Volume 124, Article No. 124026 (2019) https://doi.org/10.6028/jres.124.026
Journal of Research of the National Institute of Standards and Technology

The Use of Correlated Binomial Distribution in Estimating Error Rates for Firearm Evidence Identification

Nien Fan Zhang

National Institute of Standards and Technology,
Gaithersburg, MD 20899, USA

nien-fan.zhang@nist.gov

In the branch of forensic science known as firearm evidence identification, estimating error rates is a fundamental challenge. Recently, a new quantitative approach known as the congruent matching cells (CMC) method was developed to improve the accuracy of ballistic identifications and provide a basis for estimating error rates. To estimate error rates, the key is to find an appropriate probability distribution for the relative frequency distribution of observed CMCs overlaid on a relevant measured firearm surface such as the breech face of a cartridge case. Several probability models based on the assumption of independence between cell pair comparisons have been proposed, but the assumption of independence among the cell pair comparisons from the CMC method may not be valid. This article proposes statistical models based on dependent Bernoulli trials, along with corresponding methodology for parameter estimation. To demonstrate the potential improvement from the use of the dependent Bernoulli trial model, the methodology is applied to an actual data set of fired cartridge cases.

Key words: ballistic signatures; Bernoulli trials; beta-binomial distribution; forensic science; maximum likelihood estimation; nonlinear regression.

Accepted: September 9, 2019

Published: October 2, 2019

https://doi.org/10.6028/jres.124.026

1. Introduction

In firearm evidence analysis, the parts of the firearm that make forcible contact with the bullets or cartridge cases when fired create characteristic tool marks on their surface called “ballistic signatures” [1]. These signatures can be used for firearm evidence identifications. In general, tool marks have so-called “class characteristics” that are common to certain brands or models of firearms and individual characteristics arising from random variation in firearm manufacturing and wear. Figure 1 (from Ref. [2]) shows topography images of breech face impressions obtained from a pair of cartridge cases ejected from the same firearm slide. The slide is a component of a semiautomatic pistol firing mechanism that absorbs the recoil impact of the cartridge case on its breech face. The image pair has several features in common.

How to cite this article:
Zhang NF (2019) The Use of Correlated Binomial Distribution in Estimating Error Rates for Firearm Evidence Identification. *J Res Natl Inst Stan* 124:124026. https://doi.org/10.6028/jres.124.026

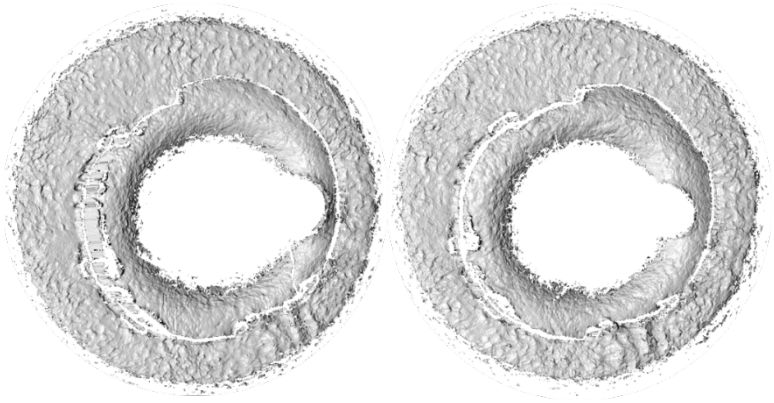

Fig. 1. Topography images of breech face impressions obtained from a pair of cartridge cases fired from the same firearm slide [2].

In investigations of crimes involving firearms, a challenge for firearms examiners is to determine whether a questioned cartridge case, typically recovered from a crime scene, and a known cartridge case, typically shot by investigators from a suspect firearm, were shot from the same firearm. Common automated ballistics identification systems are primarily based on comparison of two-dimensional (2D) images using optical microscopy and some correlation measure used to compare the image pair. 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 potentially improved basis for estimating error rates [3, 4]. To estimate the expected error rates of the results using the CMC method, two common probability models have been proposed in Ref. [2]. However, the assumption of independence among the cell pair comparisons upon which these models rely may not be valid.

In Appendix B of Ref. [2], a model for dependent Bernoulli trials and its applications to the CMC method for ballistic identification were briefly mentioned without any detail. This article provides a comprehensive discussion on that statistical model and its extension and proposes corresponding statistical methodology for parameter estimation. In Sec. 2, the CMC method is briefly described. The details of the CMC method can be found in Ref. [4] and Ref. [2]. In Sec. 3, the correlated binomial distribution based on dependent Bernoulli trials is introduced, and its properties are discussed. In Sec. 4, maximum likelihood estimators of the parameters of the correlated binomial distribution and estimators based on nonlinear regression models are proposed. In Sec. 5, the methodology is applied to a data set of fired cartridge cases to illustrate the performance of the different models. In Sec. 6, the correlated binomial distribution is combined with the beta distribution to create a compound probability distribution called the beta-correlated binomial distribution.

2. Congruent Matching Cells (CMC) Method for Ballistic Identification

The CMC method deals with pairs of measured 2D optical or three-dimensional (3D) topography images of breech face impressions for which the similarity has been objectively quantified. For an image pair, the CMC method divides one image as reference into an array of rectangular cells (after appropriate registration) as shown in Fig. 2 (from Ref. [3]). For each reference cell, a search for a matching cell is then conducted on the compared image [2]. Figure 3 shows the correlated cell pairs located in common valid regions and invalid regions [3].

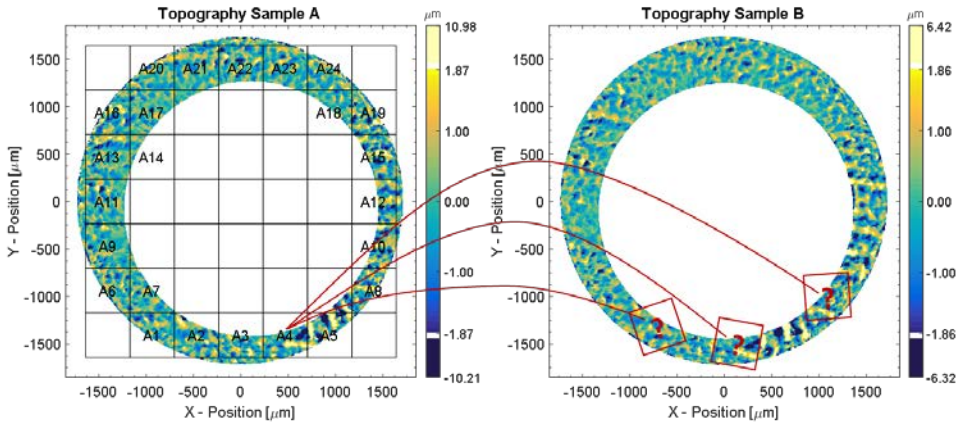

Fig. 2. Conceptual diagram of a topography image from Fig. 1 overlaid by a 7 × 7 grid, dividing the image into cells. The drag mark at the 3 o'clock position in Fig. 1 and the central hole and surrounding bulge from the firing pin impression are masked out of the images before the cell division step. Only cells with a sufficient fraction of measured pixels are used for the correlation analysis. Also shown is a schematic diagram of the automated search procedure to find an area in the compared image (right) that has a strong correlation with one of the cells in the reference image (left). Here the topography is represented by a color scale [2].

