996).

Thus we have an estimate, $\hat{\pi}$ and we can evaluate it's standard error following equation 1 assuming that the individuals tested are conditionally independent of each other. That is, we assume the X_i's are conditionally independent. The standard error of $\hat{\pi}$ is then

$$\begin{aligned} \hat{V}[\hat{\pi}] &= \hat{V}[\frac{\sum X_i}{mn}] \\ &= (mn)^{-2} \sum \hat{V}[X_i] \\ &= \frac{\hat{\pi}(1-\hat{\pi})(1+(m-1)\hat{\rho})}{mn}. \qquad (2) \end{aligned}$$

(Note that in Schuckers (2003) they refer to a quantity C which here is $(1 + (m - 1)\rho)$.) Now we can create a nominally $100 \times (1 - \alpha)\%$ confidence interval for π from this. Assuming a Gaussian distribution for the sampling distribution of $\hat{\pi}$, we get the following interval

$$\hat{\pi} \pm z_{1-\frac{\alpha}{2}} \left[\frac{\hat{\pi}(1-\hat{\pi})(1+(m-1)\hat{\rho})}{mn} \right]^{1/2} \tag{3}$$

where $z_{1-\frac{\alpha}{2}}$ represents the $1 - \frac{\alpha}{2}^{th}$ percentile of a Gaussian distribution. Here we assume Normality of the sampling distribution of $\hat{\pi}$ is assumed. That assumption relies on the asymptotic properties of these estimates, (Moore, 1986). To test the appropriateness of this we simulated data from a variety of different scenarios. Under each scenario, 1000 data sets were generated and from each data set a 95% confidence interval was calculated. The percentage of times that π is captured in these intervals is investigated and referred to as coverage. For a 95% confidence interval, we should expect the coverage to be 95%. We consider full factorial simulation study using the following value: $n = (1000, 2000)$, $m = (5, 10)$, $\pi = (0.002, 0.004, 0.008, 0.010)$, and $\rho = (0.001, 0.01, 0.1, 0.4)$. These values were chosen to determine their impact on the coverage of the confidence intervals. Specifically, these values of π were chose to be representative of possible values for a BID, while the chosen values of ρ were chosen to represent a larger range than would be expected. Performance for all of the methods given in this paper is exemplary when π is between 0.1 and 0.9. Thus we investigate how well these methods do when those conditions are not met.

To generate data from each of these scenarios, we follow Oman and Zucker (2001) who recently presented a methodology for generating correlated binary data with the specific marginal characteristics that we have here. Thus this simulation makes no assumptions about the form of the data other than that each observation is generated to have probability of error π and intra-class correlation ρ. Results from these simulations can be found for $n = 1000$ and $n = 2000$ in Table 1 and Table 2, respectively.

Table 1: Empirical Coverage Probabilities for Confidence Interval, $n = 1000$

$m = 5$

$\pi \backslash \rho$	0.001	0.01	0.1	0.4
0.002	0.924	0.921	0.891	0.862
0.004	0.943	0.943	0.931	0.914
0.008	0.940	0.947	0.951	0.931
0.010	0.952	0.954	0.947	0.928

$m = 10$

$\pi \backslash \rho$	0.001	0.01	0.1	0.4
0.002	0.944	0.950	0.908	0.855
0.004	0.942	0.948	0.916	0.904
0.008	0.947	0.941	0.941	0.928
0.010	0.953	0.966	0.937	0.936

Each line represents 1000 simulated data sets.

Before summarizing the results from these tables, it is appropriate to make some remarks concerning interpretation of these results. First, since we are dealing with simulations, we should pay attention to overall trends rather than to specific outcomes. If we ran these same simulations again, we would see slight changes in the coverages of individual scenarios but the overall trends should remain. Second, this way of evaluating how well a methodology, in this case a confidence interval, does is appropriate. Though it may seem artificial, using simulated data, is the best way to evaluate an applied confidence interval, since one can know how often the parameter, in this case π, is captured by the confidence interval. Observed test data does not give an accurate guide to how well a method performs since there is no way to know the actual error rate, π.

Turning our attention to the results in Table 1 and Table 2, several clear patterns emerge. Coverage increases as π increases, as ρ decreases, as n increases and as m increases. This is exactly as we would have expected. More observations should increase our ability to accurately estimate π. Similarly the assumption of Normality will be most appropriate when π is moderate (far from zero and far from one) and

Table 2: Empirical Coverage Probabilities for Confidence Interval, $n = 2000$

$m = 5$

$\pi \backslash \rho$	0.001	0.01	0.1	0.4
0.002	0.940	0.943	0.929	0.904
0.004	0.954	0.945	0.932	0.919
0.008	0.957	0.942	0.942	0.936
0.010	0.950	0.948	0.927	0.924

$m = 10$

$\pi \backslash \rho$	0.001	0.01	0.1	0.4
0.002	0.947	0.927	0.928	0.889
0.004	0.946	0.950	0.927	0.914
0.008	0.933	0.957	0.931	0.915
0.010	0.958	0.946	0.927	0.945

