The authors shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

Final Report
Quantifying the Effects of Database Size and Sample Quality on Measures of Individualization Validity and Accuracy in Forensics

Award Number: 2009_DN_BX_K234

Authors[1]

Donald T. Gantz, PhD, George Mason University
Christopher Saunders, South Dakota State University

March 2014

Opinions or points of view expressed are those of the authors and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

[1] Authors who compiled this Draft Report are listed alphabetically.

ABSTRACT

The grant research addressed several of the concerns detailed in *Recommendation 3* in the National Academy of Science (NAS) report: *Strengthening Forensic Science in the United States: A Path Forward*. Specifically, we have developed methods to statistically quantify 1) the random match probability (RMP) which quantifies uncertainty in measures aimed at validating a forensic discipline's basic premises (such as a uniqueness claim) and 2) the accuracy of likelihood ratio methods used in making classification/individualization conclusions.

The use of automated pairwise comparisons of biometric samples in a database is a basic element of forensic individualization determinations involving biometrics such as fingerprints, handwriting, tool marks, etc. An issue that applies to forensic individualization is that while a database of samples can be used to support individuality, it does not necessarily prove individuality. Therefore, the NAS report calls for statistically/probabilistically based statements concerning the level of support that a database of samples provides for individualization. To date, much attention has focused on how to use an automated comparison methodology applied to a database of biometric samples to estimate the RMP, which is defined as the probability of selecting two individuals at random from a population that "match" on the basis of some biometric. The RMP can be interpreted as giving the expected performance of a comparison methodology across some relevant population. Phase I of this grant focused on the RMP as a measure of the validity of a forensic individualization system. We developed theoretically sound upper confidence bounds on the RMP, which are estimated using these automated pairwise comparisons. The RMP is related to the question of whether or not we should use a given biometric modality in general.

In Phases II and III, we shifted focus to quantifying the accuracy of given forensic modalities in individual applications. The use of likelihood ratio methods in DNA analysis is well established for addressing this problem. However, research into its use in other forensic areas is not as well developed. We investigated the use of Bayes Factors and likelihood ratios in other fields, such as handwriting and glass fragments, focusing both on statistically valid quantifications of the value of the evidence. In Phase II, we focused on popular approximate procedures while in Phase III, we investigated statistically rigorous formal techniques with the main focus on Bayesian approaches to the model selection for the forensic identification of source problem.

Most of the methodologies developed in this grant will apply to any field of forensics, as RMPs and likelihood ratios are defined similarly in many of them. In Phase I, we have quantified the effect of database size and sample quality on proposed point and interval estimators. In Phase II, we have demonstrated that the three main classes of approximate LRs can have radically different quantifications of the value of the evidence, suggesting that these are not reasonable procedures for the quantification of forensic evidence. In Phase III, we have rigorously extended the Bayesian Likelihood Ratio to situations where the background population has not been accurately characterized. Finally, we have illustrated

the developed methodologies using a database of handwriting samples (which we have utilized in previous research) and on publically available glass-fragment data. We are in the process of extending our work to databases of automated comparisons of fingerprints.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Phase I

Goal: Study Interval Estimates of a measure of the validity of a forensic individualization system.

Dr. Christopher Saunders motivated the research in Phase I in his 2010 AAFS presentation which gave an overview of the goals for this research grant. During Phase I we focused on the Random Match Probability (RMP) as a measure of the validity of a forensic individualization procedure. Specifically, our research has been concerned with upper confidence bounds on measures, such as the RMP, that are estimated using automated pairwise comparisons. Pairwise comparison of samples is fundamental to forensic individualization systems. The validity of pairwise comparisons depends on the ability to effectively discriminate between samples of different origin and to accurately match samples of a common origin.

The RMP is defined as the probability of selecting two distinct sources at random from a population that "match" on the basis of some biometric sample extracted from each. The RMP can be interpreted as giving the expected performance of a comparison methodology across some relevant population. The RMP addresses the question: "In general, what is the ability of a certain biometric to match samples to source?"

A natural point estimate of the RMP is the sample proportion of matches in all pairwise comparisons; this estimate is a *U*-Statistic of degree 2. For their 2011 *Journal of Forensic Sciences* paper "Using Automated Comparisons to Quantify Handwriting Individuality," Saunders, et al. used *U*-Statistics results and adjustments to the Wald interval given in Wayman[2] to yield coverage probabilities close to the nominal confidence levels to estimate RMPs. Research using similar approaches on subsampled data from automated systems has continued under this grant. A paper, written by Drs. Davis, Saunders and Buscaglia, using modern resampling methods to estimate the RMP as a function of the quality of the samples being compared by a biometric matcher is in preparation for journal submission and is included as Phase I, Part A in the Main Section of the Final Report.

For this grant, we first completed a survey of the statistical theory for *U*-Statistics with zero-one kernels; the focus of the survey was specifically related to the behavior of *U*-Statistics when used as estimators of small probabilities. The most recent research on RMPs that we have found is by Michael E. Schuckers and is summarized in his 2010 book *Computational Methods in Biometric Authentication: Statistical Methods for Performance Evaluation.* Schuckers (and almost all other researchers in this area) have not put the estimation of random match probabilities into the context of *U*-Statistics. However, Schuckers has identified the dependency structure that arises when performing all pairwise comparisons and has incorporated this dependency structure into his confidence intervals, mainly via bootstrap methods. The non-bootstrap methods are

[2] Wayman, J. L. (2000). Confidence interval and test size estimation for biometric data. *National Biometric Center Collected Works 1997-2000*. J. L. Wayman. San Jose, CA, National Biometric Test Center: 89–99.

analogous to those used by Bickel and Wayman which we previously reviewed in our grant proposal.

Besides Dr. Saunders, three researchers on this grant worked on the estimation of confidence bounds for the RMP.

Dr. Linda Davis worked to find exact formulas for the mean and variance of the estimate of the RMP. Her development utilized structures and results from the theory of *U*-Statistics. The resulting exact formulas are only computationally tractable for very small sample sizes. Dr. Davis' work pointed out that assuming that all pairwise comparisons of samples are independent will lead to an underestimate of the variance of the estimate of the RMP. She also showed that basing the statistics on only a set of independent pairwise comparisons will lead to an overestimate of the variance of the estimate of the RMP. Dr. Davis introduced a scenario for which tighter bounds are possible for the variance of the estimate of the RMP.

Two documents prepared by Dr. Davis are attached to the Final Report in Appendices 1 and 2:

> "Link Between *U*-Statistics With 0-1 Kernels and the Union/Intersection of Events" This document presents the exact formulas for the mean and variance of the estimate of the RMP in the general case and in a special case.
>
> "RMP Confidence Interval"

These documents present issues associated with finding confidence interval bounds for estimation of the RMP and presents an approach to calculating bounds in a special case. A list of references concerning relevant statistical estimation is also included in these documents.

Dr. Davis intends to submit papers based on these two documents to research journals.

Drs. John Miller, Donald Gantz, and Christopher Saunders have developed a general parametric model for studying the distribution of pairwise comparisons of an arbitrary type tailored for small sample sizes with possibly no observed matches. The advantage of having a parametric model is that it provides an added level of structure for estimating the RMP with limited information. Furthermore, as long as the parametric model is chosen carefully, the resulting estimates appear to have a high degree of accuracy. This model is designed to incorporate the dependencies that arise in such studies with pairwise comparisons.

We introduced the parametric model to pursue our research goal of extending our *U*-Statistics based methods for estimating the RMP to the situation of small sample sizes. We are building upon the early research of Blom (1976) [3] to provide a parametric model that retains the optimal asymptotic properties of the *U*-Statistic estimate of the RMP (in the sense of being a Best Linear Unbiased Estimator of the RMP) but facilitates different estimation approaches, such as

[3] Blom, Gunnar," When is the Arithmetic Mean Blue?," *The American Statistician*, Volume 30, Issue 1, February 1976, pages 40-42.

Maximum Likelihood Estimates, Restricted Maximum Likelihood Estimates[4] (REML), and even Bayesian estimates.

The parametric model we are implementing treats the joint distribution of comparisons as a multivariate normal distribution. This approach is conceptually analogous to applying the standard Wilson Interval to estimating a proportion from a binomial random variable. This distributional assumption is only a tool used to facilitate the estimation of the RMP and is not expected to actually match the joint distribution of the discrete pairwise comparisons.
We have derived the theoretical foundation for the parametric model. We have demonstrated the use of this model in the construction of REML estimates and bounds for the RMP. We have run simulations that study the performance of the different estimates.

At the 2011 NIJ Trace Evidence Symposium, Drs. Gantz, Miller and Saunders presented their initial results on using a parametric method for estimating the RMP and constructing upper confidence bounds through non-asymptotic methods. They have recently completed a research paper on their work which has been submitted to the journal *Technometrics*. This paper studies in detail the parametric model for pairwise comparisons used in Forensic Science. It describes the eigenstructure of the covariance matrix and shows the consequences of the relations given by assuming normal distributions for the random components of the model. It shows that a closed form for an ANOVA table is possible. It shows that by using a method related to Fieller's Theorem, one can construct confidence intervals for a fixed component of the model which can then be easily turned into a confidence interval for the RMP. It also shows that two competing methods are either too conservative or just incorrect. The paper is included as Phase I, Part B of the Final Report. In the Principal Investigator's view this is a major achievement of the research in this project.

R. Bradley Patterson, a PhD Candidate supported by the grant, and Drs. Miller and Saunders authored a report that demonstrates the utility of ROC curves in forensics, where the goal is to measure the performance of methods that evaluate evidence. ROC curves offer several benefits to forensics. In contrast to the RMP, ROC curves capture the full range of error rates achievable with a method. They also depict the relative separation of the distributions of similarity scores from a given method. This then allows for comparisons of methods that produce scores on different scales. Additionally, an important characteristic for a method of evaluating pairs of evidence is the probability that a randomly selected pair from the same source would have a higher similarity score than a randomly selected pair from different sources, which the area under the curve (AUC) can estimate. To show the value of ROC curves in forensics, Patterson applied them to measuring the performance of methods of evaluating trace evidence in the form of glass fragments. The methods, based on test statistics and likelihood ratios, came from an article by Aitken and Lucy (2004). Test statistics and likelihood ratios both provide measures of association between two samples. So those values are interpreted as similarity scores, with which

[4] Restricted maximum likelihood (REML) is a particular form of maximum likelihood estimation which does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function calculated from a transformed set of data, so that nuisance parameters have no effect. (Dodge, Yadolah (2006). *The Oxford Dictionary of Statistical Terms*. Oxford [Oxfordshire]: Oxford University Press. ISBN 0-19-920613-9)

ROC curves were created for the same data as the original article. The ROC curves provided measurements of the full performance of the methods across all thresholds as well as an even basis for comparison. All of the methods performed very well. This report is included as Appendix 3.

Phase II and Phase III

Goal: Investigate Properties of Approximate Methods for Evidence Interpretations such as Score Based Likelihood Ratios

The Utilization of Data Generated through Automated Systems

Throughout the grant, the researchers have utilized forensic data generated by automated systems. For many years, the research team has played a significant role in the development of automated systems for forensic handwriting and fingerprint identification. In particular, the team has developed the scoring algorithms that exploit quantification systems for both handwriting and fingerprints. (See Saunders' and Gantz's Vitas for a complete list of these research projects.) Both Drs. Gantz and Saunders were invited presenters at the Measurement Science and Standards in Forensic Handwriting Analysis (MSSFHA) Conference, June 4 – 5, 2013. The National Institute of Standards and Technology (NIST) hosted the MSSFHA Conference which was planned and organized in collaboration with the American Academy of Forensic Sciences – Questioned Document Section, American Board of Forensic Document Examiners, American Society of Questioned Document Examiners, Federal Bureau of Investigation Laboratory, National Institute of Justice (NIJ), and Scientific Working Group for Forensic Document Examination (SWGDOC).

Attendees, both in person and via a live webcast, included representatives from the collaborating institutions as well as universities, federal agencies, forensic laboratories, and the private sector. Dr. Gantz presented the Forensic Language-Independent Analysis System for Handwriting Identification (FLASH ID) in the Advances in Measurement Science in Handwriting Session. He stressed the accuracy of the automated system which finds identifying power from measured characteristics not directly observed or addressed by examiners. Dr. Saunders spoke on Understanding Individuality of Handwriting Using Score-Based Likelihood Ratios in the Advances in Statistics for Handwriting Analysis Session-*this presentation summarized research directly funded by this research grant that is published in two papers in Forensic Science International*[5]. His examples concerning Score-Based Likelihood Ratios are based on joint work of Drs. Davis, Saunders, Hepler, and Buscaglia and were generated using data from FLASH ID. In this presentation Dr. Saunders summarized research results (from this grant) which

[5] Davis LJ, Saunders CP, Hepler A, Buscaglia J. Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. *Forensic Sci Int.* 2012 Mar 10; 216(1-3):146-57.
Hepler AB, Saunders CP, Davis LJ, Buscaglia J. Score-based likelihood ratios for handwriting evidence. *Forensic Sci Int*. 2012 Jun 10; 219(1-3):129-40.

We are including preliminary drafts of these in papers in Appendices 4 and 5.

demonstrated that common approaches to approximating the value of forensic evidence can lead to radically different values of evidence. These results are summarized in the previously mentioned papers in Forensic Science International.

To conclude the Conference, moderators led a facilitated discussion on the future state of forensic handwriting analysis, specifically focusing on the following questions: What does the future state of handwriting analysis look like; What are the barriers to implementing the future state; and what does a roadmap to achieve the future state look like? The final report summarizing the concluding discussion stated, "The future state of the discipline will incorporate the use of more quantitative analysis tools during the handwriting examination process to assess and compare handwriting characteristics. Forensic document examiners (FDEs) will employ the use of statistical models to explain the significance of their conclusions based on the uniqueness of observed and measured handwriting characteristics." Further, the report stated, "It is important to note that automated comparison systems may be considered separate from statistical models, as automated systems can facilitate the matching of a known writer with questioned documents without necessarily generating statistics. This technology provides support during the examination process and may provide new information for the human examiner to consider. FDEs can use statistics and automated systems to complement their current practices and to enhance the way they review cases, but neither can replace humans."

Drs. Gantz and Saunders presented similar messages concerning fingerprint forensics in the Statistics in Forensic Science Topic Contributed Paper Session at the Joint Statistical Meetings in Montreal in August 2013. Dr. Gantz presented his paper "A Similarity Score for Fingerprint Images." The paper co-authored with John Miller describes the scoring algorithms he developed for a totally automated innovative technology enabling the identification of crime scene fingerprints. The presentation was selected to receive an Honorable Mention in the Section on Physical and Engineering Sciences (SPES) Outstanding Presentation Awards indicating that it was among the best of the 73 talks presented in a SPES-sponsored contributed paper session. Dr. Gantz made the same statement he had made concerning automated handwriting identification systems, namely that automated systems are differentiated from statistics and that due to their accuracy and use of novel information they will impact the practice of examiners. Dr. Gantz's scoring algorithms developed for a totally automated technology enabling the identification of crime scene fingerprints are presented in some detail in the full Final Report.

In his presentation, "On Desiderata for Score-Based Likelihood Ratios for Forensic Evidence," Dr. Saunders stated opinions on the desirable features of score-based likelihood ratios (SLRs) for interpreting and presenting forensic evidence. Dr. Gantz is providing Dr. Saunders with latent print based data from automated systems for use in score-based likelihood ratio examples in future research.

Identification of Specific Source

The set of identification of source problems that we have studied considers two alternative and mutually exclusive, but non-exhaustive, propositions or models for how the forensic evidence has arisen. The first model usually corresponds to the prosecution hypothesis and states that a given specific source is the actual source of the trace of unknown origin. The second proposition

usually corresponds to the defense hypothesis and states that the actual source of the trace is not the one considered under the prosecution hypothesis, but that it originates from another, unrelated, source in a specified relevant alternative population of sources.

The evidence that we have to address the validity of the two propositions takes the following form:

1. There is a specific source of interest, from which we have a set of samples, denoted as E_s.
2. There are a set of samples of sources from a population of alternative sources, denoted as E_a.
3. A set of samples from a common, but unknown, source denoted as E_u.

The forensic scientist and statistician are then asked to quantify how much support the evidence provides for the model that E_u arose from the specific source of interest when compared to the model that E_u arose from a source in the alternative source population.

Dating back to the 1970's, this problem has been approached within the context of subjective Bayesian hypothesis testing. (See Aitken and Stoney[6]; Lindley 1978[7]; and Shafer[8]). The common approach to these problems is to assume that the problem is inherently low dimensional, the stochastic nature of the evidence can be characterized by a common parametric family of distributions, and that the evidence from the alternative source population is sufficiently precise that it completely characterizes the stochastic nature of the alternative source population. With these assumptions in hand, the forensic statistician can then provide a summary of the scientific evidence that is logical and coherent for updating a prior belief structure concerning the two competing propositions. The 'summary' is typically known as a Bayes Factor in the statistical literature (IJ Good[9]) and a 'Likelihood Ratio' in the forensic science literature. Traditionally this summary is presented as follows:

$$\underbrace{\frac{Pr\left(H_p \mid E, I\right)}{Pr\left(H_d \mid E, I\right)}}_{\text{Posterior Odds}} = \underbrace{\frac{Pr\left(E \mid H_p, I\right)}{Pr\left(E \mid H_d, I\right)}}_{\substack{\text{Bayes Factor and/or} \\ \text{Likelihood Ratio}}} \times \underbrace{\frac{Pr\left(H_p, I\right)}{Pr\left(H_d, I\right)}}_{\text{Prior Odds}},$$

where E is the evidence, H_p is the prosecution model for the stochastic nature of the evidence,

---

[6] Aitken, C. G. G., Stoney, David A., *The Use Of Statistics In Forensic Science*, CRC Press, Oct 31, 1991.
[7] Lindley,D.V. (1,977), A Problem in Forensic Science, *Biometrika* 6,4, 207-213.
[8] Glenn Shafer, Lindley's Paradox, *Journal of the American Statistical Association* , Vol. 77, No. 378 (Jun., 1982) , pp. 325-334.
[9] Good, I.J., Weight of evidence and the Bayesian Likelihood Ratio published in *The Use Of Statistics In Forensic Science*, CRC Press, Oct 31, 1991.

H_d is the defense model for the stochastic nature of the evidence and I is the relevant background information common to both models. The prior odds summarize our relative belief concerning the validity of the prosecution and defense probability models.

The Bayes Factor then allows us to update our belief and arrive at the Posterior odds concerning the relative validity of the two models. If the Bayes Factor (and the corresponding Posterior odds) is sufficiently high relative to the prior odds, then we conclude in favor of the prosecution model for the stochastic nature of the evidence; on the other hand if it is sufficiently close to zero, we conclude in favor of the defense model for the stochastic nature of the evidence. In effect the Bayes Factor is providing a numerical summary of the answer to both of these questions:

> *"What do we believe the likelihood of observing the evidence under the prosecution model is?"*
>
> vs.
>
> *"What do we believe the likelihood of observing the evidence under the defense model is?"*

An extremely important note is that, when constructing a Bayes Factor, it is necessary to use a probability measure to characterize the forensic scientist's belief about the stochastic nature of how the specific source generates evidence. The traditional default belief measure concerning the specific source is that the specific source is typical of the population of alternative sources. (Aitken and Taroni[10])

In the context of formal Bayesian Model selection, the goal of a statistical analysis is to rigorously quantify the belief concerning the validity of a given model after having observed the evidence. This type of analysis is typically decomposed into various components – the first being the prior belief concerning the relative validity of the two competing models. The second is a set of priors for prosecution and defense models that characterize the belief about the parameters of the stochastic models.

Our research program has taken two directions related to this problem of the quantification of the value of evidence. The first is concerned with various aspects the development of an approximate value of the evidence for complex evidence forms when the actual likelihood structure is intractable (the main thrust of Phase II). These approximate values of the evidence are commonly referred to as Score Based Likelihood Ratios (SLRs) in the statistical literature.

The second direction concerns the formal development of the value of evidence when the forensic scientist has to estimate the background population defined by the defense proposition or model (the focus of Phase III). This line of work has been more narrowly focused on formal Bayesian methods.

---

[10] Aitken, C. G. G., Taroni, F., *Statistics and the Evaluation of Evidence for Forensic Scientists*, Wiley, 2004, 2nd Edition.

In February and March of 2014, Dr Saunders is giving two talks, one invited presentation at Pittcon and another at the Annual Meeting of the American Academy of Forensic Sciences on Statistical Aspects of the Forensic Identification of Source Problems. These talks are presentations of the results of Phase III of this research grant. This research describes how to incorporate incomplete information about the background population into a forensic likelihood ratio in a statistically rigorous manner. We will provide an overview of these results in the Project Narrative.

In this summary we have only highlighted some of the presentations and research performed under this grant. Please see the Main Report for additional information and referenced Appendices.

We would like to acknowledge the contributions made to this Final Report through the comments of the external peer reviewers.

Main Body of the Final Technical Report

Introduction: This Report is separated into five parts. Each Part summarizes the research endeavors performed by the grantees corresponding to the major goals of the Project. Each Part is formatted as a detailed technical report. We chose to present the Report in this fashion due to the broad scope of this research Project. We have not included any previously published material in this Report even though a number of the results are published in various forms.

Phase I: (Goal) Study Interval Estimates of a measure of the validity of a forensic individualization system.

Dr. Christopher Saunders motivated the research in Phase I in his 2010 AAFS presentation which gave an overview of the goals for this research grant. During Phase I we focused on the Random Match Probability (RMP) as a measure of the validity of a forensic individualization procedure. Specifically, our research has been concerned with upper confidence bounds on measures, such as the RMP, that are estimated using automated pairwise comparisons. Pairwise comparison of samples is fundamental to forensic individualization systems. The validity of pairwise comparisons depends on the ability to effectively discriminate between samples of different origin and to accurately match samples of a common origin.

The RMP is defined as the probability of selecting two distinct sources at random from a population that "match" on the basis of some biometric sample extracted from each. The RMP can be interpreted as giving the expected performance of a comparison methodology across some relevant population. The RMP addresses the question: "In general, what is the ability of a certain biometric to match samples to source?"

A natural point estimate of the RMP is the sample proportion of matches in all pairwise comparisons; this estimate is a *U*-Statistic of degree 2. For their 2011 *Journal of Forensic Sciences* paper "Using Automated Comparisons to Quantify Handwriting Individuality," Saunders, et al. used *U*-Statistics results and adjustments to the Wald interval given in Wayman to yield coverage probabilities close to the nominal confidence levels to estimate RMPs. Research using similar approaches on subsampled data from automated systems has continued under this grant. A paper, written by Drs. Davis, Saunders and Buscaglia, using modern resampling methods to estimate the RMP as a function of the quality of the samples being compared by a biometric matcher is in preparation for journal submission and is included below as Phase I, Part A of this Final Report.

Phase I Part A:

Using Subsampling to Investigate the Dependency of Match Probabilities on the Size of Writing Samples*

Linda J. Davis, PhD [1, 2]; *Christopher P. Saunders*, PhD[1, 3]; *and JoAnn Buscaglia*, PhD[4]

[1]*Document Forensics Laboratory (MS 1G8), George Mason University, Fairfax, VA 22030*
[2]*Department of Statistics (MS 4A7), George Mason University, Fairfax, VA 22030*
[3]*Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007*
[4]*FBI Laboratory, Counterterrorism & Forensic Science Research Unit, Quantico, VA 22135*

The research detailed in this section was supported in part by Award No. 2009-DN-BX-K234 awarded by the National Institute of Justice, of Justice Programs, US Department of Justice.

The opinions and conclusions or recommendations expressed in this publication are those of the author and do not necessarily represent those of the Department of Justice.

*This work was supported in part under a Contract Award from the Counterterrorism and Forensic Science Research Unit of the Federal Bureau of Investigation's Laboratory Division. Names of commercial manufacturers are provided for information only and inclusion does not imply endorsement by the FBI. Points of view in this document are those of the authors and do not necessarily represent the official position of the FBI or the US Government.

[3]Partially supported by an IC Post Doctorial Research Fellowship, NGIA HM1582-06-1-2016

ABSTRACT

A *random match probability* (RMP) of interest in handwriting analysis is the chance of randomly selecting two different individuals from some relevant population and then randomly selecting a writing sample from each individual that are declared to "match" by a specific comparison procedure. A complementary probability, the *random non-match probability* (RNMP), is the chance of randomly selecting a single individual and then randomly selecting two writing samples from the selected individual's body of handwriting that fail to "match." In handwriting analysis, the RMP and the RNMP are standard measures of a comparison procedure's ability to discriminate among writers; both depend upon the comparison procedure used and the sizes of the writing samples being compared, as well as the

relevant population from which individuals are selected. In this study, we investigate how subsampling from available writing samples can be used to: a) investigate the dependency of the RMP and the RNMP on the sizes of the writing samples being compared; b) estimate the standard error of one estimator of the RMP (such as might be used in constructing an upper confidence bound for the RMP); and c) provide information useful for planning an empirical study of handwriting individuality.

KEYWORDS: forensic science, random match probability, handwriting individuality, writer verification, forensic document analysis

1. Introduction

One goal of a *forensic document examiner* (FDE) in evaluating a questioned document might be to determine the specific individual that wrote that specific document. One step towards this goal is the comparison of features of the questioned document with features of an exemplar writing sample from a known source. However, a perfect match between two writing samples written by the same individual is not expected.

One reason two writing samples from the same individual will not have exactly the same features is the natural variations in an individual's handwriting. As discussed by Huber and Headrick (1999, pp. 73–74), comparing writing samples ultimately is comparing writing habits across distinct individuals; characteristics of writing as measured in features or qualities are simply manifestations of habits formed over time. We will refer to the totality of accumulated habits, as reflected in one's entire body of natural handwriting, as an individual's *writing profile*. Note that an individual's writing profile is more akin to a probability distribution across documents generated by that individual than a static characteristic of an individual, such as a fingerprint or DNA (Bulacu and Schomaker, 2007).

Due to this natural variation in handwriting, a useful tool for assisting in the comparison of two writing samples might utilize some automated procedure to quantify this variability. For comparing two writing samples, such an automated comparison procedure could take (the scanned images of) two writing samples, convert these writing samples to a set of quantitative features, and then compute a *similarity score* based on these features as a measure of the similarity of the writing profiles that generated the two writing samples. With the introduction of a threshold value, a pair of writing samples can be declared to "match" (using this automated comparison procedure) if the similarity score exceeds the predefined threshold value. Otherwise, two samples are declared to "not match."

Declaring two writing samples to match (using an automated comparison procedure) provides a measure of the consistency of the writing profiles generating the two samples. However, such a match between two writing samples cannot be interpreted as proving that one individual wrote both writing

samples because there are two types of unavoidable errors associated with comparing writing samples in this manner. If two writing samples generated by different individuals are declared to match, then a *false match error* has been committed. And, if two writing samples generated by the same individual are declared to not match, then a *false no-match error* has been committed. These errors are consequences of within-writer variation and between-writer similarity (Risinger and Saks, 1996), which we are characterizing via an individual's writing profile.

The rates of these two types of errors can be used to characterize a comparison procedure's ability to discriminate among writers. One measure of the false match error rate of a comparison procedure is what we shall refer to in this paper as the *random match probability* (*RMP*). The RMP in this paper is defined as the probability of selecting two individuals at random from the relevant population and two randomly selected writing samples, one from each individual's body of handwriting, that match. It can be viewed as the rate of false match errors "averaged" over all relevant writing samples. A related quantity is the *random non-match probability* (*RNMP*); it provides one measure of the false no-match error rate of a comparison procedure. The RNMP in this paper is defined as the probability of randomly selecting an individual from the relevant population and then selecting two writing samples at random from the selected individual's body of handwriting that fail to match. It can be viewed as the rate of false no-match errors "averaged" over all relevant writing samples.

In addition to their dependence on the comparison procedure itself (i.e., the associated similarity score and threshold value used to declare a match or no match based on the similarity score), the RMP and the RNMP depend upon:

- The relevant population of individuals (more specifically, writing profiles) generating the writing samples being compared. Some individuals' writing profiles are harder to distinguish between than others.
- The sizes and content of the writing samples (such as the number of characters and distribution of letters) being compared.

Due to the potential for false match and false no-match errors, information must be available on both the associated RMP and RNMP for a comparison procedure to be of practical use. The focus of this paper is illustrating one approach to investigating the RMP and the RNMP associated with a comparison procedure, and in particular, their dependencies on the sizes of the writing samples available for comparison as measured by the number of characters in the writing samples. This dependency on sizes of writing samples is important as it relates to the ability of the comparison procedure applied to particular sizes of writing samples to distinguish between individual writers or individual writing profiles across some relevant population. Although also of interest, in this paper we do not investigate the dependencies of the RMP and the RNMP on the content of the writing samples. We assume that the content of the writing samples being compared reflect the frequency of letters as they appear in English writing.

This paper is organized as follows. First, we review the use of the RMP in other forensic settings and in particular, the relationship between the RMP and quantifying the degree of individuality of writing profiles. Then, we turn to the main focus of this paper: how simulated writing samples generated from a collection of writing samples can be used to investigate the dependency of the RMP and the RNMP on the sizes of writing samples being compared. We propose generating simulated writing samples by subsampling characters from a single writing sample available for each individual. A slight modification of this methodology is then introduced that can be used to investigate the standard error of an estimator of the RMP, which is an important component in constructing upper confidence bounds. We next describe a specific comparison procedure under investigation by the Document Forensics Laboratory at George Mason University and a set of writing samples collected by the FBI Laboratory and processed by Gannon Technologies Group (GTG). Using this comparison procedure and set of writing samples, we illustrate the methodology proposed in this paper and some of its potential applications such as providing information useful for designing an empirical study of handwriting individuality.

2. RMP and Individuality

The RMP, as defined above, appears in the general forensic literature as the probability of non-discrimination. (For an overview of this topic, see Aitken and Taroni (2004, Section 4.5).) Aitken and Taroni (2004) describe the RMP as a measure of how "good" a method is at distinguishing between biometric samples from different sources as well as a way to quantify how strong a declared match is as evidence that the two samples come from the same source. The smaller the RMP, the better the comparison procedure is for individualization, i.e., for making a positive identification of the source of some biometric sample.

As with a match itself, a small RMP in handwriting comparisons does not imply uniqueness because a small RMP does not exclude the possibility that two individuals have the same writing profile. As mentioned in Saks and Koehler (2008), infrequency cannot be equated to uniqueness. Balding (2005) uses the term "the uniqueness fallacy" to describe the fallacy in cases involving DNA evidence where a set of genetic markers that are expected to occur less than once in five billion (roughly the earth's population) are declared to be unique. Although a small RMP would be a consequence of unique writing profiles, it does not imply such even when it is smaller than one over the earth's population.

However, the RMP is related to the *degree* of individuality of writing profiles in a population. See Bolle, *et al.* (2004) and Saunders *et al.* (2011a) for a detailed discussion of this relationship. In fact, using the size of the RMP is one approach to the question of uniqueness, within the context of DNA profiles, discussed in a report from the National Research Council (1996, pp. 136–138). This report suggests that identification (beyond a reasonable doubt) may mean that the probability that there is at least one match when the DNA profiles of individuals in the population are compared is small, say 1%, or some other chosen small number. (However, the report from the National Research Council (1996) is careful to point out that it is up to the courts to decide just how small this probability should be to support individualization.)

The report from the National Research Council (1996) also includes a formula for an upper bound on this probability of at least one match (when comparing DNA profiles) using population genetics modeling, which depends on the population size and the number of loci compared in the typing. Unfortunately, to date, similar type models have not been developed to adequately characterize an individual's writing profile. In this paper, we propose an alternative approach to such modeling that provides information about the size of the RMP and its relationship to the sizes of the writing samples being compared.

3. Estimating the RMP and the RNMP

One approach to investigating the RMP and the RNMP associated with a specific comparison procedure involves estimating them from a collection of writing samples.

Consider a collection of writing samples consisting of a single writing sample (which may be composed of one or more documents collected at different times or in different environments) from each of N writers. We assume that the N writers can be considered a random sample from some relevant population of individuals.[11] We also assume that each writing sample is "representative" of its associated individual's writing profile.[12] Together, these assumptions imply that the collection of writing samples is independent and identically distributed (*iid*).

One estimator of the RMP is based on all $N(N-1)/2$ pairwise comparisons between writing samples in the collection. For $i \neq j$, let $s(D_i, D_j)$ denote a score that measures the similarity between two writing samples D_i and D_j from the i^{th} and j^{th} writers in the collection. Let τ be a threshold used to declare matching writing samples (via the comparison procedure). Then, a natural estimator of the

[11] Specifically, we assume the population of individuals is so large that it is reasonable to treat the sampled individuals as independent and identically distributed (*i.i.d.*) according to some distribution on the relevant population of individuals.

[12] That is, each writing sample can be viewed as randomly generated from that individual's writing profile.

RMP for a given comparison procedure is the proportion of pairs of writing samples that match, i.e., for which $s(D_i, D_j) > \tau$. This proportion is an unbiased estimator of the RMP. See Appendix A for more details about the properties of this estimator of the RMP, including an expression for its standard error which can be used to construct upper confidence bounds for the RMP.

Suppose now that instead of a single writing sample, the collection contains two writing samples from each writer (represented in the collection). Let $s(D_{i1}, D_{i2})$ denote the score measuring the similarity between two writing samples D_{i1} and D_{i2} from the i^{th} writer in the collection. Then, a natural estimator of the RNMP for a given comparison procedure is the proportion of pairs of writing samples from the same writer that do not match, i.e., for which $s(D_{i1}, D_{i2}) \leq \tau$. See Appendix A for more details about this estimator of the RNMP.

However, these proposed estimators of the RMP and the RNMP applied to a single collection of writing samples are of limited use when investigating the dependency of the RMP and the RNMP on the sizes of the writing samples being compared. The writing samples in a collection may be of different sizes. And, even if the available writing samples are of approximately the same size, direct comparison will only provide information about the RMP and the RNMP for that one size of writing sample. So, to investigate the dependencies of the RMP and the RNMP on sizes of writing samples being compared, one would need access to multiple writing samples of specific sizes from each of N writers randomly selected from the relevant population. Fortunately, as described in the next section, such writing samples can be "simulated" from a single collection of observed writing samples, as long as the sizes of writing samples of interest are smaller than the sizes of the observed writing samples.

4. Simulated Writing Samples

Often, one may not have access to writing samples of the types needed to study the behavior of the RMP and the RNMP associated with a specific comparison procedure. If an individual's writing profile is known, then one could generate any number of writing samples of any specified sizes by sampling from

the writing profile via Monte Carlo simulation.[13] However, as mentioned previously, an individual's writing profile is rarely, if ever, known.

Alternatively, if reasonable models for writing profiles are available, one could consider estimating the parameters associated with such models from the available writing samples and then generating any number of writing samples of any specified sizes from these fitted models. However, to date, reasonable models for writing profiles have not been developed.

Instead, we propose generating simulated writing samples by subsampling characters from a single writing sample available for each individual. The proposed methodology allows generation of writing samples of different sizes, which permits investigation of the RMP for sizes of writing samples that differ from those of the original samples. Furthermore, the proposed methodology allows generation of multiple samples from a single writing sample. So, it can be used to investigate the RNMP as well as the RMP.

Creating replicate samples from observed samples is the key idea behind many of the current resampling methods being studied in statistics: use the original data to represent the population and then generate samples from the "estimated population" (i.e., the original data) to create replicate samples. These replicate samples can then be used to estimate properties of the original population, just as if one had access to such samples from the actual population.

Most resampling methodologies are examples of the plug-in principle in statistics. Basically, the plug-in principle operates by estimating a property of a population using the statistic that is the corresponding property of the sample. Resampling substitutes the available data for the population and then draws samples (i.e., resamples) to mimic the process of building the sampling distribution.

Resampling methods still typically rely on the same Monte Carlo techniques used when the population distribution is known. In principle, one could consider all possible replicate samples that

[13] Monte Carlo simulation allows estimating properties of a distribution by generating samples from that distribution.

could be generated from the observed, but it would be too time-consuming and computer intensive. Instead, Monte Carlo resampling is used to restrict the number of replicate samples examined. The fundamental difference between Monte Carlo simulation and resampling is that, in the former, the underlying distribution from which samples are selected is assumed known, whereas, in the latter, it is assumed unknown and thus the simulation must be based on observed data.

Simulating samples via resampling from the original sample can be applied to many complicated statistical analyses. However, regardless of the application, it is very important that the simulated samples mimic the distribution of actual samples so that properties of the simulated samples provide valid estimators of the population characteristics of interest.

In generating our simulated writing samples, subsampling, i.e., sampling characters without replacement, is crucial because it produces writing samples that are distributed according to the associated underlying writing profile. Sampling characters with replacement from the observed writing sample produces writing samples distributed according to a slightly different writing profile as described in Appendix B. In particular, estimators based on simulated writing samples generated by sampling characters with replacement are not necessarily consistent as the number of writers goes to infinity while the size of writing sample remains small (i.e., contain a small number of characters).[14]

In the following sub-sections, we describe the specific details of the subsampling we propose for estimating the RMP, the RNMP, and the standard error associated with the estimator of the RMP described in Appendix A. As in Section 3, we assume the availability of an *iid* collection of writing samples consisting of a single writing sample from each of N writers.

[14] To say that an estimator is consistent in this case means that for sufficiently large number of writers, it is expected that the estimator is very close to the value for the entire population.

4.1 *Estimating RMP*

To investigate how the RMP varies as a function of sizes of writing samples, we propose the following algorithm for estimating the RMP associated with comparing two writing samples of specified (common) size.

Algorithm 4.1

1. Randomly select two writers without replacement. (This is equivalent to random sampling from all possible pairs of writers.)
2. For each selected writer, construct a simulated writing sample by selecting, without replacement, a pre-specified number, say n, of characters from that writer's total writing sample.
3. Calculate the score $s(D_1^*, D_2^*)$ where D_1^* and D_2^* are the two simulated writing samples from Step 2.

Repeat Steps 1, 2, and 3 a total of K times, for fixed size n of writing samples, resulting in a set of K scores: $\left\{ s(D_1^{*(k)}, D_2^{*(k)}) : k = 1, 2, ..., K \right\}$.

Analysis of the set of between-writer similarity scores created by this algorithm provides information about the distribution of the similarity score when applied to two writing samples, each of size n, from different individuals.

For a specified threshold τ, the proportion of pairs of simulated samples that match, i.e., for which $s(D_1^{*(k)}, D_2^{*(k)}) > \tau$, is an unbiased estimator of the RMP when comparing two writing samples each of size n. This estimator of the RMP is consistent as the number of writers increases for fixed sizes of writing samples being compared.

Note that this algorithm as stated does not allow investigating the dependency of the RMP on the content of the writing samples being compared. To do this, the algorithm would have to be modified to perform some type of stratified sampling by letter, or to select a systematic sample from the original

writing sample instead of a random sample across characters. Both of these modifications to the algorithm (i.e., stratified and systematic sampling) are currently under investigation as techniques for investigating the dependency of the RMP on the content of the writing samples being compared.

4.2 *Estimating RNMP*

A similar algorithm can be used to investigate how the RNMP varies as a function of the sizes of writing samples being compared. This algorithm differs from that for estimating the RMP because it involves pairs of simulated writing samples from a single writer, instead of simulated writing samples from a pair of writers.

Algorithm 4.2

1. Randomly select a writer.
2. For the selected writer, construct two independent simulated writing samples by selecting, without replacement, a pre-specified number, say n, of characters from that writer's total writing sample.
3. Calculate the score $s(D_1^*, D_2^*)$ where D_1^* and D_2^* are the two simulated writing samples from Step 2.

Repeat Steps 1, 2, and 3 a total of K times, for fixed size n of writing samples, resulting in a set of K scores: $\left\{ s(D_1^{*(k)}, D_2^{*(k)}) : k = 1, 2, ..., K \right\}$.

Analysis of the set of within-writer similarity scores created by this algorithm provides information about the distribution of the similarity score when applied to two writing samples, each of size n, from the same individual.

For a specified threshold τ, the proportion of pairs of simulated samples that do not match, i.e., for which $s(D_1^{*(k)}, D_2^{*(k)}) \le \tau$, is an unbiased estimator of the RNMP when comparing two writing samples each of size n. This estimator of the RNMP is consistent as the number of writers increases for fixed

sizes of writing samples being compared. As mentioned in the previous section, this algorithm as stated does not allow investigating the dependency of the RNMP on the content of the writing samples being compared; the criterion used in selecting the characters to make up the simulated samples would need to be modified.

4.3 *Estimating Standard Error*

As detailed in Appendix C, the standard error of the estimator of the RMP described in Section 3 (with details in Appendix A) depends both on the RMP and upon the probability of randomly selecting three writing samples from different individuals such that the writing sample from the first individual matches both the writing samples from the second and third individuals. We will refer to this latter probability as the *tri-match probability* (*TMP*). The form of the variance of the estimator of the RMP is given in expressions A.2 and A.3 of Appendix A. The dependence of this variance on the TMP is derived in expression C.3 of Appendix C.

As its definition suggests, estimating the TMP requires comparison of three simulated writing samples instead of two. We propose the following algorithm to investigate the TMP as a function of the sizes of writing samples being compared.

Algorithm 4.3

1. Randomly select three writers without replacement. (This is equivalent to random sampling from all possible triplets of writers.)
2. For each selected writer, construct a simulated writing sample by selecting, without replacement, a pre-specified number, say n, of characters from that writer's total writing sample.
3. Calculate the score $s(D_1^*, D_2^*)$ comparing the first simulated writing sample D_1^* vs. the second D_2^*. Calculate the score $s(D_1^*, D_3^*)$ comparing D_1^* vs. the third simulated sample D_3^*.

Repeat Steps 1, 2, and 3 a total of K times, for a fixed size n of writing samples, resulting in a set of K pairs of scores: $\left\{\left(s(D_1^{*(k)}, D_2^{*(k)}), s(D_1^{*(k)}, D_3^{*(k)})\right) : k = 1, 2, ..., K\right\}$.

Analysis of the set of pairs of scores created by this algorithm provides information about the TMP and subsequently, the standard error of an estimator of the RMP. Specifically, for a specified threshold τ, the proportion of triplets of simulated samples that match, i.e., for which $s(D_1^{*(k)}, D_2^{*(k)}) > \tau$ and $s(D_1^{*(k)}, D_3^{*(k)}) > \tau$, is an unbiased estimator of the TMP when comparing three writing samples each of size n. This estimator of the TMP is consistent as the number of writers increases for fixed sizes of writing samples being compared.

5. Applications

In this section, we illustrate how the algorithms described in the previous section can be used in a variety of applications associated with automated comparisons of writing samples. To do so, we use a specific comparison procedure under investigation by the Document Forensics Laboratory at George Mason University and a collection of research writing samples collected by the FBI Laboratory and processed by

Gannon Technologies Group (GTG). However, the algorithms themselves are "generic" — applicable to any comparison procedure and collection of writing samples.

The collection of writing samples we use was constructed from documents collected by the FBI Laboratory from volunteers at the FBI, training classes, various forensic conferences, and from friends and family members over a two-year period. These documents form a *convenience sample*, not a random sample representative of some relevant population. They are used in this study only to illustrate the algorithms described in the previous section, not to make a statement about properties of any specific population.

Each volunteer was asked to provide ten samples (five in cursive and five in hand printing) of a modified London Business Letter (Osborn, 1929), which we will refer to in this paper as the modified "London Letter". The modifications to the London Business Letter, which were made by a FDE, consisted of the addition of two sentences at the end of the London Business Letter in order to incorporate some occurrences of specific letter combinations (e.g., "ch," "qu," "ll"). The text of the modified "London Letter" is shown in Figure 1 along with an example of a cursive writing sample. The particular text of the modified "London Letter" was selected because it gives a reasonable representation of the frequencies of lowercase letters in English writing and contains at least one instance of each uppercase letter and each of the digits 0 through 9.

Following is a brief description of how the writing samples were quantified; more details about the processing can be found in Walch and Gantz (2004). Subsequent to manual character segmentation of each document, a proprietary automated process was used to represent each segmented character by a mathematical graphic isomorphism whose internal structure can be enumerated by a code, which for simplicity we refer to as an isocode. This process ultimately reduces each document to the frequency of isocodes used to write each letter, which can be represented as a cross-classified table of letter by isocode (Saunders *et al.*, 2011b).

Most cursive documents and some printed documents from each of 100 volunteers were processed for use in this study, resulting in a total of 503 documents after processing[15]. In this study, the documents from a single writer were combined, so each individual has a single writing sample in the resulting collection of writing samples.

The similarity score we consider for comparing two writing samples, the *Chi-Squared Classifier*, is based on Pearson's chi-squared statistic (Saunders *et al.*, 2011b). This similarity score is calculated as follows:

1. Conditional on each letter, calculate Pearson's chi-squared statistic on a two-way table of counts with two rows. The two rows represent the two writing samples being compared. The columns represent the various isocodes used to write a given letter in at least one of the two writing samples being compared.
2. Sum these chi-squared statistics across all letters that appear in both writing samples. Also, because the writing samples may use a different number of isocodes to represent different letters, sum the degrees of freedom associated with the different chi-squared statistics.
3. Calculate the probability that a chi-squared random variable with the summed degrees of freedom exceeds the observed value of the summed statistic. This probability is the similarity score associated with the comparison.[16]

5.1 *Determining an Appropriate Threshold Value*

The RMP and the RNMP play a role in the selection of an appropriate threshold to use with a comparison procedure for declaring a match between two writing samples. One method for selecting a threshold

---

[15] Not all individuals participating in the study provided all requested copies. Also, not all of the available documents had been processed at the time of this study.

[16] The chi-squared statistic itself could be used as the similarity score. However, conversion to a chi-squared probability at least partially normalizes comparison of writing samples for different sizes and different content.

value is to choose the threshold such that the rate of false match errors equals the rate of false no-match errors. The resulting rate, called the equal error rate (EER), is a standard method for comparing the "matching" accuracy among comparison procedures, particularly those designed for biometric authentication systems.

Alternatively, one can select the threshold to give a pre-specified rate of no-match errors, say 1%. This is more typical in forensic settings where there is an asymmetry in the severity of the two types of errors, with the false match error usually being considered the more severe.

As an application of our proposed algorithms, consider determining the threshold value that will give a RNMP of 1%. If the individual processed characters can be considered a random sample from an individual's writing profile, the similarity score associated with the Chi-Squared Classifier is related to an approximate *p*-value. So, assuming independence across characters, (theoretically) the similarity score has approximately a uniform distribution when applied to two randomly selected writing samples from the same individual, regardless of the sizes and content of the two writing samples being compared. This suggests that the 1% RNMP threshold for the Chi-Squared Classifier should be 0.01 assuming independence across characters.

However, the independence assumption is questionable. Thus, the actual 1% RNMP threshold may not be 0.01 and may vary with the sizes and content of writing samples being compared. Using Algorithm 4.2, this choice of threshold and its dependence on sizes (but not necessarily its dependence on content) of writing samples being compared can be investigated empirically by estimating the RNMP for a variety of sizes of writing samples.

We applied Algorithm 4.2 with $K = 1{,}000$. Specifically, we ran the algorithm five times with simulated writing samples of (common) sizes varying between $n = 100$ and $n = 900$ by increments of 200. This resulted in five sets of 1,000 scores, one set for each (common) size of writing sample.

Figure 2 shows the *empirical cumulative distribution function* (ECDF)[17] of the 1,000 within-writer similarity scores from Algorithm 4.2 for each of the five sizes of writing samples investigated. Each plot is overlaid with a 45-degree line, which is the *cumulative distribution function* (CDF) for the uniform distribution.

As seen in Figure 2, the behavior of the EDCF is very similar for all of the five sizes of writing samples. Specifically, the EDCF and CDF (for a uniform distribution) are not close for all values suggesting the distribution of within-writer similarity scores is not uniform. In fact, for values larger than 0.10, the ECDF is much greater than the uniform CDF. This suggests that for all five sizes of writing samples, the within-writer similarity scores tend to be more concentrated toward small values than would be expected if the similarity scores followed a uniform distribution.

However, in each of the plots in Figure 2, the ECDF is close to or below the uniform CDF for score values less than 0.10. And, the EDCF appears to be getting closer to the CDF (for score values less than 0.10) as the common size of writing sample increases. This suggests that even though the uniform approximation is not good across the entire range of values of the similarity score, the 1% RNMP threshold is close to 0.01, or perhaps slightly larger than 0.01. So, based on the simulated writing samples generated using Algorithm 4.2, 0.01 appears to be a reasonable choice (although conservative, especially for smaller sizes of writing samples) for the threshold for use with the Chi-Squared Classifier to create a comparison procedure with a pre-specified rate of no-match errors of no more than 1%. In general, using a conservative value for the RNMP threshold will result in overestimating the RMP associated with the actual 1% RNMP threshold. However, this appears to be less of an issue as the common size of writing samples being compared increases.

[17] The ECDF is a plot of a score value versus the proportion of values in the set of scores that are less than or equal to the specified score value.

5.2 *Estimating the RMP as a Function of Size of Writing Samples*

Another application of our proposed algorithm is the investigation of the RMP associated with the Chi-Squared Classifier and a threshold of $\tau = 0.01$.[18]

We applied Algorithm 4.1 (with $K = 100$)[19] 41 times with simulated writing samples of (common) size varying between $n = 50$ and $n = 450$[20] by increments of 10. This resulted in 41 sets of 100 scores, one set for each common size of writing samples. For each set of scores, we calculated the proportion of the scores that exceeded 0.01; these proportions are shown in Figure 3.

Figure 3 illustrates the dependency of the RMP on the size of the writing samples being compared. The RMP approaches zero as the common size of writing samples gets large. We used a logistic regression model cubic in size of the writing sample, which provides a reasonable fit to the observed proportions, to provide a smooth curve representing the relationship between the RMP and size of writing samples. The resulting fit is summarized in Table 1 and shown as a solid line in Figure 3. Based on the fitted logistic curve, the RMP associated with the Chi-Squared Classifier with threshold 0.01 is less than 10% when comparing writing samples each with at least 280 characters, and less than 1% for comparing writing samples each with at least 400 characters.

---

[18] As suggested by the results in the previous sub-section, this threshold corresponds to a rate of non-match errors of at most 1%.

[19] A smaller *K* is used here than was used in the previous section when investigating the RNMP. The small number of writers represented in the database of writing samples limits the study of between-writer variability. With only 100 writers, sampling of more pairs results in many pairs involving the same writer; this results in additional simulations much beyond 100 providing little additional information. The reduction in size of *K* does increase the variability in the estimated RMP.

[20] With $K = 100$, it is not possible to accurately estimate very small RMP. Thus, we considered common size of writing samples of at most 450 instead of 900 as when investigating the RNMP.