A cell is a rectangular subregion of the surface topography image that contains a sufficient quantity of distinguishing peaks, valleys, and other topographic features so that an assessment of topography similarity can be made. For example, in Fig. 3 (from Ref. [2]), if topographies A and B originating from the same firearm are registered at their position of maximum correlation, the cell pairs (A_1, B_1), (A_2, B_2), and (A_3, B_3) can be identified as correlated cell pairs. Whether the cell pair is correlated is determined by the maximum cross-correlation function [2].

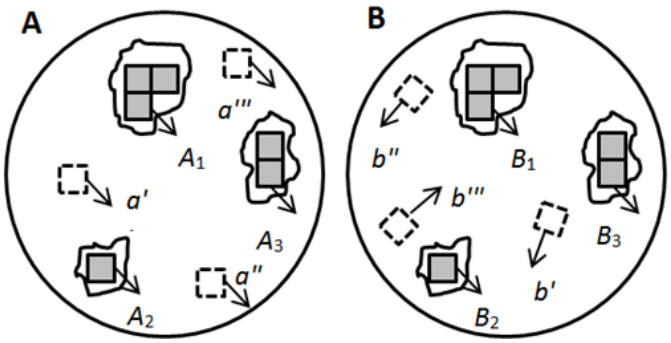

Fig. 3. Schematic diagram of topographies A and B originating from the same firearm and registered at the position of maximum correlation. The six solid cell pairs in each image are located in three valid correlated regions (A_1, B_1), (A_2, B_2), and (A_3, B_3). The dashed cell pairs (a', b'), (a'', b''), and (a''', b''') are located in the invalid correlation region [3].

Next, each correlated cell pair is determined to be a congruent matching cell (CMC) pair, or not, using four identification parameters for quantifying both the topographic similarity of the correlated cell pairs and the pattern congruency of the cell distributions. The details can be found in Ref. [2]. How many CMC pairs are required so that the two surface topographies can be identified as matching? The threshold would be determined after carefully designed experiments and error rate estimation. From a statistical point of view, the CMC method is based on the pass-or-fail tests of individual cell pairs from an image pair of breech face impressions. For a pair of images of breech face impressions, N represents the number of correlated cell pairs in the image pair. To avoid confusion with the statistical correlations among cell pair comparisons, which we will discuss later, from now on, we just call a correlated cell pair a "cell pair." For a given cell

pair, a random variable X represents the outcome of the CMC method applied to the cell pair. When the CMC method determines that a cell pair is a congruent matching cell pair, then $X = 1$; otherwise, $X = 0$. We denote the probability that $X = 1$ by p. That is, $P(X = 1) = p$, and $P(X = 0) = 1 - p$.

We sought to develop an approach for estimating the expected error rates of ballistic identification based on the CMC method. When the CMC method is applied to a set of cartridge cases, the result, in general, includes certain known matching (KM) image pair comparisons and certain known nonmatching (KNM) image pair comparisons. The false positive error rate is the rate of pairs of images incorrectly judged as matches. It represents the expected frequency or probability of obtaining an erroneous result of identification (declared match) when comparing samples from different sources (KNM). On the other hand, the false negative error rate is the rate of pairs of images incorrectly judged as nonmatches. It represents the probability of obtaining an erroneous result of exclusion (declared nonmatch) when comparing samples from the same source (KM).

To reliably estimate error rates and the associated uncertainties, the key is to find an appropriate probability distribution for the relative frequency distribution of the observed CMC results. A binomial probability distribution was proposed in Ref. [2] for the distribution of CMC measurements. Two assumptions were made there: (1) The comparisons between cell pairs are statistically independent from each other, and (2) each cell pair comparison within the image pair has the same probability, p. Under these assumptions, for an image pair with N cell pairs, we have a sequence of Bernoulli trials, $X_1, ..., X_N$. Denoting the sum of the CMC values for the comparisons of the first image pair by Y_1 with N_1 cell pairs and a sequence of Bernoulli trials, $X_{11}, ..., X_{1N_1}$, $Y_1 = \sum_{i=1}^{N} X_{1i}$. We say that Y_1 is the number of CMCs for the first image pair. In probability, $Y_1 \sim Bin(N_1, p)$ is a binomially distributed random variable with the probability mass function given by

$$P(Y_1 = k) = C_{N_1}^{k} p^k (1-p)^{N_1 - k} \text{ for } k = 0, 1, ..., N_1. \quad (1)$$

Similarly, for M image pairs, we have $Y_1, ..., Y_M$ correspondingly. Assuming $\{Y_j, j = 1, ..., M\}$ are independent from each other, we have a sequence of binomially distributed random variables, *i.e.*, $Y_j \sim Bin(N_j, p)$ for $j = 1, ..., M$, where N_j is the number of cell pairs for the jth image pair. For observed values from the CMC method, $\{y_j, j = 1, ..., M\}$, from Ref. [5], p. 56, the maximum likelihood estimator of p is given by

$$\hat{p} = \sum_{j=1}^{M} y_j \Big/ \sum_{j=1}^{M} N_j . \quad (2)$$

3. Binomial Distribution Based on Dependent Bernoulli Trials

As stated in Sec. 2, a key assumption for the proposed binomial distribution is the independence among cell pair comparisons in an image pair. However, this assumption may be invalid. For example, since the array of cells laid over the image is not precisely aligned with features in the image, neighboring cell pairs may share some individualizing features. Dependence among those neighboring cells is more likely to increase compared to cells situated relatively far apart. Reference [6], Sec. 6.1.2, discussed some practical examples indicating correlation between binary responses in biology. In these cases, we need to

consider some type of dependent Bernoulli trials, as done in other statistical research. For example, Ref. [7] and Ref. [8], p. 96–102, proposed a model for Bernoulli trials with Markov dependence. This approach assumes that the Bernoulli trials form a Markov chain and thus may not apply to the cell pair comparisons in the procedure for ballistic signatures. Reference [9] proposed a more general model for dependent Bernoulli trials, which sometimes is called the Bahadur-Lazarsfeld model.