Each line represents 1000 simulated data sets.

when ρ is small. The confidence interval performs well when $\pi > 0.002$ and $\rho < 0.1$. There is quite a range of coverages from a high of 0.966 to a low of 0.855. One way to think about ρ is that it governs how much 'independent' observations can be found in the data. Higher values of ρ indicate that there is less 'independent' information in the data. This is not surprising since data of this kind will be difficult to assess because of the high degree of correlation within an individual. This methodology would be acceptable for very large samples when the error rate, π is far from zero. However, this is often not the case for BID's. The difficulty with inference using the above methodology is that the assumption of Normality for the sample distribution of $\hat{\pi}$ is only approximate when π is moderate or when ρ is large or both. Below we present a remedy for this.

4 Transformed Confidence Intervals

As evidenced above one of the traditional difficulties with estimation of proportions near zero or one is that sampling distribution are only asymptotically Normal. One statistical method that has been used to compensate for this is to transform the calculations to another scale. Many transformations for proportions have been proposed including the logit, probit and arcsin of the square root. See Agresti (1990) for full details of these methods. Below we use the logit or log-odds transformation to create confidence intervals for the error rate π.

4.1 Logit function

The logit or log-odds transformation is one of the most commonly used transformations in Statistics. Define $\text{logit}(\pi) = log(\frac{\pi}{1-\pi})$. This is the logarithm of the odds of an error occurring. The *logit* function has a domain of $(0, 1)$ and a range of $(-\infty, \infty)$. One advantage of using the *logit* transformation is that we move from a bounded parameter space to on unbounded one. Thus, the possibility exists that the transformed sampling distribution should be closer to Normality than the untransformed one.

4.2 Confidence intervals

Our approach here is this. We transform our estimand π, create a confidence interval for the transformed quantity, then invert the transformation back to the original scale. We do this in the following manner. Let $\hat{\gamma} = logit(\hat{\pi})$, $ilogit(\gamma) = \frac{e^{\gamma}}{1+e^{\gamma}}$ and note the *ilogit* is the inverse of the *logit* function. Thus we can create a $100(1-\alpha)\%$ confidence interval using $\hat{\gamma}$. To do this we use a Delta method expansion for estimated the standard error of $\hat{\gamma}$. (The Delta method, as it is known in the statistical literature, is simply a one step Taylor series expansion. See Agresti (1990) for details.) Then our confidence interval on the transformed scale is

$$\hat{\gamma} \pm z_{1-\frac{\alpha}{2}} \left(\frac{1 + (m-1)\hat{\rho}}{\hat{\pi}(1-\hat{\pi})mn} \right)^{\frac{1}{2}} \tag{4}$$

Table 3: Empirical Coverage Probabilities for Logit Confidence Interval, $n = 1000$

$m = 5$

$\pi\backslash\rho$	0.001	0.01	0.1	0.4
0.002	0.964	0.951	0.964	0.907
0.004	0.961	0.952	0.945	0.926
0.008	0.951	0.960	0.955	0.954
0.010	0.958	0.959	0.952	0.954

$m = 10$

$\pi\backslash\rho$	0.001	0.01	0.1	0.4
0.002	0.941	0.930	0.929	0.839
0.004	0.942	0.933	0.942	0.923
0.008	0.947	0.963	0.928	0.939
0.010	0.955	0.963	0.944	0.941

Each line represents 1000 simulated data sets.

Table 4: Empirical Coverage Probabilities for Logit Confidence Interval, $n = 2000$

$m = 5$

$\pi\backslash\rho$	0.001	0.01	0.1	0.4
0.002	0.924	0.921	0.891	0.862
0.004	0.943	0.943	0.931	0.914
0.008	0.940	0.947	0.951	0.931
0.010	0.952	0.954	0.947	0.928

$m = 10$

$\pi\backslash\rho$	0.001	0.01	0.1	0.4
0.002	0.944	0.950	0.908	0.855
0.004	0.942	0.948	0.916	0.904
0.008	0.947	0.941	0.941	0.928
0.010	0.953	0.966	0.937	0.936

Each line represents 1000 simulated data sets.

We will refer to the endpoints of this interval as L and U for lower and upper respectively. The final step for making a confidence interval for π is to take the *ilogit* of both endpoints of this interval. Hence we get (*ilogit*(L),*ilogit*(U)).

The interval (*ilogit*(L),*ilogit*(U)) is asymmetric because the *logit* is not a linear transformation. Thus we do not have a traditional confidence interval that is plus or minus a margin of error. However, this interval has the same properties as other confidence intervals. To assess how well this interval performs we repeated the simulation done in section 3. Output from these simulations are summarized in Table 3 and Table 4 which contain output when $n = 1000$ and when $n = 2000$ respectively. Again coverage should be close to 95% for a 95% confidence interval. Looking at the results found in Tables 3 and 4, we note that there are very similar patterns to those found in the previous section. As before we are interested in trends rather than specific values. In general, coverage increases as π increases, as ρ decreases, as m increases and as n increases. Coverages range from a high of 0.965 to a low of 0.907. Coverage was only poor when $\pi = 0.002$ and $\rho = 0.4$. Otherwise the confidence interval based on a *logit* transformation performed quite well. Overall, coverage for the *logit* confidence interval (LCI) is higher than for the untransformed confidence interval (UCI). One possible remedy for increasing coverage slightly, when $\rho > 0.1$, would be to use a percentile from the Students t-distribution rather than a percentile from the Gaussian distribution.

