Test Sample and Size

Michael E. Schuckers[1]

St. Lawrence University, Canton, NY 13617, USA
schuckers@stlawu.edu

Synonyms

Sample Size; Crew designs

Definition

The testing and evaluation of biometrics is a complex and difficult task. The difficulties in such an endeavor include the selection of the number and type of individuals that will that will participate in this process of testing. Determining the amount of data to be collected is another important factor in this process. Choosing an appropriate set of individuals from which to collect biometrics data is another important aspect of testing a biometrics system.

Main Body Text

Introduction

The assessment of a biometric system's matching performance is an important part of evaluating such a system. A biometric implementation is an ongoing process and as such will be treated as a process in the sense of Hahn and Meeker[1]. Thus, any inference regarding that process will be analytic in nature rather than enumerative as delineated by Deming [2]:

> An enumerative study has for its aim an estimate of the number of units of a frame that belong to a specified class. An analytic study has for its aim a basis for action on the cause-system or the process, in order to improve product of the future.

Here focus is on determining the amount and type of data necessary for assessing the current matching performance of a biometrics system.

The matching performance measures that are commonly considered most important are the false match and the false non-match rate, FMR and FNMR, respectively. One of the important parts of designing a test of a biometrics system is to determine, prior to completion, the amount of testing that will be done. Below calculations that explicitly allow for determining the amount of biometric data which will be sampled are described. As with any calculations of this kind it is necessary to make some estimates about the nature of process beforehand. Without these, it is not possible to determine the amount of data to collect. These sample size calculations will be derived to achieve a certain level of sampling variability. It is important to recognize that there are other potential sources of variability in any data collection process.

Selection of the individuals from whom these images will be taken is another difficult undertaking because of the need to ensure that the images taken are representative of the matching and decision making process. The goal of any data collection should be to take a sample that is as representative as possible of the process about which inference will be made. Ideally, some probabilistic mechanism should be utilized to select individuals from a targeted population. In reality, because of limitations of time and cost, this is a difficult undertaking and often results in a convenience sample, Hahn and Meeker [1]

Test size calculation

Determining the amount of biometric information to collect is an ongoing concern for the evaluation of a biometrics device. Several early attempts to address this problem include those by Wayman [3] and [4] as well as the description in Mansfield and Wayman [5] of the "Rule of 3" and the "Rule of 30". The former is due to several authors including Louis [6] as well as Jovanovic and Levy [7], while the latter, the so-called Doddington's Rule, is due to Doddington *et al* [8]. Mansfield and Wayman note that **neither** of these approaches is satisfactory since

> they assume that error rates are due to a single source of variability, which is not generally the case with biometrics. Ten enrolment-test sample pairs from each of a hundred people is not statistically equivalent to a single enrolment-test sample pair from each of a thousand people, and will not deliver the same level of certainty in the results.

Effectively, the use of either the "Rule of 3" or the "Rule of 30" requires the assumption that the decisions used to estimate error rates are uncorrelated. More recently, Schuckers [9] provided a method for dealing with the issue of the dual sources of variability and the resulting correlations that arise from this structure.

The calculation given below is for the determination of the number of comparison pairs, n, from which samples need to be taken. Define a comparison pair, similar to the *enrolment-test sample pair* of Mansfield and Wayman [5], as a pair of possibly identical individuals from whom biometric data or images have been taken and compared. If the two individuals are the same then call the comparison pair a genuine one. If the two individuals are distinct then call the comparison pair an imposter one. In order to use this information to determine test size, it is necessary to specify some estimates of the process parameters before the data collection is complete. In order to obtain sample size calculations it is necessary to make these specifications. It is worthwhile noting here that most other biological and medical disciplines use such calculations on a regular basis and the U.S. Food and Drug Administration requires them for clinical trials. Approaches to carrying this out are discussed below.