In general, for a sequence of Bernoulli trials, we have random variables $X_1,...,X_N$, where each X_i takes the value 0 or 1, with $P(X_i=1)=p_i$ and $P(X_i=0)=1-p_i$ for $i=1,...,N$. Now, for the sequence $X_1,...,X_N$ of generalized Bernoulli trials, which may not be mutually independent, the second-order correlation between X_i and X_j, where $j\neq i$, is given by

$$r_{ij}=\frac{\mathrm{Cov}[X_i,X_j]}{\sigma_i\sigma_j}=\frac{E[(X_i-\mu_i)(X_j-\mu_j)]}{\sigma_i\sigma_j}, \tag{3}$$

where μ_i is the marginal mean, and σ_i is the marginal standard deviation for X_i. Note that $E[X_i]=\mu_i=p_i$ and $\mathrm{Var}[X_i]=p_i(1-p_i)$, for $i=1,...,N$, which are the same as in the case of independent trials. Similarly, third- and higher-order correlations, up to the Nth order correlation, are defined by

$$r_{ijk}=\frac{E[(X_i-\mu_i)(X_j-\mu_j)(X_k-\mu_k)]}{\sigma_i\sigma_j\sigma_k},\ \ldots,\ r_{12\ldots N}=\frac{E[(X_1-\mu_1)(X_2-\mu_2)\ldots(X_N-\mu_N)]}{\prod_{i=1}^{N}\sigma_i}. \tag{4}$$

That is, one can define as many as $N-1$ correlations of different order from (2, 3, …, N). The kth order correlation for $k=2,...,N$ is the correlation for any k distinct random variables of $\{X_1,...,X_N\}$. If any of these correlations is not zero, $\{X_1,...,X_N\}$ are dependent. For example, when $N=3$, for $\mathbf{X}=\{X_1,X_2,X_3\}$, there are eight possible outcomes: (0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), and (1,1,1). In this case, there are three second-order correlations: r_{12}, r_{13}, and r_{23}. There is only one third-order correlation r_{123}.

From Ref. [9], the joint probability distribution of $\mathbf{X}=\{X_1,...,X_N\}$ is expressed by

$$P(\mathbf{X})=P_{[1]}(\mathbf{X})\cdot f(\mathbf{X}), \tag{5}$$

where $P_{[1]}(\mathbf{X})$ is the joint probability distribution of $\mathbf{X}$ when the X_i values are independently distributed. That is, $P_{[1]}(X_1,...,X_N)=\prod_{i=1}^{N}p_i^{X_i}(1-p_i)^{1-X_i}$. The correction factor $f(\mathbf{X})$ due to correlations is given by

$$f(\mathbf{X})=1+\sum_{i<j}r_{ij}Z_iZ_j+\sum_{i<j<l}r_{ijl}Z_iZ_jZ_l+\ldots+r_{1\ldots N}Z_1,...,Z_N, \tag{6}$$

where $Z_i=(X_i-p_i)\big/\sqrt{p_i(1-p_i)}$. We may approximate $P(\mathbf{X})$ in Eq. (5) to any specific order k by omitting correlations higher than k. For example, the kth $(1<k\leq N)$ order approximation is given by

$$P_{[k]}(\mathbf{X})=P_{[1]}(\mathbf{X})(1+\sum_{i<j}r_{ij}Z_iZ_j+\ldots+\sum_{i<\ldots<l}r_{i\ldots l}Z_i\ldots Z_j). \tag{7}$$

Note that in the last summation in Eq. (7), the product behind the summation symbol includes k distinct Z values, and $r_{i...l}$ is the corresponding kth order correlation.

Currently, we consider only the case of symmetric distributions. That is, we assume $p_i = p$, for all $i = 1,...,N$, $r_{ij} = r_{(2)}$, for all distinct i, and j, $r_{ijk} = r_{(3)}$, *etc*., for all distinct $i, j,$ and $k,...$ For the example described earlier when $N = 3$, $r_{12} = r_{13} = r_{23} \triangleq r_{(2)}$. In particular, for each pair of X_i and X_j, the joint distribution of (X_i, X_j) is given by Eq. (6) as $N = 2$ and thus only depends on p and $r_{(2)}$. Later, we will discuss some aspects of this assumption that suggest this may not be a limiting assumption, given our purposes for the use of these models. We denote the sum of $\{X_i, i = 1,..., N\}$ by Y. Note that when $\{X_i\}$ are independent from each other as shown in Eq. (1), Y is a binomial distributed random variable. From Eq. (3),

$$\begin{aligned} \mathrm{Var}[Y] &= \sum_{i=1}^{N} \mathrm{Var}[X_i] + \sum_{i=1}^{N} \sum_{i \neq j} r_{(2)} \mathrm{Var}[X_i] \\ &= Np(1-p) + N(N-1) r_{(2)} p(1-p) \\ &= Np(1-p)\{1 + (N-1) r_{(2)}\}. \end{aligned} \tag{8}$$

When there is no correlation between pairs of $\{X_i\}$, $r_{(2)} = 0$, and $\mathrm{Var}[Y] = Np(1-p)$, which is the variance of Y when $\{X_i\}$ are independent from each other. When $r_{(2)} > 0$, from Eq. (8), $\mathrm{Var}[Y] > Np(1-p)$. Thus, the positive correlation between pairs of $\{X_i\}$ leads to a larger variance of Y. For the negative correlation, we refer to the discussion in Ref. [6], Sec. 6.3.

We denote the probability mass function of Y when $\{X_i\}$ are independent from each other, as given in Eq. (1), by $P_{[1]}(Y)$. Then, when the Bernoulli trials are dependent, and the joint distribution is symmetric, from Ref. [9], the probability mass function of Y is expressed by

$$P(Y) = P_{[1]}(Y)\{1 + \sum_{j=2}^{N} r_{(j)} g_j(Y)\}, \tag{9}$$

where $g_j(Y)$ is a polynomial in Y of degree j and also a function of p. In this case, we say Y has a correlated binomial distribution. Obviously, when the correlations for all orders are zero, the equation reduces to the binomial distribution based on independent Bernoulli trials. The explicit formulas for $g_j(Y)$ when $j = 2,3$ are given in Ref. [9]. In particular, when $j = 2$,

$$g_2(Y) = \frac{(Y - Np)^2 - (1-2p)(Y - Np) - Np(1-p)}{2p(1-p)}. \tag{10}$$

From Eq. (9), $P(Y)$ can be approximated by

$$P_{[k]}(Y) = P_{[1]}(Y)\{1 + \sum_{j=2}^{k} r_{(j)} g_j(Y)\}, \quad k = 2,...,(N-1), N, \tag{11}$$

where $P_{[k]}(Y)$ $(1 < k \le N)$ denotes the kth order approximation of the probability distribution of Y. In particular, $P_{[N]}(Y) = P(Y)$.