5.3 *Estimating the TMP and Standard Error*

One approach to constructing upper confidence bounds on the RMP is to use a Wald-type upper confidence bound[21] as described in Appendix A. This approach requires knowledge of the standard error of the estimated RMP, which in our case is unknown. However, as discussed in Appendix C, the standard error of the RMP estimator is related to the TMP, which can be estimated using Algorithm 4.3.

We applied Algorithm 4.3 with $K = 100$[22], a threshold of $\tau = 0.01$, and common size of writing samples n varying between 50 and 450 by increments of 10. This resulted in 41 sets of 100 score pairs, one set for each size of writing samples. For each set of score pairs, we calculated the proportion of the pairs for which both scores exceeded 0.01; these proportions are shown in Figure 4.

Figure 4 illustrates the dependency of the TMP on the size of the writing samples being compared. The TMP approaches zero as the common size of writing samples gets large. As with the RMP, we used a logistic regression model cubic in size of the writing samples, which provides a reasonable fit to the observed proportions, to provide a smooth curve representing the relationship between the TMP and the size of writing samples. The resulting fit is summarized in Table 2 and shown as a solid line in Figure 4.

To simplify comparison, we have also included in Figure 4 the logistic fit (Table 1) to the estimated RMP. Comparing the two curves, note that for smaller sizes of writing samples, the estimated RMP and estimated TMP are similar. However, the estimated TMP drops off more rapidly with increased (common) size of writing samples. For example, based on the fitted logistic curve, the TMP associated with the Chi-Squared Classifier with threshold 0.01 is approximately 10% when comparing writing

---

[21] A Wald-type upper confidence bound is one based on a normal approximation to the sampling distribution of an estimator. It is typically of the form of point estimator plus some number of standard errors, where the number of standard errors added to the point estimate depends on the desired confidence coefficient. For example, for a 95% upper confidence bound, one would add 1.645 times the standard error to the point estimate.

[22] We used the same value of *K* and maximum common size of writing samples as for the RMP investigation in Section 5.2.

samples each with 190 characters while the corresponding RMP is 26.0%. The estimated TMP is 1% for comparing writing samples with 300 characters while the corresponding RMP is 7.1%.

Finally, Figure 5 shows estimates of the standard error of the estimated RMP as a function of number of writers for several different sizes of writing samples. These estimates combine the two logistic curves shown in Figure 4, along with an additional logistic model[23], using the equation for the variance of the estimated RMP given in Appendix A.[24] For illustrative purposes, we have included one extrapolated value for the size of writing samples, namely $n = 600$, which is outside the range of values used in the associated logistic model fits. Note that the standard error is more sensitive to the sizes of writing samples than to the number of writers represented in the sample, although clearly affected by both.

The plots of the standard error in Figure 5 suggest that for a small size of writing samples (i.e., 150), the standard error of the estimated RMP is between 0.01 and 0.1 for up to 2000 writers. Thus, such small sizes of writing samples are probably of limited use in trying to precisely bound a very small RMP. Doubling the size of writing samples to 300 does not provide much of a reduction in the standard error. Even with writing samples of size 450 characters, the standard error is only reduced to between 0.0001 and 0.001. Only with larger sizes of writing samples, such as 600 characters, does the standard error become small enough (between 10^{-7} and 10^{-6}) to accurately bound a very small RMP.

---

[23] The logistic model for the RMP shown in Figure 3 (and repeated in Figure 4) is fit to the proportion of pairs of simulated samples that match when comparing the first and second simulated samples in the output from Algorithm 4.3. The estimates shown in Figure 5 also use the same type of logistic model, also cubic in size of writing sample, fit to the proportion of pairs of simulated samples that match when comparing the first and third simulated samples in the output from Algorithm 4.3.

[24] We are not suggesting that the best estimator of the standard error from such data is to combine fitted logistic models for the RMP and the TMP. This estimator is shown here to illustrate one approach to estimating the standard error. See Appendix C for a more detailed discussion of this issue.

5.4 *Designing a Study of Handwriting Individuality*

Selecting the number of observations is a critical part of planning a study. One procedure for selecting the number of observations is to specify a desired margin of error associated with estimating a parameter of interest and then selecting the number of observations to produce such a margin of error.

In a study of handwriting individuality, one parameter of interest is the RMP, which is related to the degree of individuality of writing profiles in a population. As mentioned in Section 2 and discussed in detail in Bolle *et al.* (2004) and Saunders *et al.* (2011a), an upper bound on the RMP is also an upper bound on the rarity of matching writing profiles. Thus, one might consider selecting the number of writers for an empirical study of the individuality of handwriting within a specific population to produce a desired upper bound on the RMP (assuming that the "true" RMP is very close to zero).

The ideal setting for obtaining a small upper bound is when there are zero observed matches. In fact, the smallest possible upper bound on the RMP (when writing samples are compared pairwise) occurs when there are no observed matches in a collection of writing samples from a large number of writers.

One could use a Wald-type upper bound (Appendix A) with an estimated standard error such as described in Section 5.3. However, a standard error estimator based on simulated samples, such as proposed in Section 5.3, requires writing samples from a large number of writers to be very precise; and typically, such a large set of writing samples is not available at the planning stages of a study. Alternatively, one can use one of the proposed estimators of the standard error based directly on an observed set of writing samples without subsequent subsampling. However, most of these proposed estimators, such as those suggested by Sen (1960), Arveson (1969), Schucany and Bankson (1989), and Wayman (2000), cannot be used when there are zero observed matches.

For interval estimation of a proportion, Agresti and Coull (1998) illustrate that an adjusted Wald interval obtained after adding two "successes" and two "failures" to the sample yields coverage probabilities close to the nominal confidence levels. We have conducted a small simulation study to investigate the coverage probability of a Wald-type upper confidence bound on the RMP when a similar

type adjustment of adding one match and one no match[25] is made to Wayman's (2000) estimate of standard error. Our preliminary investigations suggest that this adjustment yields coverage probabilities close to the nominal confidence levels.

In the case of no observed matches and adding one match and one no match, the formula for the 95% upper confidence bound using Wayman's (2000) estimate of standard error simplifies to $4.65/[(N+1)(N+2)]$, where N is the number of writers. So, the larger the number of writers in the study, the smaller the upper bound when there are no observed matches. For example, assuming no observed matches, a sample of 963 writers would yield a 95% upper confidence bound on the RMP of 5 in one million and a sample of 2,154 writers would yield a 95% upper confidence bound on the RMP on the order of 1 in one million.

Either of these upper bounds, however, assumes the ideal scenario that there are no matches observed when the collected writing samples are compared pairwise. And, for a fixed (common) size of writing samples and fixed RNMP threshold, the probability of observing a match (provided the "true" RMP is not zero) goes up as we compare writing samples from more writers. However, as shown in Figure 3, the size of the writing samples affects the RMP and thus also affects the chance of observing no matches. Therefore, the size of writing samples, in addition to the number of writers, must be considered in order to obtain a small upper bound. In other words, there must be a balance between the number of characters in the writing samples and the number of writers providing writing samples in the study.

Using the result from probability theory that the probability of a union of events is less than or equal to the sum of the probabilities of the individual events, a simplistic lower bound on the probability of no observed matches is $1 - N(N-1)\theta/2$. So for a specified "chance" $1-\beta$ of observing no matches and a

25 More specifically, the adjustment involves adding one sample that matches exactly one observed and one sample that does not match any of the other observed samples. In other words, there is exactly one match out of $(N+2)(N+1)/2$ comparisons.

given number of writers N, the RMP must be at most $2\beta/[N(N-1)]$. For example, suppose the desired 95% upper confidence bound on the RMP is 1 in one million. Then, using the formula $4.65/[(N+1)(N+2)]$, the smallest number of writers we could use to achieve this bound is 2,154. However, one then needs to determine how large a writing sample is needed to have at least a $1-\beta$ chance of achieving no matches in the 2,318,781 pairwise comparisons. In particular, for a 90% chance of observing no matches, the RMP must be at most $2\beta/[N(N-1)] = 2(0.1)/[(2154)(2153)]$ $= 4.31\times10^{-8}$. Using the parameter estimates in Table 1, such a RMP is associated with a writing sample of size 590.

6. Conclusion

The National Research Council (2009, p. 122) states:

> *The assessment of the accuracy of the conclusions from forensic analyses and the estimation of relevant error rates are key components of the mission of forensic science.*

This suggests that information concerning the RMP and the RNMP associated with a comparison procedure contributes to its practical utility in forensic science. In forensic DNA analysis, population genetics allow modeling the RMP and the RNMP as a function of population size and number of loci compared. Currently, in handwriting analysis, no comparable modeling exists.

In this paper, we have illustrated one alternative to modeling for investigation of the RMP and the RNMP associated with a comparison procedure applied to comparing writing samples. The proposed approach involves investigating the RMP and the RNMP using simulated writing samples. Specifically, we have presented algorithms for subsampling from available writing samples in a data set that can be used to consistently estimate the RMP and the RNMP as a function of the sizes of the writing samples being compared. The consistency of the subsampling estimators is dependent only on the number of writers, not the size of the writing samples. We have also described an algorithm involving subsampling

that can be used to estimate the standard error associated with an estimator of the RMP based on all pairwise comparison of samples within a data set.

All of these algorithms have been stated in terms of a common size of writing samples being compared. However, they can be trivially adapted to scenarios where the sizes of writing samples being compared are not the same for all writing samples. Such an application might arise when studying match probabilities associated with comparing very short notes, such as might be associated with bank robberies, to very large writing samples collected from potential suspects. The algorithms can also be adapted to investigate the dependency of match probabilities on criteria other than sizes of writing samples being compared. For example, the effect of content on match probabilities can be studied by changing from random sampling to stratified or systematic sampling when selecting characters to generate the simulated writing samples.

Although the main objective of this paper was to introduce this subsampling methodology, we have also shown some applications using the results of applying subsampling-based algorithms to a set of actual writing samples. For example, we have shown how the information about the RMP and how it varies with sizes of writing samples can be used when planning an empirical study of handwriting individuality within a relevant population. However, the actual values of the resulting estimates and recommendations concerning an empirical study of handwriting individuality presented in this paper must be viewed with caution. The set of writing samples used for illustrative purposes is small, including samples from only 100 individuals. This limits the ability to accurately estimate very small match probabilities. Also, the set of writing samples is a convenience sample and thus is not necessarily representative of a specific population.

Finally, although the main focus of this paper has been on match probabilities, the algorithms presented in this paper potentially have other applications in forensics. Match probabilities are utilized in studies of individuality and in validating the use of specific forensic techniques for individualization; they may not be the relevant measures for use in court (Stoney, 1984). Recently, focus has been on using the

likelihood ratio as one way to extend to other areas of forensics, such as handwriting, the DNA practice of reporting profile frequencies. The match probabilities are related to the probability that needs to be estimated for the denominator of the likelihood ratio. We are currently in the process of exploring the use of subsampling techniques detailed in this paper to estimate a likelihood ratio for handwriting.

Funding

This is publication number 10-01 of the Laboratory Division of the FBI. This work was supported in part under a Contract Award from the Counterterrorism and Forensic Science Research Unit of the FBI Laboratory. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

This work was also supported by an Intelligence Community (IC) Postdoctoral Research Fellowship [NGIA HM1582-06-1-2016 to C. S.]. Writing of this article was supported in part by Award No. 2009-DN-BX-K234 [to L. D. and C. S.] awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

Acknowledgements

The authors would like to acknowledge the help of: Donald T. Gantz regarding clarification of the original manuscript; John J. Miller for his suggestions on simulation strategies and advice on summarizing results; the computational support of Gannon Technologies Group (GTG); the FBI Laboratory for supplying the set of handwritten documents; and the scientists from the FBI Laboratory for their reviews and comments.

References

AGRESTI, A. AND COULL, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician* **52**(2): 119–126.

AITKEN, C. G. G. AND TARONI, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*, *2nd Edition*. Chichester, England, John Wiley & Sons.

ARVESON, J. N. (1969). Jackknifing *U*-statistics. *The Annals of Mathematical Statistics* **40**(6): 2076–2100.

BALDING, D. J. (2005). *Weight-of-Evidence for Forensic DNA Profiles.* Hoboken, NJ, John Wiley & Sons.

BOLLE, R. M., CONNELL, J. H., PANKANTI, S., RATHA, N. K. AND SENIOR, A. W. (2004). *Guide to Biometrics*. New York, Springer.

BULACU, M. AND SCHOMAKER, L. (2007). Text-independent writer identification and verification using textural and allographic features. *IEEE Transitions in Pattern Analysis and Machine Intelligence* **29**: 701–717.

HUBER, R. A. AND HEADRICK, A. M. (1999). *Handwriting Identification: Facts and Fundamentals*. Boca Raton, FL, CRC Press.

NATIONAL RESEARCH COUNCIL (2009). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, DC, National Academies Press.

NATIONAL RESEARCH COUNCIL (1996). *The Evaluation of Forensic DNA Evidence*. Washington, DC, National Academies Press.

OSBORN, A. S. (1929). *Questioned Documents, 2nd Edition*. Albany, NY, Boyd Printing Company.

RISINGER, D. M. AND SAKS, M. J. (1996). Science and nonscience in the courts: Daubert meets handwriting identification expertise. *Iowa Law Review* **82**(1): 21–74.

SAKS, M. J. AND KOEHLER, J. J. (2008). The individualization fallacy in forensic science evidence. *Vanderbilt Law Review* **61**(1): 199–219.

SAUNDERS, C. P., DAVIS, L. J. AND BUSCAGLIA, J. (2011a). Using automated comparisons to quantify handwriting individuality. *Journal of Forensic Sciences* **56**(3): 683–689.

SAUNDERS, C. P., DAVIS, L. J., LAMAS, A. C., MILLER, J. J. AND GANTZ, D. T. (2011b). Construction and evaluation of classifiers for forensic document analysis. *Annals of Applied Statistics* **5**(1): 381–399.

SCHUCANY, W. R. AND BANKSON, D. M. (1989). Small sample variance estimators for *U*-statistics. *Australian Journal of Statistics* **31**(3): 417–426.

SEN, P. K. (1960). On some convergence properties of *U*-statistics. *Calcutta Statistical Association Bulletin* **10**: 1–18.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York, Wiley.

STONEY, D. A. (1984). Evaluation of Associative Evidence: Choosing the Relevant Question. *Journal of the Forensic Science Society* **24**(5): 473-482.

WALCH, M. A. AND GANTZ, D. T. (2004). Pictographic matching: a graph-based approach towards a language independent document exploitation platform. *Proceedings of the 1st ACM Workshop on Hardcopy Document Processing*, K. Lubbes and M. Ronthaler. New York, Association of Computing Machinery: 53–62.

WAYMAN, J. L. (2000). Confidence interval and test size estimation for biometric data. *National Biometric Center Collected Works 1997-2000*. J. L. Wayman. San Jose, CA, National Biometric Test Center: 89–99.

Appendix A: Estimating the RMP and the RNMP

Consider an independent and identically distributed (*iid*) collection of writing samples $\{D_i : i = 1,2,...,N\}$. For $i \neq j$, let $s(D_i, D_j)$ denote the similarity score (associated with the comparison procedure) that compares the writing samples D_i and D_j of the i^{th} and j^{th} writers in the collection. Let τ be the threshold used to declare matching writing samples (via the comparison procedure).

One natural estimator of the RMP, which will be denoted as θ, for a given comparison procedure is the proportion of pairs of writing samples that match:

$$\hat{\theta} = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} m_{ij} \tag{A.1}$$

where m_{ij} equals one if D_i and D_j match, and equals zero if they do not. Using $I\{A\}$ to denote the indicator function that equals one if the event A is true and zero otherwise, $m_{ij} \equiv I\left\{s(D_i, D_j) > \tau\right\}$ for $i \neq j$. This estimator of the RMP is unbiased because in this notation, $\theta = P\left(s(D_i, D_j) > \tau\right)$ so that $\theta = E(m_{ij})$.

This estimator of the RMP is a member of the class of *U*-statistics of degree 2 (Serfling, 1980). So, under the assumption that the collection of writing samples $\{D_i : i = 1,2,...,N\}$ are *iid*, $\hat{\theta}$ has a variance of the form:

$$\mathrm{Var}(\hat{\theta}) = \frac{4(N-2)}{N(N-1)}\sigma_c^2 + \frac{2}{N(N-1)}\theta(1-\theta) \tag{A.2}$$

where

$$\sigma_c^2 \equiv \mathrm{Var}[E(m_{ij} \mid D_i)] \text{ for any } j \neq i. \tag{A.3}$$

Note that σ_c does not depend on *i* or *j* because $\{D_i : i = 1,2,...,N\}$ are assumed to be *iid*. Also, the "bar" in $E(m_{ij} \mid D_i)$ denotes conditional expectation. So, $E(m_{ij} \mid D_i)$ can be viewed as a conditional match

probability — namely, the probability that a randomly selected writing sample matches a specific (evidentiary) writing sample D_i. Note that the first term in (A.2) involving σ_c dominates $\text{Var}(\hat{\theta})$, at least for large values of N.

Based on the asymptotic distribution of a *U*-statistic, an approximate $100(1-\alpha)\%$ Wald-type upper confidence bound on the RMP is:

$$\hat{\theta} + z_\alpha \sqrt{\text{Var}(\hat{\theta})} \approx \hat{\theta} + 2 z_\alpha \sigma_c \Big/ \sqrt{N} \text{ for large } N \tag{A.4}$$

where $\text{Var}(\hat{\theta})$ (or σ_c) can be replaced by a consistent estimator, such as the one due to Bickel that is presented in Wayman (2000), and z_α is the $1-\alpha$ quantile of the standard normal distribution. Note that this upper bound depends on the sizes of the writing samples through its dependency on σ_c and also on the number of writers N.

Suppose now that instead of a single writing sample, the collection contains two writing samples from each writer (represented in the collection). In other words, consider an *iid* collection of pairs of writing samples $\{(D_{i1}, D_{i2}) : i = 1, 2, ..., N\}$. Let $s(D_{i1}, D_{i2})$ denote the similarity score (associated with the comparison procedure) that compares the two writing samples D_{i1} and D_{i2} from the i^{th} writer in the collection. Let τ be the threshold used to declare matching writing samples (via the comparison procedure).

One natural estimator of the RNMP, which will be denoted as γ, for a given comparison procedure is the proportion of pairs of writing samples from the same writer that do not match:

$$\hat{\gamma} = N^{-1} \sum_{i=1}^{N} I\left\{ s\left(D_{i1}, D_{i2}\right) \le \tau \right\}. \tag{A.5}$$

Appendix B: Subsampling vs. Resampling

In this appendix, we show that simulated writing samples generated by sampling without replacement (i.e., via subsampling) have the same distributional properties as the original writing samples, whereas those generated by sampling with replacement (i.e., via resampling) do not.

Suppose the original writing sample with V characters is represented as $\{C_1, C_2, ..., C_V\}$ where C_i denotes the features of the *i*th character in the writing sample. (In our case, the features consist of the letter being written and the isocode representing its shape.) Assume that this vector is an *iid* sample from a multinomial distribution with r categories and associated probability vector $\mathbf{p} = (p_1, p_2, ..., p_r)$, which we represent as: $\{C_1, C_2, ..., C_V\} \overset{iid}{\sim} \text{Mult}(1, \mathbf{p})$. Let $Y_j = \#\{C_i \text{ in } j\text{th category}\}$, $j = 1, 2, ..., r$. Then, the random vector of counts $(Y_1, Y_2, ..., Y_r)$ has a multinomial distribution with parameters V and $\mathbf{p}$, which we represent as: $(Y_1, Y_2, ..., Y_r) \sim \text{Mult}(V, \mathbf{p})$.

First, suppose the simulated writing sample $\{C_1^*, C_2^*, ..., C_n^*\}$ is generated by sampling $n \le V$ characters at random without replacement from the original writing sample $\{C_1, C_2, ..., C_V\}$. Since $\{C_1^*, C_2^*, ..., C_n^*\} \subseteq \{C_1, C_2, ..., C_V\}$, $\{C_1^*, C_2^*, ..., C_n^*\} \overset{iid}{\sim} \text{Mult}(1, \mathbf{p})$. So, if $Y_j^* = \#\{C_i^* \text{ in } j\text{th category}\}$, $j = 1, 2, ..., r$, then $(Y_1^*, Y_2^*, ..., Y_r^*) \sim \text{Mult}(n, \mathbf{p})$. Thus, simulated writing samples generated by sampling without replacement have the same distributional properties as the original writing sample, i.e., both are multinomial with the same probability vector $\mathbf{p}$.

Next, suppose the simulated writing sample $\{C_1^*, C_2^*, ..., C_n^*\}$ is generated by sampling $n \le V$ characters at random with replacement from the original writing sample $\{C_1, C_2, ..., C_V\}$. Let $\mathbf{Y}^* = (Y_1^*, Y_2^*, ..., Y_r^*)$, where $Y_j^* = \#\{C_i^* \text{ in } j\text{th category}\}$, $j = 1, 2, ..., r$. Now, sampling with replacement

corresponds to using the observed proportions in each category from the original sample as estimates of $\mathbf{p}$ and then sampling from the "fitted" model $\text{Mult}(1,\hat{\mathbf{p}})$. So, conditional on the observed writing sample, $\left\{C_1^*, C_2^*, ..., C_n^*\right\} \left| \{C_1, C_2, ..., C_V\} \right. \overset{iid}{\sim} \text{Mult}(1,\hat{\mathbf{p}})$ and $\mathbf{Y}^* \left| \{C_1, C_2, ..., C_V\} \right. \sim \text{Mult}(n,\hat{\mathbf{p}})$ where $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_r)$ and $\hat{p}_j = Y_j / V$, $j = 1,2,...,r$.

What is the unconditional distribution of $\mathbf{Y}^*$? For any $\mathbf{x} = (x_1, x_2, ..., x_r)$ with $x_j \in \{0,1,...,n\}$ and $\sum_{j=1}^{r} x_j = n$,

$$P(\mathbf{Y}^* = \mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{Y}^* = \mathbf{x} \mid \mathbf{Y} = \mathbf{y}) P(\mathbf{Y} = \mathbf{y})$$

where the sum is over all $\left\{\mathbf{y} = (y_1, y_2, ..., y_r) : y_j \in \{0,1,...,V\}, \sum_{j=1}^{r} y_j = V\right\}$. Substituting the multinomial probabilities,

$$\begin{aligned} P(\mathbf{Y}^* = \mathbf{x}) &= \sum_{\mathbf{y}} \left[\binom{n}{x_1, x_2, ..., x_r} \prod_{j=1}^{r} \left(\frac{y_j}{V} \right)^{x_j} \right] \left[\binom{V}{y_1, y_2, ..., y_r} \prod_{j=1}^{r} p_j^{y_j} \right] \\ &= V^{-n} \binom{n}{x_1, x_2, ..., x_r} E\left[\prod_{j=1}^{r} Y_j^{x_j} \right] \end{aligned}$$

But, $V^{-n} E\left[\prod_{j=1}^{r} Y_j^{x_j} \right] \neq \prod_{j=1}^{r} p_j^{x_j}$ for all $\mathbf{x}$. For example, for $\mathbf{x} = (n, 0, ..., 0)$,

$V^{-n} E\left[\prod_{j=1}^{r} Y_j^{x_j} \right] = V^{-n} E\left[Y_1^n \right] > p_1$ by Jensen's inequality (unless $p_1 = 1$). Thus, $\mathbf{Y}^*$ does not have a multinomial distribution with parameters n and $\mathbf{p}$. That is, a simulated sample generated by random sampling with replacement does not have the same distributional properties as the original writing sample.

Appendix C: Estimating the TMP and Standard Error

Both the variance in (A.2) of the point estimator of the RMP defined in (A.1) and the associated Wald-type upper confidence bound on the RMP defined in (A.4) are functions of the RMP as well as σ_c defined in (A.3).

Unlike the RMP, σ_c involves comparison of three writing samples instead of two. To understand why, recall the assumption that the collection of writing samples $\{D_i : i = 1, 2, ..., N\}$ are *iid*. Under this assumption, $E(m_{ij} \mid D_i)$[26] does not depend on j and $E[E(m_{ij} \mid D_i)] = E(m_{ij}) = \theta$ for any $j \neq i$.

So, for any $j \neq k \neq i$,

$$\begin{aligned} \sigma_c^2 &= \text{Var}[E(m_{ij} \mid D_i)] \\ &= E[E(m_{ij} \mid D_i)E(m_{ik} \mid D_i)] - \{E[E(m_{ij} \mid D_i)]\}^2 \\ &= E[E(m_{ij}m_{ik} \mid D_i)] - \theta^2 \\ &= E(m_{ij}m_{ik}) - \theta^2 \\ &= \text{Cov}(m_{ij}, m_{ik}) \end{aligned} \tag{C.1}$$

Since

$$m_{ij}m_{ik} = I\left\{s(D_i, D_j) > \tau\right\} I\left\{s(D_i, D_k) > \tau\right\} = I\left\{s(D_i, D_j) > \tau \text{ and } s(D_i, D_k) > \tau\right\},$$

the term $E(m_{ij}m_{ik})$ in (C.1) is just the probability of randomly selecting three individuals and then sampling one writing sample from each individual such that the writing sample from the first individual matches both the writing samples from the second and third individuals. As shown in (C.1), this probability, which we refer to as the tri-match probability (TMP), when combined with the RMP, determines σ_c.

As discussed at the end of Section 4.3, the output from Algorithm 4.3 can be used to estimate the TMP. Specifically, consider the output from Algorithm 4.3:

[26] Using $I\{A\}$ to denote the indicator function that equals one if the event A is true and zero otherwise, $m_{ij} \equiv I\left\{s(D_i, D_j) > \tau\right\}$ for $i \neq j$.

$\left\{\left(s(D_1^{*(k)}, D_2^{*(k)}), s(D_1^{*(k)}, D_3^{*(k)})\right) : k = 1, 2, ..., K\right\}$. For a fixed threshold τ, this data set can be converted into a set of pairs that flag whether each of the pairs of documents match. Defining $m_{12}^* = I\left\{s(D_1^*, D_2^*) > \tau\right\}$ and $m_{13}^* = I\left\{s(D_1^*, D_3^*) > \tau\right\}$, the output from Algorithm 4.3 can be viewed as: $\left\{(m_{12}^{*(k)}, m_{13}^{*(k)}) : k = 1, 2, ..., K\right\}$, which provides information about the dependence of the TMP and σ_c on the sizes of the writing samples.

For example, the proportion of triplets for which both match, i.e.,

$$\tilde{\psi}^*(n) \equiv K^{-1} \sum_{k=1}^{K} m_{12}^{*(k)} m_{13}^{*(k)} \tag{C.2}$$

is a consistent and unbiased estimator of the TMP as the number of writers increases for fixed sizes of writing samples. This is just the estimator described at the end of Section 4.3.

The generated data from Algorithm 4.3 can be used in several ways to estimate σ_c. For example, using the relationship in equation C.1 that $\sigma_c^2 = \text{Cov}(m_{ij}, m_{ik})$, the correlation coefficient computed on $\left\{\left(m_{12}^{*(k)}, m_{13}^{*(k)}\right) : k = 1, 2, ..., K\right\}$, i.e.,

$$\tilde{v}^*(n) \equiv (K-1)^{-1} \sum_{k=1}^{K} \left(m_{12}^{*(k)} - \bar{m}_{12}^*\right)\left(m_{13}^{*(k)} - \bar{m}_{13}^*\right) = \left(\frac{K}{K-1}\right)\left[\tilde{\psi}^*(n) - \bar{m}_{12}^* \bar{m}_{13}^*\right] \tag{C.3}$$

where $\bar{m}_{12}^*$ is the proportion of pairs of simulated samples that match when comparing the first and second simulated samples, i.e.,

$$\bar{m}_{12}^* = K^{-1} \sum_{k=1}^{K} m_{12}^{*(k)} \tag{C.4}$$

and $\bar{m}_{13}^*$ is the proportion of pairs of simulated samples that match when comparing the first and third simulated samples, i.e.,

$$\bar{m}_{13}^{*} = K^{-1}\sum_{k=1}^{K} m_{13}^{*(k)} \tag{C.5}$$

is a consistent estimator of σ_c as the number of writers increases for fixed sizes of writing samples.

Alternatively, one can estimate σ_c using the relationship in equation C.1 that $\sigma_c^2 = TMP - (RMP)^2$. If one has some other consistent estimator of the RMP, or some other information about the behavior of the RMP as a function of size of writing samples, say $\tilde{\theta}_A(n)$, then $\tilde{\psi}^*(n) - [\tilde{\theta}_A(n)]^2$ is also a consistent estimator of σ_c. One such consistent estimator of the RMP is available from the same output from Algorithm 4.3 that is used to estimate the TMP. Note that (C.4) and (C.5) are of the same form as the resulting estimator of the RMP from Algorithm 4.1. So, $\bar{m}_{12}^{*}$ and $\bar{m}_{13}^{*}$ both provide consistent estimators of the RMP; and one could combine the two generated sequences $\left\{m_{12}^{*(k)} : k = 1,2,...,K\right\}$ and $\left\{m_{13}^{*(k)} : k = 1,2,...,K\right\}$ to give a sequence of $2K$ values with which to estimate the RMP. However, for each k, $m_{12}^{*(k)}$ and $m_{13}^{*(k)}$ are correlated and thus the combined sequence is not equivalent to a sequence generated by $2K$ applications of Algorithm 4.1.

Which estimator of σ_c and subsequently the standard error of the point estimator of the RMP defined in (A.1) is the best depends upon the ultimate use of the estimator. For example, we are currently investigating estimators for use in constructing of upper confidence bounds, such as in equation A.4.

TABLE 1 Logistic regression coefficients from modeling the Random Match Probability (RMP) as a function of (common) size of writing samples (n)

Term	Coefficient	Standard Error
(Intercept)	5.52	0.52
n	-6.32×10^{-2}	8.64×10^{-3}
n^2	2.02×10^{-4}	4.34×10^{-5}
n^3	-2.72×10^{-7}	6.65×10^{-8}

Goodness of Fit Statistics

Test	*Value*	*Degrees of Freedom*	*P-Value*
Deviance Chi-Squared	33.03	37	0.66
Pearson Chi-Squared	35.42	37	0.54
Hosmer and Lemeshow	7.37	8	0.50

TABLE 2 Logistic regression coefficients from modeling the Tri-Match Probability (TMP) as a function of (common) size of writing samples (n)

Term	Coefficient	Standard Error
(Intercept)	5.35	0.64
n	-7.14×10^{-2}	1.28×10^{-2}
n^2	2.27×10^{-4}	7.72×10^{-5}
n^3	-3.32×10^{-7}	1.41×10^{-7}

Goodness of Fit Statistics

Test	*Value*	*Degrees of Freedom*	*P-Value*
Deviance Chi-Squared	25.78	37	0.92
Pearson Chi-Squared	23.93	37	0.95
Hosmer and Lemeshow	0.80	8	0.999

FIG. 1. Text of the modified "London Letter" and an example of a typical cursive writing sample.

Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott St. and then goes to Turin and Rome and will join Col. Parry and arrive at Athens, Greece, Nov. 27th or Dec. 2nd. Letters there should be addressed 3580 King James Blvd. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the "Y.X. Express" tonight. My daughter chastised me because I didn't choose a reception hall within walking distance from the church. I quelled my daughter's concerns and explained to her that it

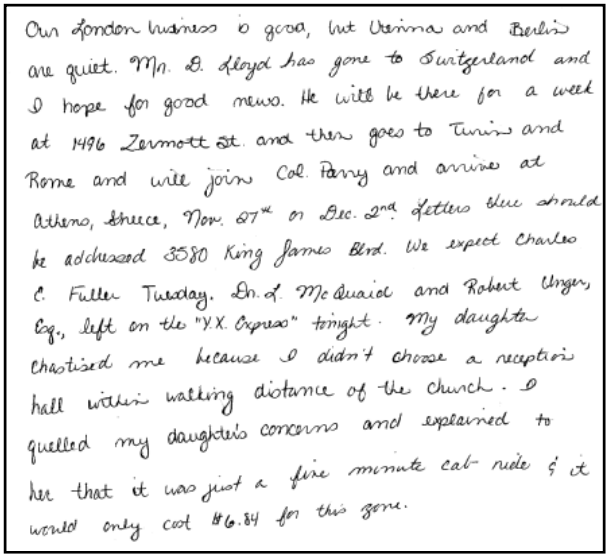


FIG. 2. Within-writer similarity scores. Empirical cumulative distribution functions (ECDFs) of within-writer similarity scores from Algorithm 4.2 (K = 1,000 for each common size n of writing samples). Diagonal (solid) line represents the theoretical cumulative distribution function (CDF) corresponding to a uniform distribution. Dotted horizontal lines are at an ECDF of 0.0 and 1.0.

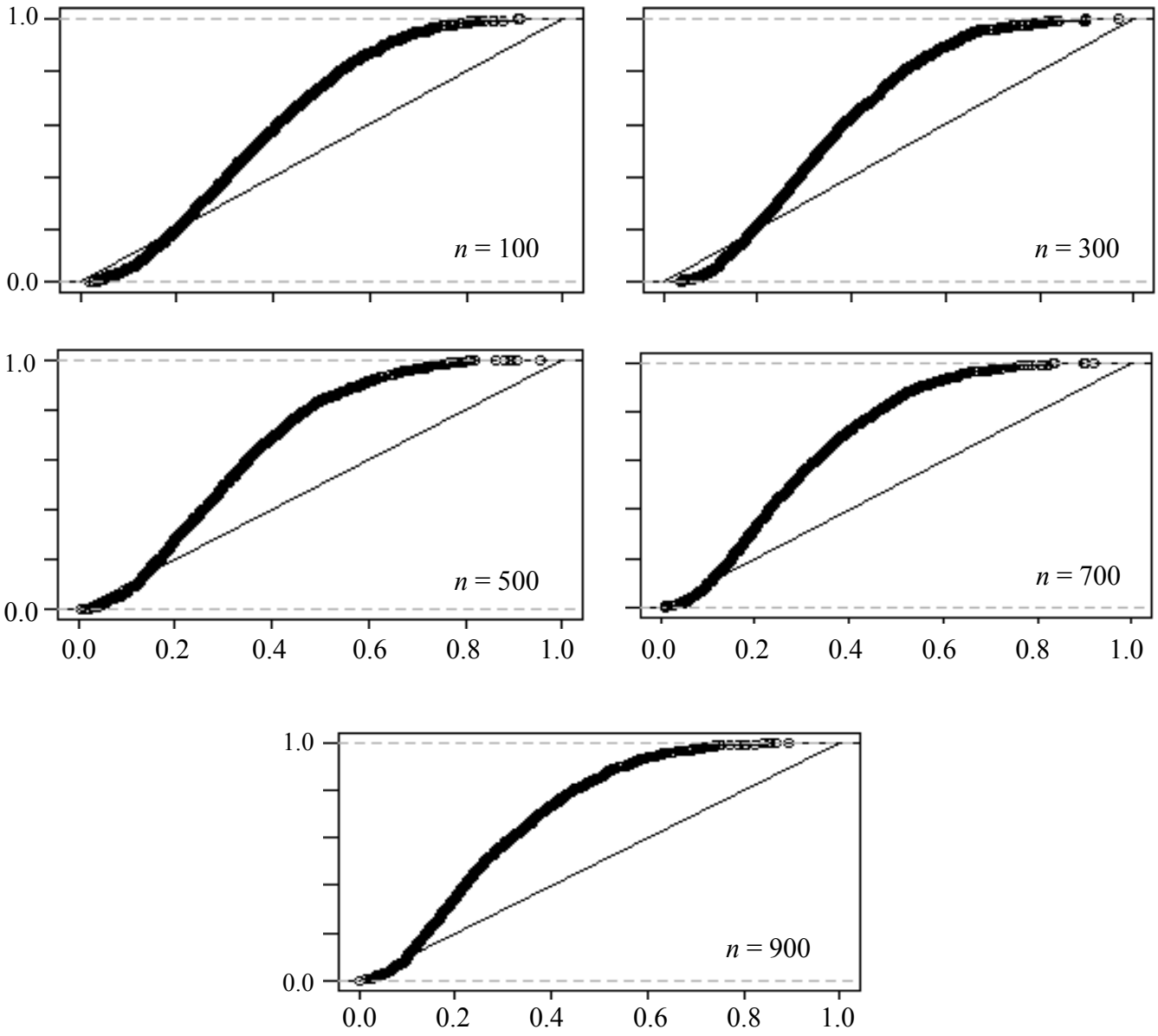


FIG. 3. Estimated random match probability (RMP) as a function of common size of writing samples. The plotted points are estimates of the RMP using Algorithm 4.1 ($K = 100$ for each common size of writing samples). The solid line is based off of a logistic fit (Table 1) modeling the RMP as a function of common size of writing samples being compared.

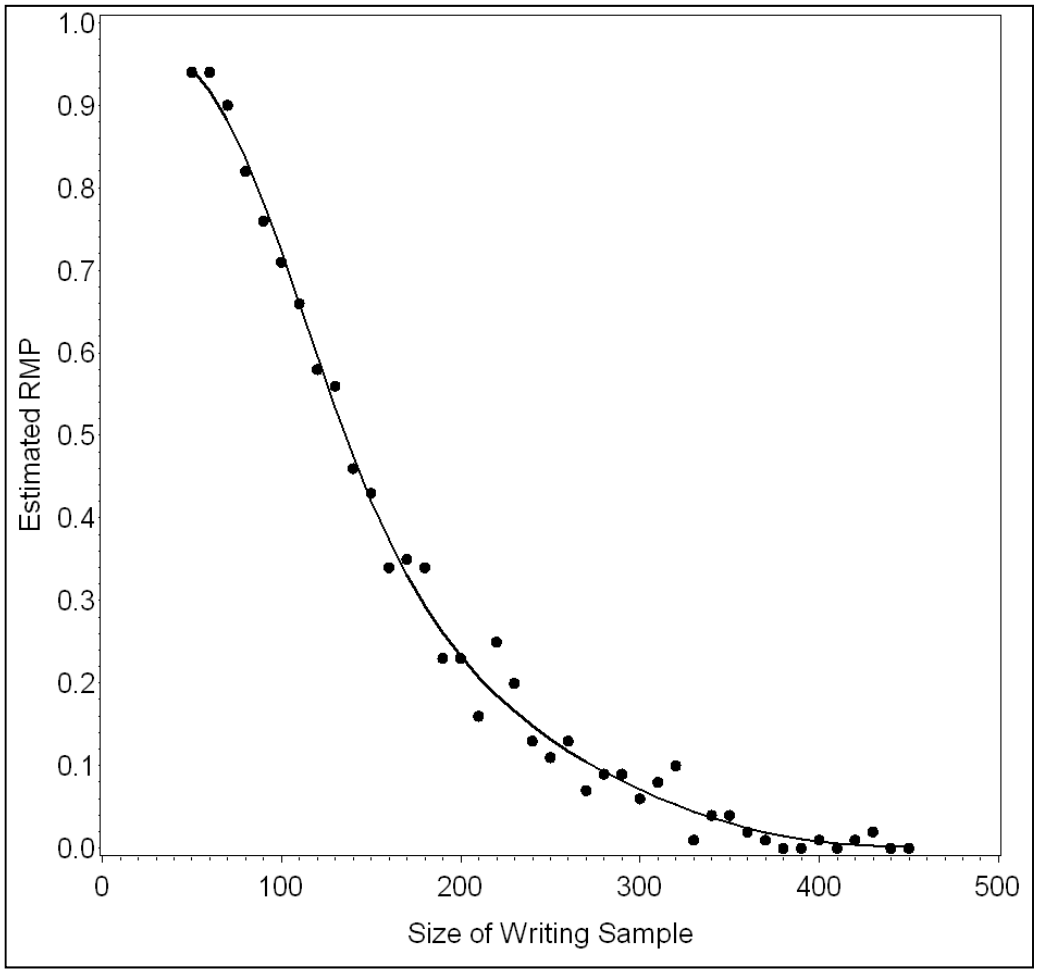


FIG. 4. Estimated tri-match probability (TMP) as a function of common size of writing samples. The plotted points are estimates of the TMP using Algorithm 4.3 ($K = 100$ for each common size of writing samples). The solid line is based off of a logistic fit (Table 2) modeling the TMP as a function of common size of writing samples being compared. The dashed line is based off of the logistic fit (Table 1) modeling the RMP as a function of size of writing samples. This logistic model is fit to the proportion of pairs of simulated samples that match when comparing the first and second simulated samples in the output from Algorithm 4.3.

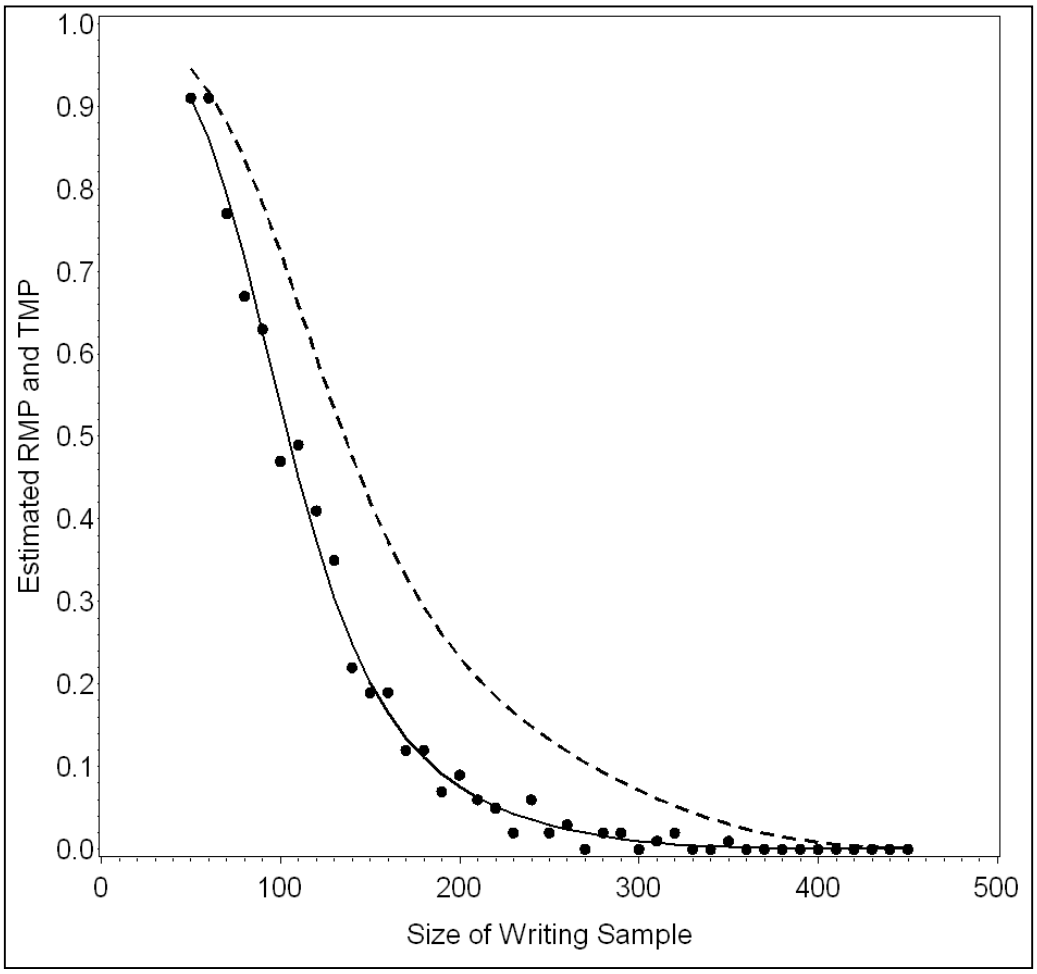
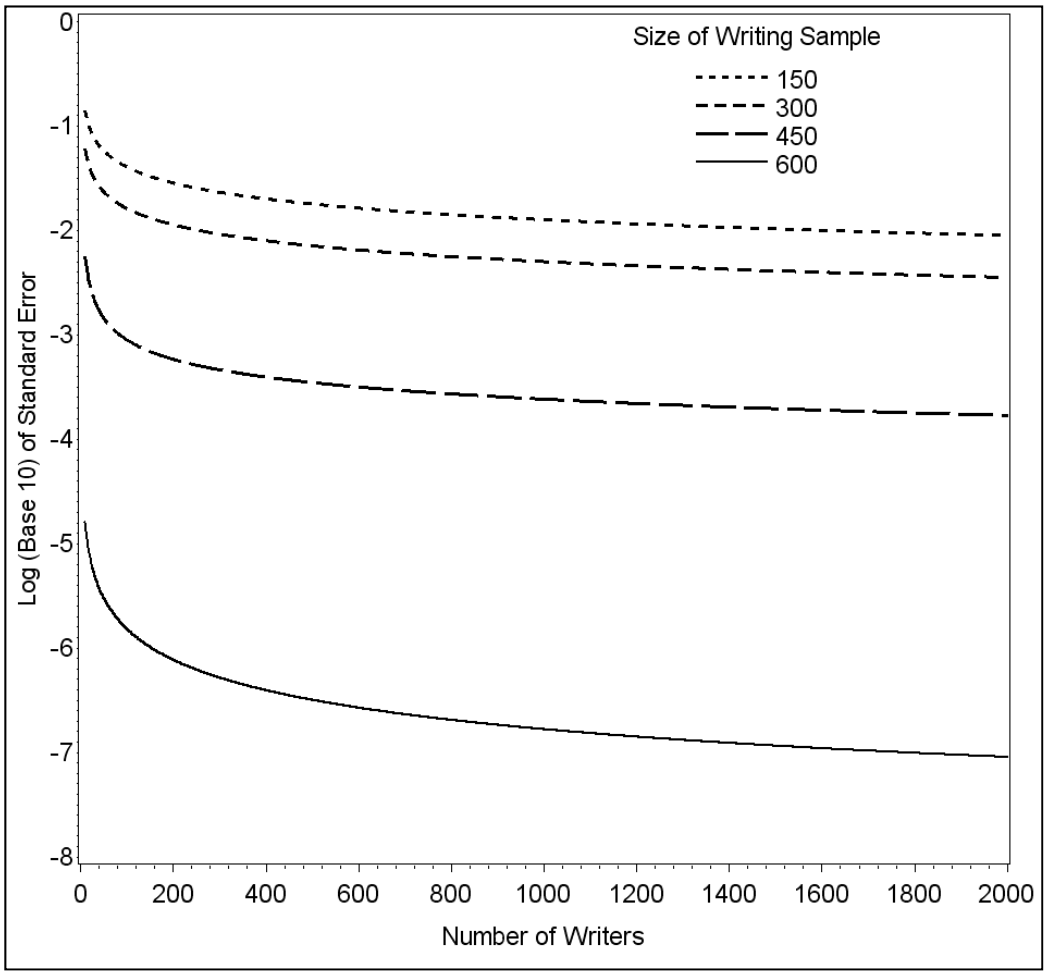


FIG. 5. Estimated standard error of the RMP estimator defined in (A.1) as a function of common size of writing samples and number of writers. The estimated standard errors are plotted on a log (base 10) scale and are based on three fitted logistic models. Two of these models are shown in Figure 4 with coefficients shown in Tables 1 and 2. The third logistic model is fit to the proportion of pairs of simulated samples that match when comparing the first and third simulated samples in the output from Algorithm 4.3.



Phase I Part b

For this grant, we first completed a survey of the statistical theory for *U*-Statistics with zero-one kernels; the focus of the survey was specifically related to the behavior of *U*-Statistics when used as estimators of small probabilities. The most recent research on RMPs that we have found is by Michael E. Schuckers and is summarized in his 2010 book *Computational Methods in Biometric Authentication: Statistical Methods for Performance Evaluation.* Schuckers (and almost all other researchers in this area) have not put the estimation of random match probabilities into the context of *U*-Statistics. However, Schuckers has identified the dependency structure that arises when performing all pairwise comparisons and has incorporated this dependency structure into his confidence intervals, mainly via bootstrap methods. The non-bootstrap methods are analogous to those used by Bickel and Wayman which we previously reviewed in our grant proposal.

Besides Dr. Saunders, three researchers (Drs. Davis, Gantz and Miller) on this grant worked on the estimation of confidence bounds for the RMP.

Dr. Linda Davis worked to find exact formulas for the mean and variance of the estimate of the RMP. Her development utilized structures and results from the theory of *U*-Statistics. The resulting exact formulas are only computationally tractable for very small sample sizes. Dr. Davis' work pointed out that assuming that all pairwise comparisons of samples are independent will lead to an underestimate of the variance of the estimate of the RMP. She also showed that basing the statistics on only a set of independent pairwise comparisons will lead to an overestimate of the variance of the estimate of the RMP. Dr. Davis introduced a scenario for which tighter bounds are possible for the variance of the estimate of the RMP.

Two documents prepared by Dr. Davis are attached to the Final Report in Appendices 1 and 2:

> "Link Between *U*-Statistics With 0-1 Kernels and the Union/Intersection of Events" This document presents the exact formulas for the mean and variance of the estimate of the RMP in the general case and in a special case.
>
> "RMP Confidence Interval"

These documents present issues associated with finding confidence interval bounds for estimation of the RMP and presents an approach to calculating bounds in a special case. A list of references concerning relevant statistical estimation is also included in these documents.

Dr. Davis intends to submit papers based on these two documents to research journals.

Drs. John Miller, Donald Gantz, and Christopher Saunders have developed a general parametric model for studying the distribution of pairwise comparisons of an arbitrary type tailored for small sample sizes with possibly no observed matches. The advantage of having a parametric model is that it provides an added level of structure for estimating the RMP with limited information. Furthermore, as long as the parametric model is chosen carefully, the resulting

estimates appear to have a high degree of accuracy. This model is designed to incorporate the dependencies that arise in such studies with pairwise comparisons.

We introduced the parametric model to pursue our research goal of extending our *U*-Statistics based methods for estimating the RMP to the situation of small sample sizes. We are building upon the early research of Blom (1976) to provide a parametric model that retains the optimal asymptotic properties of the *U*-Statistic estimate of the RMP (in the sense of being a Best Linear Unbiased Estimator of the RMP) but facilitates different estimation approaches, such as Maximum Likelihood Estimates, Restricted Maximum Likelihood Estimates (REML), and even Bayesian estimates.

The parametric model we are implementing treats the joint distribution of comparisons as a multivariate normal distribution. This approach is conceptually analogous to applying the standard Wilson Interval to estimating a proportion from a binomial random variable. This distributional assumption is only a tool used to facilitate the estimation of the RMP and is not expected to actually match the joint distribution of the discrete pairwise comparisons.
We have derived the theoretical foundation for the parametric model. We have demonstrated the use of this model in the construction of REML estimates and bounds for the RMP. We have run simulations that study the performance of the different estimates.