Let the error rate of interest, either FMR or FNMR, for a process be represented by γ and let Y_{ij} represent the decision for the j^{th} pair of captures collected on the i^{th} comparison pair, where n is the number of comparison pairs, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Thus, the number of decisions that are made for the i^{th} comparison pair is m_i, and n is the number of different comparison pairs being compared. Define

$$Y_{ij} = \begin{cases} 1 \text{ if } j^{th} \text{ decision from comparison pair } i \text{ is incorrect} \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

Assume for the Y_{ij}'s that $E[Y_{ij}] = \gamma$ and $V[Y_{ij}] = \gamma(1-\gamma)$ where $E[X]$ and $V[X]$ represent the mean and variance of X, respectively. Estimation of γ is done separately for FNMR and FMR and so there is a seperate collection of Y_{ij}'s for each. The form of the variance is a result of each decision being binary. The correlation structure for the $Y_{ij}^{\prime}s$ is

$$Corr(Y_{ij}, Y_{i'j'}) = \begin{cases} 1 \text{ if } i = i', j = j' \\ \rho \text{ if } i = i', j \neq j' \\ 0 \quad otherwise \end{cases} \tag{2}$$

This correlation structure is based upon the idea that there will only be correlations between decisions made on the comparison pair but not between decisions made on different comparison pairs. Thus, conditional upon the error rate, there is no correlation between decisions on the i^{th} comparison pair and decisions on the $i'^{\ th}$ comparison pair, when $i \neq i'$. The degree of correlation is summarized by ρ. This is not the typical Pearson's correlation coefficient, rather it is the intra-class correlation or here the intra-comparison pair correlation. More details can be found in Schuckers [10].

Derivation of sample size calculations requires an understanding of sampling variability in the estimated error rate. Thus consider

$$\hat{V}[\hat{\gamma}] = N^{-2}\hat{\gamma}(1-\hat{\gamma}) \left[N + \hat{\rho} \sum_{i=1}^{n} m_i(m_i - 1) \right] \tag{3}$$

where $N = \sum_{i=1}^{n} m_i$, and $\hat{\gamma} = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} Y_{ij}$. Fleiss et al. [11] has suggested the following moment-based estimator for ρ

$$\hat{\rho} = \left(\hat{\gamma}(1-\gamma) \sum_{i=1}^{n} m_i(m_i - 1) \right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{\substack{j'=1 \\ j' \neq j}}^{m_i} (Y_{ij} - \hat{\gamma})(Y_{ij'} - \hat{\gamma}). \tag{4}$$

Since $\hat{\gamma}$ is a linear combination, if n is large it is reasonable to assume that the central limit theorem holds, Serfling [12]. To produce a $(1-\alpha) \times 100\%$ confidence interval for γ use

$$\hat{\gamma} \pm z_{\alpha/2}\sqrt{N^{-2}\hat{\gamma}(1-\hat{\gamma})\left[N+\hat{\rho}\sum_{i=1}^{n} m_i(m_i-1)\right]} \tag{5}$$

where $z_{\alpha/2}$ represents the $1-\alpha/2^{th}$ percentile of a Gaussian distribution with mean 0 and variance 1. Further, if $m_i = m$ for all i Equation (3) simplifies to

$$V[\hat{\gamma}] = (nm)^{-1}\gamma(1-\gamma)\left[1+\rho(m-1)\right] \tag{6}$$

where N has been replaced by nm. This form will be used to derive sample size calculations.

Turning from variance estimation to sample size calculations, set the portion of Equation (6) after the $\pm$, the margin of error, equal to some desired value B and solve for n, the number of comparison pairs. Then the following sample size calculation for making a $100(1-\alpha)\%$ CI with a specified margin of error of B is obtained.