As shown in Ref. [9], an approximate distribution of $\mathbf{X}$ of order k, as given in Eq. (7), is a probability distribution as long as the corresponding $P_{[k]}(\mathbf{X})$ are nonnegative for all $\mathbf{X}$. In addition, it is shown in Ref. [9] that if all correlations of the given distribution of the Bernoulli trials are sufficiently small in absolute value, $P_{[k]}(Y)$ is a probability distribution for each k. Based on that, for an approximation of a given probability distribution of Y, all correlations of the dependent Bernoulli trials are assumed to be appropriate to make the approximation a *proper* probability distribution yielding values between 0 and 1 and with a sum equal to 1. Now, consider a more general case that assumes for all X_i the marginal probabilities are the same, *i.e.*, = p, but that the correlations are not all symmetric. From Proposition 5 in Ref. [9], an approximation based on a symmetric distribution is better than any nonsymmetric case that has same p but different correlation(s) with respect to the corresponding approximations for the probability. Therefore, in that sense, using symmetric dependent Bernoulli trials should provide better models than the case with nonsymmetric distributions, limiting the cases that need to be considered.

We now discuss some properties of the central moments of Y with respect to $P_{[k]}(Y)$ for $k = 2,..., N$. We denote the kth central moment of Y with the probability distribution of $P_{[i]}(Y)$ by $\mu_{k,P_{[i]}}(Y)$ $(i, k > 0)$. That is,

$$\mu_{k,P_{[i]}}(Y) = E_{P_{[i]}}[Y - \mu_Y]^k, \tag{12}$$

where μ_Y is the mean of Y with probability distribution of $P_{[i]}(Y)$. It is shown in the Appendix that

$$\mu_{k,P_{[i]}}(Y) = \mu_{k,P_{[k]}}(Y) \text{ when } i > k. \tag{13}$$

Thus, for the kth central moment, the mean of Y is the same with respect to $P_{[k]}(Y)$, $P_{[k+1]}(Y)$,…, $P_{[i]}(Y)$ for $N \ge i \ge k$. In particular, $E_{P_{[i]}}[Y] = E_{P_{[1]}}[Y]$ for $i \ge 1$. That is, for all $P_{[i]}(Y)$, the corresponding means of Y are the same as $\mu_Y = NP$. In addition, the variance of Y with $P_{[i]}(Y)$ when $i > 2$ is equal to that with $P_{[2]}(Y)$. For example, $\mathrm{Var}_{P_{[2]}}[Y] = \mathrm{Var}_{P_{[3]}}[Y] = ... = \mathrm{Var}_{[N-1]}[Y] = \mathrm{Var}[Y]$. Thus, from Eq. (8),

$$\mathrm{Var}_{P_{[2]}}(Y) = \mathrm{Var}_{P_{[1]}}(Y)\{1 + (N-1)r_{(2)}\}, \tag{14}$$

where $\mathrm{Var}_{P_{[2]}}(Y)$ is the variance or the second central moment of Y with respect to $P_{[2]}(Y)$. From Eq. (14), one can determine that when $r_{(2)}$ is positive, the variance of Y with respect to $P_{[2]}(Y)$ is larger than that with respect to $P_{[1]}(Y)$ and vice versa when $r_{(2)}$ is negative. In addition, from Eq. (14), $-1/(N-1) < r_{(2)} \le 1$. In addition, for $P_{[2]}(Y)$, Eq. (3.3) in Ref. [9] gave two inequalities as a function of N and p for the permissible range of $r_{(2)}$ for $P_{[2]}(Y)$ to be a probability distribution. The lower bound of $r_{(2)}$ is given by

$$r_{(2)} \geq \frac{-2}{N(N-1)} \min\left(\frac{p}{1-p}, \frac{1-p}{p}\right). \tag{15}$$

For example, when $p = 0.8$ and $N = 26$, the lower bound of $r_{(2)}$ is −0.00077, which is effectively zero. On the other hand, from Eq. (3.3) in Ref. [9], the upper bound of $r_{(2)}$ is 0.08.

We use some examples to illustrate the probability mass functions of $P_{[1]}(Y)$ and $P_{[k]}(Y)$ in Eq. (11). First, we let $p = 0.6$, $r_{(2)} = 0.02$, and $N = 26$. These probability mass functions are plotted in Fig. 4. Then, we let $p = 0.6$, $r_{(2)} = 0.05$, and $N = 26$. These probability mass functions are plotted in Fig. 5.

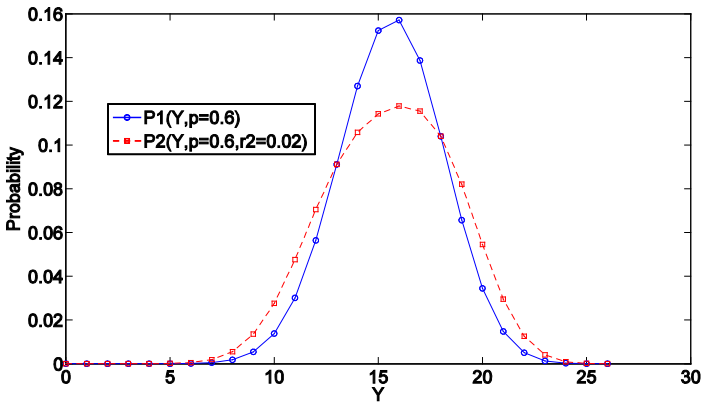

Fig. 4. Probability mass functions of $P_{[1]}(Y)$ and $P_{[2]}(Y)$ with $p = 0.6$ and $r_{(2)} = 0.02$.

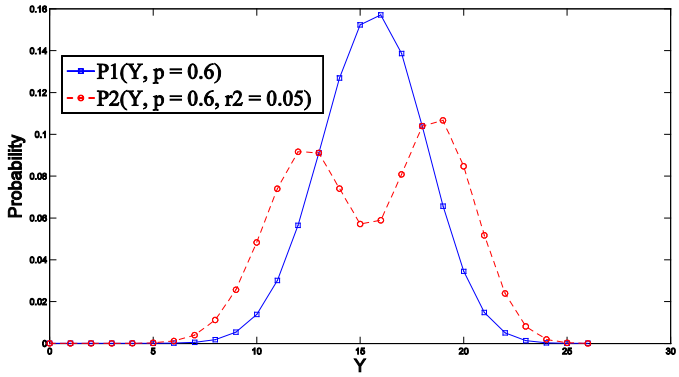

Fig. 5. Probability mass functions of $P_{[1]}(Y)$ and $P_{[2]}(Y)$ with $p = 0.6$ and $r_{(2)} = 0.05$.

Note that when $r_{(2)}$ changes from 0.02 to 0.05, the number of modes in the probability mass function $P_{[2]}(Y)$ changes from one to two. Note that the means with $P_{[1]}(Y)$ and $P_{[2]}(Y)$ with $r_{(2)} = 0.05$ are the same. Namely, from Eq. (13), $E_{P_{[1]}}(Y) = E_{P_{[2]}}(Y) = Np = 15.6$. However, the variances based on the $P_{[1]}(Y)$ and $P_{[2]}(Y)$ values are different. $\mathrm{Var}_{P_{[1]}}[Y] = Np(1-p) = 6.24$, while $\mathrm{Var}_{P_{[2]}}[Y] = 14.04$ from Eq. (14). The variance increases when $r_{(2)}$ is positive. If we let $p = 0.6$, $r_{(2)} = -0.002$, and $N = 26$, then

$\mathrm{Var}_{P_{[1]}}[Y] = Np(1-p) = 6.24$ and $\mathrm{Var}_{P_{[2]}}[Y] = 5.928$. The variance decreases due to the negative value of $r_{(2)}$. The probability mass functions for $P_{[1]}(Y)$ and $P_{[2]}(Y)$ with $p = 0.6$ and $r_{(2)} = -0.002$ are plotted in Fig. 6.