At the 2011 NIJ Trace Evidence Symposium, Drs. Gantz, Miller and Saunders presented their initial results on using a parametric method for estimating the RMP and constructing upper confidence bounds through non-asymptotic methods. They have recently completed a research paper on their work which has been submitted to the journal *Technometrics*. This paper studies in detail the parametric model for pairwise comparisons used in Forensic Science. It describes the eigenstructure of the covariance matrix and shows the consequences of the relations given by assuming normal distributions for the random components of the model. It shows that a closed form for an ANOVA table is possible. It shows that by using a method related to Fieller's Theorem, one can construct confidence intervals for a fixed component of the model which can then be easily turned into a confidence interval for the RMP. It also shows that two competing methods are either too conservative or just incorrect. The paper is included as the continuation of Phase I, Part B.

On Parametric Models for Pairwise Comparisons

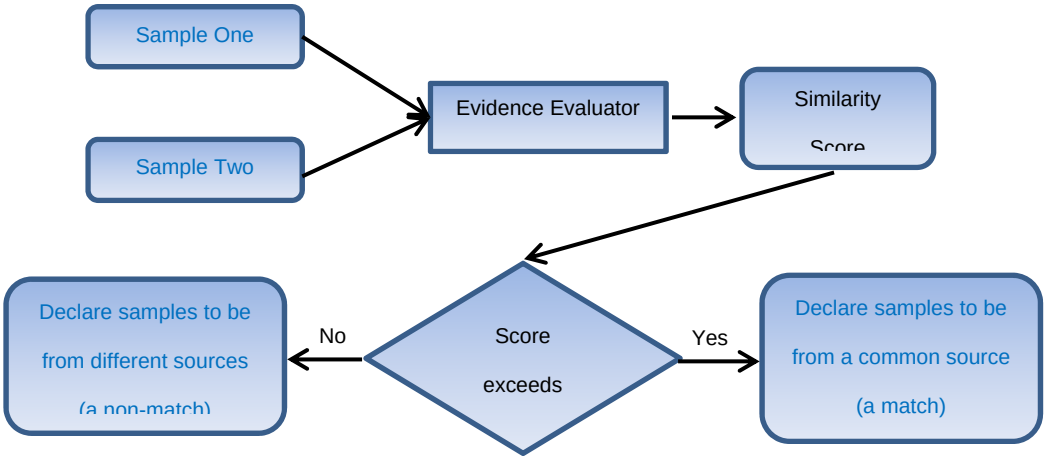
John Miller[a], Donald Gantz[b], Chris Saunders[c]

Abstract: *This paper studies a parametric model for pairwise comparisons used in Forensic Science. We describe the eigenstructure of the covariance matrix and show the consequences of the relations given by assuming normal distributions for the random components of the model. We show that a closed form for an ANOVA table is possible. We show that using a method related to Fieller's Theorem, we can construct confidence intervals for a fixed component of the model which can then be easily turned into a confidence interval for the random match probability. The random match probability is a concept which is critical for Forensic Scientists. We also show that two competing methods are either too conservative or just incorrect.*

1. Introduction:

Pairwise comparisons are a useful statistical tool in Forensic Science. We will introduce a parametric model to show critical issues associated with statistical inference based on pairwise comparisons. A common situation in forensic science is "Did two samples come from the same source?" We use a similarity score to compare two samples. Higher similarity scores are indicative of coming from a common source. We assume a population of objects which can be sampled. The random match probability, RMP, is the probability that two distinct objects selected at random from the population are erroneously judged to come from the same source.[27] We can assess the RMP by taking a random sample of n objects known to have each come from a different source and calculating the similarity score for each pair of objects. Figure 1 describes this situation.

Figure 1: Pairwise Comparison of Two Samples



[a] J. Miller, Department of Statistics, George Mason University; [b] D. Gantz, Department of Applied Information Technology, George Mason University; [c] C. Saunders, Department of Mathematics and Statistics, South Dakota State University

[27] The RMP can be thought of as the rate at which matches occur in the general population.

Write s_{ij} for the similarity score for objects i and j. Based on a sample of n objects, we get $N = n(n-1)/2$ similarity scores. A plausible correlation structure for the similarity scores is 1) If we have two similarity scores for the same pair of objects, then the scores are the same; that is the two scores have correlation one. 2) If we have two similarity scores with one of the compared samples in common, then these two similarity scores are related. The two scores have a correlation which is not zero. 3) If two similarity scores have no samples in common, then the scores are not related; that is, the correlation between the scores is zero.

The mean of the similarity scores from all possible pairs of different source objects in the population is a very important quantity in calculations of the random match probability based on scores. We will call this parameter θ. A plausible estimate for this quantity is the mean of the N scores from the pairwise comparisons in the sample. It will be important to have proper estimates of the variability of this mean in forming confidence limits for the random match probability.

Some researchers simply ignore the correlation structure and proceed as if there is a sample of N independent scores. This would yield an invalid confidence bound for the RMP. Other researchers believe that the correlation structure forces one to use only uncorrelated pairs (such as s_{12}, s_{34}, s_{56}, etc.). They use only $n/2$ similarity scores out of the possible $N = n(n-1)/2$ scores. This is too conservative. We will show that it is possible to account for the correlation structure in using data such as this.

In Section 2 we describe the parametric model; in Section 3 we describe the model in terms of its normal distribution; in Section 4 we define the random match probability (RMP); in Section 5 we describe the process of forming a confidence interval for the RMP; in Section 6 we give the results for a large simulation study of the methods described in Section 5; in Section 7 we describe two ways of forming confidence intervals based on methods which are either too conservative or just wrong; in Section 8 we state the conclusions of this paper.

2. A Parametric Model for Similarity Scores:

2.1 The Parametric Model:

One parametric model that preserves the dependency between scores is implied by a simple random sample for the original measurements. That parametric model for the similarity score s_{ij} is

$$s_{ij} = \theta + a_i + a_j + e_{ij},$$

where θ is an unknown population parameter, a_i is a random quantity dependent on object i and e_{ij} is a random error term. We assume the a_i to be i.i.d. $(0, \sigma_a^2)$ $i = 1, \ldots, n$, and the e_{ij} to be i.i.d. $N(0, \sigma_e^2)$, $i = 1, \ldots n-1; j = i+1, \ldots n$ and any a_i is independent of any e_{ij}. We will use this model to find a valid ANOVA table. We can rewrite this model in matrix terms by writing the scores (s_{ij}) in lexicographic order as a vector **y**. The errors (e_{ij}) are listed in the same order and the a's are listed in order of their subscripts. There is a design matrix **P** (for pairwise) which

has N rows and n columns. $\mathbf{P}$ is mostly composed of zeroes but has a one in the ith and jth columns for the row corresponding to s_{ij}.

Thus our model becomes

$$\mathbf{y} = \theta\mathbf{1}_N + \mathbf{Pa} + \mathbf{e},$$

where $\mathbf{y}$ and $\mathbf{e}$ are as described above, $\mathbf{a}$ is the vector of the a_i, and $\mathbf{1}_N$ is an N by 1 vector of ones.

The N by 1 expected value of the vector of scores is just $\theta\mathbf{1}_N$. The N by N covariance matrix of the vector of scores is $\mathbf{\Sigma} = \sigma_e^2\mathbf{I}_N + \sigma_a^2\mathbf{PP'}$. The form of $\mathbf{\Sigma}$ comes from the matrix formulation and the independence of the a_i and the e_{ij}, not from normality assumptions. Later the relative size of σ_a^2 to σ_e^2 will be important. We can convert the covariance matrix to a correlation matrix if we wish. The resulting correlation matrix contains mostly zeroes but has non-zero correlations of $r = \sigma_a^2/(\sigma_e^2 + 2\sigma_a^2)$ in some positions.

In Figure 2, consider an example with $n = 8$ and $N = 28$. Let $\boldsymbol{y}$ be the vector of similarity scores in lexicographic order (alphabetic order) of the subscripts i and j. $\mathbf{P}$ $(N \times n)$ is the design matrix for selecting pairs of objects. For example, the fourth line of $\mathbf{P}$ has ones for selecting objects 1 and 5. Each column has seven ones because each object is compared to the other seven objects.

Figure 2: Example with $n = 8$ and $N = 28$

$$\mathrm{y} = \begin{vmatrix} \mathtt{s12} \\ \mathtt{s13} \\ \mathtt{s14} \\ \mathtt{s15} \\ \mathtt{s16} \\ \mathtt{s17} \\ \mathtt{s18} \\ \mathtt{s23} \\ \mathtt{s24} \\ \mathtt{s25} \\ \mathtt{s26} \\ \mathtt{s27} \\ \mathtt{s28} \\ \mathtt{s34} \\ \mathtt{s35} \\ \mathtt{s36} \\ \mathtt{s37} \\ \mathtt{s38} \\ \mathtt{s45} \\ \mathtt{s46} \\ \mathtt{s47} \\ \mathtt{s48} \\ \mathtt{s56} \\ \mathtt{s57} \\ \mathtt{s58} \\ \mathtt{s67} \\ \mathtt{s68} \\ \mathtt{s78} \end{vmatrix} \qquad \mathrm{P} = \begin{vmatrix} 1&1&0&0&0&0&0&0 \\ 1&0&1&0&0&0&0&0 \\ 1&0&0&1&0&0&0&0 \\ 1&0&0&0&1&0&0&0 \\ 1&0&0&0&0&1&0&0 \\ 1&0&0&0&0&0&1&0 \\ 1&0&0&0&0&0&0&1 \\ 0&1&1&0&0&0&0&0 \\ 0&1&0&1&0&0&0&0 \\ 0&1&0&0&1&0&0&0 \\ 0&1&0&0&0&1&0&0 \\ 0&1&0&0&0&0&1&0 \\ 0&1&0&0&0&0&0&1 \\ 0&0&1&1&0&0&0&0 \\ 0&0&1&0&1&0&0&0 \\ 0&0&1&0&0&1&0&0 \\ 0&0&1&0&0&0&1&0 \\ 0&0&1&0&0&0&0&1 \\ 0&0&0&1&1&0&0&0 \\ 0&0&0&1&0&1&0&0 \\ 0&0&0&1&0&0&1&0 \\ 0&0&0&1&0&0&0&1 \\ 0&0&0&0&1&1&0&0 \\ 0&0&0&0&1&0&1&0 \\ 0&0&0&0&1&0&0&1 \\ 0&0&0&0&0&1&1&0 \\ 0&0&0&0&0&1&0&1 \\ 0&0&0&0&0&0&1&1 \end{vmatrix}$$

Figure 3 is the 28×28 correlation matrix for the vector $\mathbf{y}$ with $n = 8$. The correlations are: $r = 1$ when both similarity scores compare the same two objects; $r = \sigma_a^2/(\sigma_e^2 + 2\sigma_a^2)$ when each of the two similarity scores have one of the compared objects in common; and $r = 0$ when the two similarity scores have no compared objects in common. There are $2(n-2) = 12$ r's in each row. In this example there are about 43% of the matrix entries equal to r. For large n there are about $4/n$ of the N^2 matrix entries equal to r. For example, if $n = 100$ there are 196 r's out of 4950 elements in each row, or about 4% of the correlation entries are r's. We will show below that the correlation structure must always be accounted for to get valid statistical inference.

Figure 3: Correlation Matrix for $\mathbf{y}$ for $n = 8$ ($r = \sigma_a^2/(\sigma_e^2 + 2\sigma_a^2)$)

1	r	r	r	r	r	r	r	r	r	r	r	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	1	r	r	r	r	r	r	0	0	0	0	r	r	r	r	r	0	0	0	0	0	0	0	0	0	0	0
r	r	1	r	r	r	r	0	r	0	0	0	r	0	0	0	0	r	r	r	r	0	0	0	0	0	0	0
r	r	r	1	r	r	r	0	0	r	0	0	0	r	0	0	0	r	0	0	0	r	r	r	0	0	0	0
r	r	r	r	1	r	r	0	0	0	r	0	0	0	0	r	0	0	0	r	0	0	r	0	0	r	r	0
r	r	r	r	r	1	r	0	0	0	0	r	0	0	0	0	r	0	0	0	r	0	0	r	0	r	0	r
r	r	r	r	r	r	1	0	0	0	0	0	r	0	0	0	0	r	0	0	0	r	0	0	r	0	r	r
r	r	0	0	0	0	0	1	r	r	r	r	r	r	r	r	r	0	0	0	0	0	0	0	0	0	0	0
r	0	r	0	0	0	0	r	1	r	r	r	r	0	0	0	0	r	r	r	r	0	0	0	0	0	0	0
r	0	0	r	0	0	0	r	r	1	r	r	r	0	r	0	0	0	r	0	0	0	r	r	r	0	0	0
r	0	0	0	r	0	0	r	r	r	1	r	r	0	0	r	0	0	0	r	0	0	r	0	0	r	r	0
r	0	0	0	0	r	0	r	r	r	r	1	r	0	0	0	r	0	0	0	r	0	0	r	0	r	0	r
r	0	0	0	0	0	r	r	r	r	r	r	1	0	0	0	0	r	0	0	0	r	0	0	r	0	r	r
0	r	r	0	0	0	0	r	r	0	0	0	0	1	r	r	r	r	r	r	r	r	0	0	0	0	0	0
0	r	0	r	0	0	0	r	0	r	0	0	0	r	1	r	r	r	r	0	0	0	r	r	r	0	0	0
0	r	0	0	r	0	0	r	0	0	r	0	0	r	r	1	r	r	0	r	0	0	r	0	0	r	r	0
0	r	0	0	0	r	0	r	0	0	0	r	0	r	r	r	1	r	0	0	r	0	0	r	0	r	0	r
0	0	r	0	0	0	r	r	0	0	0	0	r	r	r	r	r	1	0	0	0	r	0	0	r	0	r	r
0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	0	1	r	r	r	r	r	r	0	0	0
0	0	r	0	r	0	0	0	r	0	r	0	0	r	0	r	0	0	r	1	r	r	r	0	0	r	r	0
0	0	r	0	0	r	0	0	r	0	0	r	0	r	0	0	r	0	r	r	1	r	0	r	0	r	0	r
0	0	r	0	0	0	r	0	r	0	0	0	r	r	0	0	0	r	r	r	r	1	0	0	r	0	r	r
0	0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	r	r	0	0	1	r	r	r	r	0
0	0	0	r	0	r	0	0	0	r	0	r	0	0	r	0	r	0	r	0	r	0	r	1	r	r	0	r
0	0	0	r	0	0	r	0	0	r	0	0	r	0	0	r	0	0	r	0	0	r	r	r	1	0	r	r
0	0	0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	r	r	0	r	r	0	1	r	r
0	0	0	0	r	0	r	0	0	0	r	0	r	0	0	r	0	r	0	r	0	r	0	r	r	0	1	r
0	0	0	0	0	r	r	0	0	0	0	r	r	0	0	0	r	r	0	0	r	r	0	r	r	r	r	1

2.2 Eigenstructure of Σ:

In order to get an ANOVA table, we need to know the eigenstructure of $\mathbf{\Sigma}$. To show what the eigenstructure of $\mathbf{\Sigma}$ is, we first begin with computing $\mathbf{P}'\mathbf{P}$, which turns out to be:

$$\mathbf{P}'\mathbf{P} = \begin{bmatrix} n-1 & 1 & \cdots & 1 \\ 1 & n-1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & n-1 \end{bmatrix} = (n-2)\mathbf{I}_n + \mathbf{E}_n = (n-2)\mathbf{I}_n + \mathbf{1}_n\mathbf{1}_n'$$

It is clear that $\mathbf{E}_n = \mathbf{1}_n\mathbf{1}_n'$ has one eigenvalue of n with eigenvector of $\mathbf{1}_n/(\mathbf{1}_n'\mathbf{1}_n)^{1/2}$ and $n-1$ eigenvalues of zero with all eigenvectors orthogonal to $\mathbf{1}_n$. This leads to the following roots of

$\mathbf{P}'\mathbf{P}$: one root of $2(n-1)$ and $n-1$ roots of $(n-2)$. Now we use the fact that the non-zero roots of $\mathbf{P}'\mathbf{P}$ are the same as the non-zero roots of $\mathbf{P}\mathbf{P}'$ to show the result below. (See Roy, 1954.) This leads to the following set of roots of $\mathbf{P}\mathbf{P}'$: one root of $2(n-1)$, $n-1$ roots of $(n-2)$, and $N-n$ roots of zero.

Therefore we can conclude that there are N eigenvectors of $\mathbf{\Sigma} = \sigma_e^2\mathbf{I}_N + \sigma_a^2\mathbf{P}\mathbf{P}'$:

- 1 eigenvector ($\mathbf{v}_1 = \mathbf{1}_N/\sqrt{N}$) with eigenvalue $\lambda_1 = \sigma_e^2 + 2(n-1)\sigma_a^2$
- $n-1$ eigenvectors ($\mathbf{v}_2$ to $\mathbf{v}_n$) with eigenvalue $\lambda_2 = \sigma_e^2 + (n-2)\sigma_a^2$
- $N-n$ eigenvectors ($\mathbf{v}_{n+1}$ to $\mathbf{v}_N$) with eigenvalue $\lambda_3 = \sigma_e^2$

Since $\sigma_e^2 > 0$, $\mathbf{\Sigma}$ has full rank. Because eigenvectors are orthogonal, we have $\mathbf{v}_k'\mathbf{1}_N = 0$ for all $k > 1$.

3. Using the Model Based on Normal Distribution:

3.1 Normality:

Under the assumption of normality for the distributions of a_i's and the e_{ij}'s, the log-likelihood can be written as

$$-2lnL = N\,ln(2\pi) + ln|\mathbf{\Sigma}| + (\mathbf{y} - \theta\mathbf{1}_N)'\mathbf{\Sigma}^{-1}(\mathbf{y} - \theta\mathbf{1}_N)$$

$$= N\,ln(2\pi) + ln\lambda_1 + (n-1)ln\lambda_2 + (N-n)ln\lambda_3$$

$$+ \frac{N(\bar{y}-\theta)^2}{\lambda_1} + \frac{\mathbf{y}'(\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k{}')\mathbf{y}}{\lambda_2} + \frac{\mathbf{y}'(\sum_{l=n+1}^{N}\mathbf{v}_l\mathbf{v}_l{}')\mathbf{y}}{\lambda_3}$$

$$= N\,ln(2\pi) + ln\lambda_1 + (n-1)ln\lambda_2 + (N-n)ln\lambda_3$$

$$+ \frac{N(\bar{y}-\theta)^2}{\lambda_1} + \frac{SS_a}{\lambda_2} + \frac{SS_e}{\lambda_3}$$

where $\bar{y} = \frac{1}{N}\mathbf{1}'\mathbf{y}$ and SS_a and SS_e are discussed further below.

Since the columns v_i are orthogonal, we obtain the following:

$$\mathbf{I} = \mathbf{v}_1\mathbf{v}_1' + \sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k' + \sum_{l=n+1}^{N}\mathbf{v}_l\mathbf{v}_l'.$$

The spectral decomposition of $\mathbf{P}\mathbf{P}'$ yields the following:

$$\mathbf{P}\mathbf{P}' = 2(n-1)\mathbf{v}_1\mathbf{v}_1' + (n-2)\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k' + 0\sum_{l=n+1}^{N}\mathbf{v}_l\mathbf{v}_l' = 2(n-1)\mathbf{v}_1\mathbf{v}_1' + (n-2)\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k'.$$

The spectral decomposition of $\mathbf{\Sigma}$ yields the following:

$$\mathbf{\Sigma} = (\sigma_e^2 + 2(n-1)\sigma_a^2)\mathbf{v}_1\mathbf{v}_1' + (\sigma_e^2 + (n-2)\sigma_a^2)\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k' + \sigma_e^2\sum_{l=n+1}^{N}\mathbf{v}_l\mathbf{v}_l'.$$

Using the spectral decomposition of $\mathbf{PP'}$, we obtain the very important result, the last statement of which can be verified by multiplying out.

$$\sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k' = (\mathbf{PP'} - 2(n-1)\mathbf{v}_1\mathbf{v}_1')/(n-2)$$

$$= \frac{1}{(n-2)}\left(\mathbf{PP'} - \frac{4}{n}\mathbf{1}_N\mathbf{1}_N'\right)$$

$$= \frac{(n-1)^2}{(n-2)}\left(\frac{1}{(n-1)}\mathbf{P} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_n'\right)\left(\frac{1}{(n-1)}\mathbf{P'} - \frac{1}{N}\mathbf{1}_n\mathbf{1}_N'\right).$$

Furthermore, it can be shown that

$\frac{1}{n-1}\mathbf{P'y} = \begin{bmatrix}\bar{y}^{(1)}\\ \bar{y}^{(2)}\\ \vdots\\ \bar{y}^{(n)}\end{bmatrix}$, where $\bar{y}^{(k)} = \frac{1}{n-1}\sum_{\substack{i=k\\ or\ j=k}} s_{ij}$, and $\frac{1}{N}\mathbf{1'y} = \bar{y}$. Hence, $\frac{1}{n-1}\mathbf{P'y} - \frac{1}{N}\mathbf{1}_n\mathbf{1}_N'\mathbf{y} = \begin{bmatrix}\bar{y}^{(1)} - \bar{y}\\ \bar{y}^{(2)} - \bar{y}\\ \vdots\\ \bar{y}^{(n)} - \bar{y}\end{bmatrix}$. Thus we see that $SS_a = \frac{(n-1)^2}{(n-2)}\sum_{k=1}^{n}(\bar{y}^{(k)} - \bar{y})^2$.

It is also true that

$$\sum_{l=n+1}^{N}\mathbf{v}_l\mathbf{v}_l' = \mathbf{I} - \mathbf{v}_1\mathbf{v}_1' - \sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k' = \mathbf{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N' - \sum_{k=2}^{n}\mathbf{v}_k\mathbf{v}_k'.$$

So if we define $SS_t = \mathbf{y'}(\mathbf{I} - \mathbf{v}_1\mathbf{v}_1')\mathbf{y}$, we then arrive at $SS_e = \mathbf{y'}\sum_{l=n+1}^{N}\mathbf{v}_l\mathbf{v}_l'\,\mathbf{y} = SS_t - SS_a$.

Because SS_a and SS_e are defined using the characteristic vectors of $\mathbf{\Sigma}$, by Cochran's Theorem, we have that SS_a and SS_e are independent, with degrees of freedom $n-1$ and $N-n$, respectively. (See Cochran, 1934.) Each of the sums of squares is independent of the sample mean. It can be shown that we also have that

$E(\bar{y}) = \theta$, $Var(\bar{y}) = \frac{\sigma_e^2}{N} + \frac{4\sigma_a^2}{n}$, $E(SS_a) = (n-1)(\sigma_e^2 + (n-2)\sigma_a^2)$, and $E(SS_e) = (N-n)\sigma_e^2$.

3.2 The ANOVA Table:

All of the above work leads to the following ANOVA table.

Source	df	Sum of Squares	Mean Square	Expected Value of Mean Square
A	$n-1$	SS_a	$MS_a = SS_a/(n-1)$	$\sigma_e^2 + (n-2)\sigma_a^2$
Error	$N-n$	SS_e	$MS_e = SS_e/(N-n)$	σ_e^2
Total	$N-1$	SS_t		

This immediately leads to unbiased estimators for all parameters in our model: First let $MS_a = SS_a/(n-1)$ and $MS_e = SS_e/(N-n)$. It then follows that

$$\hat{\theta} = \bar{y}$$

$$\hat{\sigma}_a{}^2 = \frac{MS_a - MS_e}{n-2}$$

$$\hat{\sigma}_e{}^2 = MS_e$$

These estimates are closely related to REML estimates[28]. (For future reference, let $SS_t = SS_a + SS_e$ and $MS_t = SS_t/(N-1)$.)

The variance of a randomly selected score (the similarity score for two randomly selected objects) is given by $\sigma_s^2 = \sigma_e^2 + 2\sigma_a^2$. The variance of the mean of all scores in the sample ($\bar{y}$) is given by $\sigma_{\bar{y}}^2 = \frac{\sigma_e^2}{N} + \frac{4\sigma_a^2}{n}$. We can obtain unbiased estimates for each of these quantities by plugging in the unbiased estimates of the variance components. These will be designated by "hats". Thus $\hat{\sigma}_s^2 = \hat{\sigma}_e^2 + 2\hat{\sigma}_a^2 = MS_e + 2\frac{MS_a - MS_e}{n-2} = \frac{1}{n-2}\big((n-4)MS_e + 2MS_a\big)$ and $\hat{\sigma}_{\bar{y}}^2 = \frac{\hat{\sigma}_e^2}{N} + \frac{4\hat{\sigma}_a^2}{n} = \frac{MS_e}{N} + \frac{4}{n}\frac{MS_a - MS_e}{n-2} = \frac{-2}{(n-1)(n-2)}MS_e + \frac{4}{n(n-2)}MS_a$.

4. Random Match Probabilities:

For a given cutoff τ, the random match probability is the probability that a randomly selected s_{ij} will exceed τ. That is $RMP = P\{s_{ij} > \tau\}$. For our model

$$P\{s_{ij} > \tau\} = 1 - \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) \equiv \pi$$

where Φ is the standard normal CDF[29].

[28] The difference between these estimates and the REML estimates is that in the case where using the above formulas yields a negative result for $\hat{\sigma}_a{}^2$, the new results are that $\hat{\sigma}_a{}^2 = 0$ and $\hat{\sigma}_e{}^2 = MS_t$. We can use these REML rules if we wish.

[29] The CDF (cumulative distribution function) evaluated at *x* is area under the probability density function from minus infinity up to *x*.

The following statements are equivalent for any values of B_1 and B_2 (either random or not random).

$$B_1 < \pi < B_2 \Leftrightarrow B_1 < 1 - \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) < B_2 \Leftrightarrow 1 - B_2 < \Phi\left(\frac{\tau - \theta}{\sigma_s}\right) < 1 - B_1$$

$$\Leftrightarrow L_1 \equiv \Phi^{-1}(1 - B_2) < \frac{\tau - \theta}{\sigma_s} < \Phi^{-1}(1 - B_1) \equiv L_2.$$

Thus we require a random quantities L_1 and L_2 such that

$$P\left\{L_1 < \frac{\tau - \theta}{\sigma_s} < L_2\right\} = 1 - \alpha.$$

It is straightforward to convert such an interval to an upper confidence bound for π, the RMP.

5. Computing a Confidence Interval for RMP Using a Method Based on Fieller's Theorem:

If we wish to compute a confidence interval for RMP, it turns out that creating a confidence interval for $\frac{\tau-\theta}{\sigma_s}$ is the way to go, as shown above. We can do this by using the estimates $\bar{y}$ for θ and $\hat{\sigma}_s$ for σ_s. However, there are several subtleties and approximations needed as we proceed.

The first is that we need to find the expected value of the square root of a chi-square random variable. If a random variable $W = \sigma^2 V$, where V is a chi-square random variable with ν degrees of freedom, then it turns out that $E\left[W^{\frac{1}{2}}\right] = f\sigma$, where $f = \left(\frac{2}{\nu}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$. It then follows that $Var\left[W^{\frac{1}{2}}\right] = \sigma^2(1 - f^2)$. (It is true that f is a number which is quite close to one, but never larger than one.)

The second is that for $\hat{\sigma}_s^2$ as given above, $\hat{\sigma}_s^2$ does not have an exact chi-square distribution, it being the weighted sum of two chi-squares. Fortunately, we can use an excellent approximation, given by Satterthwaite's approximation for degrees of freedom. We have that $\hat{\sigma}_s^2 = \frac{(n-4)}{(n-2)} MS_e + \frac{2}{(n-2)} MS_a$ which leads to the approximate degrees of freedom for $\hat{\sigma}_s^2$ of $\left(\frac{(n-4)}{(n-2)} MS_e + \frac{2}{(n-2)} MS_a\right)^2 / \left(\frac{\left(\frac{(n-4)}{(n-2)} MS_e\right)^2}{(N-n)} + \frac{\left(\frac{2}{(n-2)} MS_a\right)^2}{(n-1)}\right)$. This approximation is good for large values of n and hence of N.

The third is that we must have a normal distribution (or an approximation thereto) for each of a and b defined in the next paragraph. We have directly from our model that a has a normal distribution. We have that a chi-square random variable with ν degrees of freedom is the sum of ν independent chi-square random variables, each with one degree of freedom. Hence b^2 has an approximate normal distribution. It turns out by looking at normal plots from the simulations described below, that b has approximately a normal distribution for large values of n. Hence, we

at least approximately have the normal assumptions needed to use Fieller's Theorem described next.

The fourth is that we must be able to assemble the information from our model to get a confidence interval for $\frac{\tau-\theta}{\sigma_s}$. This we do by using the formulations of Fieller's Theorem. We begin by defining $a = \tau - \bar{y}$ and $b = {\hat{\sigma}_s}/{f}$. It then is true that $E[a] = \tau - \theta$ and $Var[a] = \frac{\sigma_e^2}{N} + \frac{4\sigma_a^2}{n}$. It is also true that $E[b] \approx \sigma_s$ and $Var[b] \approx \frac{\sigma_s^2}{f^2}(1-f^2)$. (The non-equal signs come from using Satterthwaite's approximation for the degrees of freedom. Again, for large values of n, the approximation works well, as we shall show below.) Finally, it is true that $Cov[a,b] = 0$ because $\bar{y}$ and $\hat{\sigma}_s^2$ have covariance zero from our model. If we now define $g = {t_{\alpha/2}^2 \hat{\sigma}_{\bar{s}}^2}/{b^2} = t_{\alpha/2}^2(1-f^2)$ and $h = {t_{\alpha/2}^2 \hat{\sigma}_{\bar{y}}^2}/{a^2}$. (These two quantities are reciprocals of assessments of whether the quantities $E[b]$ and $E[a]$ are different from zero. It will turn out to be critical that $E[b]$ is different from zero and to get a good confidence interval; it will turn out to be critical that $E[a]$ is different from zero as well. Note that requiring these numbers to be significant means that g and h must be small.)

If we define the fraction we want to form a confidence interval as $\frac{\tau-\theta}{\sigma_s}$, then we can note that $a - \frac{\tau-\theta}{\sigma_s}b$ has expected value approximately zero under our model. This allows us to form a point estimate of a/b and after some work, we derive a confidence interval for $\frac{\tau-\theta}{\sigma_s}$ of $\frac{a}{b}\left[\frac{1 \pm \sqrt{g+h-gh}}{1-g}\right]$.

6. A Study of the Method Based on Fieller's Theorem:

We ran a large simulation study based on the method illustrated in the previous section. We used four values of n: 50, 100, 500, and 1000. These values cover what could be attempted as a real study. We also used five values of $\rho = {\sigma_a}^2/{\sigma_e}^2$: 0.1, 0.5, 1.0, 2.0, and 10.0. These five values cover situations where there is very little correlation to where there is quite a bit. We feel that these values also cover what might be expected in real studies.

We ran 1,000,000 simulations in each cell and report estimates of the coverage probability, the average lower bound, and average upper bound. We realize that the lower limits are of little interest. It is getting an upper bound on the RMP that is important. However, it is worth reporting that one cannot rely on the upper intervals to have 97.5% coverage probability even though that is what we would like. It turns out for these intervals that although the coverage probabilities are very close to 95% (which was the nominal value we used), the upper limits do not get very close to 97.5% coverage probability. We are not sure why this occurs, but it surely does.

What we see is that in all the 20 cells for both an RMP of 0.001 (one in a thousand) and 0.000001 (one in a million) the coverage probabilities are very close to 0.95. The upper limits are sometimes quite far from the true values, but that is the price we pay for using a correct method.

Estimated Coverage Probability for a Method Based on Fieller's Theorem
Using Correct Method Based on New Model
(Nominal Coverage Probability 0.95; True RMP=0.001; 1,000,000 Simulations per Cell)
(Table Cell Entries: Coverage Probability, Average Lower Bound, Average Upper Bound)
$\boldsymbol{\rho = \sigma_a^2/\sigma_e^2}$

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9467 .0005 .0022	.9433 .0002 .0042	.9441 .0002 .0055	.9457 .0001 .0068	.9475 .0001 .0085
100	.9475 .0006 .0017	.9467 .0004 .0027	.9464 .0003 .0034	.9476 .0002 .0040	.9491 .0002 .0048
500	.9491 .0008 .0012	.9494 .0006 .0016	.9493 .0006 .0017	.9497 .0005 .0019	.9497 .0005 .0021
1000	.9494 .0009 .0012	.9498 .0007 .0014	.9494 .0007 .0015	.9495 .0006 .0016	.9500 .0006 .0017

Estimated Coverage Probability for a Method Based on Fieller's Theorem
Using Correct Method Based on New Model
(Nominal Coverage Probability 0.95; True RMP=0.000001; 1,000,000 Simulations per Cell)
(Table Cell Entries: Coverage Probability, Average Lower Bound, Average Upper Bound)
$\boldsymbol{\rho = \sigma_a{}^2/\sigma_e{}^2}$

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9475 .0000003 .0000050	.9419 .0000001 .0000232	.9425 .0000001 .0000464	.9443 .0000001 .0000762	.9469 .0000000 .0001274
100	.9480 .0000004 .0000027	.9458 .0000002 .0000085	.9456 .0000001 .0000148	.9470 .0000001 .0000224	.9487 .0000001 .0000350
500	.9489 .0000007 .0000015	.9492 .0000004 .0000025	.9490 .0000004 .0000032	.9496 .0000003 .0000039	.9496 .0000003 .0000049
1000	.9493 .0000008 .0000013	.9498 .0000006 .0000019	.9495 .0000005 .0000022	.9494 .0000004 .0000026	.9499 .0000004 .0000031

7. Two Bad Ways to Calculate Bounds for the RMP:

One way that has been suggested is to get around the correlation by only selecting comparisons which do not have any correlations by our model (such as s_{12}, s_{34}, s_{56}, etc.). Researchers using this method use only $n/2$ similarity scores out of the possible $N = n(n-1)/2$ scores. This is too conservative. We did a simulation study on this method and discovered that (as expected), the coverage probability had the correct value, the value of ρ did not affect the answers, but that the upper limits for the intervals were quite a bit larger than those for the correct method illustrated in the previous section. Please see the simulations given next.

Estimated Coverage Probability for a Method Based on Fieller's Theorem Using a Method Based on New Model for only $n/2$ out of $n(n+1)/2$ Observations (Nominal Coverage Probability 0.95; True RMP=0.001; 1,000,000 Simulations per Cell) (Table Cell Entries: Coverage Probability, Average Lower Bound, Average Upper Bound)

$$\rho = {\sigma_a}^2/{\sigma_e}^2$$

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9455 .0000 .0142	.9455 .0000 .0142	.9452 .0000 .0142	.9453 .0000 .0142	.9458 .0000 .0142
100	.9480 .0001 .0074	.9481 .0001 .0074	.9479 .0001 .0074	.9480 .0001 .0074	.9480 .0001 .0074
500	.9494 .0004 .0026	.9498 .0004 .0026	.9495 .0004 .0026	.9496 .0004 .0026	.9500 .0004 .0026
1000	.9496 .0005 .0020	.9492 .0005 .0020	.9499 .0005 .0020	.9499 .0005 .0020	.9498 .0005 .0020

Estimated Coverage Probability for a Method Based on Fieller's Theorem Using a Method Based on New Model for only $n/2$ out of $n(n+1)/2$ Observations (Nominal Coverage Probability 0.95; True RMP=0.000001; 1,000,000 Simulations per Cell) (Table Cell Entries: Coverage Probability, Average Lower Bound, Average Upper Bound)
$\rho = {\sigma_a}^2/{\sigma_e}^2$

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9446 .0000000 .0005140	.9447 .0000000 .0005123	.9444 .0000000 .0005124	.9445 .0000000 .0005137	.9450 .0000000 .0005111
100	.9478 .0000000 .0001169	.9478 .0000000 .0001168	.9477 .0000000 .0001168	.9475 .0000000 .0001166	.9477 .0000000 .0001166
500	.9494 .0000002 .0000096	.9498 .0000002 .0000096	.9494 .0000002 .0000096	.9497 .0000002 .0000096	.9498 .0000002 .0000096
1000	.9497 .0000003 .0000049	.9492 .0000003 .0000049	.9498 .0000003 .0000049	.9499 .0000003 .0000049	.9498 .0000003 .0000049

Some researchers simply ignore the correlation structure and proceed as if there is a sample of N independent scores. This would yield an invalid confidence bound for the RMP. The expected value of MS_t is given by $\sigma_e^2 + 2\frac{n-1}{n+1}\sigma_a^2$, which is almost the same as σ_s^2. The expected value of MS_t/N is given by $\frac{\sigma_e^2}{N} + \left(\frac{1}{n+1}\right)\frac{4\sigma_a^2}{n}$ which is not at all the same as $\sigma_{\bar{y}}^2$. If we ignored the correlation structure in this model, we could use the same formula from the previous work but substituting the values of MS_t for σ_s^2 and MS_t/N for ${\sigma_{\bar{y}}}^2$. This confidence interval has some very bad properties. Please see the simulation results which show that clearly, the coverage probability does not at all come close to 95%. Using the incorrect formulas definitely affects the coverage probability in an extremely bad way.

Estimated Coverage Probability for a Method Based on Fieller's Theorem
Using an Incorrect Method Based on New Model
(Nominal Coverage Probability 0.95; True RMP=0.001; 1,000,000 Simulations per Cell)
(Table Cell Entries: Coverage Probability)
$\boldsymbol{\rho = \sigma_a^2/\sigma_e^2}$

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9895	.8294	.7315	.6617	.5923
100	.9930	.8230	.7196	.6488	.5814
500	.9953	.8162	.7106	.6392	.5702
1000	.9955	.8158	.7095	.6371	.5689

Estimated Coverage Probability for a Method Based on Fieller's Theorem
Using an Incorrect Method Based on New Model
(Nominal Coverage Probability 0.95; True RMP=0.000001; 1,000,000 Simulations per Cell)
(Table Cell Entries: Coverage Probability)
$\boldsymbol{\rho = \sigma_a^2/\sigma_e^2}$

n	ρ =.1	ρ =.5	ρ =1	ρ =2	ρ =10
50	.9733	.7438	.6302	.5566	.4879
100	.9787	.7178	.5996	.5258	.4607
500	.9838	.6909	.5705	.4984	.4334
1000	.9841	.6875	.5664	.4938	.4298

8. Conclusions:

This paper has attempted to illustrate a model for pairwise comparisons which can be used to estimate the random match probability. The paper has shown the form of the model, $\mathbf{y} = \theta\mathbf{1}_N + \mathbf{Pa} + \mathbf{e}$ and what are the consequences of the normal distribution of the random components of the model. It has shown that there is a closed form for an ANOVA table. It has shown that there is a method for forming confidence intervals for $\frac{\tau-\theta}{\sigma_S}$ which works well based on the simulation

results. It has also shown that two other methods for making confidence intervals either fail by being too conservative or are just incorrect. We feel that the methods described here could be used by researchers working in the area of studying random match probabilities.

References:

A Useful Theorem in Matrix Theory, S. N. Roy, *Proceedings of the American Mathematical Society* , Vol. 5, No. 4 (Aug., 1954), pp. 635-638.

Cochran, William, "The distribution of quadratic forms in a normal system, with applications to the analysis of covariance", *Mathematical Proceedings of the Cambridge Philosophical Society* 30 (2): 178–191. (Apr 1934)

ROC Curves for Statistical Methods of Evaluating Evidence: Common Performance Measures Based On Similarity Scores

R. Bradley Patterson, a PhD Candidate supported by the grant, and Drs. Miller and Saunders authored a report that demonstrates the utility of ROC curves in forensics, where the goal is to measure the performance of methods that evaluate evidence. ROC curves offer several benefits to forensics. In contrast to the RMP, ROC curves capture the full range of error rates achievable with a method. They also depict the relative separation of the distributions of similarity scores from a given method. This then allows for comparisons of methods that produce scores on different scales. Additionally, an important characteristic for a method of evaluating pairs of evidence is the probability that a randomly selected pair from the same source would have a higher similarity score than a randomly selected pair from different sources, which the area under the curve (AUC) can estimate.

To show the value of ROC curves in forensics, Patterson applied them to measuring the performance of methods of evaluating trace evidence in the form of glass fragments. The methods, based on test statistics and likelihood ratios, came from an article by Aitken and Lucy (2004). Test statistics and likelihood ratios both provide measures of association between two samples. So those values are interpreted as similarity scores, with which ROC curves were created for the same data as the original article. The ROC curves provided measurements of the full performance of the methods across all thresholds as well as an even basis for comparison. All of the methods performed very well.

This report is included as Appendix 3 to this Final Report.

Phase II and Phase III

Goal: Investigate Properties of Approximate Methods for Evidence Interpretations such as Score Based Likelihood Ratios

The Utilization of Data Generated through Automated Systems

Throughout the grant, the researchers have utilized forensic data generated by automated systems. For many years, the research team has played a significant role in the development of automated systems for forensic handwriting and fingerprint identification. In particular, the team has developed the scoring algorithms that exploit quantification systems for both handwriting and fingerprints. (See Saunders' and Gantz's Vitas for a complete list of these research projects.) Both Drs. Gantz and Saunders were invited presenters at the Measurement Science and Standards in Forensic Handwriting Analysis (MSSFHA) Conference, June 4 – 5, 2013. The National Institute of Standards and Technology (NIST) hosted the MSSFHA Conference which was planned and organized in collaboration with the American Academy of Forensic Sciences – Questioned Document Section, American Board of Forensic Document Examiners, American Society of Questioned Document Examiners, Federal Bureau of Investigation Laboratory, National Institute of Justice (NIJ), and Scientific Working Group for Forensic Document Examination (SWGDOC).

Attendees, both in person and via a live webcast, included representatives from the collaborating institutions as well as universities, federal agencies, forensic laboratories, and the private sector. Dr. Gantz presented the Forensic Language-Independent Analysis System for Handwriting Identification (FLASH ID) in the Advances in Measurement Science in Handwriting Session. He stressed the accuracy of the automated system which finds identifying power from measured characteristics not directly observed or addressed by examiners. Dr. Saunders spoke on Understanding Individuality of Handwriting Using Score-Based Likelihood Ratios in the Advances in Statistics for Handwriting Analysis Session-*this presentation summarized research directly funded by this research grant that is published in two papers in Forensic Science International*. His examples concerning Score-Based Likelihood Ratios are based on joint work of Drs. Davis, Saunders, Hepler, and Buscaglia and were generated using data from FLASH ID. In this presentation Dr. Saunders summarized research results (from this grant) which demonstrated that common approaches to approximating the value of forensic evidence can lead to radically different values of evidence. These results are summarized in the previously mentioned papers in Forensic Science International.

To conclude the Conference, moderators led a facilitated discussion on the future state of forensic handwriting analysis, specifically focusing on the following questions: What does the future state of handwriting analysis look like; What are the barriers to implementing the future state; and what does a roadmap to achieve the future state look like? The final report summarizing the concluding discussion stated, "The future state of the discipline will incorporate the use of more quantitative analysis tools during the handwriting examination process to assess and compare handwriting characteristics. Forensic document examiners (FDEs) will employ the use of statistical models to explain the significance of their conclusions based on the uniqueness

of observed and measured handwriting characteristics." Further, the report stated, "It is important to note that automated comparison systems may be considered separate from statistical models, as automated systems can facilitate the matching of a known writer with questioned documents without necessarily generating statistics. This technology provides support during the examination process and may provide new information for the human examiner to consider. FDEs can use statistics and automated systems to complement their current practices and to enhance the way they review cases, but neither can replace humans."

Drs. Gantz and Saunders presented similar messages concerning fingerprint forensics in the Statistics in Forensic Science Topic Contributed Paper Session at the Joint Statistical Meetings in Montreal in August 2013. Dr. Gantz presented his paper "A Similarity Score for Fingerprint Images." The paper co-authored with John Miller describes the scoring algorithms he developed for a totally automated innovative technology enabling the identification of crime scene fingerprints. The presentation was selected to receive an Honorable Mention in the Section on Physical and Engineering Sciences (SPES) Outstanding Presentation Awards indicating that it was among the best of the 73 talks presented in a SPES-sponsored contributed paper session. Dr. Gantz made the same statement he had made concerning automated handwriting identification systems, namely that automated systems are differentiated from statistics and that due to their accuracy and use of novel information they will impact the practice of examiners. Dr. Gantz's scoring algorithms developed for a totally automated technology enabling the identification of crime scene fingerprints are presented in some detail in the full Final Report.

In his presentation, "On Desiderata for Score-Based Likelihood Ratios for Forensic Evidence," Dr. Saunders stated opinions on the desirable features of score-based likelihood ratios (SLRs) for interpreting and presenting forensic evidence. Dr. Gantz is providing Dr. Saunders with latent print based data from automated systems for use in score-based likelihood ratio examples in future research.

Phase II Part a

Scoring Algorithm for an Automated Latent Fingerprint Identification System

The first question commonly encountered by statisticians working in forensics science is the closed set identification problem.

> **Closed Set Identification:**
> In this list of k known sources, which is the source of the trace object found on the suspect?

For closed set identification, the evidence for answering this question, traditionally denoted as E, can be partitioned into two categories $\{E_S, E_U\}$ where:

1. E_S: Information from sample(s) obtained from a set of Specific sources.
2. E_U: Information from sample(s) with an Unknown source.

The samples from the specific sources in the watch list, E_S, are commonly referred to as training samples; each of these sets of samples can be used to estimate the fixed parameters necessary to describe how the corresponding source of the watch list stochastically generates evidence. Let the fixed but unknown parameters that are needed to characterize the i^{th} watch list source's sampling distribution be denoted as $\theta_{s,i}$, for $i = 1,2,\ldots,W$.

The evidence with an unknown source, E_U, is usually constant (fixed) after being discovered. However, the stochastic nature of E_U is characterized by one of the set of parameters from the sources on the watch list; i.e. $\theta_{s,1}$, $\theta_{s,2}$, ..., $\theta_{s,W}$.

Traditionally, this class of problems falls within the domain of statistical pattern recognition with the goal of generating a short list of potential sources for the trace. In most cases, an error is made when the actual source of the trace is not included in the returned short list for the trace. When the stochastic nature of the source distributions and the rates at which we encounter traces from each source are known, an optimal solution (in terms of minimizing the total number of errors made by a system) exists and is known as the Bayes Classification rule.

In a situation where the length of the shortlist of possible sources is one, the Bayes Rule for classification says that, if we are looking at traces generated by the watch list sources over a period of time, the optimal rule for identifying the source of the traces is to assign each trace to the specific source, on the watch list, that has the greatest likelihood of generating said trace. If

the likelihood model for the i^{th} suspect source on the watch list is $f\left(e \mid \theta_{s,i}\right)$, then the Bayes Rule for assigning e_u to the i^{th} source on the watch list is

$$r\left(e_u, \theta_s\right) = \left\{ \underset{i \in \{1,2,\ldots,W\}}{\arg\max}\, \pi_i f\left(e_u \mid \theta_{s,i},\right) \right\},$$

where π_i is the rate at which we observe traces from the i^{th} source and $r\left(e_u, \theta_s\right)$ is a class label from the watch list.

Unfortunately, it is extremely rare to know the exact form of $r\left(e_u, \theta_s\right)$, and it is usually necessary to statistically estimate this rule using the training samples, E_S. In the more complex evidential forms the approaches necessary to estimate r can become extremely sophisticated.

Current research on Closed Set Identification

Grossly speaking, it is our opinion that when exploring new types of forensic modalities or quantifications of forensic evidence, the best way to discover if there is any potential for the method to be applicable to the identification of sources problem is to construct statistical methods for the closed set identification problem.

The primary reason for exploring the closed set identification question first is that there are well-developed statistical tools for addressing this question. A secondary reason is that in closed set identification problems for low dimensional problems there exists a gold standard, Bayes Classification Rule, even if the rule itself needs to be estimated.

For these reasons we have pursued the development of a closed set identification algorithm with respect to a new quantification developed by Sciometrics LLC for fingerprint evidence. The system has become sufficiently accurate in terms of closed set identification error rates, and we will pursue presentation and interpretation of evidence analyzed with this new quantification procedure. The new quantification for fingerprint evidence being exploited by Drs. Saunders and Gantz is now described.

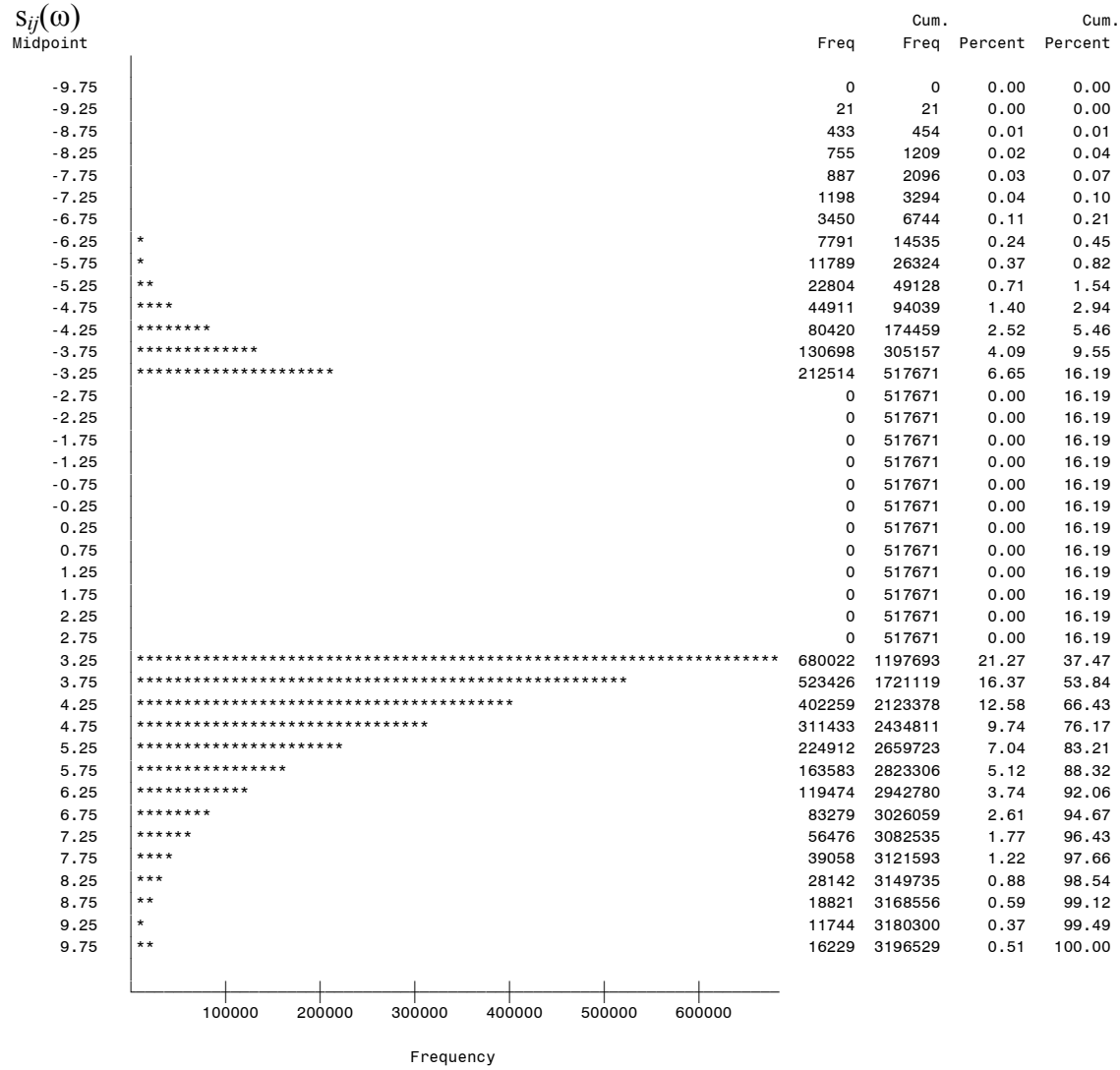
Scoring Algorithm for an Automated Latent Fingerprint Identification System

Dr. Gantz has developed the scoring algorithms for a fully automated fingerprint technology, which creates an accurate invertible Warp from a latent fingerprint to each image in a reference database of fingerprints. The Warp removes the locally nonlinear distortion from the latent. The scoring algorithm starts with a similarity score that is an assessment of the accuracy of the nonlinear, invertible Warp, which yields an overlay of the latent onto a reference image. For each ridge or furrow pixel in the thinned latent image, *pairwise comparisons* among reference prints are computed based on the physical agreement of corresponding pixels from the latent and the reference print as measured in latent space. The pairwise comparison scores are arrayed in the competitive matrix structure that has previously been introduced by Gantz for scoring within multiple biometric modalities. This structure yields scores that Gantz and Saunders are applying to score-based likelihood ratios for fingerprint evaluation. Gantz has presented on his latent fingerprint scoring algorithms at the Impression and Pattern Evidence Symposium (2011), the European Academy of Forensic Sciences (2013) and the Joint Statistical Meetings (2013).