4.3 Sample size and power calculations

One important statistical tool that is not found in the BID literature is sample size calculations. This is an extremely important tool for testing of a BID. Since the transformation to the *logit* scale give approximately correct coverage, we can use the confidence interval given in Equation 4 to determine approximate sample size calculations. The asymmetry of the *logit* interval provides a challenge relative to the typical sample size calculation. Here rather than specifying the margin of error as is typical, for example see Snedecor and Cochran (1995), we will specify the maximum value for the confidence interval. Given the nature of BID's where lower error rates are better, it seems somewhat natural to specify the highest acceptable value for the range of the interval.

Given equation 4, we can determine the appropriate sample size needed to estimate π with a certain level of confidence, $1-\alpha$, to be a specified maximum value, π_{max}. The difficulty with calculating a sample size is that there are actually two sample size components to BID testing: the number of individuals and the number of times each individual is tested. This makes a single universal formula for the sample size intractable. However, we can provide a conditional solution. We follow the following steps. First, specify appropriate values for π, ρ, π_{max}, and $1-\alpha$. As with any sample size calculations these values can be based on prior knowledge, previous studies or simply guesses. Second, fix m, the number of attempts per person. Third solve for n, the number of individuals to be tested. Find n via the following equation, given the other quantities:

$$n = \left(\frac{z^2_{1-\frac{\alpha}{2}}}{logit(\pi_{max}) - logit(\pi)} \right)^2 \frac{m\pi(1-\pi)}{(1+(m-1)\rho)}. \tag{5}$$

The above follows directly from Equation 4. For example, suppose that you wanted to estimate π to a max of $\epsilon = 0.01$ with 99% confidence. You estimate π to be 0.005 and ρ to be 0.01. If we plan on testing each individual 5 times then we would need to test, then

$$\begin{aligned} n &= \left(\frac{2.576}{logit(0.01) - logit(0.005)} \right)^2 \frac{5(0.005)(0.995)}{(1+(5-1)0.01)} \\ &= 569.14. \end{aligned} \tag{6}$$

So we would need to test 570 individuals 5 times each to achieve the specified level of confidence.

5 Discussion

For any BID, its matching performance is often critical to the success of the product from the viewpoint. The error rates, the false accept and the false reject, are crucial to this assessment. Because of variability in any estimate, it is important to be able to create inferential intervals for these error rates. In this paper we have presented two methodologies for creating a confidence interval for an error rate. The LCI, Equation 4 above, had superior performance for the simulation study described here. Though that study presented results only for 95% confidence intervals, it is reasonable to assume performance will be similar for other confidence levels. One possible improvement of coverage would be to use the appropriate percentile from a Student's t-distribution with m degrees of freedom rather than from the Gaussian distribution. For example, if we were interested in making a 95% confidence interval when $m = 5$, then we would use $t = 4.773$ in place of the z in Equations 4 and 5. Another possibility would be to use a simple function of m such as $2m$. More research is needed to understand this topic fully. This value would be conservative in that it should give coverage (and hence confidence intervals) that is larger than necessary. (An initial simulation of 1000 data sets using this methodology when $\pi = 0.002$, $\rho = 0.4$, $n = 1000$, and $m = 5$ gives coverage of 0.994.)

The most immediate consequence of the performance of the LCI is that we can 'invert' that calculation to get sample size calculations. Because of the asymmetry of this confidence interval, it is necessary to estimate the parameters of the model, π and ρ, as well as to determine the upper bound for the confidence interval. As outlined above, this now gives BID testers an important tool for determining the number of individuals and the number of attempts per individual necessary to create a confidence interval. Further work is necessary here to develop methodology that will provide power calculations and further refine this work.

In this paper we have provided simple, efficient approximations based on appropriate models that perform quite well. These are valuable initial steps. Further work is necessary to be able to compare the performance of this methodology to methodologies already in use. In addition, it is important to advance our understanding of how this model performs.

Acknowlegements: This research was made possible by generous funding from the Center for Identification Technology Research (CITeR) at the West Virginia University. The author 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.

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons, New York.

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.

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/biometrics.

Moore, D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika*, 73(3):583–588.

Moore, D. F. (1987). Modeling the extraneous variance in the presence of extra-binomial variation. *Applied Statistics*, 36(1):8–14.

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

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

Snedecor, G. W. and Cochran, W. G. (1995). *Statistical Methods*. Iowa State University Press, 8^{th} edition.

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