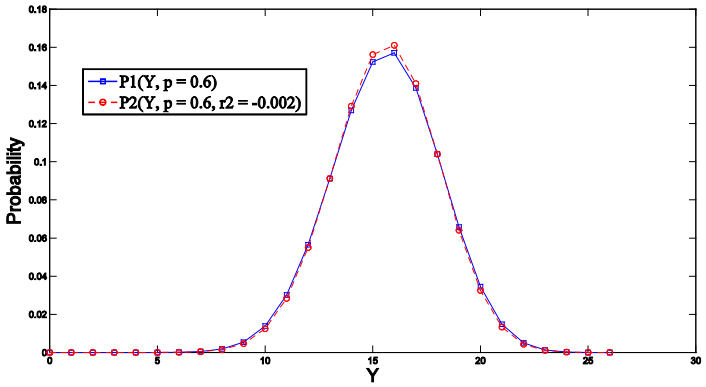

Fig. 6. Probability mass functions of $P_{[1]}(Y)$ and $P_{[2]}(Y)$ with $p = 0.6$ and $r_{(2)} = -0.002$.

In addition, we show a case with third-order correlation that has $p = 0.6$, $r_{(2)} = 0.02$, $r_{(3)} = -0.001$, and $N = 26$. These probability mass functions are plotted in Fig. 7. Again, in this case, $E_{P_{[1]}}(Y) = E_{P_{[3]}}(Y) = Np = 15.6$, and $\mathrm{Var}_{P_{[1]}}[Y] = Np(1-p) = 6.24$. From Eq. (13), $\mathrm{Var}_{P_{[3]}}[Y] = \mathrm{Var}_{P_{[2]}}[Y] = 9.36$. The third-order central moment based on $P_{[1]}(Y)$ is given by (see Ref. [5]):

$\mu_{3,P_{[1]}}(Y) = E[(Y-\mu_Y)^3] = Np(1-p)(1-2p) = -1.248$.

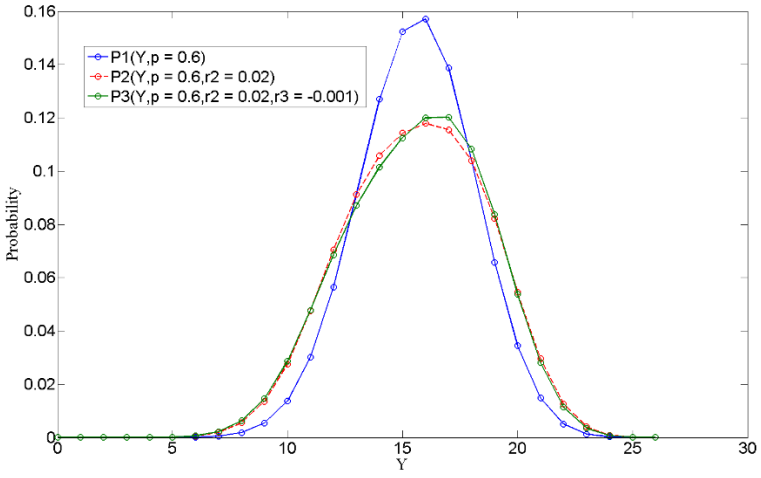

Fig. 7. Probability mass function of $P_{[1]}(Y)$ and $P_{[2]}(Y)$ with $p = 0.6$, $r_{(2)} = 0.02$, and $r_{(3)} = -0.001$.

Finally, we use an example to show the probability mass distributions for the correlated binomial distributions based on generalized Bernoulli trials. Consider $N = 5$ and $\mathbf{p} = [0.1, 0.3, 0.4, 0.6, 0.8]$. The average of $\mathbf{p}$ is 0.44. The probability mass functions of the second-order approximated correlated binomial

distributions for the symmetric case with p = 0.44, k = 2, and $r_{(2)}$ = 0.02 and −0.02 are obtained from Eq. (11) and plotted in Figs. 8 and 9, labelled as P2(Y, p=0.44, r2=0.02) and P2(Y, p=0.44, r2= −0.02), respectively. For comparison, the binomial distribution with p = 0.44 is also shown and labelled as P1(Y, p=0.44). In addition, the probability mass functions of the binomial distribution based on the generalized Bernoulli trials, with $\mathbf{p}$ labelled as P1(Y,P), as well as the second-order approximated correlated binomial distributions based on the generalized Bernoulli trials, with $\mathbf{p}$ and $r_{(2)}$ calculated from Eq. (7) and labelled as P2(Y,P,r2=0.02) and P2(Y,P,r2= -0.02), respectively, for the nonsymmetric case, are also shown in these plots.

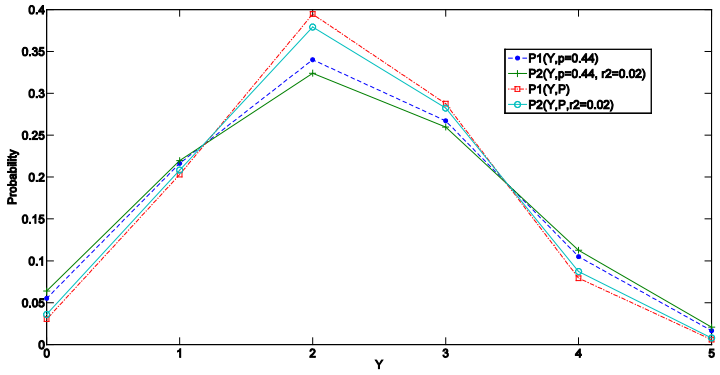

Fig. 8. Probability mass functions of $P_{[1]}(Y)$ with p = 0.44 and $\mathbf{p}$ and $P_{[2]}(Y)$ with $r_{(2)}$ = 0.02.