Two very similar scoring derivations follow this introduction. Each presents the scoring of a reference database of images against a Latent. The first is based on actual pairwise comparison scores between reference prints; the other is based on counts of wins from the pairwise comparison scores between reference prints. Each demonstrates the full details of the scoring using one NIST 27 Latent and a reference set of the true matching image and 49 other images objectively selected from a large corpus of images to be close matches to the true matching image. The limit to 49 images is only for ease of computation on a laptop computer.
The Pairwise Comparison Score $S_{ij}(\omega)$ between images *i* and *j* at latent pixel ω is defined as follows. For each pixel in the thinned latent image, the WARP pairs a short Bezier curve in the latent with a Bezier curve in the reference image. Based on this pairing, for each pixel in the thinned latent image, *pairwise comparisons* among reference prints are computed based on the physical agreement of the pixel's best latent to reference image pairs of Bezier curves as measured in latent space. For a single pixel ω, the Similarity Score for an image *i* against the Latent is $\log[d_i(\omega)]$ where $d_i(\omega)$ is the difference between the pixel's Latent Bezier curve and the Warp inversion to the Latent of the pixel's reference image Bezier. The pixel-based pairwise comparison score for images *i* and *j* is $S_{ij}(\omega) = -2\log[d_i(\omega)/ d_j(\omega)]$ when both images have the required Bezier curves defined for the Latent pixel ω. Each pixel-based pairwise comparison yields a logarithmic pixel score $s_{ij}(\omega)$ comparing reference print *i* to reference print *j*. Initially, the score is anti-symmetric in that $s_{ji}(\omega) = -s_{ij}(\omega)$.

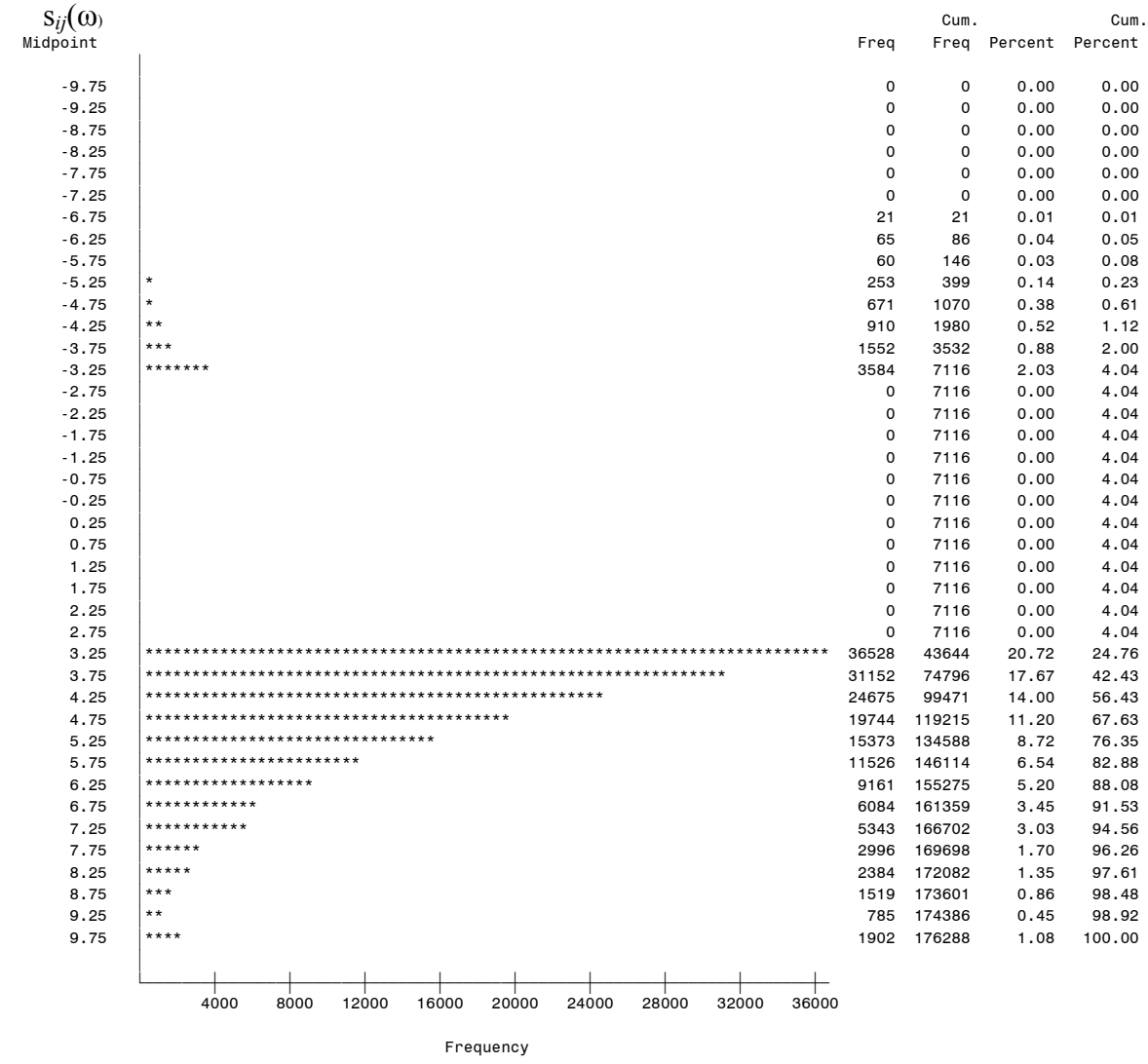
The stronger pixel scores have greater power for identification. Therefore only the pixels with scores satisfying $s_{ij}(\omega) > 3$ or $s_{ij}(\omega) < -3$ are retained for analysis. The following graph (Figure 1) displays the $s_{ij}(\omega)$ scores for the pixels that meet the filtering criteria for a reference set of 50 images, one of which is the true match to the Latent. In this graph, we also restrict $\log[d_i(\omega)] < 4$; that is, we require an accurate similarity score for image *i* for the pixel ω.

Figure 1: $s_{ij}(\omega)$ scores for the pixels that meet the filtering criteria for a reference set of 50 images, one of which is the true match to the Latent. In this graph, we also restrict $\log[d_i(\omega)] < 4$; that is, we require an accurate similarity score for image i for the pixel ω.



The following graph (Figure 2), with the same data restrictions as Figure 1, corresponds to just those pixel scores associated with the true matching reference print.

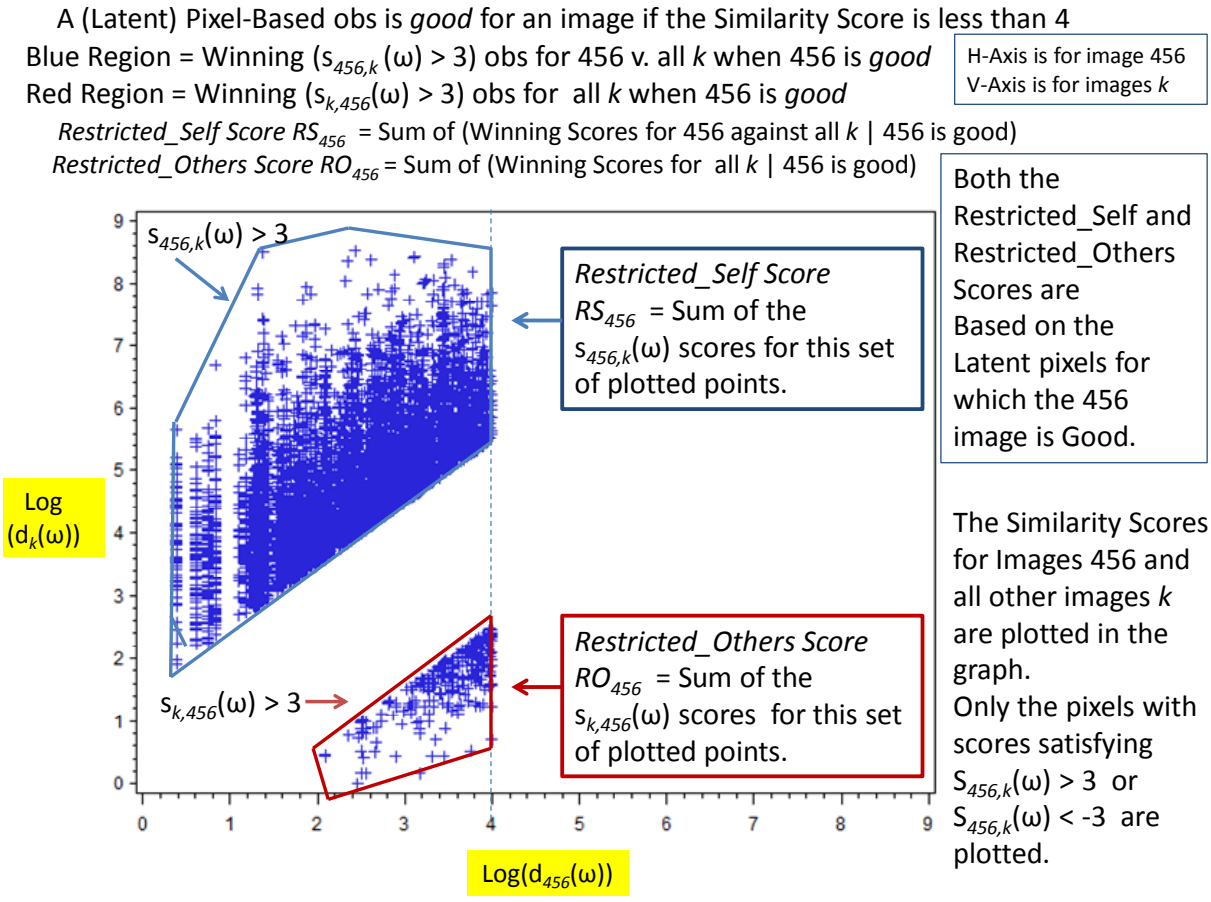
Figure 2: This graph is similar to the graph in Figure 1; however, this graph corresponds to just those pixel scores associated with the true matching reference print ($i = 456$). That is, the graph presents $s_{456,j}(\omega)$ scores for the pixels that meet the filtering criteria for the true match ($i = 456$) to the Latent. In this graph, we restrict $\log[d_{456}(\omega)] < 4$; that is, we require an accurate similarity score for image $i = 456$ for the pixel ω.



Note that 4.04 percent of the scores in this last graph are negative, but 16.19 percent of the scores in the previous graph are negative.

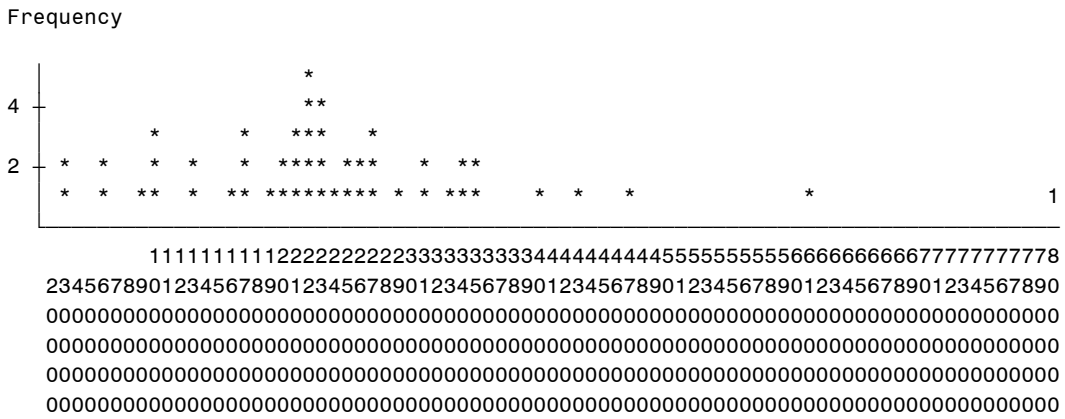
The notions of a Restricted Scores and Unrestricted Scores for an image are now defined by the following figures. The image 456 is the true matching image to the Latent. Scores for this image are used for illustration of the scoring.

Figure 3: Plot of $\log[d_k(\omega)]$ v. $\log[d_{456}(\omega)]$ for all 49 other reference images and all pixels ω satisfying $s_{456,k}(\omega) > 3$ or $s_{k,456}(\omega) > 3$.



The *Restricted_Self Score* $RS_{456} = 800{,}233$. The *Restricted_Self Scores* for all 50 reference images are displayed in the following graph (Figure 4); the *Restricted_Self Score* for image 456 is represented by the symbol '1'.

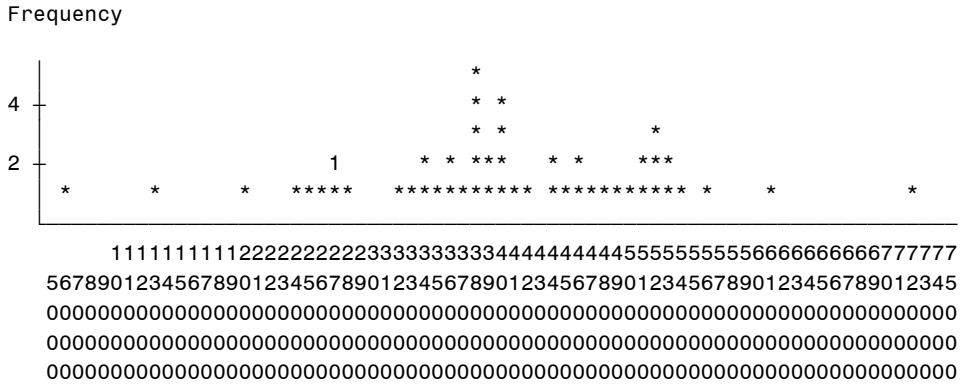
Figure 4: *Restricted_Self Scores RS_j* for all 50 reference images. The *Restricted_Self Score* for image 456 is represented by the symbol '1'.



Restricted_Self Score RS_j

The *Restricted_Others Score* RO_{456} = 26,518. The *Restricted_Others Scores* for all 50 reference images are displayed in the following graph (Figure 5); the *Restricted_Others Score* for image 456 is represented by the symbol '1'.

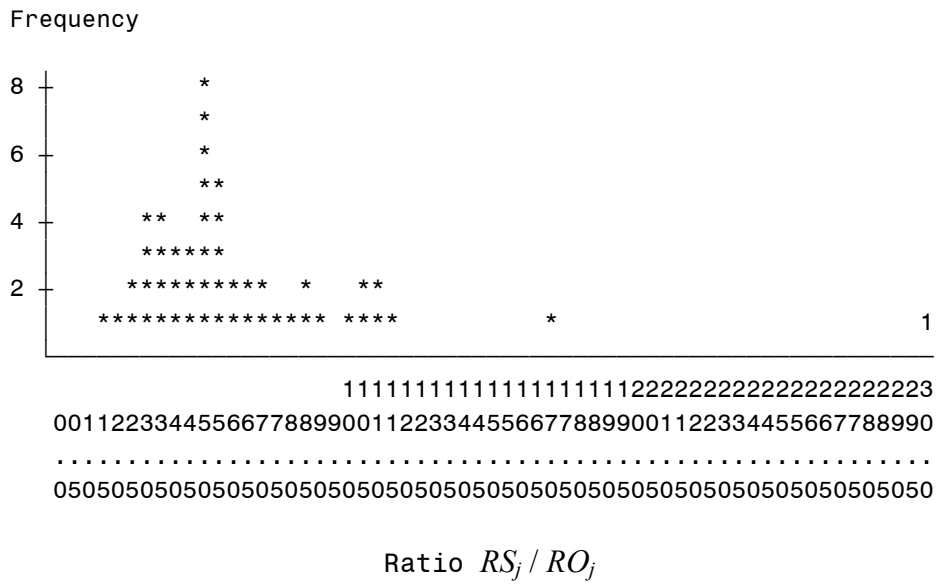
Figure 5: *Restricted_Others Scores RO_j* for all 50 reference images. The *Restricted_Others Score* for image 456 is represented by the symbol '1'.



Restricted_Others Score RO_j

We now display the ratio of the *Restricted_Self Score* to the *Restricted_Others Score* (RS_j / RO_j) for a reference image. The following graph (Figure 6) presents these ratios for the 50 reference images for the particular Latent. The ratio for image 456 is represented by the symbol '1'.

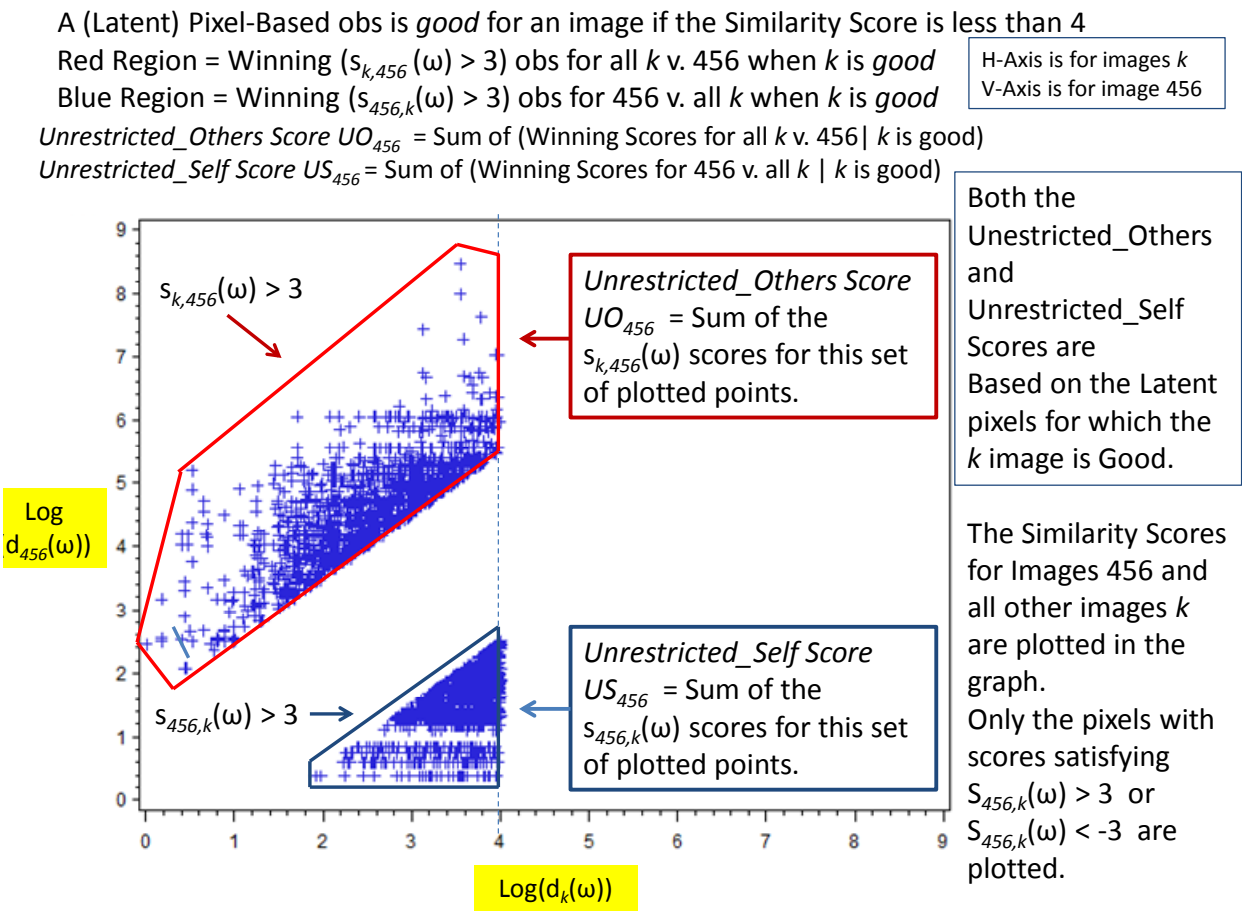
Figure 6: The ratio of the *Restricted_Self Score* to the *Restricted_Others Score* (RS_j / RO_j) for the 50 reference images for the particular Latent. The ratio for image 456 is represented by the symbol '1'.



Note that the true reference image 456 has a very much larger Ratio score (30.2) than any of the other 49 reference images. The nearest competitor image is image 487 with a Ratio Score of 17.1.

The Ratio score RS_j / RO_j measures how well Reference Image *j* competes with all of the other reference images when Reference Image *j* is good.

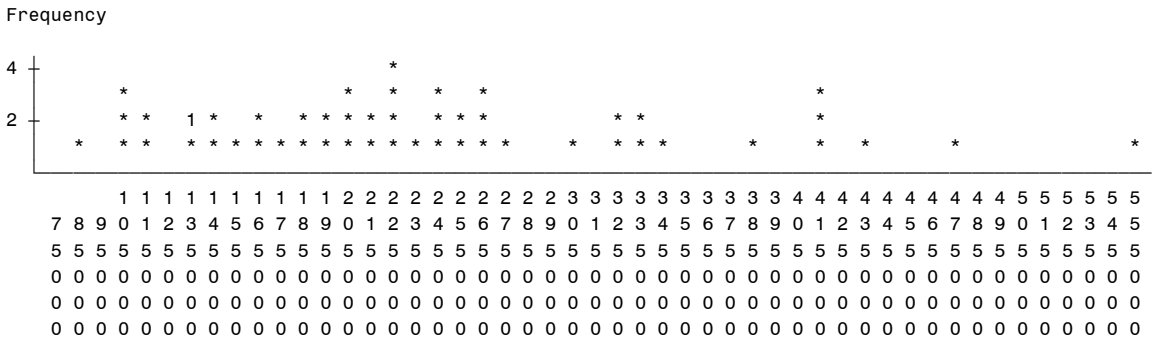
Figure 7: Plot of log[$d_{456}(\omega)$] v. log[$d_k(\omega)$] for all 49 other reference images and all pixels ω satisfying $s_{456,k}(\omega) > 3$ or $s_{k,456}(\omega) > 3$.



We now turn attention to a second Ratio score $\underline{UO_j}/US_j$ which is computed holding image *j* constant. The Ratio score $\underline{UO_j}/US_j$ measures how well Reference Image *j* competes with all of the other reference images when they are good.

Holding *j* = 456 fixed, the *Unrestricted_Others Score* UO_{456} = 130,474. The *Unrestricted_Others Scores* for all 50 reference images are displayed in the following graph (Figure 8); the *Unrestricted_Others Score* for image 456 is represented by the symbol '1'.

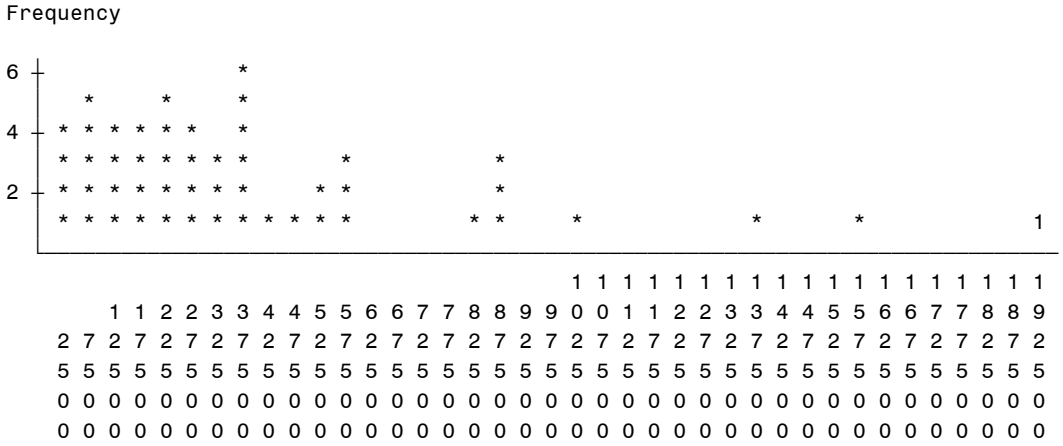
Figure 8: The *Unrestricted_Others Scores* UO_j for all 50 reference images. The *Unrestricted_Others Score* for image 456 is represented by the symbol '1'.



Unrestricted_Others Score UO_j

Holding *j* fixed, the *Unrestricted_Self Score* US_{456} = 191,690. The *Unrestricted_Self Scores* for all 50 reference images are displayed in the following graph (Figure 9); the *Unrestricted_Self Score* for image 456 is represented by the symbol '1'.

Figure 9: The *Unrestricted_Self Scores* US_j for all 50 reference images. The *Unrestricted_Self Score* for image 456 is represented by the symbol '1'.



Unrestricted_Self Score US_j

We now display the ratio of the *Unrestricted_Others Score* to the *Unrestricted_Self Score,* UO_j / US_j, for the set of reference images. The following graph (Figure 10) presents these ratios for the 50 reference images for the particular Latent. The ratio for image *j* = 456 is represented by the symbol '1'.

Figure 10: The ratio of the *Unrestricted_Others Score* to the *Unrestricted_Self Score* UO_j / US_j for the 50 reference images for the particular Latent. The ratio for image j = 456 is represented by the symbol '1'.

```
  Frequency

     |1
     |*
     |*
     |*
  10 +**
     |**
     |***
     |***
     |***
   5 +***
     |****
     |*****
     |******  *
     |****** **  * *                   *                         *                                *
     ----------------------------------------------------------------------------------------------
                                  111111111111111111111111122222222222222222222222223333333333333333333
        1112233344555667778899900111223334455566777889990011122333445556677788999001112233344555667 7
      2604826048260482604826048260482604826048260482604826048260482604826048260482604826048260482604
```

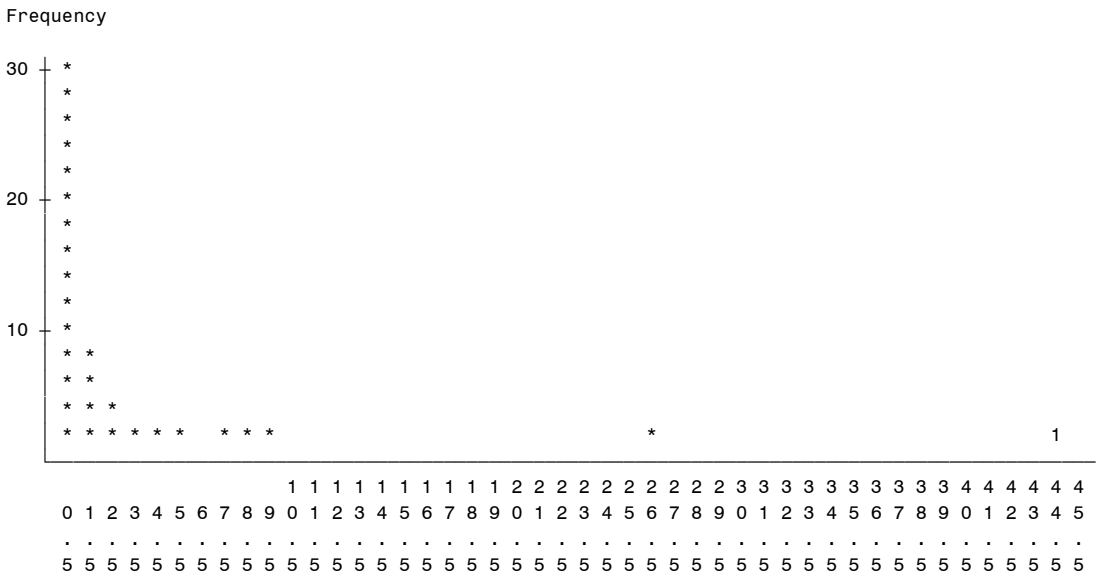
Ratio UO_j / US_j

The true reference image j = 456 has the second smallest Ratio score (0.68) among the other 50 reference images. The smallest Ratio score is for image j = 487 with a Ratio Score of 0.65. The information from all of the above scoring is summarized in the ratio of ratios:

$$(RS_j / RO_j) / (UO_j / US_j) .$$

The true reference image 456 has the value 43.3 for $(RS_j / RO_j) / (UO_j / US_j)$. Reference print 487 has the next largest value 26.2. The following graph (Figure 11) presents these ratios for the 50 reference images for the particular Latent. The ratio for image 456 is represented by the symbol '1'.

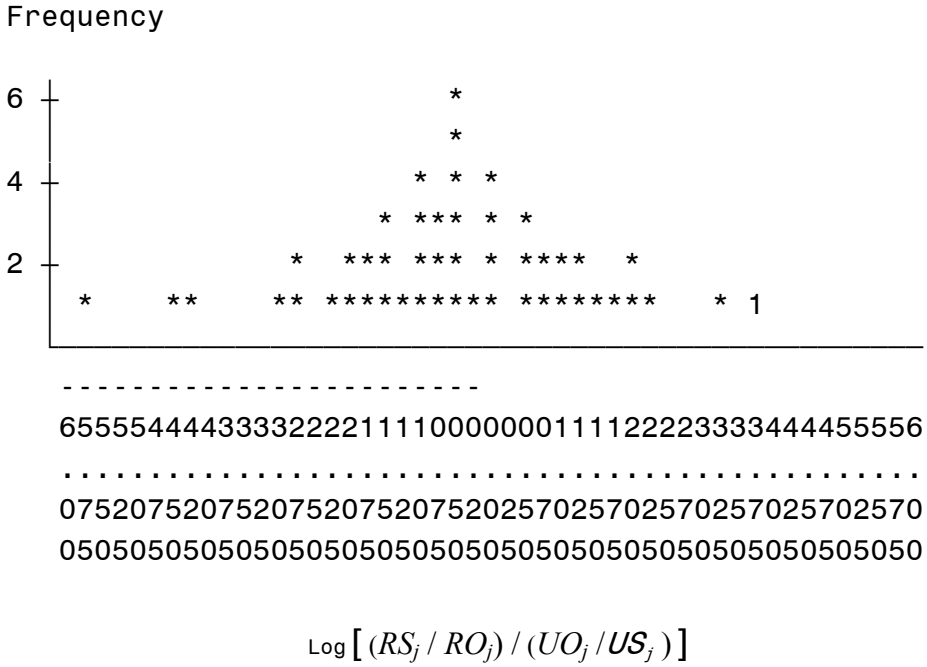
Figure 11: Ratios $(RS_j / RO_j) / (UO_j / US_j)$ for the 50 reference images for the particular Latent. The ratio for image 456 is represented by the symbol '1'.



$(RS_j / RO_j) / (UO_j / US_j)$

Taking logarithms of $(RS_j / RO_j) / (UO_j / US_j)$, we get the following graph (Figure 12).

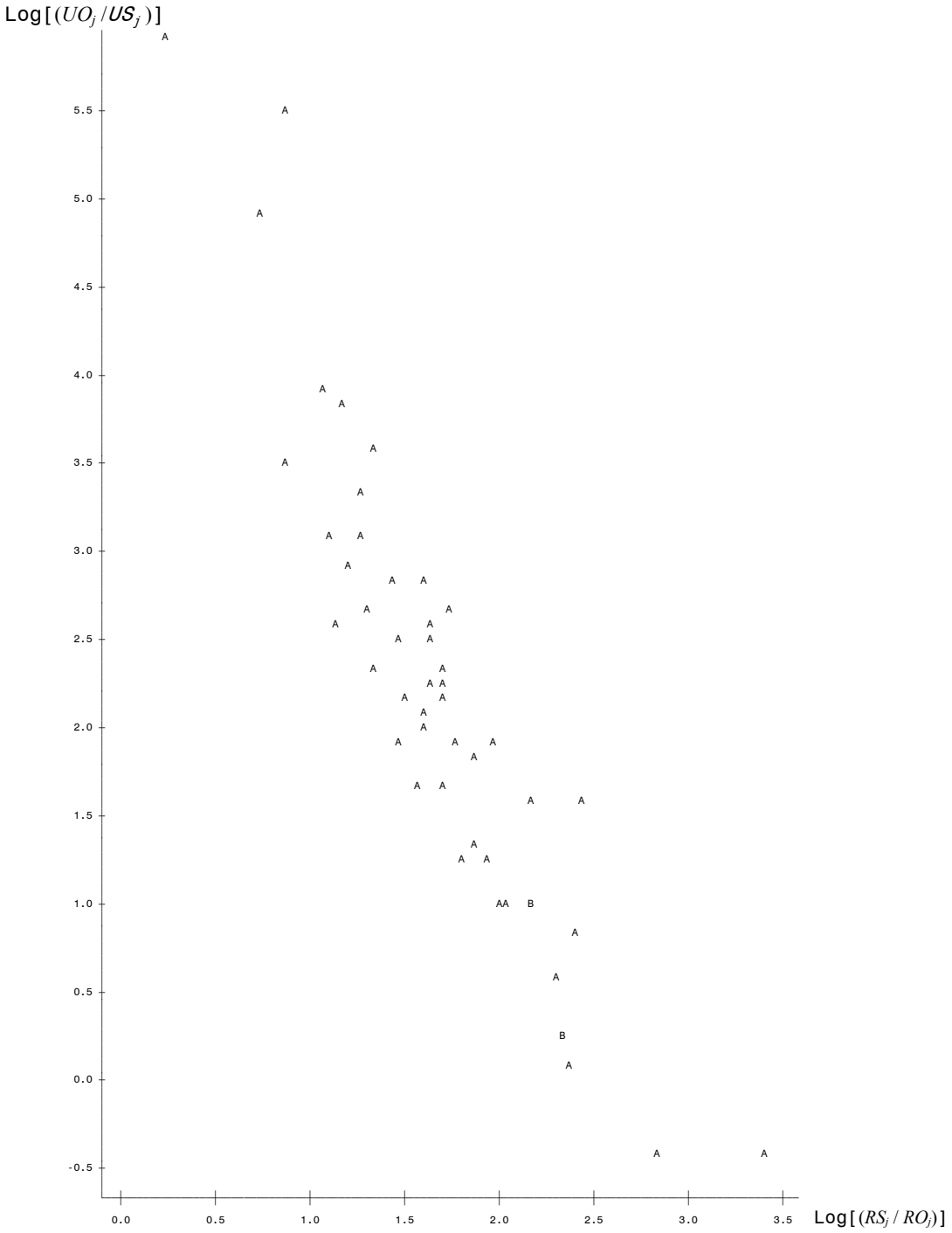
Figure 12: $\text{Log}\,[\,(RS_j / RO_j) / (UO_j / US_j)\,]$ scores for the 50 reference images. The score for the image 456 is represented by the symbol '1'.



$\text{Log}\,[\,(RS_j / RO_j) / (UO_j / US_j)\,]$

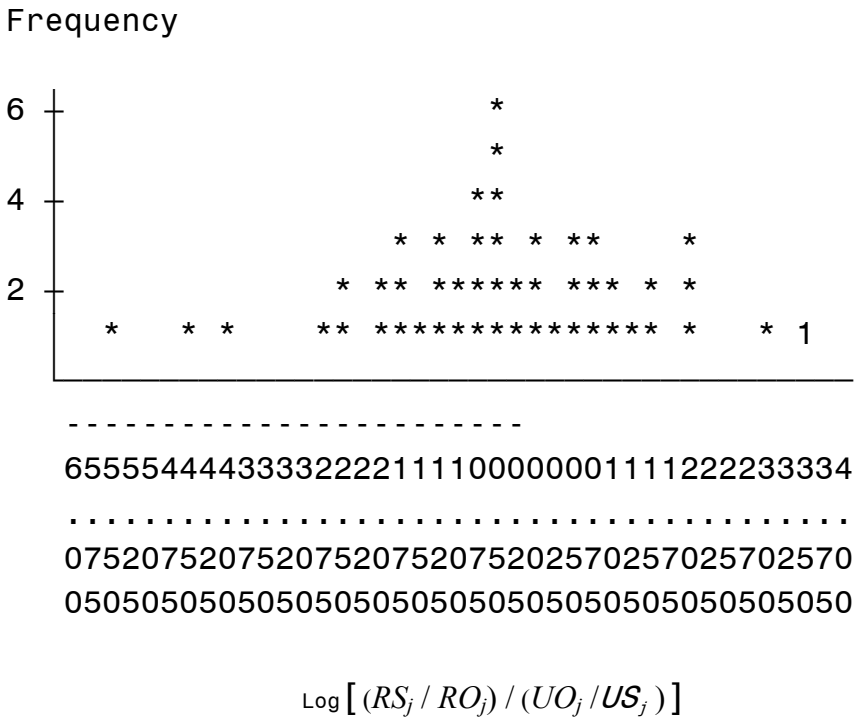
The $\text{Log}\left[(RS_j / RO_j) / (UO_j / US_j) \right]$ value for reference print 456 is 3.79 and the value for reference print 487 is 3.27. $\text{Log}\left[(RS_j / RO_j) \right]$ and $\log\left[(UO_j / US_j) \right]$ scores are highly correlated as seen in Figure 13.

Figure 13:



If we replace the actual pairwise scores $s_{ij}(\omega)$ with counts of wins, then the final ratio scores are given in the following graph (Figure 15).

Figure 15: Final ratio scores gotten by replacing the actual pairwise scores $s_{ij}(\omega)$ with counts of wins.



$$\text{Log}\left[(RS_j / RO_j) / (UO_j / US_j) \right]$$

The $\text{Log}\left[(RS_j / RO_j) / (UO_j / US_j) \right]$ value for reference print 456 is 3.60 and the value for reference print 487 is 3.12.

Utilization of the Scoring Algorithm

The scoring presented here is pixel-based. The individualizing power in a particular pixel depends on the individualizing information locally in the ridge containing the pixel. In regions with high individualizing information, the true matching reference print should accumulate good scores with more consistency than should other reference prints. In that sense, the ranking power in the ultimate score is inherently dependent on the quantity and quality of the Latent print pixels in relation to the quality regions of the true matching reference print.

Sciometrics' current technology for producing overlays of the Latent onto reference prints is very accurate, and the scoring presented here is very effective for closed set identification. The performance of the system has been thoroughly tested using standard databases such as the NIST 27 Good, Bad and Ugly data set and also against data sets prepared by various agencies.

Drs. Gantz and Saunders will soon be exploiting data from this new quantification and scoring procedure relative to the presentation and interpretation of evidence.

Identification of Specific Source

The set of identification of source problems that we have studied considers two alternative and mutually exclusive, but non-exhaustive, propositions or models for how the forensic evidence has arisen. The first model usually corresponds to the prosecution hypothesis and states that a given specific source is the actual source of the trace of unknown origin. The second proposition usually corresponds to the defense hypothesis and states that the actual source of the trace is not the one considered under the prosecution hypothesis, but that it originates from another, unrelated, source in a specified relevant alternative population of sources.

The evidence that we have to address the validity of the two propositions takes the following form:

1. There is a specific source of interest, from which we have a set of samples, denoted as E_s.
2. There are a set of samples of sources from a population of alternative sources, denoted as E_a.
3. A set of samples from a common, but unknown, source denoted as E_u.

The forensic scientist and statistician are then asked to quantify how much support the evidence provides for the model that E_u arose from the specific source of interest when compared to the model that E_u arose from a source in the alternative source population.

Dating back to the 1970's, this problem has been approached within the context of subjective Bayesian hypothesis testing. (See Aitken and Stoney[30]; Lindley 1978[31]; and Shafer[32]). The common approach to these problems is to assume that the problem is inherently low dimensional, the stochastic nature of the evidence can be characterized by a common parametric family of distributions, and that the evidence from the alternative source population is sufficiently precise that it completely characterizes the stochastic nature of the alternative source population. With these assumptions in hand, the forensic statistician can then provide a summary of the scientific evidence that is logical and coherent for updating a prior belief structure concerning the two competing propositions. The 'summary' is typically known as a Bayes Factor in the statistical literature (IJ Good[33]) and a 'Likelihood Ratio' in the forensic science literature. Traditionally this summary is presented as follows:

---

30 Aitken, C. G. G., Stoney, David A., *The Use Of Statistics In Forensic Science*, CRC Press, Oct 31, 1991.

31 Lindley,D.V. (1,977), A Problem in Forensic Science, *Biometrika* 6,4, 207-213.

32 Glenn Shafer, Lindley's Paradox, *Journal of the American Statistical Association* , Vol. 77, No. 378 (Jun., 1982) , pp. 325-334.

33 Good, I.J., Weight of evidence and the Bayesian Likelihood Ratio published in *The Use Of Statistics In Forensic Science*, CRC Press, Oct 31, 1991.

$$\underbrace{\frac{Pr\left(H_p \mid E, I\right)}{Pr\left(H_d \mid E, I\right)}}_{\text{Posterior Odds}} = \underbrace{\frac{Pr\left(E \mid H_p, I\right)}{Pr\left(E \mid H_d, I\right)}}_{\substack{\text{Bayes Factor and/or} \\ \text{Likelihood Ratio}}} \times \underbrace{\frac{Pr\left(H_p, I\right)}{Pr\left(H_d, I\right)}}_{\text{Prior Odds}},$$

where E is the evidence, H_p is the prosecution model for the stochastic nature of the evidence, H_d is the defense model for the stochastic nature of the evidence and I is the relevant background information common to both models. The prior odds summarize our relative belief concerning the validity of the prosecution and defense probability models.

The Bayes Factor then allows us to update our belief and arrive at the Posterior odds concerning the relative validity of the two models. If the Bayes Factor (and the corresponding Posterior odds) is sufficiently high relative to the prior odds, then we conclude in favor of the prosecution model for the stochastic nature of the evidence; on the other hand if it is sufficiently close to zero, we conclude in favor of the defense model for the stochastic nature of the evidence. In effect the Bayes Factor is providing a numerical summary of the answer to both of these questions:

> *"What do we believe the likelihood of observing the evidence under the prosecution model is?"*
> vs.
> *"What do we believe the likelihood of observing the evidence under the defense model is?"*

An extremely important note is that, when constructing a Bayes Factor, it is necessary to use a probability measure to characterize the forensic scientist's belief about the stochastic nature of how the specific source generates evidence. The traditional default belief measure concerning the specific source is that the specific source is typical of the population of alternative sources. (Aitken and Taroni[34])

In the context of formal Bayesian Model selection, the goal of a statistical analysis is to rigorously quantify the belief concerning the validity of a given model after having observed the evidence. This type of analysis is typically decomposed into various components – the first being the prior belief concerning the relative validity of the two competing models. The second is a set of priors for prosecution and defense models that characterize the belief about the parameters of the stochastic models.

Our research program has taken two directions related to this problem of the quantification of the value of evidence. The first, is concerned with various aspects the development of an approximate value of the evidence for complex evidence forms when the actual likelihood structure is intractable (the main thrust of Phase II). These approximate values of the evidence are commonly referred to as Score Based Likelihood Ratios (SLRs) in the statistical literature.

---

[34] Aitken, C. G. G., Taroni, F., *Statistics and the Evaluation of Evidence for Forensic Scientists*, Wiley, 2004, 2nd Edition.

Phase II Part b

Score-based likelihood ratios for handwriting evidence[35]

Amanda B. Hepler [a,*], Christopher P. Saunders [a], Linda J. Davis [b], JoAnn Buscaglia [c]

[a] Document Forensics Laboratory (MS 1G8), George Mason University, Fairfax, VA 22030, USA [b] Department of Statistics (MS 4A7), George Mason University, Fairfax, VA, 22030, USA [c] FBI Laboratory, Counterterrorism & Forensic Science Research Unit, Quantico, VA 22135, USA

Acknowledgments

Some of the results in this paper were presented at the AAFS 61st Annual Scientific Meeting, February 16-21, 2009, Denver, CO. The authors would like to acknowledge Gannon Technologies Group (GTG) for computational support and the FBI Laboratory for supplying the set of handwritten documents.

This section has been assigned a publication number of the Laboratory Division of the FBI. This work was supported in part under a Contract Award from the Counterterrorism and Forensic Science Research Unit of the FBI Laboratory. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

This article was supported in large part by Award No. 2009- DN-BX-K234 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

A. Hepler was supported by an Intelligence Community (IC) Postdoctoral Research Fellowship, NGIA HM1582-08-0036.

[35] This section has been published as Hepler AB, Saunders CP, Davis LJ, Buscaglia J. Score-based likelihood ratios for handwriting evidence. Forensic Sci Int. 2012 Jun 10; 219(1-3):129-40

Additional Background information is provided in Davis LJ, Saunders CP, Hepler A, Buscaglia J. Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. Forensic Sci Int. 2012 Mar 10; 216(1-3):146-57.

ABSTRACT: Score-based approaches for computing forensic likelihood ratios are becoming more prevalent in the forensic literature. When two items of evidential value are entangled via a score-function, several nuances arise when attempting to model the score behavior under the competing source-level propositions. Specific assumptions must be made in order to appropriately model the numerator and denominator probability distributions. This process is fairly straightforward for the numerator of the score-based likelihood ratio, entailing the generation of a database of scores obtained by pairing items of evidence from the same source. However, this process presents ambiguities for the denominator database generation – in particular, how best to generate a database of scores between two items of different sources.

Three alternatives have appeared in the literature. Denominator databases have been generated by pairing (1) the item of known source with randomly selected items from a relevant database; (2) the item of unknown source with randomly generated items from a relevant database; or (3) two randomly generated items. When the two items differ in type, perhaps one having higher information content, these three alternatives can produce very different denominator databases. While each of these alternatives has appeared in the literature, the decision of how to generate the denominator database is often made without calling attention to the subjective nature of this process.

In this paper, we compare each of the three methods (and the resulting score-based likelihood ratios), which can be thought of as three distinct interpretations of the denominator proposition. Our goal in performing these comparisons is to illustrate the effect that subtle modifications of these propositions can have on inferences drawn from the evidence evaluation procedure. The study was performed using a data set composed of cursive writing samples from over 400 writers. We found that, when provided with the same two items of evidence, the three methods often would lead to differing conclusions (with rates of disagreement ranging from 0.005 to 0.48). Rates of misleading evidence and Tippet plots are both used to characterize the range of behavior for the methods over varying sized questioned documents. The appendix shows that the three score-based likelihood ratios are theoretically very different not only from

each other, but also from the likelihood ratio, and as a consequence each display drastically different behavior.

KEYWORDS: forensic science, likelihood ratio, handwriting evidence, statistical evidence evaluation, forensic statistics, questioned documents

1. Introduction

The likelihood ratio paradigm has been proposed as a means for quantifying the strength of evidence for a variety of forensic evidence, including handwriting, speech, earmarks, glass fragments, fingerprints, footwear marks and DNA [1-9]. A body of evidence can be evaluated by calculating the likelihood ratio, which compares the probability of the "evidence" under two competing propositions (or hypotheses), often denoted as the prosecution proposition (H_p) and the defense proposition (H_d). Consider the scenario where two items of evidence are found over the course of a forensic investigation, and the following source-level hypotheses are of interest:

H_p: The two items came from the same source,

H_d: The two items came from different sources.

Let x denote[36] a measurement obtained from the *source*, or the sample with a known source (e.g., suspect's known writing samples, crime scene window). Let y denote a measurement obtained from the *trace*, or the sample with an unknown source (e.g., bank robbery note, glass fragment obtained from the suspect). If one assumes that x and y are realizations from continuous random variables X and$Y$, the likelihood ratio is defined by

$$\mathrm{LR} \equiv \frac{f(x, y|H_p, I)}{f(x, y|H_d, I)},$$

where I represents background information, and f denotes the probability distribution associated with the random variables X and Y. When x and y are discrete measurements, f is a probability; when x and y are

[36] Throughout this manuscript, the following conventions are used: uppercase bold letters denote *random* matrices or vectors; lowercase bold letters denote *observed or known* matrices or vectors; lowercase letters denote *observed or known* scalars.

continuous measurements, f is a continuous probability density function. As stated in [10], the numerator and denominator densities might be very different due to the differing conditioning arguments, but it is common practice to allow the generic symbol f to represent both functions.

In many cases (e.g., when the evidence is represented using a high-dimensional quantification technique[11]), the numerator and denominator of LR are not obtainable directly, without making (perhaps) unfounded assumptions about the underlying processes that generate the evidence [12]. A promising surrogate, which can be applied to virtually any evidence type, is a score-based approach [10, 12-18].

In this article, we critically examine three methods appearing in the literature for estimating the score-based likelihood ratio (SLR) in the specific context of natural handwriting evidence. While our illustrations focus on this modality, the concepts apply broadly to the application of these methodologies to any type of evidence for which a meaningful paired score can be defined.

Each methodology makes very different assumptions about the nature of the random variables X and Y, specifically in the denominator (under the defense's proposition). Often these are listed either as assumptions [18], an (often unstated) byproduct of database generation [10,12,16]. The intent of this paper is to illuminate, for both the statistical and non-statistical audience, the assumptions underlying the three different methodologies and how they are in fact subtle changes to the interpretation of H_d. The hope is that once these interpretations are laid bare, the forensic community can then appropriately weigh their merits and applicability. This is particularly important since, as shown in Section 5 and in Appendix A, the three methods can yield drastically different results when given the exact same evidence. It is our belief that these three score-based methods cannot gain mainstream acceptance until this denominator specification problem is resolved by the forensic community.