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}\gamma(1-\gamma)(1+(m-1)\rho)}{mB^2} \right\rceil \tag{7}$$

where $\lceil\ \rceil$ is the next largest integar or ceiling function. In order to create sample size calculations for a confidence interval, it is necessary to specify, among other things, the desired margin of error, B, for the interval. As mentioned above there are effectively two sample sizes when dealing with performance evaluation for biometric authentication devices. This derivation here is for the number of comparison pairs, n, that need to be tested and assume that the number of decisions per individual is fixed and known. This is equivalent to assuming that $m_i = m$ for all i and that m is known. In practice it will be possible to determine different values for n by varying m before proceeding with a evaluation. As with all sample size calculations it is important to note that specification of *a priori* values for the parameters in the model is necessary. In this case that means it is necessary to estimate values for γ and ρ to be able to determine the number of individuals, n. Several strategies are reasonable and have been discussed in the statistics literature for these *a priori* specifications. See, e.g., Lohr [13]. Ideally, it would be possible to make a pilot study of the process under consideration and use actual process data to estimate these quantities. Alternatively, it may be possible to use estimates from other studies perhaps done under similar circumstances or with similar devices. The last possibility is to approximate based upon prior knowledge without data. Regardless of the method used it is important to recognize that n is a function of α, B, m, γ and ρ. n varies directly with γ and ρ and inversely with α, m and B. Thus, a conservative approach to estimation of these quantities would overestimate γ and ρ and underestimate m. This will produce a value for n that is likely to be larger than required. Table 1 illustrates the use of Equation (7) It is also worth noting that most studies of this type have a not insignificant drop out rate of individuals as the data collection progresses. Thus it is adviseable to plan a collection process that assumes some attrition in the number of comparison pairs to be selected. The values of α and B are likely to be set by investigators or by standards bodies rather than the performance of the process under study.

Table 1. Illustration of the use of Equation (7)

α	B	γ	m	ρ	n
0.05	0.005	0.01	10	0.4	700
0.05	0.01	0.01	10	0.4	175
0.01	0.005	0.01	10	0.4	1209
0.05	0.005	0.02	10	0.4	1386
0.05	0.005	0.01	5	0.4	792
0.05	0.005	0.01	10	0.1	290

Equation (7) is straightforward for calculation of the number of comparison pairs that need to be tested when γ=FNMR. It is less when interest centers on γ=FMR. This is because for FNMR the number of comparison pairs translates to the number of individuals, while for FMR the number of comparison pairs is not proportional to the number of individuals. If all cross-comparisons are use to estimate FMR, then one can replace n with $n^*(n^*-1)$ in Equation (7). In that case n^* will be the number of individuals that need to be tested.

Sample selection

Once the number of individuals to be selected is determined another important step is to specify the target population of individuals to whom statistical inference will be made. Having done so, a sample would ideally be drawn from that group. However, this is often not possible. The next course of action is to specify a sample that is as demographically similar to the target population as possible. The group of individuals that will compose the sample is often referred to as the "volunteer crew" or simply the "sample crew", Mansfield and Wayman [5]. The more similar the sample crew is to the target population the more probable it will be that the estimates based upon the sample crew will be applicable to the target population. Often the sample crew is chosen to be a convenience sample, Hahn and Meeker [1]. Methodology for best selecting the sample crew is an open area of research in biometrics.

One useful tool for extrapolation from estimates based upon the "crew" is post-stratification. Poststratification is a statistical tool for weighting a samples representation after the sample has been taken so that resulting estimates reflect the known population. Suppose that there are H non-overlapping demographic groups of interest, or strata, and n_h individuals have been sampled from among the N_h total individuals in each strata. Further suppose that estimates of the error rate, $\hat{\gamma}_h$, from each of the strata are known. Then a poststratified estimate of the error rate is

$$\hat{\gamma}_{ps} = \sum_{h=1}^{H} \frac{n_h}{N_h} \hat{\gamma}_h. \tag{8}$$

An estimate of the variability of the predicted error rate is

$$\hat{V}[\hat{\gamma}_{ps}] = \sum_{h=1}^{H} \left(\frac{n_h}{N_h} \right)^2 \hat{V}[\hat{\gamma}_h] \tag{9}$$

where $\hat{V}[\hat{\gamma}_h]$ can be calculated using the equation found above. A $(1-\alpha) \times 100\%$ *poststratification* confidence interval for the process error rate can then be made using

$$\hat{\gamma}_{ps} \pm z_{\alpha/2} \sqrt{\hat{V}[\hat{\gamma}_{ps}]}. \tag{10}$$