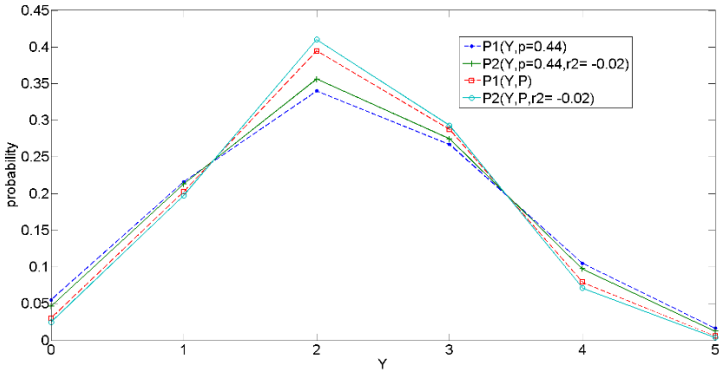

Fig. 9. Probability mass functions of $P_{[1]}(Y)$ with p = 0.44 and $\mathbf{p}$ and $P_{[2]}(Y)$ with $r_{(2)}$ = −0.02.

From these plots, we observe that (1) the tail probabilities for Y (*e.g.*, for $Y \le 1$ or $Y \ge 4$) based on the generalized Bernoulli trials with unequal marginal probabilities (P1(Y, P)) are smaller than those for the symmetric marginal probability p (P1(Y, p=0.44)); (2) when $r_{(2)}$ is positive, the tail probabilities with dependent $\mathbf{X}$ and with the symmetric marginal probability p (P2(Y, p=0.44, r2=0.02)) are larger than those with dependent $\mathbf{X}$ (P2(Y, P,r2=0.02)), and this is the same for negative $r_{(2)}$ = −0.02; (3) when $r_{(2)}$ is positive, the tail probabilities with dependent $\mathbf{X}$ and with the symmetric marginal probability p (P2(Y, p=0.44, r2=0.02)) are larger than those with independent $\mathbf{X}$ (P1(Y, p=0.44)) and vice versa for a negative $r_{(2)}$.

4. Estimating the Parameters of the Correlated Binomial Distribution

As stated earlier, when the CMC method is applied to a set of cartridge cases, the analysis, in general, includes certain KM image pair comparisons and a larger number of KNM image pair comparisons. Any image pair in the KM set is known by design to be images of two cartridge cases fired from same firearm. The image pairs in the KNM set are also known, in that sense, to be images of two cartridge cases fired from different firearms. Statistical models are fitted separately to sets of KM and KNM data, respectively, to estimate the two different types of error rates described in Sec. 2, which are of fundamental interest for characterizing matching performance.

When assessing method performance using M image pairs, the random variables for the sums of the CMC values for each image pair comparison are denoted by $Y_1,...,Y_M$. As discussed in Ref. [2], the error rates are calculated from the conditional probabilities of Y based on the sets of KM or KNM data. Again, as pointed out in Sec. 2, to estimate error rates, the key is to find an appropriate probability distribution to describe Y.

Based on the practical application described in Ref. [2], the binomial distribution fits the CMC values very well for KNM data but does not appear to fit the KM data as well. Therefore, we focus here on fitting dependent binomial models to the KM data. We assume that $Y_1,...,Y_M$ from the M KM image comparisons are independent from each other, but for each image comparison, we have a sequence of N symmetric dependent Bernoulli trials. Two approaches for estimating the parameters of the correlated binomial distribution are proposed.

4.1 Maximum Likelihood Estimator of the Parameters

In Sec. 3, approximations of the probability mass function of the correlated binomial random variable are discussed. For illustration, we will discuss the second-order approximation given in Eq. (11),

$$P_{[2]}(Y) = P_{[1]}(Y)\{1 + r_{(2)} g_2(Y, p)\}, \tag{16}$$

where g_2, given by Eq. (10), is expressed here by $g_2(Y, p)$ to show it is a function of Y as well as p. We first use maximum likelihood (ML) estimation to estimate the p and $r_{(2)}$ [10]. The CMC measurements from M independent image pair comparisons are denoted by $y_1,..., y_M$, where y_j, for $j = 1,...,M$, is the sum of the CMC results from N_j dependent Bernoulli trials for the jth image pair. N_j $(j = 1,...,M)$ is known. The likelihood function for the given p and $r_{(2)}$ is given by

$$L = \prod_{j=1}^{M} P_{[2]}(y_j \mid p, r_{(2)}) = \prod_{j=1}^{M} C_{N_j}^{y_j} p^{y_j} (1-p)^{N_j - y_j} \{1 + r_{(2)} g_2(y_j, p)\}. \tag{17}$$

The log-likelihood function is equivalent to

$$\sum_{j=1}^{M} \Big[y_j \log(p) + (N_j - y_j) \log(1-p) + \log\{1 + r_{(2)} g_2(y_j, p)\} \Big], \tag{18}$$

denoted by $\log(L)$. The maximum likelihood estimators (MLEs) of p and $r_{(2)}$ are obtained when the respective $\log(L)$ reaches its maximum using the quasi-Newton algorithm or other appropriate numerical

optimization algorithms. A *Z*-test can be used to check whether $r_{(2)}$ is significantly different from zero using an asymptotic normal distribution of the parameter estimators. See Ref. [11], p. 283–295.

4.2 Use of Nonlinear Regression Models

Similarly, we discuss the case of the second-order approximation in Eq. (16). For the CMC values from M independent image pair comparisons, each of $y_1,...,y_M$ is a sum of the results from N dependent Bernoulli trials. For $i = 0,1,...,N$, we define the frequency of $y = i$ by

$$h(i) = \frac{\{\#\, of\ \ y\ \ for\ (y = i)\}}{M}. \tag{19}$$

Since $\{Y_j, j = 1,...,M\}$ are independently and identically distributed, based on the property of the empirical distribution, $h(y)$ is an approximation of $P_{[2]}(Y)$ or $P(Y)$. In Eq. (16), approximating $P_{[2]}(Y)$ by $h(y)$, we have

$$h(y) = P_{[1]}(y)\{1 + r_{(2)} g_2(y, p)\} + \varepsilon\,, \tag{20}$$

where $g_2(y, p)$ is the nonlinear function of p given in Eq. (10), and ε is a random error with zero mean. Thus, $h(y)$ is approximated by a nonlinear function of p and $r_{(2)}$. This nonlinear regression model is fitted to $h(y)$ to find optimal estimates of p and $r_{(2)}$, which are denoted by $\tilde{p}$ and $\tilde{r}_{(2)}$, based on the criterion of minimum error sum of squares ([12], p. 21–24) using the Levenberg-Marquardt nonlinear least squares algorithm or other appropriate algorithms. The error sum of squares is defined as

$$\mathrm{ESS} = \sum_{y=0}^{N} [h(y) - P_{[1]}(y, \tilde{p})\{1 + \tilde{r}_{(2)} g_2(y, \tilde{p})\}]^2\,. \tag{21}$$

For nonlinear regression, the underlying assumption is that the nonlinear function can be approximated by a linear function. Based on that, the approximate variance-covariance matrix of the estimators of the parameters can be obtained. In addition, a *Z*-test can be used to check whether $r_{(2)}$ is significantly different from zero. Note that if $\{\varepsilon\}$ are correlated and have different variances, a generalized least squares model and the corresponding estimates can be obtained. For this, we refer the reader to Ref. [12], p. 27–30, Ref. [13], p. 96–98, and Ref. [14], p. 225–226.