The outline for this paper is as follows. Section 2 presents each method in a unified notation, while making explicit each of the underlying assumptions and their associated H_d interpretation. Section 3 briefly details the quantification technique used to quantify handwritten documents (more detailed

descriptions appear elsewhere [13,19]). Also, Section 3 details the algorithms used to obtain estimates for each SLR, denoted throughout as SLR_1, SLR_2, and SLR_3. Finally, Sections 4 and 5 detail the design and results of a comparison study showing the impact that selecting one SLR over another (i.e. one set of assumptions over another) has on the estimated SLRs.

2. Score-Based Likelihood Ratios

For many types of forensic evidence, obtaining the likelihood ratio, as defined above, has proved difficult, if not impossible [12]. For some types of evidence, it is rare that the underlying process which generates X and Y is sufficiently understood to make the assumption that the distribution is an element of some common family of distributions. For example, with certain quantifications of the elemental composition of glass fragments it is not necessarily reasonable to make the blanket assumption that X and Y follow a normal distribution [20], as is often done [21]. Even for the most basic forms of DNA evidence there were many years of research and academic discussions before reasonable distributional assumptions were known to an adequate degree of certainty that might be required in a court of law [8]. Even if the distribution is known or can reasonably be assumed, true parameter values are rarely known and are likely difficult to estimate for the more complex quantifications of the evidence. When x and y represent high dimensional measurements, as would be the case if one considers the multifaceted attributes that make up one's full body of handwriting (or *writing profile*), the problem is exacerbated as now we are faced with 1) how to probabilistically characterize each attribute individually and 2) how to capture probabilistic dependencies sure to exist among the attributes.

Score-based approaches seem able to overcome at least some of these challenges. If one can capture similarities or differences between two items via a univariate score function that illuminates as to whether or not the items have a common source, then dimensionality of the problem is greatly reduced [12,15,16]. Determining (or estimating) the probability distribution of this score function remains a challenge however, as will be highlighted throughout the remaining sections of this paper.

A brief introduction to score-based likelihood ratios is provided here. A more detailed discussion, within the context of handwritten documents can be found in [13]. Let the function which assesses the dissimilarity between x and y be denoted by $\Delta(x, y)$. The score-based likelihood can then be described as a proxy of sorts to the LR,

$$\text{LR} = \frac{f(x, y|H_p, I)}{f(x, y|H_d, I)} \approx \frac{g(\Delta(x, y)|H_p, I)}{g(\Delta(x, y)|H_d, I)}, \quad (1)$$

where g denotes the probability distribution associated with the random variable $\Delta(X, Y)$. Often in the literature the rightmost quantity is also denoted by LR [10, 12]. In the interest of transparency and clarity, in this work this quantity is denoted by SLR. Another impetus to keep these quantities distinct is that, as noted by [19], the suitability of the approximation LR $\approx$ SLR has not been investigated thoroughly. It is shown in Appendix A for a simplified scenario (where the probability distributions of $X, Y,$ and $\Delta(X, Y)$ are all known) that the three SLRs under consideration here often do not well approximate the LR.

The numerator of the leftmost expression in Equation (1) can be interpreted in layman's terms as: the likelihood of observing these two measurements if the items come from the same source. Similarly the denominator can be interpreted as the likelihood of observing these two measurements if the items come from different sources. In order to compute this quantity, statisticians typically make the assumptions 1) the marginal distribution of X is independent of whether or not H_p or H_d is true, and 2) measurements on X and Y are independent if H_d is true. Under assumptions 1) and 2), the LR reduces to:

$$\text{LR} = \frac{f(x, y|H_p, I)}{f(x, y|H_d, I)} = \frac{f(y|\, x, H_p, I) f(x|\, H_p, I)}{f(y|\, H_d, I) f(x|\, H_d, I)} = \frac{f(y|\, x, H_p, I)}{f(y|\, H_d, I)}. \quad (2)$$

The simplification achieved in Equation (2) is what drives all DNA likelihood ratio calculations, and most non-score based approaches [8,22]. Unfortunately, an analogous development for the SLR (right side of Equation (1)) is not possible since measurements from the trace and the known source are now tied together via the score function and cannot be disentangled. Conditioning on x is of no use here, since:

$$\text{SLR} = \frac{g(\Delta(x, y)|\, H_p, I)}{g(\Delta(x, y)|\, H_d, I)} = \frac{\int g(\Delta(x, y)|\, x, H_p, I) f(x|H_p, I) dx}{\int g(\Delta(x, y)|\, x, H_d, I) f(x|H_d, I) dx}, \quad (3)$$

and, in general, Equation (3) cannot be simplified in a straightforward manner, if at all. The simplifications leading to Equation (2) no longer hold – the conditioning on x must remain in the denominator, and the marginal distribution of x no longer cancels out as it appears inside separate integrals in the numerator and denominator.

Despite the fact that the SLR cannot be simplified in any meaningful way to facilitate computation, several score-based methods have emerged in the literature. Many make, either explicitly or implicitly, simplifying assumptions in order to estimate the SLR. The body of literature here is growing, and we restrict our attention to three such methods which serve as a continuation of our work in [13]. Each SLR method makes use of a similar numerator estimation technique previously reviewed in [13], while differing in their approach to estimating the denominator.

The numerator of the simplified LR appearing in Equation (2) can be interpreted in layman's terms as: the likelihood of observing the trace measurement if it came from the known source. The denominator can be interpreted as: the likelihood of observing the trace measurement if it came from a different source. To compute the denominator directly, an additional assumption must be made regarding the alternate source. The most common, often referred to as the "random man" assumption, is that the source of the trace is randomly selected from some "relevant population" of sources [22]. This leads to the following *statistical interpretation* of the denominator: the likelihood that the trace measurement came from a random source in a relevant population.

The interpretation of the numerator for the SLR is slightly different from that of the LR: the likelihood of observing *this score between the trace and the known source* if they came from the same source. The interpretation of the denominator is: the likelihood of observing *this score between the trace and the known source* if they came from different sources. When one tries to be more specific about the denominator in order to obtain probability distributions, ambiguity arises. As above, some notion of "random source" must come in, but there is subjectivity in how to proceed. Distinct interpretations of H_d motivate the three SLRs under consideration in this paper. The first method contends that the known

source is a random selection from the relevant population; the second contends that the source of the trace is a random selection from the relevant population; and the third contends that both the trace and the known source are randomly selected from the relevant population.

2.1 Score-based Numerator

All three methods we consider here have considered the following interpretation of the SLR numerator: the likelihood of observing this score if the known source measurement is paired with measurements taken from traces randomly drawn from the known source population. The new specification of the hypothesis being entertained is:

H_p: $\Delta(x, y)$ arises from the distribution of scores obtained by pairing x with a randomly generated Y, where both x and y arise from the same distribution.

While this hypothesis is not necessarily reasonable from the perspective of a prosecution attorney, it is in fact the hypothesis under consideration when one reports one of the three SLRs in court. For clarity, we will refer to the type of proposition, which fully specifies the desired probability distribution as a *statistical proposition*, whereas *forensic propositions* refer to those of direct interest to the courts. We prefer this approach over relegating these specifications to the background information or enumerating them as assumptions because we feel those approaches lack transparency and/or clarity, particularly for non-statisticians.

This new specification introduces conditioning upon x the numerator of the SLR, that is $g\big(\Delta(x,y)\big|\, H_p, I\big) \approx\ g\big(\Delta(x,y)\big|\, x, H_p, I\big)$. From Equation (3) it is clear that this is indeed an approximation. The impact this type of approximation has on the resultant score-based likelihood ratios is investigated in Appendix A for a simplified scenario where all distributions are known.

2.2 SLR_1: Trace-anchored

Some researchers [14-16] have considered the following interpretation of the SLR denominator: the likelihood of observing this score if *the* trace measurement is paired with measurements taken from random sources in some relevant population. The statistical proposition being entertained is:

H_{d1}: $\Delta(x, y)$ arises from the distribution of scores obtained by pairing y with a randomly selected x from the relevant population.

This new interpretation of the denominator of the SLR actually changes the specification of the SLR denominator, $g(\Delta(x,y)|\, H_d, I) \approx\ g(\Delta(x,y)|\, y, H_d, I)$. Noting that conditioning on y in Equation (3) (rather than x) would also not lead to any simplification, it is clear that these two quantities are not in fact equal. Using this approximation, the first score-based likelihood ratio under consideration is

$$\mathrm{SLR}_1 = \frac{g\big(\Delta(x,y)\big|\, x, H_p, I\big)}{g(\Delta(x,y)|\, y, H_d, I)}.$$

Whether or not SLR_1 serves as a reasonable proxy for LR is an open question. The example in Appendix A is aimed at informing this debate.

One issue with conditioning on y in the denominator is that it is asymmetric, in the sense that the numerator and denominator are conditioning on different quantities. Another conceptual issue with SLR_1 is that, in the case of glass evidence (or any type of evidence where the item of unknown source is taken from the suspect), the conditioning in the denominator is on measurements taken from the suspect. Specific properties of the crime scene window are ignored entirely, and it is therefore less informative than if those characteristics had been accounted for [19]. However, in the case of handwriting this type of conditioning seems more plausible, as specific properties of the bank robbery note are informing the denominator probability distribution.

One also might consider the recommended conditioning rules provided in [23]. They advocate conditioning on the sample with greater *information content*, which in the case of handwriting would be y (the suspect's known writing samples). However, for glass the desired conditioning would be x (the window at the scene) which again leads to ambiguous notions of the "correct" conditioning. It should be noted that in [23] this conditioning strategy was aimed at simplifying the computation (much like the arguments in Equation (2)). This computational advantage is lost for the SLR, as illustrated above in Equation (3).

2.3 SLR_2: Source-anchored

Others [10] have proceeded with following interpretation of the SLR denominator: the likelihood of observing this score if trace measurements taken from randomly selected sources from a relevant population are paired with *the* measurement taken from the known source. This is somewhat analogous to the LR denominator interpretation in that the trace measurement now comes from a random source. This interpretation again changes the specification of the SLR denominator:

$$\mathrm{SLR}_2 = \frac{g\left(\Delta(x,y)\middle|\, x, H_p, I\right)}{g(\Delta(x,y)|\, x, H_d, I)}.$$

We now have symmetric conditioning, on X in both the numerator and denominator. Again, whether or not this quantity serves as reasonable proxy for LR is considered in Appendix A. This development leads to the following denominator proposition:

H_{d2}: $\Delta(x,y)$ arises from the distribution of scores obtained by pairing x with a randomly selected Y from the relevant population.

This approach succumbs to some of the same criticisms as SLR_1, but in reverse: for glass evidence this conditioning seems reasonable in that specific characteristics of the crime scene evidence are directly relevant to the denominator distribution, but for handwriting specific characteristics of the bank robbery note are ignored (i.e. the writer used all capital letters).

2.4 SLR_3: General Match

The final SLR approach considered in this work, often used in biometrics [24], applies the following interpretation of the denominator: the likelihood of observing this score if a trace measurement taken from a randomly selected source from a relevant population is paired with a measurement taken from a different source randomly selected from a relevant population. This makes no changes to the SLR denominator:

$$\mathrm{SLR}_3 = \frac{g\left(\Delta(x,y)\middle|\, x, H_p, I\right)}{g(\Delta(x,y)|\, H_d, I)}.$$

Again, whether or not this serves as a reasonable proxy for LR is considered in Appendix A. This development leads to the following denominator proposition:

H_{d3}: $\Delta(x, y)$ arises from the distribution of scores obtained by pairing a randomly selected X from the relevant population with a randomly selected Y from that same relevant population.

This SLR is far less informative in that the denominator distribution depends neither on specific characteristics trace nor on characteristics of the known source [19]. That is, the denominator distribution would remain unchanged if a different trace were observed, or if a different known source is considered.

The next section of the paper shows how to generate each SLR for a specific quantification of handwritten documents.

3. Estimating SLRs in handwriting.

Handwriting-specific definitions of the evidence (following the notation introduced in [13]) are as follows: E_S denotes a collection of writings known to have originated from the suspect (henceforth *suspect's template*) and x represents some quantification obtained from those writings. E_U denotes a handwritten questioned document (QD) found at the scene of unknown source, and y represents some quantification obtained from that document. E_A denotes a collection of writing samples taken from alternative sources.

3.1 Handwriting Quantification

Selection of an appropriate score will depend heavily on the numeric representation, or quantification technique used to describe a handwritten document. The quantification method used here, developed by Gannon Technologies Group, first scans and skeletonizes the document, which has been manually parsed into characters, as shown for the word "London" in Figure 1. Subsequent to this segmentation, a proprietary, automated process was used to represent each parsed character's skeleton by an isomorphic class of graphs (a geometric form that remains invariant under certain transformations, e.g. bending or stretching), referred to as an *isocode*. Details of this process are described at length elsewhere [13,25] however a schematic depicting the method appears in Figure 1.

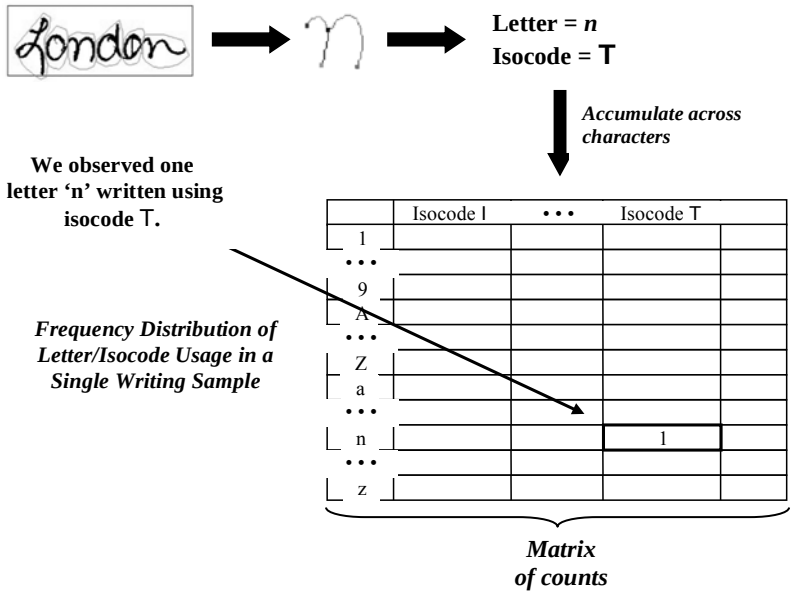


Figure 1. ***Schematic of the Quantification Process***

Define a *writing profile* as the entire body of writing that a writer has written or will ever write. Define a writer's *template* as a collection of writing samples from an individual assumed to be sufficiently rich for characterizing an individual's writing profile. Using this quantification method, E_S is reduced to the matrix of counts computed by combining counts over a large collection of known writing samples obtained from a suspect (*suspect's template*), represented by the random variable $\boldsymbol{X}$. E_U is reduced to the matrix of counts computed from a questioned document, represented by the random variable $\boldsymbol{Y}$.

3.2 Estimating the SLR

3.2.1 Dissimilarity Score

We first define a dissimilarity statistic (or score) that can be computed for two documents (or collection of documents). We selected the Kullback-Leibler divergence [26] to capture the difference between the observed matrices of counts for two writing samples, row by row (i.e. letter by letter). These divergences are combined over letters using a weighted average, ensuring that frequently observed letters (across both documents) contribute more to the dissimilarity score. Details appear in Appendix B.

At this point, it is important to emphasize that the procedure that follows does not depend on the selection of this particular score. A multitude of scoring methods can be used in its place (e.g. see [13] for a similar analysis using a similarity score based on Pearson's chi-squared statistic).

3.2.2 Database Generation

To estimate the numerator and denominator densities of the SLRs we need to obtain databases of scores generated in several ways. For the numerator, we need a database of scores where both x and y were obtained from documents written by the suspect. This is a fairly straightforward matter in our case, and the reader is referred to [13] for specific details. For the denominator, we need a database of scores where x and y are generated from different sources, according to the conditioning assumptions of SLR_1, SLR_2, and SLR_3.

Numerator Database

Ideally, a database would exist consisting of scores obtained by comparing "QD-like" documents to the suspect's template. It is unreasonable to expect a large number of "QD-like" documents to be discovered over the course of the investigation. For example, if the QD is a bank robbery note, only in extremely rare cases would, a priori, a collection of such bank robbery notes exist. One might suggest requesting the generation of a collection of "QD-like" documents from the suspect; however, this might not result in the most representative sample, especially in cases where the suspect is indeed the culprit, as there is motivation to disguise his or her writing style. In addition, the number of samples needed to accurately estimate the distribution of scores would be prohibitive.

In light of these challenges, [25] proposed a method of obtaining an arbitrarily large database (size denoted by *N*) of 'within' scores using a subsampling algorithm. Noting that n_U represents the number of characters in QD and n denotes the total number of characters in the suspect's template, the details of the slightly modified[37] algorithm employed appears below:

[37] In [***25***], a random selection of n_U characters was chosen, whereas here n_U consecutive characters were chosen. We feel that the use of consecutive characters best aligns with the natural writing that might appear in a QD.

Subsampling Algorithm for Generating Numerator Database

For $i = 1 \ldots N$, where N denotes a sufficiently large number of iterations,

1. Randomly divide the suspect's template into two subsets, with character counts n_U and $n - n_U$ respectively. This is done by randomly selecting a (starting) character from the first $n - n_U$ characters. The selected character along with the next $n_U - 1$ characters is defined as the *pseudo-QD* and from it we obtain the matrix of counts $\boldsymbol{y}_i$. The remaining characters form a *pseudo-template*, from which we obtain the matrix of counts $\boldsymbol{x}_i$.
2. Compare the two simulated writing samples, recording the resultant score: $\Delta(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

Denominator Databases

Before the detailed algorithms are presented, we first must address the challenge of obtaining a representative collection of writing templates from potential alternate sources. Recall, above we denoted this collection of templates by E_A as it is considered part of the evidence collected which may differ from case to case and which, especially when E_A is of limited size, will have a significant impact on the estimation of the score-based likelihood ratio. We make the simplifying assumption that a large, representative collection of templates exists. In future work, we intend to examine more practical scenarios, and investigate the impact typical violations of these assumptions have on the estimation procedure.

Once a large, representative collection of templates E_A is established, the mechanics of generating between scores for each of the three denominator SLR interpretations can be detailed.

$\mathbf{SLR_1}$: The trace-anchored interpretation of H_d, tailored to handwriting evidence, is "the evidence score arises from the distribution of scores obtained by pairing *the* QD with *a* template written by a random individual." A detailed illustration of an adaptation of this method for the analysis of handwriting

evidence can be found in [13]. The specific algorithm appears below.

Trace-anchored Algorithm for Generating Denominator Database

Obtain a matrix of counts from the QD, denoted by $\boldsymbol{y}_U$. Then, for $i = 1 \ldots N_A$, where N_A represents the number of writers in E_A,

1. Select the i^{th} writer from E_A and obtain a matrix of counts from that individual's template, denoted by $\boldsymbol{x}_i$.
2. Compare the two writing samples, recording the resultant score: $\Delta(\boldsymbol{x}_i, \boldsymbol{y}_U)$.

$\mathbf{SLR_2}$: The source-anchored interpretation is "the evidence score arises from the distribution of scores obtained by pairing *a* QD written by a random individual with *the* template written by the suspect." The specific algorithm appears below.

Source-anchored Algorithm for Generating Denominator Database

Obtain a matrix of counts from the suspect's template, denoted by $\boldsymbol{x}_S$. Then, for $i = 1 \ldots N$, where N denotes a sufficiently large number of iterations,

1. Randomly select a writer from E_A, and randomly select n_U characters to serve as the pseudo-QD. Obtain the matrix of counts, denoted by $\boldsymbol{y}_i$.
2. Compare the two writing samples, recording the resultant score: $\Delta(\boldsymbol{x}_S, \boldsymbol{y}_i)$.

It should be noted that while [10] does hold $\boldsymbol{x}_S$ fixed, they do not proceed with their database generation in exactly the same manner. They introduce an extra layer of complexity by generating (what would be the equivalent of) multiple pseudo-QDs from every writer in E_A in order to generate N_A different writer-specific databases. Here, due to computational constraints, only one pseudo-QD is generated per writer.

$\mathbf{SLR_3}$: The final interpretation considered, which avoids anchoring all together, is "the evidence score arises from the distribution of scores obtained by pairing *a* QD written by a random individual with *a* template written by a different random individual." The specific algorithm appears below.

General Match Algorithm for Generating Denominator Database

For $i = 1 \ldots N$, where N denotes a sufficiently large number of iterations,

1. Randomly select writer 1 from E_A and randomly select a document of size n_U from his/her template to obtain a pseudo-QD. Obtain the matrix of counts, denoted by $\boldsymbol{y}_i$.
2. Randomly select writer 2 (distinct from writer 1) from E_A, and obtain a matrix of counts from his/her template, denoted by $\boldsymbol{y}_i$.
3. Compare the two writing samples, recording the resultant score: $\Delta(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

3.2.3 Distribution Estimation

Assuming one of the three denominator algorithms is selected, two collections of scores have been obtained, one under the prosecution's hypothesis and one under the selected defense hypothesis. The probability densities of those scores are rarely known exactly and must be estimated. Denote those estimated densities by $\hat{g}$. Normal probability plots of the "numerator scores" and the three sets of "denominator scores" indicated a normal approximation was reasonable (results not shown). After obtaining the sample mean and variance of our N (or N_A for the trace-anchored approach) generated observations, $\hat{g}$ is defined to be a normal distribution centered at the sample mean, with variance equal to the sample variance estimate. Other methods were considered (e.g. kernel density estimation, as employed in [13] and histogram estimators) but both methods have been shown to poorly model the tail behavior, leading to unwarranted extreme values for the estimated SLR, denoted by $\widehat{\text{SLR}}$, both when H_p is true and when H_d is true. The true distributions of scores appear to have light left tails and heavy right tails. Thus, the normal approximation seems the choice of least harm, as it tends to arrive at conservative[38] estimates for $\widehat{\text{SLR}}$. Again, it is important to emphasize that the procedure which follows does not depend on the selection of this particular estimation technique for the probability distribution of the scores.

[38] Conservative in the sense that it protects against Type I errors (errs on the side of innocence) as the estimated SLRs tend to be smaller than the true SLRs.

3.2.4 Computing $\widehat{\text{SLR}}$

The evidence score, δ, is obtained by comparing the actual QD (specifically the observed matrix of counts denoted by $\boldsymbol{y}_U$), with the suspect's template (specifically the observed matrix of counts denoted by $\boldsymbol{x}_S$), using the modified Kullback-Leibler divergence as detailed in Appendix B, $\Delta(\boldsymbol{x}_S, \boldsymbol{y}_U) = \delta$. The final step is to evaluate the estimated distributions at that score: $\hat{g}(\delta|H_p, I)$ and the correct corresponding denominator, $\hat{g}(\delta|x, H_p, I)$, $\hat{g}(\delta|y, H_p, I)$, or $\hat{g}(\delta|H_p, I)$, and then taking their ratio to obtain the estimated score-based likelihood ratio, $\widehat{\text{SLR}}$. The next section illustrates that, as expected from the results shown in Appendix A, very different results are obtained for each method.

4. Comparative Study

In summary, three methods have been presented for obtaining denominator databases used to estimate the SLR: trace-anchored, source-anchored, and general match. These three databases will necessarily result in three different estimates of SLR, denoted[39] by SLR_1, SLR_2, and SLR_3. It seems prudent to investigate whether or not, given the exact same evidence, the three estimates would differ substantially. To that end, a comparative study was performed.

4.1 Writing Samples

The set of writing samples used in the comparative study are those described in detail in [25], collected by the FBI Laboratory over a two-year period. Samples were collected from about 500 different writers. Each writer was asked to provide 10 samples (5 in print and 5 in cursive) of a modified "London Letter" [27] paragraph (533 characters long). In this study, only writing samples in which the writer submitted all five cursive paragraphs were included. This restriction results in 424 writers for a total of 2,120 London Letter paragraph writing samples.

4.2 Simulation Design

We performed the following simulation:

[39] The 'hat' notation is suppressed for ease of presentation; however the reader should be mindful that these are estimates of the true values of SLR_1, SLR_2, and SLR_3.

1. Randomly select two of the 424 writers, denoted by w_1 and w_2. Define E_A to be the remaining 422 writers in the database.
2. Obtain SLR_1, SLR_2, and SLR_3 for two scenarios.

   H_p True: The suspect is the culprit[40] (w_1= suspect = culprit). One of the five paragraphs written by w_1 is randomly selected, from which a string of size n_U is randomly extracted to serve as QD. We varied n_U to be 20, 40, 60, 80, 100, and 150. The number of scores, N, generated to estimate the numerator distribution was set to 500.

   H_dTrue: The suspect is not the culprit (w_1= suspect, w_2 = culprit). QD is obtained in the same manner as the first scenario, except taken from w_2's template rather than w_1's. The number of scores, N, generated to estimate the denominator distribution for SLR_2 and SLR_3 was set to 500.

Repeat steps 1 & 2, 200 times, a computationally feasible number of repetitions.

5. Results and Discussion

The estimates obtained for the three methods were highly variable. To illustrate, for one iteration of the above simulation where H_p is true, values obtained were SLR_1 = 1858, SLR_2 = 1701, SLR_3 = 15. Another iteration resulted in the values SLR_1 = 2370, SLR_2 = 6, SLR_3 = 19.

This trend continues over many runs, which are summarized for the H_p true scenario in Table 1. To facilitate the discussion, we arbitrarily assigned a cutoff so that any SLR estimate greater than 100 leads to the conclusion "supports H_p"[41]. Similarly for any SLR estimate less than 1/100, we conclude "supports H_d". Finally, for any intermediate values, no conclusion is reached. Results are presented in Table 1. For a QD with 80 characters, we observed a high rate (0.43) of disagreement among the three methods. That is, 43% of the time at least one of the three methods disagreed with the others as to whether or not the evidence supports H_p, supports H_d, or is inconclusive.

[40] Throughout, c*ulprit* refers to the individual who actually wrote the QD.
[41] The authors are not implying that this is, in any way, a *meaningful* cutoff.

	Agreement			Disagreement
	Supports H_p	*Inconclusive*	*Supports* H_d	
n_U	SLR > 100	1/100 < SLR ≤ 100	SLR ≤ 1/100	
20	0.000	0.830	0.005	0.165
40	0.070	0.610	0.005	0.315
60	0.125	0.395	0.000	0.480
80	0.200	0.370	0.000	0.430
100	0.240	0.305	0.000	0.455
150	0.330	0.215	0.000	0.455
	Correct			

Table 1. Rates of agreement and disagreement for the three SLR estimates when H_p is true. To disagree, at least one of the three reached a different conclusion. Rates sum to 1 across each row of the Table.

Disagreement rates generally increase as the QD gets larger, an indication that most of the agreement that does occur for smaller QDs is due to the majority of values falling in the inconclusive range. More agreement occurs when H_d is true as seen in Table 2, although there is still some disagreement (3% for n_U = 80). From the results in Table 1, it is clear the methods are differing substantially in terms of the conclusions one would draw in cases where H_p is true, and a more detailed analysis of the results is warranted.

	Agreement			Disagreement
	Supports H_p	*Inconclusive*	*Supports* H_d	
n_U	SLR > 100	1/100 < SLR ≤ 100	SLR ≤ 1/100	
20	0.000	0.760	0.200	0.040
40	0.000	0.515	0.450	0.035
60	0.000	0.395	0.600	0.005
80	0.000	0.240	0.750	0.010
100	0.000	0.205	0.765	0.030
150	0.000	0.120	0.865	0.015
			Correct	

Table 2. Rates of agreement for the three SLR estimates when H_d is true. To disagree, at least one of the three reached a different conclusion. Rates sum to 1 across each row of the Table.

Tippet plots (following the conventions described in [28] are shown for all three SLRs, on the natural log scale, in Figure 2. The three methods can be compared by the measurement of two "error rates"[42] as described in [12]: RMEP ≡ the rate of misleading evidence in favor of the prosecution, i.e., when H_p is true ($|ln(\text{SLR})| < 0$) and RMED ≡ the rate of misleading evidence in favor of the defense, i.e., when H_d is true ($|ln(\text{SLR})| > 0$).

[42] One common critique of likelihood methods is that there is no "error rate" one can report for a given case, as is required by the *Daubert* standard. This is due to the fact that source attribution is not typically reported when LRs are employed. However, in a simulated setting overall error rates can be computed by selecting interval values between which match (or no match) statements might be made.

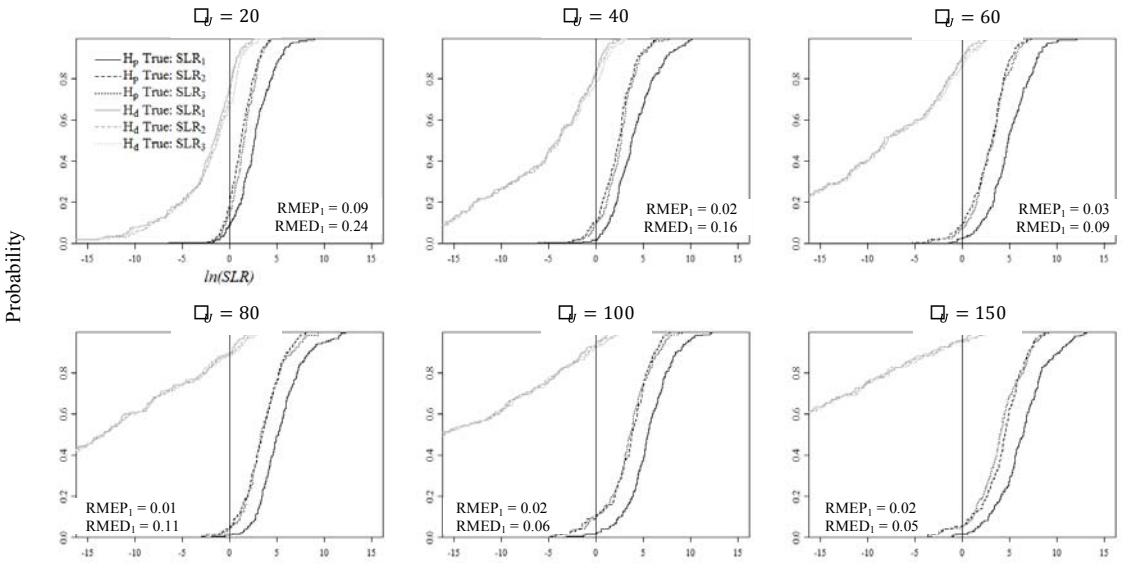


Figure 2. Tippet Plots for three SLR approaches, under two scenarios: $\square_p$ true (black lines) or $\square_d$ true (grey lines). Rates of misleading evidence are reported for SLR_1, the method exhibiting the smallest rates.

Before proceeding, the reader is reminded that samples were obtained by convenience, all consisting of the exact same cursive text, and under particularly mundane circumstances. These facts most certainly prohibit generalization of results. In addition, the reader must be mindful that selection of a different score or a different distribution estimation technique may lead to very different performances of the three methods. The authors are currently investigating the robustness of the three approaches to alternate scoring and estimation methods.

For each scenario considered, the rates of misleading evidence for SLR_1 were far lower than the other two methods. A full listing of the error rates appears in Table 3. Rates for SLR_2 and SLR_3 are nearly indistinguishable. As expected, the rates decrease as the size of QD increases.

	RMEP			RMED		
n_U	SLR_1	SLR_2	SLR_3	SLR_1	SLR_2	SLR_3
20	0.090	0.200	0.150	0.240	0.290	0.330
40	0.015	0.105	0.095	0.160	0.180	0.220
60	0.025	0.095	0.075	0.090	0.105	0.130
80	0.010	0.045	0.045	0.105	0.115	0.125
100	0.015	0.095	0.105	0.055	0.070	0.080
150	0.015	0.055	0.055	0.045	0.050	0.045

Table 3. Rates of misleading evidence in favor of the prosecution (RMEP) and in favor of the defense (RMED).

The reporting of this type of error rate is less than ideal, as the possibility of an inconclusive determination is fully ignored. An approach that is more representative of the realities of forensic casework is to impose symmetric cutoffs (e.g., η and $-\eta$ on the natural log scale) so that three intervals are created (e.g. $(-\infty,-\eta]$, $(-\eta, \eta)$, and $[\eta, \infty)$), corresponding to the three common conclusions: exclusion, inconclusive, and source attribution (or match). For a QD with 80 characters, these rates for all three methods and both scenarios are presented in Table 4, for $\eta = 4.61$ (corresponding to SLR ≈ 100).

	H_p true			H_d true		
	Exclusion	*Inconclusive*	*Match*	*Exclusion*	*Inconclusive*	*Match*
	$(-\infty,-4.61]$	$(-4.61, 4.61)$	$[4.61, \infty)$	$(-\infty,-4.61]$	$(-4.61, 4.61)$	$[4.61, \infty)$
SLR_1	0.000	0.415	0.585	0.760	0.240	0
SLR_2	0.000	0.710	0.290	0.750	0.250	0
SLR_3	0.000	0.715	0.285	0.750	0.250	0
			Correct	***Correct***		

Table 4. Rates of exclusion, inconclusive, and match conclusions.

The results in Table 4 illustrate that additional information is gained from looking at all three intervals, compared to simply reporting RMEP and RMED. The results show that when H_p is true, both SLR_2 and SLR_3 tend toward the inconclusive range far more often than SLR_1 [43].

6. Conclusions

Several methods for obtaining a score-based likelihood ratio for handwriting evidence were illustrated, based on a categorical representation of the feature data produced by the proprietary quantification method developed by Gannon Technologies Group. Regardless of the method selected, the results from Table 4 indicate extremely low false match and false exclusion rates are attained when a moderate conclusion threshold is set ($|ln(\text{SLR})| \leq 4.61$). Since the categorical representation is an extreme simplification of the entire set of feature data generated by Gannon's quantification method (which includes more detailed information, e.g. segment lengths, angles, etc.), it may be that incorporating this additional information would lead to improved performance. However, preliminary investigations indicate that generating a score that makes use of the full set of high-dimensional data and is also highly discriminating is an elusive task (results not shown). While we feel that these types of quantitative analyses may prove fruitful for document examiners at some point, they should only be employed after careful consideration of the inherently subjective decisions the statistical analyst must make in order to calculate such quantities.

Indeed, the primary purpose of this work is to highlight to the forensic community at large, through an empirical study, that score-based likelihood ratios are not the same as, and cannot be interpreted as, the likelihood ratio. Although one should also note that the comparison of equations (2) and (3) suggest that there is a more basic conflict between the two approaches for calculating the "value" of the evidence. This point has been largely ignored in existing literature. Their interpretations must differ as SLRs are considerably more subjective than LRs, in that an analyst must select and defend 1) the similarity (or

[43] This trend can also be gleaned from careful consideration of the Tippet plots in Figure 1. The authors are simply cautioning against reporting RMEP and RMED as the "error rate" for any likelihood ratio method and illustrating a more meaningful alternative.

dissimilarity) score, 2) the appropriate interpretation of the denominator, and 3) the technique relied upon to estimate the numerator and denominator distributions. Due to these points of subjectivity, SLR values must be interpreted with far more caution than the LR based on a well-defined and known probability model (e.g., simple one-contributor DNA LRs)[44].

Some conclusions could be drawn from the various results presented above as to the best SLR technique to use; however, the authors resist as varying any of the subjective factors enumerated above may affect the outcome. Also, innovative score-based approaches have appeared in the literature since this work was undertaken that also should receive consideration [18]. Due to the nature of density estimation, the performance of all methods will heavily depend on the size and representativeness of the database E_A. To date, no such handwriting database exists. The samples used here are not representative of the general population and the simulated evidence documents are not typical of QDs and templates that might be obtained in real casework. As mentioned earlier, our intention is to simply illustrate the feasibility of obtaining an SLR for handwriting evidence, and to emphasize the ambiguities that arise when calculating this value.

Bibliography

[1] C.G.G. Aitken, D. Stoney, The use of statistics in forensic science, Chichester, England, Ellis Horwood Limited, 1991.

[2] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Communication. 31:2-3 (2000) 193-203.

[3] C. Champod, I.W. Evett, B. Kuchler, Earmarks as evidence: A critical review, Journal of Forensic Sciences. 46:6 (2001) 1275-1284.

[4] D.V. Lindley, A problem in forensic science, Biometrika. 64 (1977) 207-213.

[44] Often when calculating a LR, the probability distribution is unknown and must be estimated. In these cases, this estimation process induces subjectivity, just as when estimating the SLR.

[5] J.M. Curran, The statistical interpretation of forensic glass evidence, International Statistical Review. 71:3 (2003) 497-520.

[6] C. Champod, I.W. Evett, A probabilistic approach to fingerprint evidence, Journal of Forensic Identification. 51:2 (2001) 101-122.

[7] I.W. Evett, J.A. Lambert, J.S. Buckleton, A Bayesian approach to interpreting footwear marks in forensic casework, Science & Justice. 38:4 (1998) 241-247.

[8] I.W. Evett, B.S. Weir, Interpreting DNA Evidence, Sunderland, MA, Sinauer, 1998.

[9] S. Bozza, F. Taroni, R. Marquis, M. Schmittbuhl, Probabilistic evaluation of handwriting evidence: Likelihood ratio for authorship, Applied Statistics. 57:3 (2008) 329-341.

[10] A. Nordgaard, T. Höglund, Assessment of approximate likelihood ratios from continuous distributions: A case study of digital camera identification, Journal of Forensic Science. 56 (2011) 390-402.

[11] M.A. Walch, D.T. Gantz, J.J. Miller, L.J. Davis, C.P.Saunders, M.L. Lancaster, A.C. Lamas, J. Buscaglia, Evaluation of the individuality of handwriting using FLASH ID – A totally automated, language independent system for handwriting identification, in: Proc. of the 2008 AAFS Annual Meeting, Washington, DC, 2008.

[12] C.Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of three minutiae, Journal of Forensic Science. 51 (2006) 1255-1266.

[13] L.J. Davis, C.P. Saunders, A.B. Hepler, J. Buscaglia, Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios, Forensic Science International (2011), doi: 10.1016/j.forsciint.2011.09.013.

[14] D. Meuwly, Forensic individualisation from biometric data, Science & Justice. 46 (2006) 205-213.

[15] J. Gonzalez-Rodriguez, D. Ramos, Forensic automatic speaker classification in the "Coming Paradigm Shift", in: C. Müller (Ed.), Speaker Classification I, Springer Berlin / Heidelberg, 2007, vol. 4343, pp. 205-217.

[16] N. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems—Modelling within finger variability, Forensic Science International. 167 (2007) 189-195.

[17] C. Neumann, P. Margot, New perspectives in the use of ink evidence in forensic science Part III: Operational applications and evaluation, Forensic Science International. 192 (2009) 29-42.

[18] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm, Journal of the Royal Statistical Society: Series A. 175:2 (2012) 1-26.

[19] C. Neuman, "New Perspectives in the Use of Ink Evidence in Forensic Science," University of Lausanne, PhD Thesis 2008.

[20] E.J. Garvin, R.D. Koons, Evaluation of match criteria used for the comparison of refractive index of glass fragments, Journal of Forensic Sciences. 56:2 (2011) 491-500.

[21] C.G.G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, Applied Statistics. 53:1 (2004) 109-122.

[22] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., Chichester, UK, John Wiley and Sons, 2004.

[23] C. Champod, I.W. Evett, G. Jackson, Establishing the most appropriate databases for addressing source level propositions, Science & Justice. 44:3 (2004) 156-64.

[24] A. Ross, K. Nandakumar, A. Jain, Handbook of Multibiometrics, New York, NY, Springer, 2006.

[25] C.P. Saunders, L.J. Davis, A.C. Lamas, J.J. Miller, D.T. Gantz, Construction and evaluation of classifiers for forensic document analysis, Annals of Applied Statistics. 5:1 (2011) 381-399.

[26] S. Kullback, R.A. Leibler, On information and sufficiency, Annals of Mathematical Statistics. 22 (1951) 79-86.

[27] A.S. Osborn, Questioned Documents, Albany, NY, Boyd Printing Company, 1929.

[28] P. Gill, J. Curran, C. Neumann, Interpretation of complex DNA profiles using Tippett plots, Forensic Science International: Genetics Supplement Series. 1 (2008) 646-648.

Appendix A: Score-based LRs with known distributions

In this segment, we intend to illustrate the theoretical differences between the three SLRs and the LR by way of a simple illustration. Suppose we have two items of evidence: x, a sample of known source (e.g., suspect's writing template, crime scene window) and y, a sample of unknown source (e.g., bank robbery note, glass fragment obtained from the suspect). Suppose it is known, as a general rule, that samples of this type follow a normal distribution with some mean parameter. Assume the variance parameter representing the within source variability for samples of this type, denoted by σ_w^2, is fixed and known. Also, assume the variance parameter for representing the between source variability for samples of this type, denoted by σ_b^2, is fixed and known.

In this example, we consider the sample x to be one observation from a random process. Let X denote the random variable associated with samples of this type, arising from this specific known source (e.g., writing samples obtained from the suspect, fragments obtained from the crime scene windows). For this illustration, suppose X follows a normal (Gaussian) distribution with mean μ_X, denoted $X \sim N(\mu_X, \sigma_w^2)$.

We also consider the sample y to be one observation from a random process. Let Y denote the random variable associated with samples of this type, arising from this specific unknown source (e.g., writing samples the culprit could have left at the scene, fragments from a specific, but unknown, window found on the suspect). Suppose $Y \sim N(\mu_Y, \sigma_w^2)$.

One final distribution must be defined, that of samples of this type taken from some broader, 'relevant' population denoted by A. For this illustration, suppose these arise from a normal distribution: $N(\mu_A, \sigma_A^2)$ where $\sigma_A^2 = \sigma_b^2 + \sigma_w^2$.

Suppose we are interested in evaluating the evidence in relation to the following two hypotheses:

H_p: x and y arise from the same source

H_d: x and y arise from different sources

Likelihood Ratio

The likelihood ratio, assuming x and y are continuous measurements, is defined by

$$\text{LR} \equiv \frac{f(x, y|H_p)}{f(x, y|H_d)},$$

where f denotes the joint probability density function for the random variables X and Y. The assumptions above imply this will be a bivariate normal density. Thus, in this scenario, we can obtain a closed-form solution for the likelihood ratio.

Numerator

Under the numerator hypothesis H_p, the source of x and y are the same (e.g., the suspect wrote the bank robbery note, the fragment found on the suspect is from the crime scene window).

Thus x and y are random (independent) draws from the same distribution, so that $\mu_Y = \mu_X$. Therefore,

$$X \sim N(\mu_X, \sigma_w^2)$$
$$Y \sim N(\mu_X, \sigma_w^2)$$

Noting that the joint density for two independent normal random variables is simply the product of their respective densities, we have

$$f(x, y|H_p) = \frac{1}{\sigma_w^2} \phi\left(\frac{x - \mu_X}{\sigma_w}\right) \phi\left(\frac{y - \mu_X}{\sigma_w}\right),$$

where ϕ denotes the standard normal probability density function.

Denominator

Under the denominator hypothesis H_d, the source of x and y are different (e.g., someone else wrote the bank robbery note, the fragment found on the suspect is from another window). A common assumption made in the forensic literature is that the source of y is a random individual selected from some relevant population, so that $\mu_Y = \mu_A$. Therefore, $Y \sim N(\mu_A, \sigma_A^2)$.

Typically X and Y are assumed to be independent – that is, information about the known source provides no additional information about the unknown source. Therefore, the joint density is again the product of their respective densities,

$$f(x, y|H_d) = \frac{1}{\sigma_A \sigma_w} \phi\left(\frac{x - \mu_X}{\sigma_w}\right) \phi\left(\frac{y - \mu_A}{\sigma_A}\right).$$

Taking the ratio of $f(x, y|H_p)$ and $f(x, y|H_d)$, and noting the second term of each cancel, we find

$$LR = \frac{\sigma_A \phi\left(\frac{y - \mu_X}{\sigma_w}\right)}{\sigma_w \phi\left(\frac{y - \mu_A}{\sigma_A}\right)}.$$ [45]

Score-Based Likelihood Ratios

We now would like to compare the behavior of this likelihood ratio with that of the three SLRs in the ideal case, where we have databases that were of sufficiently large as to completely characterize the relevant probability distributions. Before defining a (dissimilarity) score we first note desired properties:

- If x and y are measurements from the same source, we expect the score to be close to zero.
- If x and y are measurements from different sources, we expect the score to be large.

One reasonable such score for two normal random variables, X and Y, is the square of their differences. Thus define the random variable $\Delta(X, Y) = (X - Y)^2$. Another added advantage of this particular score is that we can exploit the following relationship between squared normal distributions and a chi-squared (χ^2) distribution to obtain exact expressions each SLR.

[45] This is a true likelihood ratio when the nuisance parameters are known under each of the competing propositions. See Chapter 6 of *Asymptotic Statistics* by A. W. van der Vaart (2000, Cambridge University Press) for details.

Property 1. Squared Normal Distributions
If $T \sim N(\mu, \sigma^2)$, then

$$\frac{T^2}{\sigma^2} \sim \chi^2_{1,\lambda}$$

where $\chi^2_{1,\lambda}$ denotes a non-central chi-squared distribution with one degree of freedom and non-centrality parameter $\lambda = \frac{\mu^2}{\sigma^2}$. It is also true that for any random variable R with probability density function (pdf) f_R and scalar $c > 0$, the random variable $S = cR$ has pdf $f_S = \frac{1}{c} f_R\left(\frac{s}{c}\right)$. Therefore

$$f_{T^2}(t) = \frac{1}{\sigma^2} \chi^2_{1,\lambda}\left(\frac{t}{\sigma^2}\right),$$

with non-centrality parameter $\lambda = \frac{\mu^2}{\sigma^2}$.

To evaluate the evidence, now reduced to $\Delta(x, y) = (x - y)^2 = \delta$, via likelihood ratio, in light of the two hypotheses defined above, H_p and H_d, we are interested in

$$SLR \equiv \frac{g(\delta | H_p)}{g(\delta | H_d)},$$

where g denotes the probability density function for the random variable $\Delta(X, Y) = (X - Y)^2$.

Numerator

All three SLRs make the exact same assumption regarding the numerator probability distribution, namely Y represents an additional independent draw from the distribution associated with the known source. Thus to evaluate the numerator, we need to derive the distribution of $(X - Y)^2$ conditional on $X = x$, where Y follows a $N(\mu_X, \sigma_w^2)$ distribution. For the difference we find:

$$[(X - Y) | X = x] \sim N(x - \mu_X, \sigma_w^2).$$

Per Property 1, the numerator is then

$$g_{\Delta|X}(\delta|x, H_p) = \frac{1}{\sigma_w^2}\chi^2_{1,\lambda}\left(\frac{\delta}{\sigma_w^2}\right),$$

where $\lambda = \frac{(x-\mu_X)^2}{\sigma_w^2}$. The denominator of SLR will vary, depending on which method you choose $(\mathrm{SLR}_1, \mathrm{SLR}_2, \mathrm{SLR}_3)$.

SLR_1 Denominator

The method used to arrive at SLR_1 assumes x is a randomly selected sample from some relevant population. That is,

$$X \sim N(\mu_A, \sigma_A^2).$$

The method of SLR_1 also assumes sample y is fixed and known. Therefore we need to find the distribution $g_{\Delta|Y}(\delta|y, H_d)$.

We find

$$[(X - Y)|Y = y] \sim N(\mu_A - y, \sigma_A^2).$$

Per Property 1, we find the denominator for SLR_1 is

$$g_{\Delta|Y}(\delta|y, H_d) = \frac{1}{\sigma_A^2}\chi^2_{1,\lambda_1}\left(\frac{\delta}{\sigma_A^2}\right),$$

where $\lambda_1 = \frac{(\mu_A - y)^2}{\sigma_A^2}$. Thus,

$$\mathrm{SLR}_1 = \frac{\sigma_A^2\ \chi^2_{1,\lambda}\left(\frac{\delta}{\sigma_w^2}\right)}{\sigma_w^2\ \chi^2_{1,\lambda_1}\left(\frac{\delta}{\sigma_A^2}\right)}.$$

SLR_2 Denominator

The method used to arrive at SLR_2 assumes sample y is a randomly selected sample from relevant population. That is,

$$Y \sim N(\mu_A, \sigma_A^2).$$

The method of SLR_2 also assumes sample x is fixed and known. Therefore we need to find the distribution $g_{\Delta|X}(\delta|x, H_d)$.

We find,

$$[(X - Y)|X = x] \sim N(x - \mu_A, \sigma_A^2).$$

Per Property 1, we find the denominator for SLR_2 is

$$g_{\Delta|X}(\delta|x, H_d) = \frac{1}{\sigma_A^2} \chi^2_{1,\lambda_2}\left(\frac{\delta}{\sigma_A^2}\right).$$

where $\lambda_{2=} \frac{(x-\mu_A)^2}{\sigma_A^2}$. Thus,

$$SLR_2 = \frac{\sigma_A^2 \chi^2_{1,\lambda}\left(\frac{\delta}{\sigma_w^2}\right)}{\sigma_w^2 \chi^2_{1,\lambda_2}\left(\frac{\delta}{\sigma_A^2}\right)}.$$

$\mathbf{SLR_3}$ Denominator

Here, we neither condition on x or y, and assume that x and y are independent draws from the distribution associated with the relevant population. Thus, both X and Y follow $N(\mu_A, \sigma_A^2)$, with X and Y independent. Therefore, their differences are distributed as:

$$X - Y \sim N(0, 2\sigma_A^2).$$

Per Property 1, we find the denominator for SLR_3 is

$$g_{\Delta}(\delta|H_d) = \frac{1}{2\sigma_A^2} \chi_1^2\left(\frac{\delta}{2\sigma_A^2}\right).$$

where χ_1^2 denotes the *central* chi-squared distribution ($\lambda_3 = 0$). Therefore,

$$\text{SLR}_3 = \frac{2\sigma_A^2 \chi^2_{1,\lambda}\left(\frac{\delta}{\sigma_w^2}\right)}{\sigma_w^2 \chi_1^2\left(\frac{\delta}{2\sigma_A^2}\right)}.$$

It is very important to note that each of the SLRs have a *different* functional form. While here we are making many simplistic and unrealistic assumptions, it stands to reason that the different methods will *necessarily* provide different answers, providing some insight into the results found in this work. SLR_1 and SLR_2 differ only in their non-centrality parameters in the denominator.

We have laid out a framework where we can easily compare the three SLRs to the LR, a luxury that is not possible in most realistic applications. To help the reader comprehend the differences among the SLRs themselves, and to highlight the deviations of each from the LR (in this contrived example), a graphical illustration is provided below.

