Using statistical models in estimating error rates for forensic firearm identification

N F Zhang
National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD 20899, USA
E-mail: nien-fan.zhang@nist.gov

Abstract. Estimating error rates for firearm evidence identification is a fundamental challenge in forensic science. The recently developed Congruent Matching Cells (CMC) method provides applications to firearm evidence identification. To estimate error rates, appropriate statistical models are needed for the CMC values. In this paper, in addition to the binomial probability distribution the correlated binomial distribution is proposed. For an image comparison, the correlated binomial distribution can be applied to the cell pairs from CMC method. An application to an actual data set demonstrates that the correlated binomial distribution fits the relative frequency distribution of CMC values much better than the binomial distribution.

1. Introduction

In firearm evidence identification, when bullets and cartridge cases are fired or ejected from a firearm, the parts of the firearm create characteristic tool marks called ballistic signatures. In general, tool marks have so called “class characteristics” that are common to certain brands of firearms and individual characteristics arising from random variation in firearm manufacturing. Recently, a quantitative approach known as the Congruent Matching Cells (CMC) method was developed to improve the accuracy of ballistic identifications and to provide a base for estimating error rates [1]. This paper proposes statistical models and the corresponding methodology for estimating the model parameter and error rates. In Section 2, the CMC method is described. In Section 3, the correlated binomial distribution based on dependent Bernoulli trials is introduced. In Section 4, maximum likelihood estimators of the parameters of correlated binomial distribution is proposed and applied to an actual data set followed by the conclusions.

2. CMC methods for ballistic identification and statistical models

The CMC method deals with pairs of measured 2D or 3D topography images of breech face impressions whose similarity we wish to quantify. The CMC method divides each image into a rectangular array of cells. For each cell, a search for a matching cell is then made on a compared image [2]. Figure 1 shows the correlated and uncorrelated cell pairs in an image pair. The cell-by-cell analysis is done to determine whether each cell pair is a correlated call pair or not.

1 To whom any correspondence should be addressed.
Figure. 1. Schematic diagram of topographies A and B originating from the same firearm and registered at the position of maximum correlation. The six solid cell pairs in each image are located in three valid correlated regions \((A_1, B_1), (A_2, B_2),\) and \((A_3, B_3)\). The dotted cell pairs \((a', b'), (a'', b''),\) and \((a''', b''')\) are located in the invalid correlation region.

Congruent matching cell pairs, or CMCs are determined by three sets of identification parameters for quantifying both the topography similarity of the correlated cell pairs and the pattern congruency of the cell distributions. From a statistical point of view, the CMC method is based on pass-or-fail tests of individual cell pairs comprising an image pair of breech face impression. For a pair of images of breech face impressions, \(N\) represents the number of correlated cell pairs in the image pair. For a given correlated cell pair, a random variable \(X\) represents the outcome of the CMC method applied to the correlated cell pair. When the CMC method determines that the cell pair is a congruent matching cell pair, e.g., \((A_i, B_i)\), then \(X = 1\); otherwise \(X = 0\). Symbol \(P\) represents probability in general and the symbol \(p\) represents the probability that \(X = 1\). That is \(P(X = 1) = p\), and \(P(X = 0) = 1 - p\).

An approach was developed in [2] for estimating the expected error rates of ballistic identifications based on the CMC method. Error rates are discussed in detail in [2]. To estimate error rates, a key is to find the best probability distribution for the relative frequency distribution of the observed CMC values.

3. The binomial distribution based on dependent Bernoulli trials

In [2], for CMC method, the random variable \(X\) is assumed to be a Bernoulli trial. Namely, the trials results \(\{X_1, ..., X_N\}\) are \(N\) dichotomous items. Namely, the comparisons between cell pairs are independent from each other and with a common probability \(p\). Denote the sum of the CMC values for the comparisons of the first image pair by \(Y_1\) with \(N\) cell pairs. \(Y_1 = \sum_{i=1}^{N} X_{i1}\). In probability, \(Y_1 - Bin(N, p)\) is a binomial distributed random variable (r.v.) with the probability mass function
\[
P_{1_{Y}}(Y = k) = C_k^N p^k (1 - p)^{N-k} \quad \text{for} \quad k = 0, 1, ..., N.
\]

Similarly, for the \(M\) image pairs, we have \(Y_1, ..., Y_M\) correspondingly. When \(\{Y_j, j = 1, ..., M\}\) are independent from each other, we have a sequence of independently binomial distributed r.v.’s. That is, \(Y_j = \sum_{i=1}^{N} X_{ji}\) for \(j = 1, ..., M\) and \(Y_j - Bin(N, p)\). For observed values from the CMC method, \(\{y_j, j = 1, ..., M\}\), the maximum likelihood estimator of \(p\) is given by \(\hat{p} = \frac{\sum_{j=1}^{M} y_j}{MN}\).