5. Estimating the Parameters of the Correlated Binomial Distribution for the Weller Data Set

The Weller data set of cartridge cases was obtained from a set of 11 firearm slides produced by the same manufacturer using the same process. The data set has 370 KM and 4095 KNM pairs. The details of the data set can be found in Ref. [15] and Ref. [2]. We applied the two approaches in Sec. 4 to the KM data set. For the KM data, N = 47 and M = 370. Under the assumption of a binomial distribution, $\hat{p}$ = 0.7864 is obtained from Eq. (2).

If we only include the second-order correlation, the MLE estimates from Eq. (17) for p are $\hat{p}$ = 0.7823 and $\hat{r}_{(2)}$ = 0.0191. Using nonlinear regression, for the second-order correlation model in Eq. (16), the least squares (LS) estimates are: $\tilde{p}$ = 0.7979 and $\tilde{r}_{(2)}$ = 0.0149. The standard deviations of the estimators are: $\hat{\sigma}_{\tilde{p}}$ = 0.0032 and $\hat{\sigma}_{\tilde{r}_{(2)}}$ = 0.0013. Parameters p and $r_{(2)}$ are significantly different from zero based on the *Z*-test.

The third-order correlation was also considered. However, it seems that including $r_{(3)}$ makes only a minor difference. If we include the third-order correlation term, the LS estimates are given by $\tilde{p}$ = 0.7997, $\tilde{r}_{(2)}$ = 0.0145, and $\tilde{r}_{(3)}$ = 0.0005. The standard deviations of the estimators are: $\hat{\sigma}_{\tilde{p}}$ = 0.0031, $\hat{\sigma}_{\tilde{r}_{(2)}}$ = 0.0013, and $\hat{\sigma}_{\tilde{r}_{(3)}}$ = 0.0004. The parameter $r_{(3)}$ is not significantly different from zero and thus would likely be omitted in the model. Figure 10 demonstrates that the correlated binomial distribution model with parameters estimated using nonlinear regression fits the data much better than the simpler model based on the binomial distribution.

For completeness, we also applied the model based on the correlated binomial distribution to the KNM data set with N = 42 and M = 4095. If we only include the second-order correlation, the LS estimates of p and $r_{(2)}$ are 0.0011116 and 0.00023, respectively, with both parameters determined to be significantly different from zero. However, in this case, the results are not distinctly different from those when the model is based on the binomial distribution, which produces an estimate of $\hat{p}$ = 0.0011105 .

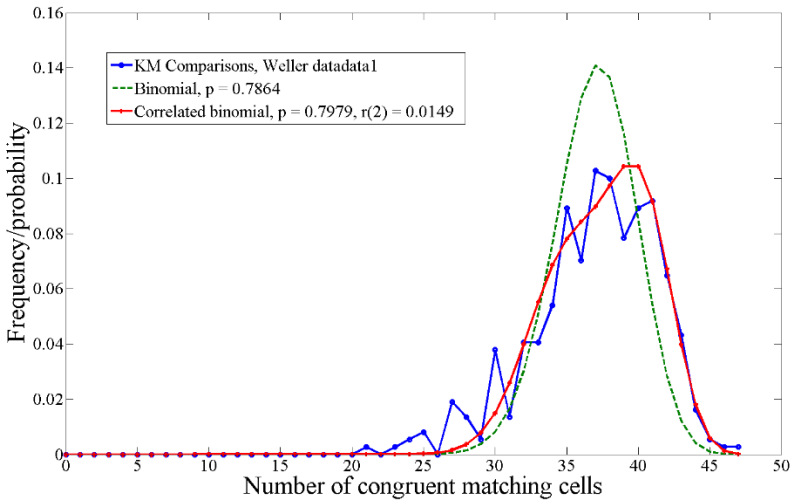

Fig. 10. Frequency distribution of CMC numbers of KM image pairs of the Weller data set with binomial and correlated binomial mass probability functions.

6. Use of the Beta-Correlated Binomial Distribution for CMC Measurements

In Ref. [2], it was proposed to relax the assumption of a fixed probability of congruency for the binomial distribution when modeling the CMC measurements. This revised model allows one to vary p for different image pair comparisons. In this case, we assume that within one image pair comparison, the probability p for all the Bernoulli trials is the same, while for different image pair comparisons, p varies. See Ref. [16]. As in the framework of Bayesian statistics, we assume that the parameter p is a random

variable with a beta distribution. For the first image pair with N cell pairs, we have a sequence of Bernoulli trials, $X_{11},...,X_{1N}$, which are independent from each other and have a common probability of $p = p_1$. The sum of the CMC values for the first image pair is Y_1, which for a given p_1 has a binomial distribution. Namely, $Y_1 \mid p_1 \sim Bin(N, p_1)$. In general, for M image pairs, we have $Y_j \mid p_j \sim Bin(N, p_j)$, $j = 1,...,M$, where p has a beta distribution, *i.e.*, $p \sim Beta(\alpha, \beta)$, with positive α and β as parameters to be fitted to the data. The probability mass function of the beta-binomial random variable Y for given N, α, and β is then given by

$$\begin{aligned} P(Y = k \mid N, \alpha, \beta) &= \int_0^1 \frac{P_{[1]}(k, p)}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{C_N^k}{B(\alpha, \beta)} \int_0^1 p^{k+\alpha-1}(1-p)^{N-k+\beta-1} dp \\ &= C_N^k \frac{B(k+\alpha, N-k+\beta)}{B(\alpha, \beta)}, \end{aligned} \tag{22}$$

where $B(\alpha, \beta)$ is a beta function with parameters α and β, and $P_{[1]}(k, p)$ is the binomial probability mass function in Eq. (1) when $Y = k$.

In Ref. [2], comparisons were made of the fits of the beta-binomial probability model and the binomial probability model for different data sets, including the Weller data set for cartridge cases. Here, we need to emphasize that although use of the beta-binomial distribution can relax the assumption of the same p for all image pairs, it still assumes that within each image pair, all cell pair comparisons are independent from each other. We checked this assumption by considering correlations among cell pair comparisons.