Comparison

Rather than inspect the rather complex functional forms of each ratio, we have deferred to illustrating their differences graphically. In Figure A1, we have plotted the values of SLR_1, SLR_2, SLR_3 and LR given by the formulas above, for various values of σ_b^2 and σ_w^2, and for different μ_A and μ_Y. The x-axis represents a range of possible measurements on sample y. For clarity, we have eliminated one source of variability by making the (unrealistic) assumption x always equals μ_X (i.e., the measurement taken from the known source is always equal to the true mean of its distribution).

Consider the first plot appearing in Figure A1. Here, the mean of the distribution from which the known source sample arises is 0 (i.e., $\mu_X = 0$). The mean of the distribution from the relevant population is -8 (i.e. $\mu_A = -8$). The black line represents the likelihood ratio. As expected, the likelihood ratio takes on positive values as the measurement from the unknown sample (y) approaches the mean of the known source. It continues to increase as the value of x increases, up until it becomes less and less likely to have come from the known source.

Moving on to the SLRs, it is important to note that the functional form of SLR_1 is changing along with y, as the non-centrality parameter in the denominator changes with y. This is stated here to emphasize that we are not looking at the functional form of SLR_1, just the evaluation of SLR_1 at each point y. Each of the SLRs peak at $y = 0$, which is a marked deviation from the LR. There is one segment ($y < -5$) where both SLR_1 and SLR_3 are in fact larger than the LR implying that under these conditions, those SLRs are overstating the value of the evidence in favor of the prosecution (though all values are extremely small, thus providing very strong support the defense hypothesis). However, in most other places where $\log(LR) > 0$ ($LR > 1$) the SLRs are understating the value of the evidence, in some cases drastically so. For example, when $y = 2.5$, we find $LR = 3.05 \text{ x } 10^{13}$ ($\log(LR) = 13.48$) whereas $SLR_1 = 1.23 \text{ x } 10^{7}$, $SLR_2 = 1.60 \text{ x } 10^{2}$, and $SLR_3 = 1.34 \text{ x } 10^{-2}$. This shows that, at least in this contrived situation, some amount of evidential value is not being adequately captured by these three methods. It is interesting to note that SLR_2 closely approximates (although slightly overestimating) the LR when $y < 0$, and this property is evident in each graph.

The properties displayed in the first graph are certainly the most extreme. In most cases, the SLRs are fairly well behaved in comparison to the LR, particularly so when the between variability is much larger than the within variability (looking down the rows in Figure A1). The approximations are also well behaved when the alternate population mean approaches the mean of the known source (looking across the columns in Figure A1). In general, SLR_3tends to underestimate the value of the evidence, very rarely producing log values greater than zero. Several additional interesting features can be observed in these plots, and rather than enumerate them here, the reader is encouraged to study them closely.

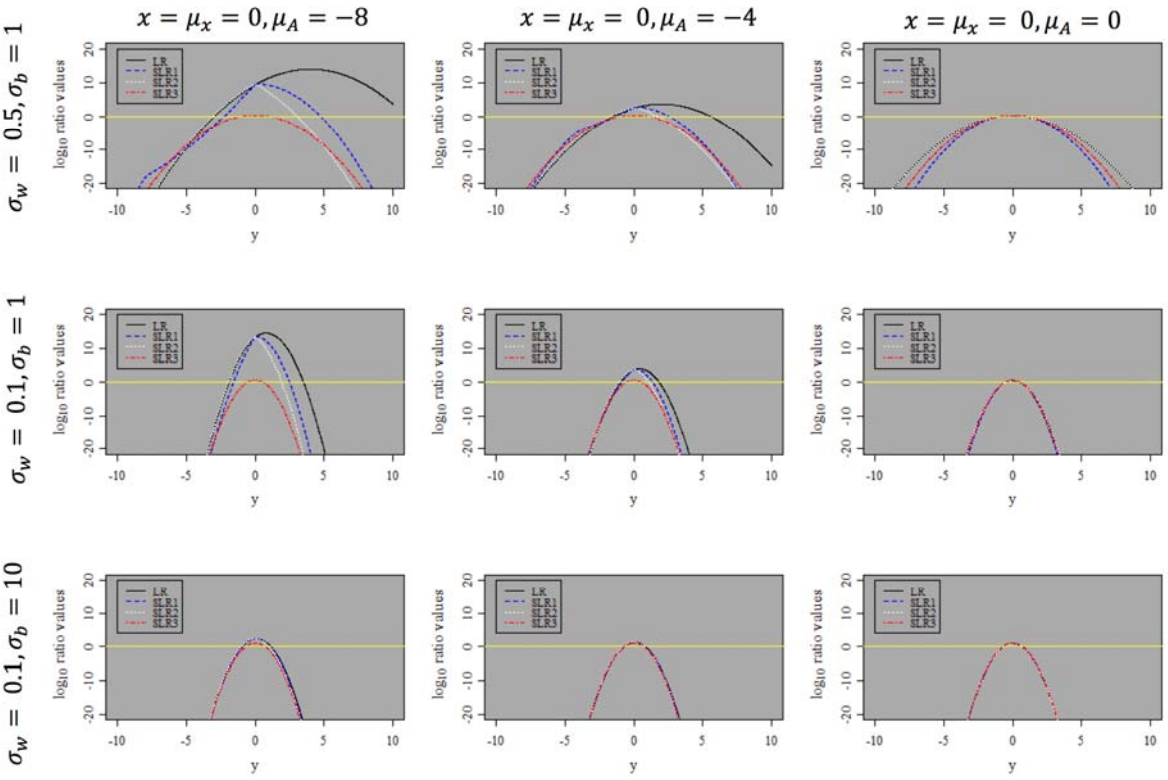


Figure A1

Appendix B: Dissimilarity Score

In this segment, we describe in detail the dissimilarity score used in this study. Suppose we have observed matrices of counts for two writing samples, denoted by $\boldsymbol{x}$ and $\boldsymbol{y}$. For a given letter l (or a given row of $\boldsymbol{x}$ and $\boldsymbol{y}$), define

$$\nu_{li} = \frac{x_{li} + \frac{1}{I_l}}{x_{l\cdot} + 1} \quad \text{and} \quad \tau_{li} = \frac{y_{li} + \frac{1}{I_l}}{y_{l\cdot} + 1},$$

where $x_{l\cdot} = \sum_{i=1}^{I_l} x_{li}$, $y_{l\cdot} = \sum_{i=1}^{I_l} y_{li}$, and $i = 1, \ldots I_l$ indexes the distinct isocodes used to write the l^{th} letter in either $\boldsymbol{x}$ or $\boldsymbol{y}$. Then the dissimilarity score for a given letter l is defined as

$$\Delta(\boldsymbol{x}_l, \boldsymbol{y}_l) \equiv \sum_{i=1}^{I_l} \tau_{li} \ln\left(\frac{\tau_{li}}{\nu_{li}}\right),$$

except when $I_l = 1$ (i.e., when only one isocode is used to write letter l in either $\boldsymbol{x}$ or $\boldsymbol{y}$), in which case $\Delta(\boldsymbol{x}_l, \boldsymbol{y}_l) \equiv 0$.

To combine across all letters, $l = 1, \ldots L$, define a set of weights,

$$\lambda_l \propto \begin{cases} \dfrac{1}{\sqrt{\dfrac{1}{x_{l\cdot}}} + \sqrt{\dfrac{1}{y_{l\cdot}}}}, & \min\left(x_{l\cdot}, y_{l\cdot}\right) \geq 1 \\ 0, & \text{otherwise}, \end{cases}$$

such that $\sum_{i=1}^{L} \lambda_l = 1$. Thus, a letter only receives weight when it appears at least once in both $\boldsymbol{x}$ and $\boldsymbol{y}$. The combined score over all letters is then

$$\Delta(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{L} \lambda_l \Delta(\boldsymbol{x}_l, \boldsymbol{y}_l).$$

Phase III

Identification of Specific Source (Continued)

Our second direction concerns the formal development of the value of evidence when the forensic scientist has to estimate the background population defined by the defense proposition or model (the focus of Phase III). This line of work has been more narrowly focused on formal Bayesian methods.

In February and March of 2014, Dr Saunders is giving two talks, one invited presentation at Pittcon and another at the Annual Meeting of the American Academy of Forensic Sciences on Statistical Aspects of the Forensic Identification of Source Problems. These talks are presentations of the results of Phase III of this research grant. This research describes how to incorporate incomplete information about the background population into a forensic likelihood ratio in a statistically rigorous manner. We provide an overview of these results in the following materials based, in part, on a poster presentation at the Joint Statistical Meetings in 2012. Dr. Saunders has been continuing this research activity through the end of the grant.

Investigation into Formal Bayesian Methods for incorporating Uncertainty about the Background Population

The Effect of Uncertainty About the Alternative Source Population on the Assessment of the Value of Forensic Evidence[46]

Christopher P. Saunders, PhD

Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007

A goal in the forensic interpretation of scientific evidence is to make an inference about the source of a trace of unknown origin; the inference usually concerns two propositions. The first proposition is usually referred to as the prosecution hypothesis and states that a given specific source is the actual source of the trace of unknown origin. The second, usually referred to as the defense hypothesis, states that the actual source of the trace of unknown origin is randomly selected from a relevant alternative source population. The evidence a forensic scientist is given for deciding between these two propositions is: (a) the trace of unknown origin, (b) a sample from the specific source specified by the prosecution hypothesis, and (c) a collection of samples from the alternative source population. One common approach is to assume that the alternative source population is completely known and rely on a Bayes Factor for deciding between the competing hypotheses. In this presentation we will relax this assumption and explore some of the resulting issues from the estimation of the alternative source population. We will illustrate the resulting effects on the calculation of the Bayes Factors with a well-studied collection of samples relating to glass fragments described above.

Background and Conventions

Let $E = \{E_s, E_u, E_a\}$ be a random element that represents the evidence available in a specific case for distinguishing between the defense proposition and the prosecution proposition; where E_s is the evidence about a specific source, E_u is the evidence from an unknown source, and E_a is the evidence from possible alternative sources. We assume that E_s, E_u, and E_a are three independent samples drawn in the following way:

[46] This section is based, in part, on a poster presentation at the Joint Statistical Meetings in 2012. Dr. Saunders is continuing this research activity through the end of the grant; he will deliver an invited talk at Pittcon in 2014 and a presentation at the 2014 AAFS Annual Meeting.

1. E_s is a simple random sample from a given specific source determined by H_p. Let θ_{s_0} denote the fixed parameters necessary to describe this sampling induced distribution.
2. E_a is constructed by first taking a simple random sample of sources from a given relevant population of possible sources; then from each sampled source we have a simple random sample. This collection of samples is E_a. Let θ_{a_0} denote the fixed parameters necessary to describe this sampling induced distribution.
3. E_u is a simple random sample from a single source. It is unknown whether the source of E_u is the specific source determined by H_p or if the source of E_u is randomly selected from the given relevant population of the possible sources in (2). The sampling distribution of E_u is characterized by either the parameters θ_{s_0} or θ_{a_0}.

We will follow these conventions for distinguishing between sampling-induced probability and probability used as a measure of belief:

1. Latin letters denote sampling induced probability measures; for example, $f(e_s \mid \theta_{s_0})$ denotes the likelihood of observing the realized value of the sample from the specific source given the actual value of the specific source distribution parameters.
2. Greek letters denote a probability measure that is a measure of belief; for example, $\pi(\theta_s \mid e_s)$ denotes the posterior density of $\theta_s \mid e_s$, which describes our belief about the value of θ_{s_0} after observing a sample $E_s \mid \theta_{s_0}$.
3. When combining a belief with a sampling induced probability through Bayes theorem, we end up with another belief that is informed or updated by the observed sample. We denote the resulting distribution with a π.

In this setting, the stochastic nature of the evidence E is characterized by an unknown but fixed parameter θ_0. However, θ_0 is usually of interest only in so far as knowledge of its value facilitates the quantification of support that E provides for either the prosecution model of define of the evidence. In this sense, having to estimate θ_0 is a nuisance, and hence in the statistical nomenclature these parameters in this situation are known as a '*nuisance parameters*'. To deal with these nuisance parameters in a formal a Bayesian manner, we need to characterize our belief concerning their likely values (and hopefully, eventually update that belief with empirical evidence!). In these situations we have studied, we typically will need two sets of prior beliefs; one summarizing our belief about how the specific source generates evidence (θ_{s_0}) and another summarizing our prior belief about how the alternative source population stochastically generates evidence (θ_{a_0}). Unfortunately, to be statistically rigorous in our application of Bayesian methods to this problem we need to specify these priors before we look at the evidence, including the evidence from the alternative source population.

As an example of (3), say I am interested the "*likelihood of observing e_u, if is from the same source distribution as e_s*". Using a Bayesian method would provide the following answer "*I believe the likelihood of observing e_u, if it is from the same source distribution as e_s is ...* ". This posterior belief is known as the posterior predictive distribution for e_u given e_s and is calculated as follows:

$$\begin{aligned}\pi\left(e_u \middle| e_s\right) &= \frac{\pi(e_u, e_s)}{\pi(e_s)} \\ &= \frac{\int f(e_u, e_s \mid \theta_s)\pi(\theta_s)d\theta_s}{\int f(e_s \mid \theta_s)\pi(\theta_s)d\theta_s} \\ &= \frac{\int f(e_u \mid \theta_s) f(e_s \mid \theta_s)\pi(\theta_s)d\theta_s}{\int f(e_s \mid \theta_s)\pi(\theta_s)d\theta_s} \\ &= \int f(e_u \mid \theta_s)\frac{f(e_s \mid \theta_s)\pi(\theta_s)}{\int f(e_s \mid \theta_s)\pi(\theta_s)d\theta_s}d\theta_s \\ &= \int f(e_u \mid \theta_s)\pi(\theta_s \mid e_s)d\theta_s.\end{aligned}$$

Known Alternative Source Population Parameters

In this section we are assuming that we have a well-studied alternative source population with known parameters, i.e. θ_{a_0} is known. The only unknown parameters that are contributing to the uncertainty about the value of the evidence are the ones associated with the specific source, θ_{s_0}.

Let $e = \{e_s, e_u, e_a\}$ represent the realization of the random element E for a specific case at hand. Since θ_{a_0} is known, e_a is irrelevant to the value of the evidence. Let $\theta_0 = \{\theta_{s_0}, \theta_{a_0}\}$ and $\pi(\theta) = \pi(\theta_s)$ be a probability distribution used to describe our prior belief about θ_0.

Following [1], define the value of the evidence as

$$V = \frac{\pi(e \mid H_p, I)}{\pi(e \mid H_d, I)},$$

where

$$\pi\left(e \mid H_p, I\right) = \int f(e_s \mid \theta_s) f(e_u \mid \theta_s) f(e_a \mid \theta_{a_0}) d\pi(\theta)$$
$$= f(e_a \mid \theta_{a_0}) \int f(e_u \mid \theta_s) f(e_s \mid \theta_s) d\pi(\theta_s),$$

and

$$\pi(E \mid H_d, I) = \int f(e_s \mid \theta_s) f(e_u \mid \theta_{a_0}) f(e_a \mid \theta_{a_0}) d\pi(\theta)$$
$$= f(e_u \mid \theta_{a_0}) f(e_a \mid \theta_{a_0}) \int f(e_s \mid \theta_s) d\pi(\theta_s).$$

Now we can rewrite V as

$$V = \frac{\pi(e \mid H_p, I)}{\pi(e \mid H_d, I)}$$
$$= \frac{\int f(e_u \mid \theta_s) f(e_s \mid \theta_s) d\pi(\theta_s)}{f(e_u \mid \theta_{a_0}) \int f(e_s \mid \theta_s) d\pi(\theta_s)}$$
$$= \frac{\pi(e_u \mid e_s, H_p, I)}{f(e_u \mid \theta_{a_0})}.$$

By assuming we know (or we are just certain that we know) the value of θ_{a_0}, the denominator reduces to evaluating the sampling distribution at e_u; in effect the denominator does not contain any belief measures when is θ_{a_0} is known.

Unknown Alternative Source Population Parameters

Let $e = \{e_s, e_u, e_a\}$ represent the realization of the random element E for a specific case at hand. Let $\theta_0 = \{\theta_{s_0}, \theta_{a_0}\}$ and $\pi(\theta) = \pi(\theta_s)\pi(\theta_a)$ be a probability distribution used to describe our prior belief about θ_0. We are choosing to restrict ourselves to priors on θ_{s_0} and θ_{a_0} that are independent of each other.

The value of the evidence is now

$$V = \frac{\pi(e \mid H_p, I)}{\pi(e \mid H_d, I)},$$

where

$$\pi(e \mid H_p, I) = \int f(e_s \mid \theta_s) f(e_u \mid \theta_s) f(e_a \mid \theta_a) d\pi(\theta)$$
$$= \int f(e_a \mid \theta_a) d\pi(\theta_a) \int f(e_u \mid \theta_s) f(e_s \mid \theta_s) d\pi(\theta_s),$$

and

$$\pi(e \mid H_d, I) = \int f(e_s \mid \theta_s) f(e_u \mid \theta_a) f(e_a \mid \theta_a) d\pi(\theta)$$
$$= \int f(e_u \mid \theta_a) f(e_a \mid \theta_a) d\pi(\theta_a) \int f(e_s \mid \theta_s) d\pi(\theta_s).$$

Now we can rewrite V as

$$
\begin{aligned}
V &= \frac{\pi(e \mid H_p, I)}{\pi(e \mid H_d, I)} \\
&= \frac{\int f(e_a \mid \theta_a) d\pi(\theta_a)}{\int f(e_u \mid \theta_a) f(e_a \mid \theta_a) d\pi(\theta_a)} \times \frac{\int f(e_u \mid \theta_s) f(e_s \mid \theta_s) d\pi(\theta_s)}{\int f(e_s \mid \theta_s) d\pi(\theta_s)} \\
&= \frac{\pi(e_u \mid e_s, H_p, I)}{\pi(e_u \mid e_a, H_d, I)}.
\end{aligned}
$$

Glass Data Example:

In [3] a collection of glass fragments is analyzed. This dataset consists of three classes of windows, with 16, 16, and 30 windows in each class. There are 5 glass fragments per window. Following [3], we consider the logarithm of measurements of elemental ratios on each glass fragment: $log(Ca / K)$ (V2), $log(Ca / Si)$ (V3), and $log(Ca / Fe)$ (V4).

As an illustrative example, we consider the first group of windows, letting the 4^{th} window take the role of the hypothesized specific source. The first three fragments from window 4 will serve as e_s and the last two fragments from window four will serve as e_u. A second example will be constructed where two fragments from the 2^{nd} window serve as e_u. The remaining 70 glass fragments divided among the 14 windows will serve as e_a.

Model H_p

We will assume that the glass fragments composing e_s are an $i.i.d.$ sample from a multivariate normal with a mean vector μ_s and covariance Σ_s. Let X_i denote the vector of measurements on the i^{th} fragment for $i = 1,2,3$, then $X_i \sim MNV\left(\mu_s, \Sigma_s\right)$.

We will use a normal prior on μ_s, centered at the zero vector with a diagonal covariance matrix with diagonal elements equal to 10^3 and an inverse gamma prior for σ_s centered at a diagonal covariance matrix with diagonal elements equal to .05, .00005, .0005 and one degree of freedom. The marginal 95% credible intervals for the mean vector of the specific source are $\mu_{s(v2)}$:$(2.7428, 6.0518)$, $\mu_{s(v3)}$:$(-0.3736, -0.2305)$, and $\mu_{s(v3)}$:$(2.3399, 2.9879)$. The numerators of the values evidence are

$$
\begin{array}{lll}
 & Exp1 & Exp2 \\
\pi\left(e_u \middle| e_s\right) = & 62760.38 & 1942.637
\end{array}
$$

Model H_d

We assume that the glass fragments composing e_a follow a hierarchical multivariate normal model with the assumption that all windows in the alternative source population have a common within-covariance matrix, Σ_w, and a between source mean μ and covariance Σ_b.

Let Y_{ij} denote the vector of measurements on the j^{th} fragment, for $i = 1,2,\ldots m_i$ fragment from the i^{th} window, for $i = 1,2,\ldots n$. The hierarchical multivariate model in this case is the same as a simple random effects model: $Y_{ij} = \mu + a_i + w_{ij}$, where a_i are $i.i.d.$ multivariate normal random vectors with a mean vector of zero and a covariance matrix of Σ_b. The w_{ij} are assumed to be $i.i.d.$ multivariate normal vectors with a mean vector of zero and a covariance matrix of Σ_w.

Our prior for Σ_w is the same that is used for Σ_s and the prior for μ is the same as that used for μ_s. We use an inverse gamma prior for $\Sigma_b$centered at the identity covariance matrix with one degree of freedom.

To calculate the denominator likelihood under the assumption that the alternative source population is known, we used the estimates suggested in [3] as plug-in values for the parameters. All posterior predictive distributions are fit using [6]. The denominators of the value of evidence being

	$Exp1$	$Exp2$
$\pi\left(e_u \middle\| e_a\right)$	445.65	592.1348
$f\left(e_u \middle\| \hat{\theta}_{a_0}\right)$	556.35	10931.68

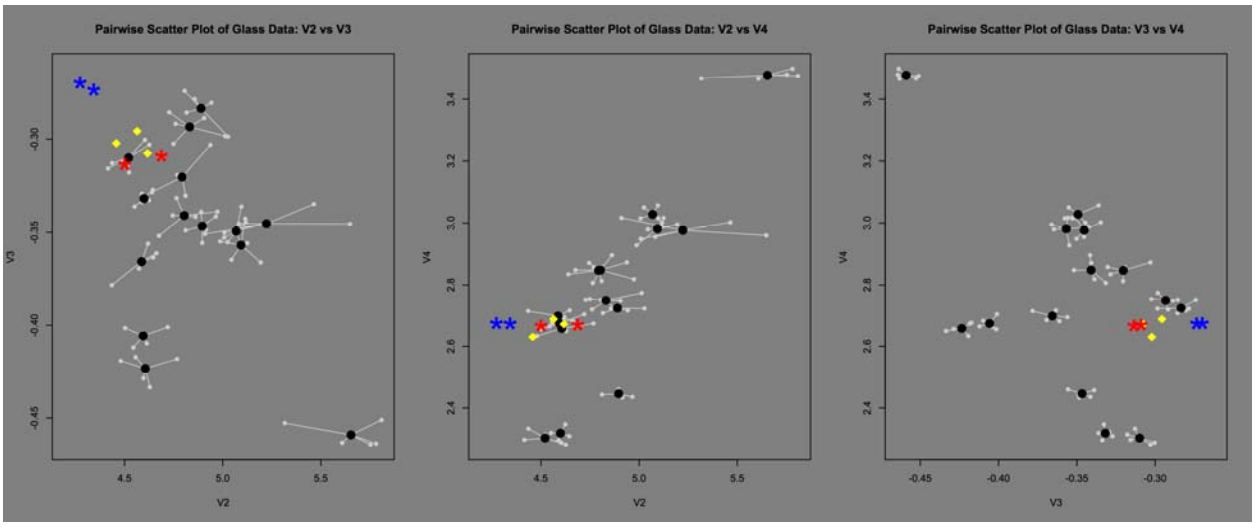


Figure 1: Pairwise Scatter Plots of Glass Data for v2 versus v3, v2 versus v4, and v3 versus v4. The gray points are e_a; the samples from the background population with the source (window) sample means denoted in black. The yellow diamonds are the 3 glass fragments from the specific source, e_s. The blue and red *'s are the glass samples from an unknown source, e_u, under two different conditions. The red *'s are fragments that are actually from the specific source. The blue *'s are fragments for e_u that are from a window that is not the specific source.

Conclusions and Current Research

In the two examples we have worked with a traditional forensic dataset, we have found that there can be a rather dramatic effect in the value of evidence when we incorporate the uncertainty about the background population.

It should be noted that the priors we have used are not the ideal choice in forensic science. The common approach in forensic statistics, dating back to [8] and covered in great detail in [1], for determining a prior for θ_{s_0} is to rely on the ``random man distribution''. The basic idea is that before we observe anything from the specific source, we believe that the specific source is similar to a source that is randomly selected from the alternative source population. Our current research is focused on using this type of prior while incorporating E_a. The difficulty is that we would like to use part of the evidence to suggest the prior for the specific source parameters, which leads to an empirical Bayes approach.

References

[1] Aitken and Taroni (2004), Statistics and the Evaluation of Evidence for Forensic Scientists (2nd ed).

[2] Bozza et al. (2008), "Probabilistic Evaluation of Handwriting Evidence: Likelihood Ratio for Authorship", App. Stats. 57(3),329-341.

[3]Aitken, C. G. G. and Lucy, D, "Evaluation of trace evidence in the form of multivariate data", Journal of the Royal Statistical Society: Series C (Applied Statistics) Vol. 53 (2004) 109-122.

[4] DasGupta (2008), Asymptotic Theory of Statistics and Probability. [5] Gelman et al. (2004), Bayesian Data Analysis (2nd ed).

[6] Hadfield, J.D. (2010) “MCMC methods for Multi-response Generalized Linear Mixed Models: The MCMCglmm R Package". Journal of Statistical Software, Vol. 33, Issue 2, Feb 2010.

[7] F. Taroni, R. Marquis, M. Schmittbuhl, A. Biedermann, A. Thiery, S. Bozza, “The use of the likelihood ratio for evaluative and investigative purposes in comparative forensic handwriting examination”, Forensic Science International 214 (2012) pp. 189-194.

[8] D. V. LINDLEY , “A problem in forensic science”, Biometrika 64(2) (1977): pp. 207-213.

Conclusions and Impact

The research performed during the Phase I, Part A of this project has provided two statistically rigorous methodologies for estimating the Random Match Probabilities of a forensic matching system. As is noted in The National Research Council (2009, p. 122):

> *The assessment of the accuracy of the conclusions from forensic analyses and the estimation of relevant error rates are key components of the mission of forensic science.*

This suggests that information concerning the RMP and the RNMP associated with a comparison procedure contributes to its practical utility in forensic science. In this section we have illustrated one alternative to modeling for investigation of the RMP and the RNMP associated with a comparison procedure applied to comparing writing samples. We have also described an algorithm involving subsampling that facilitates the studying of various properties of the RMP as a function of the size of the documents being compared. All of these algorithms have been stated in terms of a common size of writing samples being compared. However, they can be trivially adapted to scenarios where the sizes of writing samples being compared are not the same for all writing samples. Such an application might arise when studying match probabilities associated with comparing very short notes, such as might be associated with bank robberies, to very large writing samples collected from potential suspects. The algorithms can also be adapted to investigate the dependency of match probabilities on criteria other than sizes of writing samples being compared. For example, the effect of content on match probabilities can be studied by changing from random sampling to stratified or systematic sampling when selecting characters to generate the simulated writing samples. Finally, although the main focus of this section has been on match probabilities, the algorithms have other applications in forensics. These algorithms have facilitated the development of new strategies for the interpretation and presentation of forensic handwriting evidence discussed in Phase II.

In addition to the nonparametric methods described in Part A of Phase I, we have also developed in Phase I, Part B a parametric methodology for estimating the RMP. This approach hinges on the modeling the dependency structure among a set of pairwise comparisons. The parametric methods facilitate the estimation of RMPs when there are no 'matches' between pairwise comparisons; a situation that commonly occurs when there are a small number of samples to compare. It has been shown that there is a closed form for an ANOVA table. It has been shown that there is a method for forming confidence intervals for RMPs, which works well based on the simulation results. It has also been shown that two other methods for making confidence intervals either fail by being too conservative or are just incorrect. We feel that the methods described here could be used by researchers working in the area of studying random match probabilities. We also expect that these models will support the development of new statistical methods for the interpretation and presentation of forensic evidence for which it is only possible to compare pairs of objects.

In Part A of Phase II, we discussed the development of a novel quantification system for fingerprint evidence. In this section, our focus was on the closed set identification problem, where it was quickly discovered that there was need to have a sophisticated statistical algorithm to extract the relevant information. The resulting approach allows for a comparison between

pairs (or sets) of fingerprints. When this quantification is applied to forensic evidence, the resulting scores will naturally require a score based likelihood ratio approach. Extending the research performed in this section will be one of the main focuses of our ongoing research program.

In Part B of Phase II, several methods for obtaining a score-based likelihood ratio for handwriting evidence were illustrated and explored using the subsampling approaches developed in Phase II. All of the methods explored with the available research data set indicate that extremely low false match and false exclusion rates are attained when a moderate conclusion threshold is set ($|\ln(SLR)| \leq 4.61$). While we feel that these types of quantitative analyses may prove fruitful for document examiners at some point, they should only be employed after careful consideration of the inherently subjective decisions the statistical analyst must make in order to calculate such quantities.

The main achievement of this work has been to demonstrate, both empirically and theoretically, that score-based likelihood ratios are not the same as, and cannot be interpreted as, a traditional forensic likelihood ratio. This point has been largely ignored in the existing literature. Their interpretations must differ as SLRs are considerably more subjective than *LR*s, in that an analyst must select and defend 1) the similarity (or dissimilarity) score, 2) the appropriate interpretation of the denominator, and 3) the technique relied upon to estimate the numerator and denominator distributions. Due to these points of subjectivity, SLR values must be interpreted with far more caution than the LR based on a well-defined and known probability model (e.g., simple one-contributor DNA LRs). Our current research focus in this area has been to develop a set of reasonable standard properties that a *SLR* should possess. The preliminary results of this research were discussed at the Joint Statistical Meetings in 2013. This research is ongoing, with a PhD Candidate at South Dakota State University developing the embryonic stages of the research to date. Due to the complexity of modern forensic techniques, we do expect that score-based likelihood ratios will be one of the main tools available for the presentation and interpretation of complex evidence forms.

In final section, Phase III, of this report we have summarized our preliminary work on a statistically rigorous Bayesian approach to the specific source identification problem when the background population is not known with certainty. We have used a traditional data set concerning glass fragments and worked two examples. As expected, we have found that there can be a rather dramatic effect in the value of evidence when we incorporate the uncertainty about the background population.

It should be noted that the priors we have used are not the ideal choice in forensic science. It is common practice to choose the prior for the specific source distribution parameters by thinking of the specific source as typical of the population of alternative sources specified by the defense proposition. The basic idea is that before we observe anything from the specific source, we believe that the specific source is similar to a source that is randomly selected from the alternative source population. The difficulty is that we would like to use part of the evidence to suggest the prior for the specific source parameters, which leads to an empirical Bayes approach. This is of concern because empirical Bayesian methods do not possess the properties normally

associated with formal Bayesian methods. Phase III is active research, with Dr. Saunders presenting a series of talks in the winter and early spring of 2014. After these presentations are complete, the researchers will submit the resulting papers to the appropriate journals. A master's student at South Dakota State University is currently exploring computational issues associated with this research. The resulting work will become her Master's thesis.

DISSEMINATION

Two referred journal articles on the grant research were published to date under this grant.

Davis LJ, Saunders CP, Hepler A, Buscaglia J. Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. Forensic Sci Int. 2012 Mar 10; 216(1-3):146-57.

Hepler AB, Saunders CP, Davis LJ, Buscaglia J. Score-based likelihood ratios for handwriting evidence. Forensic Sci Int. 2012 Jun 10; 219(1-3):129-40.

The grant researchers gave numerous conference presentations under this grant. Many presentations or abstracts are available online at conference web sites. The presentations are listed below followed by full abstracts.

Abstracts of Presentations delivered by the Grant Researchers at:

AAFS 2010 Annual Meeting, February 2010, Seattle, WA
The NIJ 2011 Trace Evidence Symposium, August 2011, Kansas City, MO
The NIJ 2012 Impression and Pattern Evidence Symposium, August 2012, Clearwater, FL
The 2012 Joint Statistical Meetings, July-August 2012, San Diego, CA
EAFS 2012, August 20-24, 2012, The Hague
The Measurement Science and Standards in Forensic Handwriting Analysis (MSSFHA) Conference, June 4 – 5, 2013 at NIST
The 2013 Joint Statistical Meetings, August 2013, Montreal, Canada
Pittcon 2014 (Application: Homeland Security/Forensics, Primary Focus: Methodology), March 2014, Chicago, Il
AAFS 2014 Annual Meeting, February 2014, Seattle, WA,

All Titles & Authors are listed first. Full Abstracts follow the listings.

Abstracts of Presentations delivered at the AAFS 2010 Annual Meeting

Abstract 1: A Comparison Between Different Likelihood Ratios for Assessing Handwriting Evidence

Authors: Amanda Hepler, PhD, George Mason University, Document Forensics Lab, Department of Applied Information Technology, 4400 University Drive, Fairfax, VA 22030; and*

Christopher Saunders, PhD, George Mason University, Document Forensics Lab, 4400 University Drive, Fairfax, VA 22033

Abstract 2: New Results for Addressing the Open Set Problem in Automated Handwriting Identification

Authors: Donald T. Gantz, PhD, John J. Miller, PhD*, and Christopher Saunders, PhD, Intelligence and Security Research Center (MS 1G8), George Mason University, Fairfax, VA 22030; Mark A. Walch, March, MPH, The Gannon Technologies Group, 7600 Colshire Drive, McClean, VA 22102; and JoAnn Buscaglia, PhD, FBI Laboratory, Counterterrorism & Forensic Science Research Unit, Quantico, VA 22135*

Abstracts of Presentations delivered at the NIJ 2011 Trace Evidence Symposium

Abstract 1: ROC Curves for Methods of Evaluating Evidence: A Common Performance Measure Based on Similarity Scores

Authors: R. Bradley Patterson, John Miller, Christopher P. Saunders
Video and Slides Available at http://projects.nfstc.org/trace/2011/agenda.htm

Abstract 2: Predictive modeling for determining the discriminative power of trace glass evidence as a function of the number of sampled glass fragments.

Authors: Eric Kalendra, Christopher P. Saunders
Video and Slides Available at http://projects.nfstc.org/trace/2011/agenda.htm

Abstract 3: On Parametric Models for Pairwise Comparisons with Applications to the Estimation of Random Match Probabilities.

Authors: Donald Gantz, Christopher P. Saunders
Video and Slides Available at http://projects.nfstc.org/trace/2011/agenda.htm

Abstracts of Presentations delivered at the NIJ 2012 Impression and Pattern Evidence Symposium

Abstract 1 (Poster): Automated Statistically Ranked Latent-to-Reference Print Overlays

Authors: Donald T. Gantz, George Mason University, JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division, Mark A. Walch, The Gannon Technologies Group,

Maria Antonia Roberts, Federal Bureau of Investigation Laboratory Division, Daniel T. Gantz, The Gannon Technologies Group

Abstract 2: (Poster) A Note on the Value of Forensic Evidence for Sparse Categorical Tables

Authors: Krista M. Heim, George Mason University, Christopher P. Saunders, South Dakota State University, and JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division.

Abstract 3: The Effect of the Order of Suspect and Background Population Samples on the Assessment of the Value of Evidence

Authors: Eric Kalendra, George Mason University, Christopher P. Saunders, South Dakota State University, JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division.

Abstract 4 in the Workshop: Guidelines for a Successful Research Project: *The story of an Academic/Commercial Partnership developing a product for the Forensic Community*

Panelists & Presenters: Donald Gantz, George Mason University, and Mark Walch, The Gannon

Abstracts of Presentations delivered at the 2012 Joint Statistical Meetings, August 2012

Abstract 1: (Poster) "The effect of uncertainty about the Alternative Source population on the value of Forensic evidence."

Author: Christopher P. Saunders, South Dakota State University

Abstracts of Presentations delivered at EAFS 2012, The Hague, August 20-24, 2012

Abstract 1: The Effect of Uncertainty About the Background Population on the Forensic Value of Evidence

Author: Christopher P. Saunders, South Dakota State University

Abstract 2: Ridge Specific Markers for Latent Fingerprint Identification

Authors: Donald T. Gantz, George Mason University, JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division, Mark A. Walch, The Gannon Technologies Group, Maria Antonia Roberts, Federal Bureau of Investigation Laboratory Division, Daniel T. Gantz, The Gannon Technologies Group

Abstracts of Presentations delivered at The Measurement Science and Standards in Forensic Handwriting Analysis (MSSFHA) Conference, June 4 – 5, 2013 at NIST.

Abstract 1: The Forensic Language-Independent Analysis System for Handwriting Identification (FLASH ID)

Authors: Mark Walch, Gannon Technologies Group and Donald Gantz, George Mason University

Abstract 2: Understanding Individuality of Handwriting Using Score-Based Likelihood Ratios

Author: Christopher Saunders, PhD, Mathematical Statistician, South Dakota State University

Abstracts of Presentations delivered at The 2013 Joint Statistical Meetings, August 6, 2013, Sponsor: Committee of Representatives to AAAS

Abstract 1: A Similarity Score for Fingerprint Images

Authors: Donald T. Gantz, PhD, John J, Miller, PhD, George Mason University, Fairfax, VA; Mark A. Walch, Daniel T. Gantz, Gannon Technologies Group, Alexandria, VA

Abstract 2: On Desiderata for Score-Based Likelihood Ratios for Forensic Evidence

Authors: Christopher Saunders, South Dakota State University and John J. Miller, PhD, George Mason University

Abstracts of Presentation to be delivered at Pittcon 2014 (Application: Homeland Security/Forensics, Primary Focus: Methodology)
The Pittcon talk is an overview of the Statistical methods used in the identification of source problems.

Abstract 1: Statistical Aspects of the Forensic Identification Source Problem

Author: Christopher Saunders, South Dakota State University

Abstracts of Presentation to be delivered at AAFS 2014 Annual Meeting
The AAFS 2014 presentation is mainly focused on current research related to Phase III of the grant proposal.

Abstract 1: Statistical Aspects of the Forensic Identification Source Problem

Author: Christopher Saunders, South Dakota State University, Joshua R. Dettman and JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division

Abstracts of Presentations delivered at the AAFS 2010 Annual Meeting

Abstract 1: A Comparison Between Different Likelihood Ratios for Assessing Handwriting Evidence

Authors: Amanda Hepler, PhD, George Mason University, Document Forensics Lab, Department of Applied Information Technology, 4400 University Drive, Fairfax, VA 22030; and Christopher Saunders, PhD, George Mason University, Document Forensics Lab, 4400 University Drive, Fairfax, VA 22033*

After attending this presentation, attendees will understand how sometimes subtle changes to the prosecution and defense propositions can have a large effect upon the corresponding likelihood ratio. These impacts will be illustrated using an automated handwriting system developed and applied to handwriting samples collected by the FBI laboratories. This presentation will impact the forensic science community by illustrating the effects of modifying the prosecution and defense propositions when interpreting handwriting evidence. The ultimate goal of the court (and/or jury) is to make a decision concerning a specific suspect's guilt given the evidence, which in the likelihood ratio paradigm for presenting evidence is usually expressed as the posterior odds in favor of the suspect's guilt. In this paradigm, the court (and/or jury) is usually responsible for prior beliefs about guilt while the forensic scientist is responsible for providing the likelihood of the evidence when the suspect is guilty (the prosecution proposition) vs. when the suspect is not guilty (the defense proposition). In this presentation, various sets of prosecution and defense propositions (and the resulting likelihood ratios), which have appeared in the literature, will be compared and contrasted. The goal in performing these comparisons is to illustrate the effect that subtle modifications of these propositions can have on the resulting likelihood ratios. In addition, the practical and logical implications of each variation will be discussed. This study will be performed using a dataset of bank robbery notes and a reference database composed of writing samples from over 400 writers.

Abstract 2: New Results for Addressing the Open Set Problem in Automated Handwriting Identification

Authors: Donald T. Gantz, PhD, John J. Miller, PhD*, and Christopher Saunders, PhD, Intelligence and Security Research Center (MS 1G8), George Mason University, Fairfax, VA 22030; Mark A. Walch, March, MPH, The Gannon Technologies Group, 7600 Colshire Drive, McClean, VA 22102; and JoAnn Buscaglia, PhD, FBI Laboratory, Counterterrorism & Forensic Science Research Unit, Quantico, VA 22135*

The Open Set Problem involves making a two-stage decision when attempting to ascertain whether a questioned document was written by some individual in a reference collection (for which training material exists for each writer in the reference collection). The first step is to decide whether the document was written by any writer in the reference collection and the second step is to decide which writer in the reference collection is the most likely writer of the questioned document (or to give a "short list" of likely writers), presuming that the decision is

that some writer in the reference collection was the writer of the questioned document. At AAFS 2009, we presented some results for this problem that were generated using the FLASH ID software system. Those results used the difference between the aggregated score (totaled overall graphemes in the questioned document) for the first place writer and the aggregated score for the second place writer as the basis for the "in the reference collection" decision. In this paper, we will give some results for an improved open set decision based on a combination of the original criterion with a new criterion based on a "Vector of Counts" (VOC) methodology described below.

The VOC methodology is a way to obtain categorical type feature data by using the FLASH ID system with continuous feature data. It works in the following manner. First, we obtain a "base set" of writers, who are not in the reference collection or likely to be among writers of any questioned documents we observe. We obtain writing samples from these individuals and using FLASH ID create a trained system of the same sort as is used for the reference collection. We use this base set to analyze any document by recording for each grapheme in that document, which writer in the base set is most likely to have written that grapheme. In this way, a vector of counts for the document can be developed by counting how many graphemes are assigned to each writer in the base set.

Next, we take the training writings for each writer in the reference collection and obtain a VOC for each of those writers. When a questioned document is analyzed, we obtain its VOC as well. Then, we can compare the VOC for the questioned document with the VOC for the first place writer when writers in the reference collection are assigned questioned document scores by FLASH ID. One way to do this comparison of VOCs is using a chi-squared statistic. Since large values of chi-squared would indicate a relative mismatch between the questioned document and the first place writer and since small values of the previously used difference of first and second place writer scores would also indicate a poor match, taking the ratio of these two criteria can be an effective tool for improvement of the open set decision. We give numerical results based on extensive simulations to illustrate the improvement.

Abstracts of Presentations delivered at the NIJ 2011 Trace Evidence Symposium

Abstract 1: ROC Curves for Methods of Evaluating Evidence: A Common Performance Measure Based on Similarity Scores

Authors: R. Bradley Patterson, John Miller, Christopher P. Saunders
Video and Slides Available at http://projects.nfstc.org/trace/2011/agenda.htm

Many forensic methods produce a numerical value that indicates the degree of association between two pieces of evidence. We may treat such a number as a similarity score providing a univariate measure of association between two observations. High similarity scores support the

hypothesis that the observations come from the same source; low similarity scores support the hypothesis that they come from different sources. A method's performance depends on its capability of supporting the hypothesis that corresponds to the truth. In measuring the performance of such methods, most often single values (e.g., the 0.05 significance level for test statistics and the number one for likelihood ratios) serve as fixed cutoffs on the similarity scores. These fixed values lead to single sets of errors (i.e., false positives and false negatives) as measures of performance. Comparing sets of just two numbers may make interpreting performance ambiguous. However, techniques that consider all possible cutoffs and thus the full range of performance exist.

In this work, we demonstrate the use of receiver operating characteristic (ROC) curves in measuring the performance of methods that evaluate trace evidence and discuss the benefits of ROC curves to forensics. An ROC curve is a plot of the true positive rate (the complement of the false negative rate) versus the false positive rate for all possible thresholds on similarity scores. ROC curves present a complete picture of error rates achievable with a method. Each point on an ROC curve gives the true positive and false positive rate for a particular threshold on similarity scores. So instead of choosing arbitrarily, we could pick a threshold based on the error rates. Furthermore, because the relative ordering of similarity scores for pairs of observations from different sources and for pairs from the same source determines the ROC curve, we may use ROC curves to compare methods of evaluating evidence which generate similarity scores on different scales.

In addition, the area under the ROC curve (AUC) provides more insight into performance. The empirical AUC estimates the probability that a randomly selected pair from the same source would have a higher similarity score than a randomly selected pair from different sources. It also gives the mean true positive rate averaged uniformly across the false positive rate.

To show the value of ROC curves in forensics, we applied them to comparing the performance of four methods of evaluating trace evidence in the form of glass fragments. The forensic question was whether fragments recovered from a suspect had the same source as fragments found at a crime scene. The methods, based on test statistics and likelihood ratios, come from an article by Aitken and Lucy (2004, JRSS-C, vol. 53(1), pp. 109-122) in which the authors reported false positives and false negatives at nominal cutoffs. Test statistics and likelihood ratios both provide measures of association between two samples. So we interpreted those values as similarity scores.

The data set utilized in this study is the same as reported on in the article by Aitken and Lucy (2004) and is publicly available on The Royal Statistical Society Website. It includes elemental composition measurements for each of five fragments from 62 windowpanes. We applied the methods of evaluating trace evidence with different allotments of fragments to the control and recovered samples. By depicting all possible error rates for the methods, the ROC curves made comparisons of performance easier and allowed for the choice of threshold based on error rates.

The results for the glass data evaluated with methods from Aitken and Lucy (2004) indicate that all methods perform very well. All four methods had very high accuracy. The AUC values were all above 0.988 and within less than 0.002 across the methods for a given allotment of window fragments. In regard to choosing a threshold, nearly overlapping ROC curves showed that test-statistic and likelihood ratio methods could achieve comparable error rates.

In summary, ROC curves offer a common measure of performance for methods of evaluating forensic evidence. We noted that the false positive and false negative rates at nominal thresholds on output from the methods made assessing their performance unclear. Treating the output from the methods as similarity scores allowed us to analyze the methods with ROC curves. The ROC curves showed the methods' capability of discriminating between true positives and true negatives more completely. They also allowed for different thresholds on similarity scores for achieving error rates with each method. The results for the glass data evaluated with methods from Aitken and Lucy (2004) indicated that all methods perform very well. Applying ROC curves to different methods of evaluating trace evidence or to the same methods but with different data would provide very interesting future research.

Abstract 2: Predictive modeling for determining the discriminative power of trace glass evidence as a function of the number of sampled glass fragments.

Authors: Eric Kalendra, Christopher P. Saunders
Video and Slides Available at http://projects.nfstc.org/trace/2011/agenda.htm

The four elements silicon (Si), potassium (K), calcium (Ca), and iron (Fe) are typically used to characterize the composition of glass for forensic trace evidence. As a byproduct of the manufacturing process, panes of glass will inherently have some small amount of variation of the elemental constituents.

Using the composition from glass shards from multiple locations, a common hypothesis is that the shards came from the same window versus the shards came from different windows. In general the glass shards are collected from two locations, the control (crime scene) and the recovered data (glass fragment(s) on the suspect). The competing hypotheses are: the recovered glass fragments came from the same window as the control and the recovered glass fragments did not come from the same window as the control.

To evaluate the hypothesis, multiple methods have been proposed.

However, in this presentation we use a full Bayesian predictive model to estimate the discriminative power of the trace evidence as a function of the number of sampled glass fragments. The model we use is a multivariate two level normal model with various priors for the within and between window covariance structures. Using the two level normal model allows for the proper characterization of the sources of variability under the hypotheses that the two samples of glass fragments arise from different randomly selected windows.

We use the glass data studied in Aitken and Lucy (2004) as a collection of glass fragments to estimate the within and between window covariances. Using a Bayesian approach we simulate new glass fragments from random windows with the posterior predictive distribution. By simulating new glass fragments from the posterior predictive distribution, we can estimate the discriminative power for any sample size of the control and recovered data while taking into account uncertainty in the within and between window covariance structures. In effect by drawing samples from the posterior predictive distribution and applying the matching criteria to

these samples we are able to draw a sample from the posterior distribution of the discriminating power given our empirical data.

This approach provides a useful Monte Carlo tool for deciding on the number of observations that should be used to achieve a desired level of discriminating power for a given comparison methodology.

Abstract 3: On Parametric Models for Pairwise Comparisons with Applications to the Estimation of Random Match Probabilities.

Authors: Donald Gantz, Christopher P. Saunders
Video and Slides Available at http://projects.nfstc.org/trace/2011/agenda.htm

This presentation concerns Recommendation 3 in the National Academy of Science (NAS) report: Strengthening Forensic Science in the United States: A Path Forward. Specifically, we present a new of class of methods to statistically quantify the uncertainty in measures aimed at validating a forensic discipline's basic premises (such as a uniqueness claim).

An issue that applies to forensic individualization is that while a database of samples can be used to support individuality, it cannot directly prove individuality. Therefore, the NAS report calls for statistically/probabilistically based statements concerning the level of support that a database of samples provides for individualization. To date, much attention has focused on how to use an automated comparison methodology applied to a database of samples to estimate the random match probability (RMP), which is defined as the probability of selecting two individuals at random from a population that "match" on the basis of some biometric. The RMP can be interpreted as giving the expected performance of a comparison methodology across some relevant population. During phase I of our NIJ Grant award we have focused on the RMP as a measure of the validity of a forensic individualization procedure. Specifically, our research has been concerned with upper confidence bounds on measures, such as the RMP, that are estimated using these automated pairwise comparisons.

The use of automated pairwise comparisons of biometric samples in a database is a basic element of forensic individualization determinations involving biometrics such as fingerprints and handwriting. In this presentation, we introduce a general parametric model for studying the distribution of pairwise comparisons of an arbitrary type. The advantage of having a parametric model is that it provides an added level of structure for estimating the RMP with limited information. Furthermore, as long as the parametric model is chosen carefully, the resulting estimates appear to have a high degree of accuracy. This model is designed to incorporate the dependencies that arise in such studies.

The common method to estimate the random match probability (or discriminating power) of a comparison procedure is to take a large simple random sample and perform all pairwise comparisons between the sample observations. The proportion of these comparisons that 'match' with respect to the comparison procedure is an estimate the random match probability. In most situations the resulting estimate is a U-statistic of degree 2. (See Saunders et al. (2011) for an overview.)

A common problem with the above methods is that, due to the non-parametric nature of the estimators, they are unable to accurately be used with a small sample size (where small is

determined relative to how close the true RMP is to zero or one). One of the ongoing goals of our research group at GMU is the extension of U-statistic based methods for estimating the RMP to the situation of small sample sizes.

In this presentation we build upon the early research of the Blom (1976) to provide a parametric model that retains the optimal asymptotic properties of the U-statistic estimate of the RMP but facilitates different estimation approaches, such as Maximum Likelihood Estimates, Restricted Maximum Likelihood Estimates, and Bayesian estimates.

The parametric model we implement treats the joint distribution of comparisons as a multivariate normal distribution. This approach is conceptually analogous to applying the standard Wilson Interval to estimating a proportion from a binomial random variable. This distributional assumption is only a tool used to facilitate the estimation of the RMP and is NOT expected to actually match the joint distribution of the discrete pairwise comparisons.

We will demonstrate the use of this model in the construction of REML and Bayesian estimates and bounds for the RMP. We will also present the results of simulations that study the performance of the different estimates. We will apply these results to the glass data studied in Aitken and Lucy (2004).