However, the assumption of independence among cell pair comparisons may be invalid. For various reasons, for example, the physical similarity between the cell pairs may lead that the comparisons
considered are not be statistically independent with each other in general. In addition, two cell pairs may have a duplicate cell, for example, \((A_1, B_1)\) vs. \((A_2, B_4)\). In this case, we consider a model for dependent Bernoulli trials proposed by Bahadur [3], which sometimes is called Bahadur-Lazarsfed model. Similar to the Bernoulli trials, for a sequence of \(\{X_1,...,X_N\}\), with each \(X_i\) equal to 0 or 1 with \(P(X = 1) = p\), and \(P(X = 0) = 1 - p\) for \(i = 1,...,N\). However, \(\{X_1,...,X_N\}\) may not be mutually independent. We define the second order correlation between \(X_i\) and \(X_j\) \((i,j = 1,...,N)\) by

\[
    r_{ij} = \frac{\text{Cov}[X_i, X_j]}{\sigma_i^2} = \frac{E[(X_i - \mu)(X_j - \mu)]}{\sigma_i^2},
\]

where \(\mu\) is the marginal mean and \(\sigma\) is the marginal standard deviation of \(X_i\). Higher order correlations are similar. For simplicity, we only consider the second order correlation and assume that the correlations are symmetric [3], which means \(r_{ij}\) \((i,j = 1,...,N)\) or in (2) are the same and denoted by \(r(2)\). In this case. The probability mass function of the sum of \(\{X_1,...,X_N\}\) denoted by \(Y\) can be approximated by

\[
    P_{[2]}(Y) = P_{[1]}(Y)[1 + r(2)g_2(Y, p)]
\]

where \(P_{[1]}(Y)\) is the probability mass function when \(\{X_j, i = 1,...,N\}\) is a sequence of independent Bernoulli trials shown in (1), \(r(2)\) is the second order correlation, and the function \(g_2(Y, p)\) is a second order polynomial in \(Y\). In this case, we say the r.v. \(Y\) has a correlated binomial distribution. Figure 2 shows the probability mass functions of \(P_{[1]}(Y)\) and \(P_{[2]}(Y)\) with \(N = 26\), \(p = 0.6\), and \(r(2) = 0.02\). The two r.v.’s have the same mean = 15.6 while the r.v. with a correlated binomial distribution has a variance = 9.36, which is larger than the variance of 6.24 for the r.v. with the binomial distribution.

![Figure 2. Probability mass functions for binomial (blue) and correlated binomial distribution (red) with \(N = 26\).](image)

4. Estimating the parameters of the correlated binomial distribution

When the CMC method applies to a set of cartridge cases, the result, in general, includes certain known matching (KM) image pair comparisons and certain known non-matching (KNM) image pair comparisons. In [2], Statistical models are fitted to the cases of KM and KNM, respectively. In either case, we assume that for \(M\) image pairs, the random variables for the sums of the CMC values for each image comparison are denoted by \(Y_{1},...,Y_M\). As discussed in Section 3, we assume that \(Y_{1},...,Y_M\) are independent from each other while for each image comparison, we have a sequence of \(N\) dependent
Bernoulli trials. Maximum likelihood is used to estimate the parameters of the correlated binomial distribution. The likelihood function for given $p$ and $r(2)$ is given by

$$L = \prod_{i=1}^{M} p_{i(2)}(y_i | p, r(2)) = \prod_{i=1}^{M} C_N^{y_i} p^{y_i} (1 - p)^{N-y_i} [1 + r(2) g_2(y_i, p)]. \quad (4)$$

The maximum likelihood estimator (MLE) of $p$ and $r(2)$ are obtained when the respected $\log(L)$ reaches the maximum. We evaluated the models on a set of cartridge cases created by Weller et al. [4]. The cartridge cases were obtained from a set of eleven slides produced by the same manufacturer. The data set includes 370 KM image pairs. For illustration, based on the KM data set, the MLE of the parameters of the correlated binomial distribution are obtained. For comparison, the parameter $p$ for the binomial distribution is also obtained. Figure 3 shows the relative frequency distribution of the observed CMC numbers and the probability mass functions for the binomial and correlated binomial distributions with the corresponding estimates of the parameters, respectively. It is clear that the correlated binomial is a much better fit to the CMC values than that for the binomial distribution.

![Figure 3. Relative frequency distribution in blue of CMC numbers for KM image pairs and the green and red curves represent the binomial and correlated binomial distributions for the KM data.](image)

5. Conclusions

Estimating error rates is an important part for firearm identifications. To evaluate error rates, the key is to determine the appropriate statistical model for the CMC values for image comparisons. The proposed correlated binomial distribution is reasonable for the actual image comparisons and demonstrates a good fit to the CMC data.

References

[1] Song J 2015 Proposed “Congruent matching cells (CMC)” method for ballistic identification and error rate estimation. *AFTE J.* 47 (3) 177-85.

[2] Song J, Vorburger T V, Chu W, Yen J, Soons J A, Ott D B, and Zhang N F 2017 Estimating error rates for firearm evidence identification in forensic science. *Forensic Science International* (284) 15-32.

[3] Bahadur R R 1961 A representation of the joint distribution of the response to n dichotomous items, In: H. Solomon (Ed.), *Studies in Item Analysis and Prediction* (Stanford: Stanford University Press) 158-68.

[4] Weller T J, Zhang A, Thompson R, and Tulleners F 2012 Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides, *J. Forensic Sci.* 57 (4) 912-17.