Now instead of the independent Bernoulli trials, we assume that the cell pair comparisons within each image pair are dependent Bernoulli trials. The corresponding probabilities of the sum are approximated by $P_{[2]}(Y)$ as given by Eq. (16) when only the second-order correlation with a constant $r_{(2)}$ is assumed. Assume that p in the correlated binomial distribution is random with a beta distribution. Namely, $Y_j \mid p_j, r_{(2)} \sim corr.Bin(N, p_j, r_{(2)})$, for $j = 1,...,M$, where p has a beta distribution, *i.e.*, $p \sim Beta(\alpha, \beta)$. Similar to Eq. (22), the probability mass function of Y for given N, α, β, and $r_{(2)}$ is given by

$$\begin{aligned} &P(Y = k \mid N, \alpha, \beta, r_{(2)}) \\ &= \int_0^1 \frac{P_{[2]}(k, p)}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{C_N^k}{B(\alpha, \beta)} \int_0^1 p^k (1-p)^{N-k} \{1 + r_{(2)} g_2(k, p)\} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{C_N^k}{B(\alpha, \beta)} \int_0^1 p^{k+\alpha-1}(1-p)^{N-k+\beta-1} \{1 + r_{(2)} g_2(k, p)\} dp, \end{aligned} \tag{23}$$

where $g_2(k, p)$ is given by Eq. (10). In this case, the marginal probability $P(Y = k \mid N, \alpha, \beta, r_{(2)})$ for $k = 0,1,...,N$ has no explicit expression. However, it can be calculated by numerical integration. In this

case, the random variable Y has a compound probability distribution called a beta-correlated binomial distribution.

7. Conclusions

The recently proposed CMC method improves the accuracy of ballistic identification. From a statistical point of view, the CMC results are based on pass-or-fail tests for comparison of individual cell pairs taken from an image pair of breech face impressions. To estimate the expected error rates of the CMC measurements, several probability models have been proposed. However, the assumption of independence among the cell pair comparisons from the CMC method required by some models may not be valid. To relax the assumption of independence, we propose using correlated binomial and beta-correlated binomial probability distribution models to fit the CMC measurements. Application to practical data demonstrates that the correlated binomial probability distribution model fits the data much better than the binomial probability distribution model. The statistical models proposed in this article can be applied to other types of data with N dichotomous comparisons as well.

8. Appendix: Proof of Eq. (13)

The proof is as follows.

For $Y = 0,1,...,N$,

$$\begin{aligned}\mu_{k,P_{[i]}}(Y) &= E_{P_{[i]}}[(Y-\mu_Y)^k] \\ &= \sum_{m=0}^{N}(m-\mu_Y)^k P_{[1]}(Y=m)\{1+\sum_{j=2}^{k} r_{(j)} g_j(Y=m)+\sum_{j=k+1}^{i} r_{(j)} g_j(Y=m)\} \\ &= E_{P_{[k]}}[(Y-\mu_Y)^k]+\sum_{m=0}^{N}(m-\mu_Y)^k P_{[1]}(Y=m)\sum_{j=k+1}^{i} r_{(j)} g_j(Y=m) \\ &= \mu_{k,P_{[k]}}(Y)+\sum_{m=0}^{N}(m-\mu_Y)^k P_{[1]}(Y=m)\sum_{j=k+1}^{i} r_{(j)} g_j(Y=m).\end{aligned}$$

From [9], Eq. (4.3), p. 164–165, $\{g_j(Y), j = 0,1,...,N\}$ are orthogonal associated with $P_{[1]}(Y)$. Namely, the inner product $(g_i(Y), g_j(Y)) = \sum_{m=0}^{N} g_i(Y=m) * g_j(Y=m) * P_{[1]}(Y=m) = 0$ for $i \neq j$ and $\| g_j(Y) \|^2 = \binom{N}{j}$ for $j = 0,1,...,N$; see Eq. (4.8) in Ref. [9]. Thus, $\sum_{m=0}^{N}(m-\mu_Y)^k P_{[1]}(Y=m)\sum_{j=k+1}^{i} r_{(j)} g_j(Y=m) = 0$ because $(Y-\mu_Y)^k$ can be expressed as a linear combination of $\{g_j(Y), j = 0,1,...,k\}$ in the corresponding inner product space with an orthogonal basis of $\{g_j(Y) \Big/ \binom{N}{j}^{0.5}, j = 0,1,...,N\}$. This implies $\mu_{k,P_{[i]}}(Y) = \mu_{k,P_{[k]}}(Y)$ when $i > k$.

Acknowledgments

The author thanks M. Henn for his assistance on computation. The author also thanks T. V. Vorburger, W. F. Guthrie, J. Lu, Z. Chen, and three reviewers for their helpful comments.

9. References

[1] Scientific Working Group for Firearms and Toolmarks (SWGGUN) (2016) *The Foundation of Firearm and Toolmark Identification*. http://www.nist.gov/sites/default/files/documents/2016/11/28swggun_foundational_report.pdf
[2] Song J, Vorburger TV, Chu W, Yen J, Soons JA, Ott DB, Zhang NF (2018) Estimating error rates for firearm evidence identification in forensic science. *Forensic Science International* 284:15–32. https://doi.org/10.1016/j.forsciint.2017.12.013
[3] Song J (2015) Proposed NIST ballistics identification system (NBIS) based on 3D topography measurements on correlated cells. *Journal of the Association of Firearm and Tool-Mark Examiners* 45(2):184–189. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=910868
[4] Song J (2015) Proposed "congruent matching cells (CMC)" method for ballistic identification and error rate estimation. *Journal of the Association of Firearm and Tool-Mark Examiners* 47(3):177–185. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=911193
[5] Johnson NI, Kotz S(1969) *Discrete Distributions* (Houghton Mifflin Company, Boston, MA).
[6] Collett D (2003) *Modeling Binary Data* (Chapman & Hall/CRC, Boca Raton, FL).
[7] Klotz J (1978) Statistical inference in Bernoulli trials with dependence. *The Annuals of Statistics* 1(2):373–379. https://www.jstor.org/stable/2958025
[8] Cox DR, Snell EJ(1989) *Analysis of Binary Data* (Chapman and Hall, London), 2nd Ed.
[9] Bahadur RR (1961) A representation of the joint distribution of the response to *n* dichotomous items. *Studies in Item Analysis and Prediction*, ed Solomon H (Stanford University Press, Stanford, CA).
[10] Scholz FW (1985) Maximum likelihood estimation. *Encyclopedia of Statistical Sciences*, Vol. 5 (Wiley, New York).
[11] Cox DR, Hinkley DV (1974) *Theoretical Statistics* (Chapman and Hall, London).
[12] Seber GAF, Wild CJ (1989) *Nonlinear Regression* (Wiley, New York).
[13] Rao CR, Toutenburg H (1995) *Linear Models—Least Squares and Alternatives* (Springer, New York).
[14] Baltagi BH (2008) *Economics* (Springer, London), 4th Ed.
[15] Weller TJ, Zheng A, Thompson R, Tulleners F (2012) Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides. *Journal of Forensic Sciences* 57(4):912–917. https://doi.org/10.1111/j.1556-4029.2012.02072.x
[16] Wilcox RR (1981) A review of the beta-binomial model and its extensions. *Journal of Educational and Behavioral Statistics* 6(1):3–32. https://doi.org/10.3102/10769986006001003

About the author: *Nien Fan Zhang is a mathematical statistician in the Statistical Engineering Division of the NIST Information Technology Laboratory. The National Institute of Standards and Technology is an agency of the U.S. Department of Commerce.*