Aitken, C.G.G and Lucy, D., 2004. "The evaluation of trace evidence in the form of multivariate data." Applied Statistics, 53, 109-12

Saunders, C.P., Davis, L.J., Buscaglia, J. (2011). "A Comparison between Biometric and Forensic Handwriting Individuality". Accepted for publication in Journal of Forensic Sciences.

Blom, G. (1976). "When is the Arithmetic Mean Blue?". The American Statistician, Vol.30, No. 1, pp.40-42.

Abstracts of Presentations delivered at the NIJ 2012 Impression and Pattern Evidence Symposium

Abstract 1 (Poster): Automated Statistically Ranked Latent-to-Reference Print Overlays

Authors: Donald T. Gantz, George Mason University, JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division, Mark A. Walch, The Gannon Technologies Group, Maria Antonia Roberts, Federal Bureau of Investigation Laboratory Division, Daniel T. Gantz, The Gannon Technologies Group

Many latent fingerprints confound conventional means of automated identification because they lack sufficient minutiae (ridge bifurcations and endings) to support matching by existing AFIS technology. Poorly recorded/captured exemplar prints, due to the collection method and/or limitations in the friction ridge skin, may exhibit many of the same problems as latent prints. Even in the absence of traditional minutiae, these problematic prints contain very important information in their ridges that permit the automated matching by a new approach described herein. This approach creates surrogates for minutiae by using ridge geometry to create a new class of feature that supplements the lack of bifurcations and ridge endings. These new "ridge-specific features" can be reliably associated with a specific section of a ridge using the geometric information available from the ridge. A stable ridge feature should be functionally equivalent to a traditional minutiae point. A method for capturing ridges is found in Bezier-based curve descriptors, a particular type of smooth mathematical curves that can be used to approximate the path of a ridge. Because they can be precisely fitted into the curvature of ridges, Bezier descriptors can be used to "mark" positions on the ridges creating "minutiae" where traditional minutiae do not exist. The resultant Bezier approximations of ridge curvature and the use of this information to "mark" specific positions on ridges create a new set of reference points for fingerprints. As is the case with minutiae, the power of these new ridge-based reference points is derived when they are taken in concert. By using Bezier curves as ridge descriptors, our automated process produces very accurate overlays of the latent onto a reference print. No print orientation or information beyond the Bezier ridge descriptors is required for the overlays. The latent-to-reference print overlays are the basis for a scoring algorithm that statistically ranks the reference prints according to the likelihood of being a true match to the latent print. The overlay is an invertible nonlinear mapping that associates a Bezier curve in the latent print to a Bezier curve in the reference print. The nonlinearity accounts for local distortions in the images. Beziers in the reference print are inverse mapped to latent space where corresponding Beziers are compared. Bezier-based scores yield a ranking of reference prints in the database relative to the accuracy of the latent overlays onto the database reference prints.

Abstract 2: (Poster) A Note on the Value of Forensic Evidence for Sparse Categorical Tables

Authors: Krista M. Heim, George Mason University, Christopher P. Saunders, South Dakota State University, and JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division.

We focus our attention on evaluating forensic handwriting evidence under two competing hypotheses in the context of a high dimensional quantification of handwritten documents. The first hypothesis, which is usually referred to as the *prosecution hypothesis*, states that a suspected writer is the actual writer of the document of unknown origin (henceforth referred to as the questioned document). The second hypothesis, usually referred to as the *defense hypothesis*, states that that a randomly selected writer from a relevant alternative population of writers is the actual writer of the questioned document. In particular, we discuss the issues surrounding the evaluation of count data that is in the form of sparse categorical tables, using Bayesian estimation methods to handle nuisance parameters.

In Bayesian estimation methods, the value of evidence is typically measured by calculating the likelihood of observing the questioned document if it was written by the suspected writer and comparing it with the likelihood of observing the questioned document if it was written by a randomly selected writer. In this context, the standard evidence statement used for presentation of the value of evidence in a likelihood ratio format is: the likelihood of observing the evidence (which includes the questioned document and the samples known to come from the suspect writer) is k times more likely if the suspect writer actually wrote the questioned document than if a randomly selected writer (from the population of alternative writers) wrote the questioned document. When using Bayesian methods to account for nuisance parameters, the resulting value of the evidence usually takes the form of a Bayes Factor; in the forensic science literature, a Bayes Factor is sometimes referred to as a likelihood ratio. Unfortunately, in problems where the evidence has a high dimensional quantification, it is usually technically difficult to calculate a Bayes Factor that is meaningful for distinguishing between the two competing hypotheses.

One strategy for dealing with high dimensional sparse categorical tables is to use the table of counts associated with the smaller document, which is usually the questioned document, to reduce the dimensionality of the problem. We discuss the impact this dimension reduction strategy has on the interpretation of evidence; in effect, by using part of the evidence to determine the dimensionality of the problem, we will need to use an alternative evidence statement in place of the standard one mentioned above. The new evidence statement is with respect to the smaller of the documents (which we will assume is the questioned document) and is as follows: the likelihood of observing the *questioned document* is k times more likely if the suspected writer actually wrote the questioned document than if a randomly selected writer from the population of alternative writers wrote the questioned document. The new evidence statement suggests an estimation problem in place of the standard inference problem that usually arises in forensic statistics. We will discuss this difference between the estimation and inference approaches in assessing the value of forensic pattern evidence. Finally, point estimates and credible intervals of the likelihood ratio from the alternative evidence statement are calculated for a collection handwriting data collected from over 400 writers for research purposes. In this presentation, we focus on evaluating forensic evidence under two competing hypotheses with respect to handwriting evidence. This is measured by calculating the likelihood that the suspect is the source of a questioned document. In particular, we discuss the issues surrounding the evaluation of count data that is in the form of sparse categorical tables. We use Bayesian estimation to handle nuisance parameters and discuss the choice of prior and the effect on our probability model under this construct. We show the impact of reducing the dimensionality by

fixing on the number of categories of the questioned document, which will require an alternative statement of the hypotheses.

Abstract 3: The Effect of the Order of Suspect and Background Population Samples on the Assessment of the Value of Evidence

Authors: Eric Kalendra, George Mason University, Christopher P. Saunders, South Dakota State University, JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division.

Categorical based pattern evidence is commonly available, such as in handwriting. In this presentation we focus on the interpretation of the evidence when the rate of only a single feature is available. For example, how the probability of a coin observing heads or the probability of the letter "e" taking a particular graph structure. Our goal is the forensic interpretation of the evidence with respect to the prosecution hypothesis, the suspect gave rise to the evidence, and defense hypothesis, a random member of the population gave rise to the evidence. In addition to the collected sample, we will typically have a sample from the suspect and samples from the background population. Using all the available information, Bayes Factors can be calculated for the competing hypotheses using a simultaneous or a two-stage procedure. The simultaneous method would typically be used in instances where the suspect's sample and background population samples are available concurrently, and the two-stage procedure would be used if the samples become available sequentially. Either the suspect's sample or the background population may be available first. The difference in calculation arises from the information available when inference is conducted. For example, in the two-stage calculation when the suspect's sample is available first, the prior distribution does not include information from the background population as no information is known about the background population in the first stage. By changing the order of the information used, the assessment of the value of the evidence is also affected. We will illustrate the resulting effects of information order on the calculation of the Bayes Factors with an example using a population of coins with varying probabilities of observing heads. While simple in design, the illustrated effects of information order hold in general. Although the evidence itself remains the same, the assessed forensic value will change depending on how inference is conducted.

Abstract 4 in the Workshop: Guidelines for a Successful Research Project: *The story of an Academic/Commercial Partnership developing a product for the Forensic Community*

Panelists & Presenters: Donald Gantz, George Mason University, and Mark Walch, The Gannon Technologies Group.

A vision that a "lights out" automated system for handwriting identification is possible was kick-started by Mark Walch's experience with Intelligent Character Recognition. For several years, Walch had used a special method for quantifying writing for purposes of character recognition. It was believed by Walch and his sponsors that this method would be relevant as the foundation for a writer identification system. He had unsuccessfully approached a number of premier statistics departments with the proposition of exploiting his handwriting quantification for writership

identification. However, his overture to Don Gantz at the George Mason University (GMU) Applied & Engineering Statistics launched a successful partnership. Walch and Gantz began the development of FLASH ID by exploiting Gannon Technologies' considerable experience in the quantification of handwriting—originally developed for handwriting recognition. This ability to segment and quantify handwriting into a graph-based data structure became the foundation of the handwriting biometric capability. Initially, the handwriting biometric researched focused on individual characters as the basic units for biometric analysis. Ultimately, these characters were replaced by graphemes which made possible a totally language independent system. Encoded graphemes are converted into numeric data suitable for statistical analysis. The Gannon/GMU Alliance flourished through nurturing mutually beneficial opportunities. Gannon's sponsors funded a multi-year research effort in document forensics at GMU that has supported the development and application of the Forensic Language-Independent Analysis System for Handwriting Identification (FLASH ID). This research partnership continues and has branched into the area of latent fingerprint identification.

Abstracts of Presentations delivered at the 2012 Joint Statistical Meetings, August 2012

Abstract 1: (Poster) "The effect of uncertainty about the Alternative Source population on the value of Forensic evidence."

Author: Christopher P. Saunders, South Dakota State University

A goal in the forensic interpretation of scientific evidence is to make an inference about the source of a trace of unknown origin; the inference usually concerns two propositions. The first proposition is usually referred to as the prosecution hypothesis and states that a given specific source is the actual source of the trace of unknown origin. The second usually referred to as the defense hypothesis, states that the actual source of the trace of unknown origin is randomly selected from a relevant alternative source population. The evidence a forensic scientist is given for deciding between these two propositions is: (a) the trace of unknown origin, (b) a sample from the specific source specified by the prosecution hypothesis, and (c) a collection of samples from the alternative source population. One common approach is to assume that the alternative source population is completely known and rely on a Bayes Factor for deciding between the competing hypotheses. In this presentation we will relax this assumption and explore some of the resulting issues from the estimation of the alternative source population. We will illustrate the resulting effects on the calculation of the Bayes Factors with a well-studied collection of samples relating to glass fragments.

Abstracts of Presentations delivered at EAFS 2012, The Hague, August 20-24, 2012

Abstract 1: The Effect of Uncertainty About the Background Population on the Forensic Value of Evidence

Author: Christopher P. Saunders, South Dakota State University

A goal in the forensic interpretation of scientific evidence is to make an inference about the source of a trace of unknown origin; the inference usually concerns two propositions. The first proposition is usually referred to as the prosecution hypothesis and states that a given specific source is the actual source of the trace of unknown origin. The second usually referred to as the defense hypothesis, states that the actual source of the trace of unknown origin is randomly selected from a relevant alternative source population; i.e. the *background population*. The evidence that a forensic scientist is given for deciding between these two propositions is: (a) the trace of unknown origin, (b) a sample from the specific source specified by the prosecution hypothesis, and (c) a collection of samples from the alternative source population. One common approach is to assume that the collection of samples from the alternative source population is sufficiently large as to completely specify the alternative source population and to rely on a value of evidence for deciding between the competing hypotheses, as described in Lindley (1977).

In this presentation, we present our construction of a Bayes Factor for deciding between the prosecution and defense hypotheses when the collection of samples from the alternative source population is not sufficiently large to completely characterize the alternative source population. We argue that the resulting Bayes Factor should be considered the Value of the Evidence and discuss its relationship to the standard value of evidence as developed by Lindley and presented in Aitken and Taroni (2004). We conclude with a discussion of some of our concerns about the effect of prior choice for the nuisance parameters in the alternative and specific source distributions on the resulting Bayes Factor.

We will illustrate the construction of the Bayes Factors with a well-studied collection of samples relating to glass fragments under the assumption of a hierarchical normal model.

Lindley, D. V. (1977). A problem in forensic science. Biometrika. 64 (2): 207-213.

Aitken, C. G. G. and Taroni, F. (2004), F. Statistics and the Evaluation of Evidence for Forensics Scientists. 2nd Edition, John Wiley and Sons.

Abstract 2: Ridge Specific Markers for Latent Fingerprint Identification

Authors: Donald T. Gantz, George Mason University, JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division, Mark A. Walch, The Gannon Technologies Group, Maria Antonia Roberts, Federal Bureau of Investigation Laboratory Division, Daniel T. Gantz, The Gannon Technologies Group

Many latent fingerprints confound conventional means of automated identification because they lack sufficient minutiae (ridge bifurcations and endings) to support matching by existing AFIS technology. Poorly recorded/ captured exemplar prints, due to the collection method and/or limitations in the friction ridge skin, may exhibit many of the same problems as latent prints.

Even in the absence of traditional minutiae, these problematic prints contain very important information in their ridges that permit the automated matching by a new approach described herein. This approach creates surrogates for minutiae by using ridge geometry to create a new class of feature that supplements the lack of bifurcations and ridge endings. These new "ridge-specific features" can be reliably associated with a specific section of a ridge using the geometric information available from the ridge. A stable ridge feature should be functionally equivalent to a traditional minutiae point. A method for capturing ridges is found in Bezier-based curve descriptors, a particular type of smooth mathematical curves that can be used to approximate the path of a ridge. Because they can be precisely fitted into the curvature of ridges, Bezier descriptors can be used to "mark" positions on the ridges creating "minutiae" where traditional minutiae do not exist. The resultant Bezier approximations of ridge curvature and the use of this information to "mark" specific positions on ridges create a new set of reference points for fingerprints. As is the case with minutiae, the power of these new ridge-based reference points is derived when they are taken in concert. By using Bezier curves as ridge descriptors, our automated process produces very accurate overlays of the latent onto a reference print. No print orientation or information beyond the Bezier ridge descriptors is required for the overlays. The latent-to-reference print overlays are the basis for a scoring algorithm that statistically ranks the reference prints according to the likelihood of being a true match to the latent print. The overlay is an invertible nonlinear mapping that associates a Bezier curve in the latent print to a Bezier curve in the reference print. The nonlinearity accounts for local distortions in the images. Beziers in the reference print are inverse mapped to latent space where corresponding Beziers are compared. Bezier-based scores yield a ranking of reference prints in the database relative to the accuracy of the latent overlays onto the database reference prints.

Abstracts of Presentations delivered at The Measurement Science and Standards in Forensic Handwriting Analysis (MSSFHA) Conference, June 4 – 5, 2013 at NIST.

Abstract 1: The Forensic Language-Independent Analysis System for Handwriting Identification (FLASH ID)

Authors: Mark Walch, Gannon Technologies Group and Donald Gantz, George Mason University

FLASH ID is a fully functional software application that automatically identifies writers by their handwriting. FLASH ID works by maintaining a database of information derived from *reference* handwriting and determining whether a new, unidentified writing specimen such as a questioned document matches any of the writings in the database. FLASH ID operates on a conventional personal computer platform—including laptops. Questioned documents subjected to biometric analysis are scanned and passed to FLASH ID as image files. Once the image has been captured, FLASH ID distills the biometric content from the handwriting, compares this content to reference samples stored in a database, computes scores representing biometric similarity and compiles the results in a ranked list of all writers from the database. The writer at the top of this list bears the strongest similarity to the writer of the captured specimen.

Functionally, FLASH ID finely segments the writings within the loaded images. Adjacent segments are combined into *graphemes*, which are the bases for analysis. Graphemes may be parts of characters, whole characters or groups of characters. Graph-matching algorithms at the core of the technology classify the graphemes, first, by their *topology* and, second, by their *geometric features*. Topology includes the structure of graphs in terms of their edges and vertices—links and nodes—and their quantity and connectivity. Geometric features address the shapes of curves. Physical measurements on the graphemes make up associated feature vectors. The power to distinguish an individual's writing from that of other writers is derived from statistical analysis of the topological and geometric characteristics of the individual's writings as well as from the statistical analysis of feature vectors within each particular topology and geometry combination.

FLASH ID consists of two modules: (1) the Database Builder and (2) the Matcher. The FLASH ID Database Builder pre-processes reference writing samples and builds an identification database from statistical analyses of the topology, geometry and features of graphemes. The FLASH ID Matcher uses this database to process questioned documents and identify the writer in the database who best matches a questioned document. Both FLASH ID modules are designed to scale across multiple computers to handle higher volumes.

There are two operational versions of FLASH ID. The document examiner version provides specific visual grapheme level feedback concerning similarities of writings and supports exporting of annotated images and reports. The web services version can be integrated into other systems for the analysis of handwriting and provides writer identification results for submitted writing samples.

Abstract 2: Understanding Individuality of Handwriting Using Score-Based Likelihood Ratios

Author: Christopher Saunders, PhD, Mathematical Statistician, South Dakota State University

Recent studies in automated forensic handwriting identification have shown that for a given value of the evidence, subtle changes in conditioning arguments regarding the defense proposition can often lead to radically different values of the so called Score-Based-Likelihood-Ratios (SLR). Within the forensic literature there are three general classes of SLR's. While each of the proposed SLRs has advantages and disadvantages; they are at best only approximations to a true LR in a Bayesian decision theoretic sense. In our estimation, it is best to resist the idea of a "universally correct" SLR. This presentation will review the different types of SLR's currently being used in forensic science, discuss strategies for implementing them for forensic handwriting analysis, and review some of the problems related to the interpretation of the resulting SLRs.

Abstracts of Presentations delivered at The 2013 Joint Statistical Meetings, August 6, 2013, Sponsor: Committee of Representatives to AAAS

Abstract 1: A Similarity Score for Fingerprint Images

Authors: Donald T. Gantz, PhD, John J, Miller, PhD, George Mason University, Fairfax, VA; Mark A. Walch, Daniel T. Gantz, Gannon Technologies Group, Alexandria, VA

This talk presents a similarity score for a pair of fingerprint images that is based on a novel quantification of the images. The processes presented are those used in an automated system that provides a latent fingerprint examiner with an accurate overlay of a latent (crime scene) print to each of the fingerprints that have been returned by an AFIS (Automated Fingerprint Identification System) search of a database. The similarity score is an assessment of the accuracy of the nonlinear, invertible Warp which yields the overlay of the latent onto each returned print. The similarity score provides a prioritized ranking of all AFIS returned prints. This process is also applied independently to small snippets from a latent fingerprint and the resulting similarity scores are fused to yield an overall score for the latent fingerprint. In support of latent print examination, the system makes possible a much greater number of AFIS returned prints and provides a substantial starting point for latent to reference print examination.
In order to create a *Warp* between two fingerprint images, a new concept of *Ridge Specific Markers* (RSMs) is introduced to serve as landmarks for constructing a Warp. High contrast images are created with black ridges and white furrows. The boundaries between black and white are covered redundantly with cubic Bezier curves of multiple lengths to "mark" specific positions in the image. These Bezier curves are the RSMs. A Warp between images is grown incrementally from a seed of RSMs. Hundreds of Warps are generated between the latent and each reference print, and a best Warp for each reference print is selected. Given the best Warp, an algorithm symmetrically associates RSMs in the latent with RSMs in a reference print. For each *pixel* in an image there is a *best pair* of associated Bezier Curves. The Similarity Score measures how accurately the Warp associates the pixel-based pairs of Bezier Curves.

Abstract 2: On Desiderata for Score-Based Likelihood Ratios for Forensic Evidence

Authors: Christopher Saunders, South Dakota State University and John J. Miller, PhD, George Mason University

This presentation offers some opinions on the desirable features of score-based likelihood ratios (SLRs) for interpreting and presenting forensic evidence. Let E denote all of the available evidence, with decomposition $E = \{ E_s, E_u\}$ where E_s denotes the evidence sample(s) obtained from a suspect, E_u denotes the evidence sample of unknown source obtained. We consider an arbitrary, but fixed score function, s, serving to reduce the evidence to the following form $E' = \{s(E_s, E_u)\}$. Several score-based interpretations of the likelihood ratio have appeared in the literature providing a method for evaluating the weight of E' in light of two competing hypotheses, H_p and H_d. Recent studies in writer identification have shown that when E' is held constant, subtle changes in conditioning arguments regarding the defense proposition often lead to radically different values of the SLR. Each proposed SLR has advantages and disadvantages and, in our estimation, it is best to resist the idea of a "universally correct" SLR. We instead have concentrated our efforts on enumerating desirable theoretical properties for SLRs in general and the evaluation of proposed SLRs against each property.

Abstracts of Presentation to be delivered at Pittcon (Application: Homeland Security/Forensics, Primary Focus: Methodology)

Abstract 1: Statistical Aspects of the Forensic Identification Source Problem

Author: Christopher Saunders, South Dakota State University

In 1977, Lindley and Evett, introduced modern Bayesian methods for forensic evidence interpretation to the forensic science community. This and related approaches have dominated the academic research related to the interpretation and presentation of forensic evidence. However, in recent years there have been number debates, in both academic and forensic communities, related to the applicability of these methods in the U.S. judicial system.

Broadly speaking, these methods require the explicit statement of two mutually exclusive, but non-exhaustive, models about how the evidence in a given situation has arisen; one usually corresponding to a defense model and one corresponding to a prosecution model. Once these models have been defined and the evidence collected, the forensic science expert is then required to present the evidence in a concise and transparent manner so that a decision maker can ultimately decide between the two proposed models of how the evidence has arisen.

The evidence that a forensic scientist has available to evaluate between the two models is generally composed of the following components: (1) a trace of unknown origin; (2) a sample from the specific source specified by the prosecution model and (3) a collection of samples from the alternative source population specified by the defense model. In certain applications, the choice of the alternative source population will be mandated by available databases or, in extreme situations, there will be no such samples available.

We will review some of the common sets of probability models and statistical approaches that forensic scientists use to characterize the support that the evidence provides for deciding between the prosecution and defense models. We will also discuss how the various sets of competing models can be addressed with the commonly available evidence. The general approach will be illustrated with an example of the trace element analysis of high purity copper evidence.

Abstracts of Presentation to be delivered at AAFS 2014 Annual Meeting

Abstract 1: Statistical Aspects of the Forensic Identification Source Problem

Author: Christopher Saunders, South Dakota State University, Joshua R. Dettman and JoAnn Buscaglia, Federal Bureau of Investigation Laboratory Division

It is expected that the attendees will have a greater understanding of the current trends in statistical evidence interpretation, which will foster better communication between statisticians, evidence interpretation experts, and the broader forensic science community. Improving

communication between these experts should assist in the development of statistically sound, rigorous methods of interpretation that are appropriate to the diverse needs of the U.S. forensic science and legal communities.

In 1977, Dennis Lindley, with Ian Evett, introduced modern Bayesian methods for forensic evidence interpretation to the forensic science community. This and related approaches have dominated the academic research related to the interpretation and presentation of forensic evidence. However, in recent years there have been number debates, in both academic circles and forensic communities, related to the applicability of these methods in the U.S. judicial system.

Broadly speaking, these methods require the explicit statement of two mutually exclusive, but non-exhaustive, propositions about how the evidence in a given situation has arisen; one usually corresponding to a defense proposition and one corresponding to a prosecution proposition. Using this approach, once these propositions have been defined and the evidence has been collected, the forensic science expert is then required to present the evidence in a concise and transparent manner so that a decision maker can ultimately decide between the two proposed models of how the evidence has arisen.

Commonly, the evidence that a forensic scientist has available to evaluate between the two propositions is generally of one of the following forms: (1) a trace of unknown origin; (2) a sample from the specific source specified by the prosecution hypothesis and (3) a collection of samples from the alternative source population specified by the defense proposition. In certain applications, the choice of the alternative source population will be mandated by available databases or, in extreme situations, there will be no such samples available.

In this presentation, we will review some of the common sets of propositions and statistical approaches that forensic scientists use to characterize the support that the evidence provides for deciding between the prosecution and defense propositions. We will also discuss how the various sets of competing propositions can be addressed with the commonly available evidence. The general approach will be illustrated with an example of the trace element analysis of high purity copper evidence.

Key Personnel and CVs

- ***Donald T. Gantz, PhD (PI from 2012 to end date of Grant, Vita Attached)***

  Director, Document Forensic Laboratory
  Chair, Department of Applied Information Technology, George Mason University
  Full Professor of Statistics, George Mason University

  Under his direction, George Mason partnered with Gannon Technologies Group to develop cutting edge methodologies for the quantification and analysis of handwriting. This effort continues in CTAF, which has been applying these methodologies to multi-language document exploitation and biometric identification. CTAF has government funding and is staffed by five statistics professors, one research professor, one postdoctoral research fellow, and graduate research assistants.

  Member of the IEEE Certified Biometric Professional Exam Specifications Committee.

- ***Christopher P. Saunders, PhD (Initial PI 2009-2012, Ongoing Collaborator, Vita Attached)***

  Department of Mathematics and Statistics, South Dakota State University

  Dr. Saunders received his Ph.D. in statistics from the University of Kentucky in 2006 under the direction of Dr. Constance L. Wood. The focus of his dissertation was the application of statistical approximation theory to the testing of the distribution assumption of multivariate normality. While working on his Ph.D., he was a member of the Microarray Research Center at the University of Kentucky Medical Center, and was also awarded and completed an NIH training grant with Professor Thomas Getchell's Laboratory in the Sanders-Brown Center on Aging.

  Since completing his dissertation, Dr. Saunders has focused on providing statistical support to the Intelligence Community (IC) — first as an Intelligence Community Post Doctoral Research Fellow and then as a Research Assistant Professor with the Document Forensics Research Laboratory at George Mason University, Assistant Professor of Statistics at South Dakota State University, and Visiting Scientist at the FBI Laboratory.

- ***Linda J. Davis, PhD (Co-PI)***

  Document Forensic Laboratory & Department of Statistics, George Mason University

  Worked as a member of the technical staff at TRW Inc. and Northrup Grummen for over twenty years before transitioning to a faculty position at George Mason University. While working for TRW Inc. and Northrup Grummen, she supported a number of projects for a variety of government agencies. She also assisted in development of program performance evaluation and quantification databases.

  Experienced with comparative analysis of handwriting samples using discrete features.

  Current research interests include categorical data analysis and biometric identification.

- ***Amanda B. Hepler, PhD***

  Document Forensic Laboratory, George Mason University

Experienced consultant with extensive training and expertise in forensic statistics, statistical genetics, population genetics, and computational statistics.

Current Research Interests: (1) Statistics and the law; (2) Use of Bayesian belief networks (probabilistic expert systems) to present genetic evidence in court; (3) Incorporating population relatedness into forensic genetic calculations.

- ***John J. Miller, PhD***

  Document Forensic Laboratory & Department of Statistics, George Mason University

  Experienced with using handwriting as a biometric and with computational statistics.

  Has general experience in litigation (not necessarily involving forensics), which gives him perspectives on the use of statistics in the legal arena.

Students Associated with the Grant -South Dakota State University
- Under the Supervision of Dr. Saunders

Danica M. Ommen- Pursuing a Master's in Statistics- Master's Thesis is focused on computational issues associated with forensic likelihood ratios. (Not directly funded from grant.)

Austin F. O'Brien- Pursuing a Ph. D. in Computational Statistics, Directly supported. Has provided computational support for various aspects of Phase I and III of the research program.

Douglas Armstrong- Pursuing a Ph.D. in Statistics, His Ph.D. dissertation proposal will focus on various aspects of the score based Likelihood Ratios.

Students Associated with the Grant -South Dakota State University
- Under the Supervision of Dr. Gantz and Miller

Krista Heim- Pursuing a PhD in statistics. Provided support on computational aspects of Phase III of the research program.

R. Brad Patterson- Completed a PhD in statistics. Provided support for data analysis and developed new methods for assessing the accuracy of LRs. His dissertation was funded in part by a separate NIJ award directly to Dr. Patterson for research started under this award.

VITAS
DONALD T. GANTZ, PhD

Education

Fordham University, Bronx, N.Y.	Mathematics	A.B., 1966
University of Rochester, Rochester, N.Y.	Mathematics	M.A., 1971
University of Rochester, Rochester, N.Y.	Mathematics	Ph.D., 1974

Appointments

1974 - Present: Professor, George Mason University, Fairfax, VA.

2004 - Present: Chair of the Department of Applied Information Technology, Volgenau School of Engineering, George Mason University, Fairfax, VA

Dr. Gantz is the founding Chair of the Applied Information Technology (AIT) Department. AIT's Bachelor of Science in Applied Information Technology Degree enrolls more than 1,300 students and the Master of Science in Applied Information Technology Degree enrolls close to 200 students. Dr. Gantz was Interim Associate Dean for Undergraduate Studies in the Volgenau School during the Spring 2003 and Fall 2003 semesters. He is the Director of the Document Forensics Laboratory.

He is a Full Professor of Statistics. As an applied statistician, Dr. Gantz has developed cutting edge methodologies for the quantification and analysis of handwriting and is applying these methodologies to multi-language document exploitation and biometric identification. He has developed a new technology for the analysis of latent fingerprints; he has lectured internationally on this technology. He has used statistical and geographic information system methods to analyze the relationship between TB incidence and socioeconomic factors, in particular the patient's national origin. He has lectured on statistical methods for surveillance systems to detect levels of infection due to a natural epidemic or bioterrorism threat. He has been an active researcher and practitioner in the application of geographic information systems, modeling systems and decision support systems to transportation demand management and traffic mitigation. He has done considerable work, research, and lecturing in computer performance evaluation and capacity planning. He has worked on the application of estimation and control methods to the analysis of flight test data. Throughout his years as an applied statistician, he has been involved with survey design, analysis and reporting. He has considerable experience in the development of management decision systems and in litigation related analyses.

Document Forensics: Dr. Gantz directs the Document Forensics Laboratory (DFL), which has partnered with Gannon Technologies Group to develop cutting edge methodologies for the quantification and analysis of handwriting. They are applying these methodologies to multi-language document exploitation and biometric identification. The DFL has government funding and is staffed by statistics professors, research professors, postdoctoral research fellows and graduate research assistants. DFL research has been reported at the 1st ACM Workshop on Hardcopy Document Processing 2004; the FBI Laboratories Forensics Lecture Series 2005; SDIUT 2005 The 2005 Symposium on Document Image Understanding Technology; SACH06 Summit on Arabic and Chinese Handwriting; AAAS 2006 Annual Meeting; EAFS 2006, EAFS 2009 and EAFS 2012 European Academy of Forensic Science Meetings; AAFS 2008, 2009 and 2010 Annual Meetings of the American Academy of Forensic Sciences, IAFS 2008 Triennial Meeting of the International Association of Forensic Sciences; ICFIS08 The Seventh International Conference on Forensic Inference and Statistics; and ROBUST2008 Robust Biometrics: Understanding Science & Technology held in Honolulu, Hawaii, November 2-5, 2008. DFL researchers have a three-year grant (2010-2013) from the National Institute of Justice: Quantifying the Effects of Database Size and Sample Quality on Measures of Individualization Validity and Accuracy in Forensics. DFL researchers gave three presentations on statistical methods at the National Institute of Justice (NIJ) Trace Evidence Symposium, Kansas City, August 2011. The presentations were: "Predictive Modeling for Determining the

Discriminative Power of Trace Glass Evidence as a Function of the Number of Sampled Glass Fragments"; "ROC Curves for Methods of Evaluating Evidence: A Common Performance Measure Based on Similarity Scores"; and "On Parametric Models for Pairwise Comparisons with Applications to Estimation of Random Match Probabilities." DFL researchers gave four presentations on statistical methods at the National Institute of Justice (NIJ) Impression and Pattern Evidence Symposium (IPES 2012), Clearwater, FL, August 2012. The presentations were: "The story of an Academic/Commercial Partnership: developing a product for the Forensic Community"**;** "The Effect of the Order of Suspect and Background Population Samples on the Assessment of the Value of Evidence"; "Automated Statistically Ranked Latent-to-Reference Print Overlays"; and "A Note on the Value of Forensic Evidence for Sparse Categorical Tables." Dr. Gantz was an invited presenter at the Measurement Science and Standards in Forensic Handwriting Analysis (MSSFHA) Conference, June 4 – 5, 2013. The National Institute of Standards and Technology (NIST) hosted the MSSFHA Conference which was planned and organized in collaboration with the American Academy of Forensic Sciences – Questioned Document Section, American Board of Forensic Document Examiners, American Society of Questioned Document Examiners, Federal Bureau of Investigation Laboratory, National Institute of Justice (NIJ), and Scientific Working Group for Forensic Document Examination (SWGDOC).

Latent Fingerprint Research: Dr. Gantz has been working with the Gannon Technologies Group since 2009 on FBI-sponsored research projects to apply graph-based technologies to latent fingerprint examination. The goal of the research has been to provide an Examiner with a prioritized ranking of each AFIS returned print based on a Warp of the latent onto each returned print. This research makes possible a greater number of AFIS returned prints and provides a substantial starting point for latent to reference print examination. This research introduced *Ridge Specific Markers* (RSMs) as landmarks for constructing a Warp. The Warp is an invertible nonlinear mapping that transforms any point within the Latent to an associated point in a reference print. Latent-to-reference print Warps provide a visual frame for an examiner and also are the basis for a scoring algorithm that ranks the reference prints according to the accuracy of the match to the latent print. This research has been reported at the National Institute of Justice (NIJ) Impression and Pattern Evidence Symposium (IPES 2012), Clearwater, FL, August 2012 (poster "Automated Statistically Ranked Latent-to-Reference Print Overlays"); at the EAFS 2012 European Academy of Forensic Science Conference, The Hague, The Netherlands, August 2012 ("Ridge Specific Markers for Latent Fingerprint Identification"); and at the USACIL RTD+E Working Group, Atlanta, GA, November 2012 ("Fingerprint Fragment Fusion"). Dr. Gantz presented his algorithms for fingerprint forensics in the Statistics in Forensic Science Topic Contributed Paper Session at the Joint Statistical Meetings in Montreal in August 2013. Dr. Gantz presented his paper "A Similarity Score for Fingerprint Images." The paper co-authored with John Miller describes the scoring algorithms he developed for a totally automated innovative technology enabling the identification of crime scene fingerprints. The presentation was selected to receive an Honorable Mention in the Section on Physical and Engineering Sciences (SPES) Outstanding Presentation Awards indicating that it was among the best of the 73 talks presented in a SPES-sponsored contributed paper session.

Selected Publications and Presentations

"Structuring and analyzing competing hypotheses with Bayesian networks for intelligence analysis," Karvetski, C. W, Olson, K. C., Gantz, D. T., Cross, G. A., 2013. Accepted to EURO Journal on Decision Processes, Special Issue on Risk Management.

"The Forensic Language-Independent Analysis System for Handwriting Identification (FLASH ID)," presented at the Measurement Science and Standards in Forensic Handwriting Analysis Conference & Webcast," National Institute of Standards and Technology, U.S. Department of Commerce, June 4-5, 2013, Gaithersburg, MD.

"Ridge Specific Markers for Latent Fingerprint Identification," Presentation at the triennial 2012 European Academy of Forensic Science (EAFS) Conference, The Hague, The Netherlands, August 20-24.

"An Academic/Commercial Partnership: developing a product for the Forensic Community," presented in a Workshop on Guidelines for a Successful Research Project at the 2012 National Institute of Justice Impression and Pattern Evidence Symposium, *Recognize, Develop, and Implement: Building on our Foundations*, August 5-9, 2012, Clearwater, FL.

"Construction and Evaluation of Classifiers for Forensic Document Analysis," Christopher P. Saunders, Linda J. Davis, Andrea C. Lamas, John J. Miller, and Donald T. Gantz, Annals of Applied Statistics, 2011, Vol. 5, No. 1, 381–399.

"On Parametric Models for Pairwise Comparisons with Applications to Estimation of Random Match probabilities," presented at the 2011 National Institute of Justice Trace Evidence Symposium, August 8-11, 2011, Kansas City, MO.

"New Results for Addressing the Open Set Problem in Automated Handwriting Identification," Donald T. Gantz, John J. Miller, Christopher P. Saunders, Mark A. Walch and JoAnn Buscaglia, Proceedings of the American Academy of Forensic Sciences Annual Scientific Meeting, Seattle, WA, February 22-27, 2010, pages 431-432.

"Training the Architects of the Networked Future: How a public/private partnership is benefiting students, an institution, and the local economy" *University Business*, September, 2010.

"An Approach to a Capstone Curriculum," Robert T. Quinn and Donald T. Gantz, *Proceedings of the 2009 ACM Information Technology Education Conference*, Fairfax, Virginia, October 22-24, 2009, pages 150-154.

"Combining Academic Studies with IT Certifications: Becoming a Cisco Regional Academy," Louis R. D'Alessandro and Donald T. Gantz, *Proceedings of the 2009 ACM Information Technology Education Conference*, Fairfax, Virginia, October 22-24, 2009, pages 209-214.

"Evaluation of the Language-Independent Process in the FLASH ID System for Handwriting Identification," Mark A. Walch, Donald T. Gantz, John J. Miller and JoAnn Buscaglia,

Proceedings of the American Academy of Forensic Sciences Annual Scientific Meeting, Denver, CO, February 16-21, 2009, pages 381-382.

"Statistical Characterization of Writers for Identification," Donald T. Gantz, John J. Miller, Christopher P. Saunders, Mark J. Lancaster and JoAnn Buscaglia, Proceedings of the American Academy of Forensic Sciences Annual Scientific Meeting, Washington, DC, February 18-23, 2008, pages 390-391.

Christopher P. Saunders, Ph.D.

Professional Preparation

California State University, Chico	Mathematics	B.S., 1996-2000
University of Kentucky	Statistics	M.S., 2000-2002
University of Kentucky	Statistics	Ph.D., 2002-2006
George Mason University	Intelligence Community	Fellow, 2006-2008
Visiting Scientist with the FBI Labs	Forensic Science Research	Summer 2013

Appointments

Assistant Professor of Statistics, Department of Mathematics and Statistics, South Dakota State University, 2012–Present.

Lead Signal Processing Engineer, Washington Signal Processing Department, The MITRE Corporation, 2011–Present.

Associate Research Professor, Applied Information Technology, George Mason University, effective August 2012. (Left GMU before promotion.)

Assistant Research Professor, Document Forensics Laboratory, George Mason University, 2008– June 2012.

IC Postdoctoral Research Fellow, Document Forensics Laboratory, George Mason University, 2006–2008.

Research Assistant, Department of Statistics, University of Kentucky, 2003–2006.

Teaching Assistant/Instructor, Department of Statistics, University of Kentucky, 2001–2003.

Teaching Assistant, Department of Statistics, University of Kentucky, 2000–2001.

Publications

Mallory, J.C., Crudden, G., Oliva, A., Saunders, C., Stromberg, A., and Craven, R.J. (2005). A novel group of genes regulate susceptibility to anti-neoplastic drugs in highly tumorigenic breast cancer cells. *Mol Pharmacol*. Sep 8.

Ebersole, J., Meka, A., Stromberg, A., Saunders, C., and Kesavalu, L. (2005). Host Gene Expression in Local Tissues in Response to Periodontal Pathogens. *Oral Biosciences & Medicine* 2 (2/3), 175–184.

Liu, H., Saunders, C.P., Borders, A.S., Getchell, T.V., Getchell, M.L., Bathke, A., and Stromberg, A.J. (2006). Statistical and graphical identification of functional gene categories in microarray experiments. *Proceedings of the 2006 Joint Statistical Meetings*, Alexandria, VA. American Statistical Association, 262–269.

Getchell, T. V., Kwong, K., Saunders, C. P., Stromberg A. J., and Getchell M. L. (2006). Leptin regulates olfactory-mediated behavior in ob/ob mice. *Physiol Behav.* 30; 87(5):848-56. Epub 2006 Mar 20.

Balko, J.M., Potti, A., Saunders, C., Stromberg, A., Haura, E.B. and Black, E.P. (2006). Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. *BMC Genomics* 7, 289.

Huang, L. Zhu, W., Saunders, C.P., MacLeod, J.N., Zhou, M., Stromberg, A.J. and Bathke, A.C. (2008). A novel application of quantile regression for identification of biomarkers exemplified by equine cartilage microarray data. *BMC Bioinformatics* 9, 300.

Saunders, C.P., Davis, L.J., Lamas, A.C., Miller, J.J., Gantz, D.T. Construction and Evaluation of Classifiers for Forensic Document Analysis. *Annals of Applied Statistics.* 2011. 5-1.

Saunders, C.P., Davis, L.J., Buscaglia, J., Using Automated Comparisons to Quantify Handwriting Individuality. J Forensic Sci. 2011. May; 56-3.

Davis LJ, Saunders CP, Hepler A, Buscaglia J. Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. Forensic Sci Int. 2012 Mar 10; 216(1-3):146-57.

Hepler AB, Saunders CP, Davis LJ, Buscaglia J. Score-based likelihood ratios for handwriting evidence. Forensic Sci Int. 2012 Jun 10; 219(1-3):129-40.

Scott L. Rosen, Christopher P. Saunders, Samar K. Guharay, Metamodeling of simulations consisting of time series inputs and outputs. Proceedings of the Winter Simulation Conference 2012, 211.

Rosen, S.L.; Saunders, C.P.; Tierney, M.; Guharay, S.K., Configuration of a standoff detection system via rapid, model-based systems engineering, System of Systems Engineering (SoSE), 2013 8th International Conference on, vol., no., pp.52,57, 2-6 June 2013.

Published Abstracts

Saunders, C.P., Hepler, A.B., Davis, L. J., Buscaglia, J. (2010). Estimation of likelihood ratios for forensic handwriting analysis. *Science & Justice*, *Volume 50, Issue 1*, March 2010, Page 32.

Hepler, A.B. and Saunders, C.P. (2010). A Comparison between different likelihood ratios for assessing handwriting evidence. *2010 Proceedings of American Academy of Forensic Sciences*, Extended Abstract.

Gantz, D., Miller, J., Saunders, C., Walch, M, and Buscaglia, J. (2010). New results for addressing the open set problem in automated handwriting identification. *2010 Proceedings of American Academy of Forensic Sciences*, Extended Abstract.

Saunders, C., Buscaglia, J., Davis, L., and Lancaster, J. (2009). Handwriting Individuality: Probability Models, Subsampling Routines, and Implications. *2009 Proceedings of American Academy of Forensic Sciences*, Extended Abstract.

Saunders, C., Davis, L., Lamas, A., and Buscaglia, J. (2008). A Comparison Between Biometric and Forensic Handwriting Individuality. *2008 Proceedings of American Academy of Forensic Sciences*, Extended Abstract.

Walch, M., Gantz, D., Miller, J., Davis, L., Saunders, C., Lancaster, M., Lamas, A., and Buscaglia, J. (2008). Evaluation of the Individuality of Handwriting Using FLASH ID –A Totally Automated, Language Independent System for Handwriting Identification. *2008 Proceedings of American Academy of Forensic Sciences*, Extended Abstract.

Gantz, D., Miller, J., Saunders, C., Lancaster, M., and Buscaglia, J. (2008). Statistical Characterization of Writers for identification. *2008 Proceedings of American Academy of Forensic Sciences*, Extended Abstract.

Service

Referee for the *Journal of Forensic Science, Forensic Science International, Annals of Applied Statistics, Journal of Statistical Computation and Simulation, and various IEEE publications*.

Provides support to the FBI Research Laboratory as well as the broader Intelligence Community by reviewing and commenting on the statistical components of various research projects.

Provides support to the Latent Fingerprint Unit of the FBI Research Laboratory by reviewing courses and literature relevant to the interpretation of latent fingerprint evidence and ensuring this information is relevant to the National Academy of Science (NAS) report entitled: *Strengthening Forensic Science in the United States: A Path Forward*.

Served as an instructor in a statistical course for the Latent Print unit at the FBI Laboratory Division.

Serves as a member of a panel for the Division Mathematical Sciences at the National Science Foundation.

Served as a discussant for the Scientific Working Group for Shoeprint and Tire Tread Evidence September 2011 meeting.

Serves as a Technical Advisor to the IARPA.

Currently serving on the Standing Scientific Review Panel for National Institute of Justice.

Currently serving on the ad-hoc Advisory Committee on Forensic Science for the American Statistical Society. This includes additional support to the FBI labs with respect to statistical aspects of forensic science.

Provided statistical and engineering support to the Veteran's Administration as part of a team of MITRE engineers prototyping software and algorithms that support situational awareness for crisis management and disaster preparedness. I received a commendation for my contribution to this effort.

Presentations

Jan., 2005 "The Weak Convergence of Empirical Processes from Multivariate Normal Vectors," Hawaii International Conference on Statistics, Honolulu, HI., Contributed Talk.

Aug., 2005 "The Weak Convergence of Empirical Processes from Multivariate Normal Vectors for Goodness-of-fit Tests," Joint Statistical Meeting, Minneapolis, MN., Contributed Talk.

Jan., 2006 "Transcriptional profiling of equine chondrocytes under hypoxic culture conditions," Plant and Animal Genome XIV. 2006.1, Contributed Poster, with Miura N. (Presenter), Huang L, Stromberg AJ, MacLeod JN.

Mar., 2006 "The Asymptotic Distribution of Modified Shapiro-Wilk Statistics for Testing Multivariate Normality," The Eastern North American Region of the International Biometric Society Spring Meeting, Tampa, FL., Contributed Talk.

Jan., 2007 "Identification of genes with a cartilage-restricted pattern expression." Plant and Animal Genome XV. 2007.1, Contributed Poster, with Zhu W (Presenter), Huang L, Saunders CP, Bathke A, Stromberg AJ, MacLeod JN.

Mar., 2007 "Empirical Processes for Estimated Projections of Multivariate Normal Vectors with Applications to E.D.F. and Correlation type Goodness-of-Fit tests," George Mason University, Statistics Department, Invited Talk.

April, 2007 "A Categorical Based Approach to Biometric Handwriting Identification," Seventh Annual IC Postdoctoral Research Fellowship Colloquium, Chantilly, Va., Invited Poster.

April, 2007 "FDR-Based Ensemble Learners for High Dimensional Data," Seventh Annual IC Postdoctoral Research Fellowship Colloquium, Chantilly, Va., Invited Poster, with Mark J. Lancaster (Presenter).

Aug., 2007 "Classifiers from Categorical Data for Forensic Document Analysis," 2007 Joint Statistical Meetings, Salt Lake, UT, Contributed Poster, with Andrea C. Lamas, Linda J. Davis, and John J. Miller.

Aug., 2007 "A Novel Application of Quantile Regression for Identification of Cartilage Biomarkers in Equine Microarray Data," 2007 Joint Statistical Meetings, Salt Lake, UT, Contributed Poster, with Liping Huang (Presenter), Wenying Zhu, James N. MacLeod, Arnold J. Stromberg, Arne Bathke.

Feb., 2008 "A Comparison Between Biometric and Forensic Handwriting Individuality," 2008 American Academy of Forensic Sciences Annul Meeting, Contributed Poster with Davis, L., Lamas, A., and Buscaglia, J.

Feb., 2008 "Evaluation of the Individuality of Handwriting Using FLASH ID –A Totally Automated, Language Independent System for Handwriting Identification," 2008 American Academy of Forensic Sciences Annul Meeting, Contributed Talk, Mark A. Walch, Donald T. Gantz, John J. Miller, Linda J. Davis, Christopher P. Saunders, Mark J. Lancaster, Andrea Lamas, and JoAnn Buscaglia.

Feb., 2008 "Statistical Characterization of Writers for Identification," 2008 American Academy of Forensic Sciences Annul Meeting, Contributed Talk, Donald T. Gantz, John J. Miller, Christopher P. Saunders, Mark J. Lancaster, and JoAnn Buscaglia.

Nov., 2008 "A Comparison Between Biometric and Forensic Handwriting Individuality." The ROBUST Biometrics Conference. Contributed Poster with Davis, L., Lamas, A., and Buscaglia, J.

Aug., 2008 "Modeling the Relationship Between Random Match/Non-Match Probabilities and the Sizes of Writing Samples." The Seventh International Conference on Forensic Inference and Statistics. Contributed talk with Davis, L. and Buscaglia, J.

Feb., 2009 "Handwriting Individuality: Probability Models, Subsampling Routines, and Implications." American Academy of Forensic Sciences. Contributed Talk with Buscaglia, J., Davis, L., Hepler, A., and Lancaster, J.

Sep., 2009 "Estimation of likelihood ratios for forensic handwriting analysis." 5th European Academy of Forensic Science Conference. Submitted as a Poster, *invited to present as a talk*. Saunders C. P.

Feb, 2010 "A Comparison between different likelihood ratios for assessing handwriting evidence." American Academy of Forensic Sciences. Contributed Talk, Hepler A. and Saunders C.

Feb, 2010 "New results for addressing the open set problem in automated handwriting identification." American Academy of Forensic Sciences. Contributed Talk, Gantz D. T., Miller J. J., Saunders C P., Walch M. A., and Buscaglia J.

Feb, 2010 "Quantifying the effects of database size and sample quality on measures of individualization validity and accuracy in forensics." 2010 General Forensics R&D Grantees Meeting. Saunders C.P, *invited talk.*

Aug., 2010 "Estimation of likelihood ratios for forensic handwriting analysis." NIJ Impression and Pattern Evidence Symposium. submitted poster, Saunders C. P. and Hepler A. H.

Aug., 2010 "Quantifying the effects of database size and sample quality on measures of individualization validity and accuracy in forensics." NIJ Impression and Pattern Evidence Symposium. Saunders C.P, *invited poster.*

Aug., 2011 "Predictive Modeling for Determining the Discriminative Power of Trace Glass Evidence as a Function of the Number of Sampled Glass Fragments." NIJ Trace Evidence Symposium. Kalendra, E. (Presenter) Saunders C.P., *invited talk.*

Aug., 2011 "ROC Curves for Methods of Evaluating Evidence: A Common Performance Measure Based on Similarity Scores." NIJ Trace Evidence Symposium. R. Bradley Patterson (Presenter), John J. Miller, Christopher P. Saunders, *invited talk.*

Aug., 2011 "On Parametric Models for Pairwise Comparisons with Applications to Estimation of Random Match Probabilities." NIJ Trace Evidence Symposium. Donald T. Gantz (Presenter), John J. Miller, Christopher P. Saunders, *invited talk.*

Aug., 2012 "The Effect of the Order of Suspect and Background Population Samples on the Assessment of the Value of Evidence." NIJ Impression and Pattern Evidence Symposium- Invited Talk. Eric Kalendra (Presenting and supported travel), Buscaglia J. and Saunders C.P.

Aug., 2012 "A note on the value of forensic evidence for sparse categorical tables." NIJ Impression and Pattern Evidence Symposium- Poster. Krista Heim (Presenting and supported travel), Buscaglia J. and Saunders C.P.

Aug., 2012 "The Effect of Uncertainty About the Alternative Source Population on the Assessment of the Value of Forensic Evidence." Joint Statistical Meetings. Poster Presentation, Eric Kalendra, Buscaglia J. and Saunders CP (Presenting).

Aug., 2012 "The Effect of Uncertainty About the Alternative Source Population on the Assessment of the Value of Forensic Evidence." EAFS2012. Presentation, Eric Kalendra, Buscaglia J. and Saunders CP (Presenting).

Nov., 2012 "Algorithm for Spectroscopic Data Analysis and Outlier Detection." NSF Algorithms for Threat Detection Workshop. Presentation, Numerous authors, the talk was focused on algorithms I developed.