As above, use of the Gaussian distribution here is justified by the fact that the estimated error rate, $\hat{\gamma}_{ps}$, is a linear combination of random variables.

Summary

Testing and evaluation of biometric devices is a difficult undertaking. Two crucial elements of this process are the selection of the number of individuals from whom to collect data and the selection of those individuals. Determining the number of individuals to test can be calculated based on Equation (7). To obtain the number of individuals that need to be tested, some process quantities need to be specified. These specification can be based on previous studies, pilot studies or on qualified approximations. Selection of the "crew" for a study is a difficult process. Ideally a sample from the target population is best, but a demographically similar "crew" is often more attainable. The inference from a demographically similar crew can be improved by use of poststratification.

Related Entries

Performance Evaluation (Overview), Performance Measures, Performance Testing Methodology Standardization, Influential Factors to Performance

References

1. Hahn, G.J., Meeker, W.Q.: Statistical Intervals: A Guide for Practioners. John Wiley & Sons (1991)
2. Deming, W.E.: On probability as a basis for action. The American Statistician **29**(4) (1975) 146–152
3. Wayman, J.L.: Confidence interval and test size estimation for biometric data. In: National Biometrics Test Center, Collected Works 1997-2000. www.engr.sjsu.edu/biometrics/nbtccw.pdf (2000) 89–99

4. Wayman, J.L.: Confidence interval and test size estimation for biometric data. In: Proceedings of IEEE AutoID ’99. (1999) 177–184
5. Mansfield, T., Wayman, J.L.: 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)
6. Louis, T.A.: Confidence intervals for a binomial parameter after observing no successes. The American Statistician **35**(3) (1981) 154
7. Jovanovic, B. D., Levy, P. S.: A look at the rule of three. The American Statistician **51**(2) (1997) 137–139
8. Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A.: The NIST speaker recognition evaluation: overview methodology, systems, results, perspective. Speech Communication **31**(2-3) (2000) 225–254
9. Schuckers, M.E., Sheldon, E., Hartson, H.: When enough is enough: early stopping of biometrics error rate testing. In: Proceedings of the IEEE Workshop on Automatic Identification Advanced Technologies (AutoID). (2007)
10. Schuckers, M.E.: Estimation and sample size calculations for correlated binary error rates of biometric identification rates. In: Proceedings of the American Statistical Association: Biometrics Section [CD-ROM], Alexandria, VA, American Statistical Association (2003)
11. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. John Wiley & Sons, Inc. (2003)
12. Serfling, R.J.: Contributions to central limit theory for dependent variables. The Annals of Mathematical Statistics **39**(4) (1968) 1158–1175
13. Lohr, S.L.: Sampling: Design and Analysis. Duxbury Press (1999)

Definitional Entries

Analytic Study

An analytic study is one where the goal is the utilization of the information gathered for improvement of the process going forward. This is in contrast to an enumerative study which

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for some parameter θ is a range of values (L, U) such that $P(\theta \in (L, U)) = 1-\alpha$ where L and U are random variables.

Volunteer Crew

The volunteer crew for a biometrics test is the individuals that participate in the evaluation of the biometric and from whom biometric samples are taken.

Poststratification

Poststratification is a statistical technique that forms strata of observations after the data has been collected to better inform statistical inference.

Convenience Sample

A convenience sample is a sample that uses individuals or sample units that are readily available rather than those that are selected to be representative or selected via a probabilistic mechanism.