Nov., 2012 "Computer Vision and Statistical Learning on a Budget." NSF Algorithms for Threat Detection Workshop. Presentation, Lancaster ML and Saunders CP (Co-Presenting).

June, 2013 "Understanding Individuality of Handwriting Using Score-Based Likelihood Ratios." NIST, Measurement Science and Standards in Forensic Handwriting Analysis. Invited Presentation.

June, 2013 "Scale Invariant Feature Transform (SIFT): Writer Recognition by Computer Vision." NIST, Measurement Science and Standards in Forensic Handwriting Analysis. Invited presentation with J. Woodard and L. Lancaster, Saunders Presenting.

July, 2013 "Statistical Aspects of the Forensic Identification of Source Problem." Invited Seminar Lecture at the FBI Laboratory, Quantico, VA.

Aug., 2013 "On Desiderata for Score-Based Likelihood Ratios for Forensic Evidence." Joint Statistical Meetings 2013, Montreal Canada, Christopher Saunders (Presenting) and John J. Miller.

Funded Research Support and Awards

Funded as a research assistant professor by Gannon Technologies (with Donald T. Gantz-PI and John J. Miller-co-PI), 2008-2009, George Mason University.

Intelligence Community Postdoctoral Research Fellowship Award, Donald T. Gantz (PI), $240,000, 2006-2008, George Mason University.

Graduate Research Assistant in Thomas Getchell's Lab in the Sanders Brown Center on Aging, National Institute of Health Training Grant, 2005, University of Kentucky.

Microarray Core Facility Graduate Research Assistant, Arnold Stromberg (co-PI), 2003-2006 University of Kentucky.

Grants and Awards (Saunders as P.I. or co. P.I.)

Intelligence Community Postdoctoral Research Fellowship Award, Eric Kalendra as Research Fellow, Christopher Saunders PI, **$240,000**, 2010-2012.

Quantifying the Effects of Database Size and Sample Quality on Measures of Individualization Validity and Accuracy in Forensics, National Institute of Justice Grant Award, Christopher Saunders PI, Linda Davis co-PI**, $974,981**, 2010-2014.

Decomposition-Based Information Elicitation & Aggregation, Aggregative Contingent Estimation (ACE) Program, Broad Agency Announcement IARPA-BAA-10-05, Charles Twardy PI, Kathryn B Laskey co-PI, Christopher Saunders co-PI, Approximately **$8,200,000** over four years. (I had to step down as co-PI to accept a position with MITRE).

ORISE Visiting Scientist Award, summer funding to support the work of the Counterterrorism and Forensic Science Research Unit of the FBI Labs, Summer 2013.

Contact Information

Christopher P. Saunders
Department of Mathematics and Statistics
Harding Hall 213
South Dakota State University
Brookings, SD 57007

Cell: 703-944-6043
Email: Christopher.saunders@sdstate.edu

Appendix 1

Link Between *U*-Statistics With 0-1 Kernels and Union/Intersection of Events

By: Linda J. Davis

ABSTRACT: This paper illustrates a connection between the distribution of a *U*-statistic with a 0-1 kernel of degree 2 and the probability formulas for unions and intersections of events. This connection provides one way to derive formulas for expectations and means as well as bounds on probabilities related to the distribution of a *U*-statistic with a 0-1 kernel.

1. Introduction

The distribution of a random count depends upon how the count is generated.

One common example of a random count involves a fixed number n of independent trials on each of which some event either occurs or does not occur. The distribution of the number of occurrences is binomial with parameter n and π, where π is the probability of occurrence of the event on a single trial (which is assumed constant across trials).

One way to represent such a count is via a set of n random variables $\{X_1, X_2, ..., X_n\}$ and a binary function of one variable $\omega(\cdot)$: $\omega(X_i) = 1$ if the event does occur (often called a success) on the *i*th trial and $\omega(X_i) = 0$ if the event does not occur (often called a failure) on the *i*th trial. Then, $\sum_{i=1}^{n} \omega(X_i)$ equals the number of successes across n trials and has a binomial distribution provided the set of n random variables are independent and identically distributed (*iid*).

There are a number of ways to "complicate" this scenario while still being interested in the number of occurrences across some fixed number of trials. First, one could envision that the probability of occurrence on a single trial varies from trial to trial. The effect on the distribution of the number of occurrences depends upon how the probabilities vary. Alternatively, one could consider situations in which the trials are not independent, so called correlated binomial trials. In this scenario, the effect on the distribution of the number of occurrences depends upon how the trials are related.

In this paper, I consider a set of correlated trials where the correlation is of a specific type – namely that induced by individual "trials" involving pairs of independent random variables. For example, one might envision comparing the two random variables within a pair and deciding on the basis of some criteria that some event either occurs or does not occur for the pair. Blom and Holst (1989) use the term *similar pair* to describe a pair for which some event does occur. One may then be interested in the distribution of the number of similar pairs, that is, of the number of pairs for which some event occurs.

One way to represent such a count is via a binary function on two variables $\psi(\cdot,\cdot)$: $\psi(X_i, X_j) = 1$ if the (i, j)th pair is similar and $\psi(X_i, X_j) = 0$ if the (i, j)th pair is not similar. Then, the number of similar pairs among a set of n *iid* random variables $\{X_1, X_2, ..., X_n\}$ where $n \geq 2$ can be represented as:

$$Y_n \equiv \sum_{i<j} \psi(X_i, X_j)$$

where $\sum_{i<j}$ represents $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n}$ or equivalently, $\sum_{j=2}^{n} \sum_{i=1}^{j-1}$ and the proportion of similar pairs is:

$$U_n \equiv N^{-1} Y_n = N^{-1} \sum_{i<j} \psi(X_i, X_j)$$

where $N \equiv \binom{n}{2}$, i.e., the number of pairs of integers (i, j) with $1 \leq i < j \leq n$. Note that U_n is a *U*-statistic with a 0-1 kernel $\psi(\cdot,\cdot)$ of degree 2 (Lee, 1990, p. 8); the relationship to a *U*-statistic can be exploited to establish both distributions and bounds on probabilities.

The question I address in this paper is: What is the distribution of Y_n (or equivalently, U_n)? My interest in this question arose out of studying the inferences that can be made from the number of "matches" observed in databases of forensic evidence, such as fingerprints and handwritten documents.

This paper is organized as follows. First, I list formulas for the moments of Y_n that follow directly from the relationship of Y_n to a *U*-statistic. These formulas relate the moments of Y_n to two probabilities:

$$\delta_1 \equiv P[\psi(X_1, X_2) = 1]$$

$$\delta_2 \equiv P[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1]$$

Here, δ_1 is the probability that two independent random variables are similar, where similar is defined as $\psi(X_1, X_2) = 1$. δ_2 is the probability that three independent random variables include at least two similar pairs, namely the probability that two independent random variables X_2 and X_3 are each similar to a third independent random variable X_1, i.e., $\psi(X_1, X_2) = 1$ and $\psi(X_1, X_3) = 1$.

Next I provide some formulas for the probability mass function of Y_n that are related to the probability formulas for unions and intersections of events. Using the relationship to such formulas, I derive bounds on the probability mass function for Y_n related to moments of Y_n.

I conclude this paper with some applications of the developed formulas to specific distributions of the X's.

2. Moments of the Proportion of Similar Pairs

Because $\psi(\cdot,\cdot)$ takes on only the values of 0 and 1, δ_1 and δ_2 can be represented as expectations:

$$E\left[\psi^2(X_1,X_2)\right] = E\left[\psi(X_1,X_2)\right] = P\left[\psi(X_1,X_2)=1\right] = \delta_1$$

$$E\left[\psi(X_1,X_2)\psi(X_1,X_3)\right] = P\left[\psi(X_1,X_2)=1,\psi(X_1,X_3)=1\right] = \delta_2 \ .$$

So, in terms of δ_1 and δ_2, the first two moments of $\psi(X_1,X_2)$ are:

$$E\left[\psi(X_1,X_2)\right] = \delta_1$$
$$\mathrm{Var}\left[\psi(X_1,X_2)\right] = E\left[\psi^2(X_1,X_2)\right] - \left\{E\left[\psi(X_1,X_2)\right]\right\}^2 = \delta_1 - \delta_1^2 = \delta_1(1-\delta_1) \ .$$

Note that from the properties of *U*-statistics (Lee, 1990, p. 8), U_n is an unbiased estimator of δ_1, i.e.,

$$E(U_n) = \delta_1 \ . \tag{1}$$

Also, since $\{X_1, X_2, ..., X_n\}$ are assumed to be *iid*, the variance of U_n is (Lee, 1990, p. 12)[1]:

$$\mathrm{Var}(U_n) = N^{-1}\left[2(n-2)\xi_1 + \xi_2\right] \tag{2}$$

where

$$\begin{aligned} \xi_1 &\equiv \mathrm{Cov}\left[\psi(X_1,X_2),\psi(X_1,X_3)\right] \\ &= E\left[\psi(X_1,X_2)\psi(X_1,X_3)\right] - \left\{E\left[\psi(X_1,X_2)\right]\right\}^2 \\ &= \delta_2 - \delta_1^2 \end{aligned} \tag{3}$$

$$\begin{aligned} \xi_2 &\equiv \mathrm{Cov}\left[\psi(X_1,X_2),\psi(X_1,X_2)\right] \\ &= \mathrm{Var}\left[\psi(X_1,X_2)\right] \\ &= \delta_1(1-\delta_1) \end{aligned} \tag{4}$$

Substituting (3) and (4) into (2),

$$\begin{aligned} \mathrm{Var}(U_n) &= N^{-1}\left[2(n-2)(\delta_2-\delta_1^2) + \delta_1(1-\delta_1)\right] \\ &= N^{-1}\left[2(n-2)\delta_2 + \delta_1 - (2n-3)\delta_1^2\right] \\ &= N^{-1}\left[2(n-2)(\delta_2-\delta_1) + (2n-3)\delta_1(1-\delta_1)\right] \end{aligned} \tag{5}$$

Relationships involving ξ_1 and ξ_2 can be used to derive relationships involving δ_1 and δ_2. First, $\xi_1 \geq 0$ (Lee, 1990, p. 10). That is, the covariance of $\psi(X_1,X_2)$ and $\psi(X_1,X_3)$ is non-negative, as expected due to the common term X_1. However, as shown in (3), $\xi_1 = \delta_2 - \delta_1^2$. Thus, $\xi_1 \geq 0$ implies $\delta_1^2 \leq \delta_2$.

Second, $\xi_1 \leq \xi_2/2$ (Lee, 1990, p. 15). So, substituting the formulas for ξ_1 and ξ_2 given in (3) and (4), $2(\delta_2 - \delta_1^2) = 2\xi_1 \leq \xi_2 = \delta_1(1-\delta_1)$. Rewriting, $\delta_2 \leq \delta_1(1+\delta_1)/2$. Also, since δ_1 is a probability, $\delta_1 \leq 1$ which

[1] In (Lee, 1990), ξ_1 is denoted by σ_1^2 and ξ_2 is denoted by σ_2^2.

implies $(1+\delta_1)/2 \le 1$ and thus that $\delta_1(1+\delta_1)/2 \le \delta_1$. So, viewing the proportion of similar pairs as a *U*-statistic leads to the following bounds on δ_2 in terms of δ_1.

$$\delta_1^2 \le \delta_2 \le \frac{\delta_1(1+\delta_1)}{2} \le \delta_1 \tag{6}$$

These relationships involving δ_1 and δ_2 lead to insightful bounds on the variance of U_n. Applying the relationship $\delta_1^2 \le \delta_2$ to (5),

$$\begin{aligned}\mathrm{Var}(U_n) &= N^{-1}\left\lfloor 2(n-2)(\delta_2 - \delta_1^2) + \delta_1(1-\delta_1)\right\rfloor \\ &\ge N^{-1}\left[2(n-2)(\delta_1^2 - \delta_1^2) + \delta_1(1-\delta_1)\right] = \frac{\delta_1(1-\delta_1)}{N}\end{aligned}$$

Thus, the variance of U_n is bounded below by the variance of the proportion of successes in a binomial distribution with the number of trials equal to N and success probability on each trial equal to δ_1. In other words, the variance of the proportion of similar pairs is bounded below by the variance of a sample proportion associated with treating the $N \equiv \binom{n}{2}$ pairwise comparisons as N independent trials, each with probability of success δ_1.

Note that the variance of U_n equals the binomial variance when $\delta_2 = \delta_1^2$, i.e., when

$$P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1\right] = \delta_2 = \delta_1^2 = P\left[\psi(X_1, X_2) = 1\right] \times P\left[\psi(X_1, X_3) = 1\right].$$

This shows that treating the number of similar pairs in N pairwise comparisons as having a binomial distribution underestimates the variance whenever X_1 being similar to one other observation, say X_2, increases its chances of being similar to yet another observation, say X_3.

Applying the relationship $\delta_2 \le \delta_1(1+\delta_1)/2$ to (6),

$$\begin{aligned}\mathrm{Var}(U_n) &= N^{-1}\left\lfloor 2(n-2)(\delta_2 - \delta_1^2) + \delta_1(1-\delta_1)\right\rfloor \\ &\le N^{-1}\left\{2(n-2)\left[\delta_1(1+\delta_1)/2 - \delta_1^2\right] + \delta_1(1-\delta_1)\right\} = \frac{\delta_1(1-\delta_1)}{n/2}\end{aligned}$$

Thus, the variance of U_n is bounded above by the variance of a binomial distribution with number of trials equal to $n/2$ and success probability on each trial equal to δ_1. This corresponds to $n/2$ independent trials, which would occur if the $X's$ are paired (assuming *n* is even) and only the resulting $n/2$ pairs were compared. This shows that treating the number of similar pairs in N pairwise comparisons as $n/2$ independent trials overestimates the variance.

3. Distribution of Number of Similar Pairs

Consider the discrete random variables $Y_n = \sum_{i<j} \psi(X_i, X_j) = NU_n$, which takes on values $0,1,...,N$. Recall that Y_n is the number of similar pairs, i.e., the number of (i, j) pairs $(1 \le i < j \le n)$ with $\psi(X_i, X_j) = 1$.

3.1 Moments of Y_n

From (1) and (5),

$$E(Y_n) = N\delta_1 \tag{7}$$

$$\begin{aligned} \mathrm{Var}(Y_n) &= N\left\lfloor 2(n-2)\delta_2 + \delta_1 - (2n-3)\delta_1^2 \right\rfloor \\ &= N\left[2(n-2)(\delta_2 - \delta_1) + (2n-3)\delta_1(1-\delta_1)\right] \end{aligned} \tag{8}$$

and

$$\begin{aligned} E(Y_n^2) &= \mathrm{Var}(Y_n) + [E(Y_n)]^2 \\ &= N\left[2(n-2)\delta_2 + \delta_1 + 0.5(n-2)(n-3)\delta_1^2 \right] \\ &= 6\binom{n}{3}\delta_2 + N\delta_1 + 6\binom{n}{4}\delta_1^2 \end{aligned} \tag{9}$$

(Note: An alternative derivation of the formulas for the mean and variance of Y_n appear in Blom and Holst (1989); they use different notation: $\delta_1 \Rightarrow p$, $N \Rightarrow M$, and $\delta_2 - \delta_1^2 \Rightarrow c$.)

The expression for $E(Y_n^3)$ is more complex than that for the first and second moments. A technique similar to that used in Lee (1990, p.12) to derive a general formula for the variance of a *U*-statistic, can be used to derive a formula for the third non-central moment:

$$\begin{aligned} E(Y_n^3) &= N\left\lfloor 0.25(n-2)(n-3)(n-4)(n-5)\delta_1^3 + 1.5(n-2)(n-3)\delta_1^2 + \delta_1 \right. \\ &\quad +3(n-2)(n-3)(n-4)\delta_1\delta_2 + 2(n-2)(3\delta_2 + \delta_{2b}) \\ &\quad \left. +2(n-2)(n-3)(3\delta_{3a} + \delta_{3b})\right] \\ &= 90\binom{n}{6}\delta_1^3 + 18\binom{n}{4}\delta_1^2 + N\delta_1 + 180\binom{n}{5}\delta_1\delta_2 + 6\binom{n}{3}(3\delta_2 + \delta_{2b}) + 24\binom{n}{4}(3\delta_{3a} + \delta_{3b}) \end{aligned} \tag{10}$$

where for $n \ge 3$[2]:

$$\delta_{2b} \equiv P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1, \psi(X_2, X_3) = 1\right] = E\left[\psi(X_1, X_2)\psi(X_1, X_3)\psi(X_2, X_3)\right]$$

and for $n \ge 4$[3]:

$$\delta_{3a} \equiv P\left[\psi(X_1, X_2) = 1, \psi(X_3, X_4) = 1, \psi(X_1, X_3) = 1\right] = E\left[\psi(X_1, X_2)\psi(X_3, X_4)\psi(X_1, X_3)\right]$$

---

[2] For $n < 3$, $\delta_2 \equiv 0$ and $\delta_{2b} \equiv 0$.

[3] For $n < 4$, $\delta_{3a} \equiv 0$ and $\delta_{3b} \equiv 0$.

$$\delta_{3b} \equiv P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1, \psi(X_1, X_4) = 1\right] = E\left[\psi(X_1, X_2)\psi(X_1, X_3)\psi(X_1, X_4)\right].$$

δ_{2b}, δ_{3a}, and δ_{3b} can be interpreted as probabilities of certain numbers of similar pairs. Recall that δ_1 is the probability that two independent random variables are similar; and δ_2 is the probability that three independent random variables include at least two similar pairs.

Similarly, δ_{2b} is the probability of three similar pairs among three independent random variables. From the interpretation (or the definition), it is clear that $\delta_{2b} \leq \delta_2$ with equality if and only if two pairs being similar implies the third pair also must be similar.

The interpretations for δ_{3a} and δ_{3b} are more difficult to put in words as they involve four independent random variables. δ_{3a} is the probability that a specific pair (say, (X_1, X_3)) of the four independent random variables is similar and the remaining two random variables (say, X_2 and X_4) are each similar to a different member of this specific pair (say, X_2 is similar to X_1, and X_4 is similar to X_3). δ_{3b} is the probability that a specific one of the four independent random variables is similar to the other three.

Expressions for higher-order moments can in principle be computed using the technique shown in Lee (1990, p. 12). However, as suggested by the expression for $E\left(Y_n^3\right)$, these do not assume a simple form.

3.2 Probability Distribution of Y_n

To derive expressions for the distribution of Y_n, I relate its distribution to probability formulas associated with the union and intersection of events.

Define

$$B_{ij} \equiv \left\{\psi(X_i, X_j) = 1\right\} \quad \text{for } i < j, \ 1 \leq i < j \leq n \ .$$

In terms of these events,

$$Y_n = \sum_{i<j} I\{B_{ij}\}$$

where $I\{\cdot\}$ is the indicator function:

$$I\{B\} = \begin{cases} 1 & \text{if event } B \text{ occurs} \\ 0 & \text{if event } B \text{ does not occur} \end{cases} .$$

In other words, Y_n equals the number of the N events $\left\{B_{ij}, 1 \leq i < j \leq n\right\}$ that occur. Viewing Y_n in this manner allows relating the computation of the distribution of Y_n to probability formulas for unions and intersections of events.

It is convenient to map the index "pair" (i, j) to a single index k via say:

$$k(i,j) \equiv (j-1)(j-2)/2 + i \text{ for } i < j .^{4} \tag{11}$$

This allows working with a set of events with a single index k: $A_k \equiv A_{k(i,j)} = B_{ij}$. So, in the following, I consider a set of N events $\{A_1, A_2, \ldots, A_N\}$ with:

$$Y_n = \sum_{k=1}^{N} I\{A_k\}$$

A key issue to keep in mind in the following discussion is that the N events $\{A_1, A_2, \ldots, A_N\}$ are not independent. So, for example $P\left(A_{k_1} A_{k_2}\right)$ [5] may or may not equal $P\left(A_{k_3} A_{k_4}\right)$.

For $1 \le r \le N$, define[6]:

$$S_r \equiv \sum_{(N,r)} P\left(A_{k_1} A_{k_2} \cdots A_{k_r}\right) \tag{12}$$

where the sum $\sum_{(N,r)}$ is taken over all $\binom{N}{r}$ r-tuples $(k_1, k_2, \ldots, k_r)$ of integers between 1 and N inclusive with $1 \le k_1 < \cdots < k_r \le N$ [7]. By definition, $S_0 \equiv 1$ and $S_r \equiv 0$ for $r > N$. Examples for some other values of r include:

$$S_1 = \sum_{k=1}^{N} P(A_k)$$

$$S_2 = \sum_{k_1 < k_2} P\left(A_{k_1} A_{k_2}\right)$$

$$\vdots$$

$$S_N = P\left(\bigcap_{k=1}^{N} A_k\right)$$

Consider the following result from Feller (1968, p. 99) [8]: The probability P_1 of the realization of at least one among the events $A_1, A_2, \ldots, A_N$ equals:

$$P_1 = S_1 - S_2 + S_3 - S_4 + \cdots \pm S_N .$$

But, $P_1 = P(Y_n \ge 1)$. So,

---

[4] The actual mapping used is not important.

[5] For two events E and F, the notation EF is used for the intersection of the two events.

[6] Much of the following notation is adapted from Chapter IV of Feller (1968).

[7] The sum notation is adapted from Lee (1990, p. 7).

[8] Most of the results quoted from Feller are not Feller's original work. References to the original work are available within Feller. I chose to reference Feller because Chapter IV of his book provides a useful summary of the formulas that has been developed concerning combinations of events, and some of the original work is difficult to acquire.

$$P(Y_n \geq 1) = S_1 - S_2 + S_3 - S_4 + \cdots \pm S_N = \sum_{r=1}^{N} (-1)^{r+1} S_r$$

and using $S_0 \equiv 1$,

$$P(Y_n = 0) = 1 - P(Y_n \geq 1) = 1 - \sum_{r=1}^{N} (-1)^{r+1} S_r = \sum_{r=0}^{N} (-1)^r S_r$$

Formulas for $P(Y_n = m)$ for values of m greater than 0 can be derived from the following result in Feller (1968, p. 106): For any integer m with $1 \leq m \leq N$, the probability $P_{[m]}$ that exactly m among the N events $A_1, A_2, \ldots, A_N$ occur simultaneously is:

$$P_{[m]} = S_m - \binom{m+1}{m} S_{m+1} + \binom{m+2}{m} S_{m+2} - \cdots \pm \binom{N}{m} S_N = \sum_{r=0}^{N-m} (-1)^r \binom{m+r}{m} S_{m+r} .$$

But, $P_{[m]} = P(Y_n = m)$. So, for $1 \leq m \leq N$,

$$P(Y_n = m) = \sum_{r=0}^{N-m} (-1)^r \binom{m+r}{m} S_{m+r} .$$

Note that this formula also holds for $m = 0$.

Finally, formulas for $P(Y_n \geq m)$ for values of m greater than 1 can be derived using the following formula from Feller (1968, p. 109). The probability P_m that m or more of the events $A_1, A_2, \ldots, A_N$ occur simultaneously is:

$$P_m = S_m - \binom{m}{m-1} S_{m+1} + \binom{m+1}{m-1} S_{m+2} - \binom{m+2}{m-1} S_{m+3} + \cdots \pm \binom{N-1}{m-1} S_N .$$

So,

$$P(Y_n \geq m) = P_m = \sum_{r=0}^{N-m} (-1)^r \binom{m-1+r}{m-1} S_{m+r} .$$

Note that this formula also holds for $m = 1$.

In summary, the distribution of Y_n can be expressed in terms of $\{S_0, S_1, S_2, \ldots, S_N\}$. For any integer m with $0 \leq m \leq N$, the probability mass function of Y_n is:

$$P(Y_n = m) = \sum_{r=0}^{N-m} (-1)^r \binom{m+r}{m} S_{m+r} \qquad (13)$$

and the complementary cumulative distribution function of Y_n is:

$$P(Y_n \geq m) = \sum_{r=0}^{N-m} (-1)^r \binom{m-1+r}{m-1} S_{m+r} \quad [9] \tag{14}$$

3.3 Probability Bounds

The form of the expression for $P(Y_n = m)$ in (13) and for $P(Y_n \geq m)$ in (14) allows some Bonferroni-type inequalities. As stated in Feller, (1968, p. 110), if one approximates $P(Y_n = m)$ (or $P(Y_n \geq m)$) by dropping the terms involving $\{S_{m+t}, S_{m+t+1}, \ldots, S_N\}$ in either (13) or (14), then the sign of the "error" (i.e., true value minus approximation) is that of the first omitted term. Specifically, for any integer m with $0 \leq m \leq N$ and any <u>even</u> integer t with $0 \leq t \leq N-m-1$,

$$\sum_{r=0}^{t+1} (-1)^r \binom{m+r}{m} S_{m+r} \leq P(Y_n = m) \leq \sum_{r=0}^{t} (-1)^r \binom{m+r}{m} S_{m+r} \tag{15}$$

$$\sum_{r=0}^{t+1} (-1)^r \binom{m-1+r}{m-1} S_{m+r} \leq P(Y_n \geq m) \leq \sum_{r=0}^{t} (-1)^r \binom{m-1+r}{m-1} S_{m+r} \tag{16}$$

In particular, for $t = 0$,

$$S_m - (m+1)S_{m+1} \leq P(Y_n = m) \leq S_m$$

$$S_m - mS_{m+1} \leq P(Y_n \geq m) \leq S_m$$

and for $t = 2$,

$$S_m - (m+1)S_{m+1} + \binom{m+2}{m} S_{m+2} - \binom{m+3}{m} S_{m+3}$$
$$\leq P(Y_n = m) \leq S_m - (m+1)S_{m+1} + \binom{m+2}{m} S_{m+2}$$

$$S_m - mS_{m+1} + \binom{m+1}{m-1} S_{m+2} - \binom{m+2}{m-1} S_{m+3}$$
$$\leq P(Y_n \geq m) \leq S_m - mS_{m+1} + \binom{m+1}{m-1} S_{m+2}$$

The special case of $m = 0$ leads to a bound on the probability of zero similar pairs. Substituting $m = 0$ in (15), for any <u>even</u> integer t with $0 \leq t \leq N-1$,

$$\sum_{r=0}^{t+1} (-1)^r S_r \leq P(Y_n = 0) \leq \sum_{r=0}^{t} (-1)^r S_r$$

---

[9] For the case $m = 0$, I am using the convention that $\binom{a}{b} \equiv 0$ whenever a is a nonnegative integer and b is a negative integer; and the convention that $\binom{-1}{-1} \equiv 1$.

So,

$$P(Y_n = 0) \geq \sum_{r=0}^{t} (-1)^r S_r \quad \text{for } t \text{ odd, } 1 \leq t \leq N-1$$

$$\leq \sum_{r=0}^{t} (-1)^r S_r \quad \text{for } t \text{ even, } 0 \leq t \leq N-1$$

$$= \sum_{r=0}^{t} (-1)^r S_r \quad \text{for } t = N$$

Substituting $t = 1$ and $t = 2$, and recalling that $S_0 = 1$,

$$1 - S_1 \leq P(Y_n = 0) \leq 1 - S_1 + S_2 \text{ (provided } N \geq 2).$$

Substituting $t = 3$ and $t = 4$, and recalling that $S_0 = 1$,

$$1 - S_1 + S_2 - S_3 \leq P(Y_n = 0) \leq 1 - S_1 + S_2 - S_3 + S_4 \text{ (provided } N \geq 4).$$

3.4 Values of S_r

The probability bounds in the previous section require knowing the values of S_r. Recall that $S_r \equiv \sum_{(N,r)} P\left(A_{k_1} A_{k_2} \cdots A_{k_r}\right)$ for any integer r with $1 \leq r \leq N$. Thus, one could compute S_r directly from the set of probabilities $P\left(A_{k_1} A_{k_2} \cdots A_{k_r}\right)$ for all $\binom{N}{r}$ r-tuples $(k_1, k_2, \ldots, k_r)$ of integers between 1 and N inclusive with $1 \leq k_1 < \cdots < k_r \leq N$.

As pointed out earlier, the A_k's are not independent. So, the terms $P\left(A_{k_1} A_{k_2} \cdots A_{k_r}\right)$ are not all equal. The assumed independence of the underlying $\{X_1, X_2, ..., X_n\}$ means that subsets of these probabilities are the same. However, counting the number of $P\left(A_{k_1} A_{k_2} \cdots A_{k_r}\right)$ with the same value complicates the computation of S_r, particularly for larger values of r.

For $r = 1$, the computation is straightforward. For any k, which corresponds via (11) to a unique index pair (i, j), $P(A_k) = P\lfloor \psi(X_i, X_j) = 1 \rfloor = \delta_1$. So, $S_1 = N\delta_1$.

For $r = 2$, the computation is not as straightforward and serves to illustrate the combinatorics involved in computing the S_r's for $r > 1$. Each of the terms $P\left(A_{k_1} A_{k_2}\right)$ is associated via (11) with two index pairs (i_1, j_1) and (i_2, j_2);

$$P\left(A_{k_1} A_{k_2}\right) = P\left[\psi(X_{i_1}, X_{j_1}) = 1, \psi(X_{i_2}, X_{j_2}) = 1\right]$$

and

$$S_2 = \sum_{\substack{i_1<j_1,\, i_2<j_2 \\ k(i_1,j_1)<k(i_2,j_2)}}{}' P\left[\psi(X_{i_1}, X_{j_1}) = 1, \psi(X_{i_2}, X_{j_2}) = 1\right]$$

where $k(i, j)$ is defined in (11).

The difficulty in computing this sum is that the value of $P\left[\psi(X_{i_1}, X_{j_1}) = 1, \psi(X_{i_2}, X_{j_2}) = 1\right]$ depends upon how many distinct values there are among the four indices $\{i_1, j_1, i_2, j_2\}$. For example, $P[\psi(X_1, X_2) = 1, \psi(X_3, X_4) = 1]$ where all indices are distinct is not equal to $P[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1]$ where exactly two of the indices are equal. Also, there are restrictions on these four indices: $i_1 < j_1$, $i_2 < j_2$, and $(j_1 - 1)(j_1 - 2)/2 + i_1 = k(i_1, j_1) < k(i_2, j_2) = (j_2 - 1)(j_2 - 2)/2 + i_2$. These restrictions complicate counting the number of sets of four indices $\{i_1, j_1, i_2, j_2\}$ where all indices are distinct vs. exactly two of the four indices are equal.

An alternative approach to the computation of S_r is via the relationship of S_r to the moments of Y_n. Consider the following formula from Feller (1968, p. 110):

$$S_r = \sum_{m=r}^{N} \binom{m}{r} P_{[m]}$$

As noted earlier, $P_{[m]} = P(Y_n = m)$. So, for any integer r with $0 \le r \le N$,

$$S_r = \sum_{m=r}^{N} \binom{m}{r} P(Y_n = m) = E\left[\binom{Y_n}{r}\right]$$

using the convention that $\binom{a}{b} \equiv 0$ for integers a and b with $0 \le a < b$. In particular,

$$S_0 = E\left[\binom{Y_n}{0}\right] = 1$$

$$S_1 = E\left[\binom{Y_n}{1}\right] = E(Y_n) = N\delta_1$$

Consider now S_2.

$$S_2 = E\left[\binom{Y_n}{2}\right] = \frac{1}{2}E\left[Y_n(Y_n - 1)\right] = \frac{1}{2}\left\{E(Y_n^2) - E(Y_n)\right\}$$

Substituting (7) and (9),

$$S_2 = \frac{1}{2}\left[6\binom{n}{3}\delta_2 + N\delta_1 + 6\binom{n}{4}\delta_1^2 - N\delta_1\right] = 3\binom{n}{3}\delta_2 + 3\binom{n}{4}\delta_1^2.$$

Consider now S_3.

$$S_3 = E\left[\binom{Y_n}{3}\right] = \frac{1}{6}E\left[Y_n(Y_n-1)(Y_n-2)\right] = \frac{1}{6}\left\{E(Y_n^3) - 3E(Y_n^2) + 2E(Y_n)\right\} \tag{17}$$

Substituting (7), (9), and (10) into (17),

$$S_3 = 15\binom{n}{6}\delta_1^3 + 30\binom{n}{5}\delta_1\delta_2 + \binom{n}{3}\delta_{2b} + 4\binom{n}{4}(3\delta_{3a} + \delta_{3b})$$

So, in summary, $\{S_0, S_1, S_2, \ldots, S_N\}$ can be expressed in terms of moments of the distribution of Y_n:

$$S_r = E\left[\binom{Y_n}{r}\right] \text{ for } r = 0,1,\ldots,N$$

and in particular,

$$\begin{aligned} S_0 &= 1 \\ S_1 &= N\delta_1 = \binom{n}{2}\delta_1 \\ S_2 &= 3\binom{n}{4}\delta_1^2 + 3\binom{n}{3}\delta_2 \\ S_3 &= 15\binom{n}{6}\delta_1^3 + 30\binom{n}{5}\delta_1\delta_2 + \binom{n}{3}\delta_{2b} + 4\binom{n}{4}(3\delta_{3a} + \delta_{3b}) \end{aligned} \tag{18}$$

Conversely, one can express the moments of Y_n in terms of $\{S_0, S_1, S_2, \ldots, S_N\}$. In particular:

$$\begin{aligned} E(Y_n) &= S_1 \\ E(Y_n^2) &= 2S_2 + S_1 \\ \mathrm{Var}(Y_n) &= 2S_2 + S_1(1 - S_1) \\ E(Y_n^3) &= 6(S_3 + S_2) + S_1 \end{aligned} \tag{19}$$

4. Examples – General Case

In this section, I illustrate and verify the formulas for some small values of *n* by expressing the sampling distribution and moments of Y_n in terms of the S_r's.

4.1 $n = 2$

When $n = 2$, $Y_2 \equiv \psi(X_1, X_2)$. So, the support of Y_2 is $\{0,1\}$ which implies Y_2 is Bernoulli with parameter δ_1. So, the probability mass function is:

$$P(Y_2 = 0) = 1 - \delta_1$$

$$P(Y_2 = 1) = \delta_1$$

And, the formulas for the following moments are easily derived from the probability mass function:

$$E(Y_n) = \delta_1$$
$$E(Y_n^2) = \delta_1$$
$$\text{Var}(Y_n) = \delta_1(1-\delta_1)$$
$$E(Y_n^3) = \delta_1$$

These can also be derived from the formulas in this paper. For $n=2$, $N=\binom{n}{2}=1$. So, $S_1 = N\delta_1 = \delta_1$ and $S_r = 0$ for $r \geq 2$. Substituting into (13),

$$P(Y_n = 0) = \sum_{r=0}^{1}(-1)^r \binom{r}{0} S_r = S_0 - S_1 = 1-\delta_1$$

$$P(Y_n = 1) = \sum_{r=0}^{0}(-1)^r \binom{1+r}{1} S_{1+r} = S_1 = \delta_1$$

And, substituting into (19):

$$E(Y_n) = S_1 = \delta_1$$
$$E(Y_n^2) = 2S_2 + S_1 = S_1 = \delta_1$$
$$\text{Var}(Y_n) = 2S_2 + S_1(1-S_1) = S_1(1-S_1) = \delta_1(1-\delta_1)$$
$$E(Y_n^3) = 6(S_3 + S_2) + S_1 = S_1 = \delta_1$$

4.2 $n=3$

When $n=3$, $Y_3 \equiv \psi(X_1, X_2) + \psi(X_1, X_3) + \psi(X_2, X_3)$. So, the support of Y_3 is $\{0,1,2,3\}$. The corresponding probability mass function can be derived directly from the form of Y_3. However, it is easier to derive using the formulas in this paper.

For $n=3$, $N=\binom{n}{2}=3$. Also, recall that $\delta_{3a} = \delta_{3b} = 0$ for $n<4$. So, using the formulas in (18), $S_1 = 3\delta_1$, $S_2 = 3\delta_2$, and $S_3 = \delta_{2b}$; and by definition, $S_r = 0$ for $r \geq 4$.

Substituting into (13),

$$P(Y_3 = 0) = \sum_{r=0}^{3}(-1)^r S_r = S_0 - S_1 + S_2 - S_3 = 1 - 3\delta_1 + 3\delta_2 - \delta_{2b}$$

$$P(Y_3 = 1) = \sum_{r=0}^{2}(-1)^r \binom{1+r}{1} S_{1+r} = S_1 - 2S_2 + 3S_3 = 3\delta_1 - 6\delta_2 + 3\delta_{2b}$$

$$P(Y_3 = 2) = \sum_{r=0}^{1}(-1)^r \binom{2+r}{2} S_{2+r} = S_2 - 3S_3 = 3(\delta_2 - \delta_{2b})$$

$$P(Y_3 = 3) = \sum_{r=0}^{0}(-1)^r \binom{3+r}{3} S_{3+r} = S_3 = \delta_{2b}$$

Substituting into (19),

$$E(Y_3) = 3\delta_1$$

$$E(Y_3^2) = 3\delta_1 + 6\delta_2$$

$$\mathrm{Var}(Y_3) = 3\delta_1(1-3\delta_1) + 6\delta_2$$

$$E(Y_3^3) = 3\delta_1 + 6(3\delta_2 + \delta_{2b})$$

4.3 Larger n

For larger values of n, use of the expressions in this paper is not helpful in general. However, these expressions can still be useful in some special cases, as shown in the next section.

5. Examples – Special Case

As a special case, consider a population Ω with $K+1$ groups of objects $\{\Omega_k, k = 0,1,...,K\}$ satisfying the following assumptions:

a) $\Omega = \bigcup_{k=0}^{K} \Omega_k$ and $\Omega_i \bigcap \Omega_j = \varnothing$ for $i \neq j$ (i.e., groups are mutually exclusive and exhaustive).

b) When an object is selected at random from this population, π_k is the probability that the object belongs to Group Ω_k.

c) Any object from Group Ω_0 is not similar to any other object.

d) For Groups Ω_1 through Ω_K:

   i. Any two objects from the same group are similar.

   ii. Any two objects from different groups are not similar.

Suppose one is interested in the distribution of the number of pairs of similar objects among *n* randomly selected objects from this population. This number of similar pairs can be represented as follows.

Let the random variable X denote the group to which a randomly selected object belongs. Then, the distribution of X is discrete with support $\{0,1,2,...,K\}$ and probability mass function $\{\pi_0, \pi_1, ..., \pi_K\}$, i.e., $P(X_i = k) = \pi_k$. And, the group membership of a random sample of size *n* of objects can be represented as $\{X_1, X_2, ..., X_n\}$.

Next, consider the specific binary function:

$$\psi(X_i, X_j) = \begin{cases} 1 & \text{if } X_i = X_j \neq 0 \\ 0 & \text{if } X_i \neq X_j \text{ or } X_i = 0 \text{ or } X_j = 0 \end{cases} \tag{20}$$

Then, the number of similar pairs is:

$$Y_n \equiv \sum_{i<j} \psi(X_i, X_j) \ .$$

So, in this special case, the distribution of the number of pairs of similar objects among *n* randomly selected objects is just the distribution of Y_n, which is the focus of this paper.

One major simplification in this scenario is that for three independent random variables X_1, X_2, and X_3,

$$P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1, \psi(X_2, X_3) = 0\right] = 0$$

In other words, if two objects x_2 and x_3 are each similar a third object x_1, then necessarily, x_2 and x_3 are also similar to each other. Another way to state this is:

$$P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1, \psi(X_2, X_3) = 1\right] = P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1\right] \ .$$

As a consequence, some of the expected values take on simpler forms. For example,

$$\delta_1 \equiv P\left[\psi(X_1, X_2) = 1\right] = P\left[X_1 = X_2 \neq 0\right] = \sum_{k=1}^{K} \pi_k^2$$

$$\delta_2 \equiv P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1\right] = P\left[X_1 = X_2 = X_3 \neq 0\right] = \sum_{k=1}^{K} \pi_k^3$$

$$\delta_{2b} \equiv P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1, \psi(X_2, X_3) = 1\right] = \delta_2 = \sum_{k=1}^{K} \pi_k^3$$

$$\delta_{3a} \equiv P\left[\psi(X_1, X_2) = 1, \psi(X_3, X_4) = 1, \psi(X_1, X_3) = 1\right] = P\left[X_1 = X_2 = X_3 = X_4 \neq 0\right] = \sum_{k=1}^{K} \pi_k^4$$

$$\delta_{3b} \equiv P\left[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1, \psi(X_1, X_4) = 1\right] = P\left[X_1 = X_2 = X_3 = X_4 \neq 0\right] = \delta_{3a} = \sum_{k=1}^{K} \pi_k^4$$

Defining $\delta_3 \equiv \sum_{k=1}^{K} \pi_k^4$,

$$\delta_j \equiv \sum_{k=1}^{K} \pi_k^{j+1} \tag{21}$$

for $j = 1, 2, 3$.

There are a few simplifications for the small values of *n* studied in the last section. For $n = 2$, the only simplification is that $\delta_1 = \sum_{k=1}^{K} \pi_k^2$. For $n = 3$, making the substitution that $\delta_2 = \delta_{2b}$ and $\delta_3 = \delta_{3a} = \delta_{3b}$,

$$P(Y_3 = 0) = 1 - 3\delta_1 + 2\delta_2$$

$$P(Y_3 = 1) = 3(\delta_1 - \delta_2)$$

$$P(Y_3 = 2) = 0$$

$$P(Y_3 = 3) = \delta_2$$

and

$$E(Y_3) = 3\delta_1$$

$$E(Y_3^2) = 3\delta_1 + 6\delta_2$$

$$\text{Var}(Y_3) = 3\delta_1(1-3\delta_1) + 6\delta_2$$

$$E(Y_3^3) = 3\delta_1 + 24\delta_2$$

Consider now larger values of n. As mentioned previously, use of the expressions in this paper is not helpful in general because of the difficulty of finding expressions for S_r for $r \geq 4$. However, in this special case, expression can be derived directly from the definition of S_r given in (12):

$$S_r \equiv \sum_{(N,r)} P\left(A_{k_1} A_{k_2} \cdots A_{k_r}\right)$$

for specific values of n because, in this simplified scenario, one of the key properties is that the formula is driven by the amount of overlap in the indices of the $\{X_1, X_2, ..., X_n\}$ involved in the expression.

So, let's consider $n = 4$. With $n = 4$, there are a total of $N = \binom{4}{2} = 6$ different pairs of indices that can be selected from $\{1,2,3,4\}$.

For $r = 4$, there are 4 pairs of indices $\{(i_1, j_1),(i_2, j_2),(i_3, j_3),(i_4, j_4)\}$ associated with each term $P\left(A_{k_1} A_{k_2} A_{k_3} A_{k_4}\right)$ in S_4, and the value of $P\left(A_{k_1} A_{k_2} A_{k_3} A_{k_4}\right)$ depends upon the amount of overlap among these indices. With $n = 4$, regardless of which 4 pairs are chosen from the 6 possible, all four of the indices $\{1,2,3,4\}$ appear. So, for each of the $\binom{6}{4} = 15$ terms in S_4,

$$P\left(A_{k_1} A_{k_2} A_{k_3} A_{k_4}\right) = P(X_1 = X_2 = X_3 = X_4 \neq 0) = \delta_3$$

So,

$$S_4 = 15\delta_3 \tag{22}$$

For $r = 5$, there are 5 pairs of indices associated with each $P\left(A_{k_1} A_{k_2} A_{k_3} A_{k_4} A_{k_5}\right)$. With $n = 4$, regardless of which 5 pairs are chosen, all four of the indices $\{1,2,3,4\}$ appear. So, for each of the $\binom{6}{5} = 6$ terms in S_5,

$$P\left(A_{k_1} A_{k_2} A_{k_3} A_{k_4} A_{k_5}\right) = P(X_1 = X_2 = X_3 = X_4 \neq 0) = \delta_3$$

So,

$$S_5 = 6\delta_3 \tag{23}$$

Finally, for $r = 6$,

$$S_6 \equiv P\left(A_1A_2A_3A_4A_5A_6\right) = P(X_1 = X_2 = X_3 = X_4 \neq 0) = \delta_3 \tag{24}$$

In summary, combining equations (22), (23), and (24) with equations in (18) evaluated at $n = 4$:

$$\begin{aligned}
S_0 &= 1 \\
S_1 &= 6\delta_1 \\
S_2 &= 3\delta_1^2 + 12\delta_2 \\
S_3 &= 4\delta_2 + 16\delta_3 \\
S_4 &= 15\delta_3 \\
S_5 &= 6\delta_3 \\
S_6 &= \delta_3
\end{aligned}$$

Substituting these values for S_r into (13):

$$\begin{aligned}
P(Y_4 = 0) &= \sum_{r=0}^{6} (-1)^r S_r = S_0 - S_1 + S_2 - S_3 + S_4 - S_5 + S_6 \\
&= 1 - 6\delta_1 + 3\delta_1^2 + 12\delta_2 - 4\delta_2 - 16\delta_3 + 15\delta_3 - 6\delta_3 + \delta_3 \\
&= 1 - 6\delta_1 + 3\delta_1^2 + 8\delta_2 - 6\delta_3
\end{aligned}$$

$$\begin{aligned}
P(Y_4 = 1) &= \sum_{r=0}^{5} (-1)^r \binom{1+r}{1} S_{1+r} = S_1 - 2S_2 + 3S_3 - 4S_4 + 5S_5 - 6S_6 \\
&= 6\delta_1 - 2(3\delta_1^2 + 12\delta_2) + 3(4\delta_2 + 16\delta_3) - 4(15\delta_3) + 5(6\delta_3) - 6(\delta_3) \\
&= 6\delta_1 - 6\delta_1^2 - 12\delta_2 + 12\delta_3
\end{aligned}$$

$$\begin{aligned}
P(Y_4 = 2) &= \sum_{r=0}^{4} (-1)^r \binom{2+r}{2} S_{2+r} = S_2 - 3S_3 + 6S_4 - 10S_5 + 15S_6 \\
&= 3\delta_1^2 + 12\delta_2 - 3(4\delta_2 + 16\delta_3) + 6(15\delta_3) - 10(6\delta_3) + 15\delta_3 = 3\delta_1^2 - 3\delta_3
\end{aligned}$$

$$\begin{aligned}
P(Y_4 = 3) &= \sum_{r=0}^{3} (-1)^r \binom{3+r}{3} S_{3+r} = S_3 - 4S_4 + 10S_5 - 20S_6 \\
&= 4\delta_2 + 16\delta_3 - 4(15\delta_3) + 10(6\delta_3) - 20\delta_3 = 4\delta_2 - 4\delta_3
\end{aligned}$$

$$P(Y_4 = 4) = \sum_{r=0}^{2} (-1)^r \binom{4+r}{4} S_{4+r} = S_4 - 5S_5 + 15S_6 = 15\delta_3 - 5(6\delta_3) + 15\delta_3 = 0$$

$$P(Y_4 = 5) = \sum_{r=0}^{1} (-1)^r \binom{5+r}{5} S_{5+r} = S_5 - 6S_6 = 6\delta_3 - 6\delta_3 = 0$$

$$P(Y_4 = 6) = \sum_{r=0}^{0} (-1)^r \binom{6+r}{6} S_{6+r} = S_6 = \delta_3$$

References

Feller, William. 1968. *An Introduction to Probability Theory and Its Applications*. 3d ed. Vol. I. 2 vols. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Lee, A. J. 1990. *U-Statistics: Theory and Practice*. Statistics, Textbooks and Monographs v. 110. New York: M. Dekker.
Serfling, Robert J. 1980. *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Appendix 2

RMP Confidence Interval

1. Setup

In this paper, we will be considering a set of correlated trials where the correlation is of a specific type – namely from individual trials involving pairs of independent random variables. Specifically, consider a set of n *r.v.* $\{X_1, X_2, ..., X_n\}$. Assume this set of n *r.v.*'s are independent and identically distributed (*iid*).

Define a binary function on two variables $\psi(\cdot,\cdot)$; assume $\psi(x,x)=1$. Associated with this binary function are two probabilities:

$$\delta_1 \equiv P[\psi(X_1,X_2)=1] = E[\psi(X_1,X_2)]$$

$$\delta_2 \equiv P[\psi(X_1,X_2)=1, \psi(X_1,X_3)=1] = E[\psi(X_1,X_2)\psi(X_1,X_3)]$$

Here, δ_1 is the probability that two random variables match, where a match is defined as $\psi(X_1,X_2)=1$. In this paper, we refer to this probability as the random match probability (RMP). Similarly, δ_2 is the probability that two random variables $\{X_2, X_3\}$ each match a third X_1, i.e., $\psi(X_1,X_2)=1$ and $\psi(X_1,X_3)=1$. We will refer to this probability as the random tri-match probability (TMP).

In terms of δ_1 and δ_2, the first two moments of $\psi(X_1,X_2)$ are:

$$\begin{aligned} E[\psi(X_1,X_2)] &= \delta_1 \\ \mathrm{Var}[\psi(X_1,X_2)] &= E\left[\psi^2(X_1,X_2)\right] - \left\{E[\psi(X_1,X_2)]\right\}^2 \\ &= \delta_1 - \delta_1^2 = \delta_1(1-\delta_1) \end{aligned}$$

A natural point estimate of δ_1 is the *U*-statistic of degree 2 (Serfling, 1980):

$$U_n \equiv N^{-1}\sum_{i<j}\psi(X_i,X_j)$$

where $N \equiv \binom{n}{2}$, i.e., the number of pairs (i, j) with $1 \le i < j \le n$ and $\sum_{i<j}$ is used to represent $\sum_{j=2}^{n}\sum_{i=1}^{j-1}$ or equivalently, $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}$. Note that U_n is the sample proportion of matches in N pairwise comparisons.

From the properties of *U*-statistics, U_n is an unbiased estimator of δ_1, i.e.,

$$E(U_n) = \delta_1 . \qquad (1)$$

Furthermore, when $\{X_1, X_2, ..., X_n\}$ are *iid*, the variance of U_n is (Lee 1990, p. 12)[1]:

[1] In Lee (Lee, 1990), ξ_1 is denoted by σ_1^2 and ξ_2 is denoted by σ_2^2.

$$\mathrm{Var}(U_n) = N^{-1}\left[2(n-2)\xi_1 + \xi_2\right] \tag{2}$$

where

$$\begin{aligned}\xi_1 &= \mathrm{Cov}\left[\psi(X_1, X_2), \psi(X_1, X_3)\right] \\ &= E\left[\psi(X_1, X_2)\psi(X_1, X_3)\right] - \left\{E\left[\psi(X_1, X_2)\right]\right\}^2 \\ &= \delta_2 - \delta_1^2\end{aligned} \tag{3}$$

$$\begin{aligned}\xi_2 &= \mathrm{Cov}\left[\psi(X_1, X_2), \psi(X_1, X_2)\right] \\ &= \mathrm{Var}\left[\psi(X_1, X_2)\right] \\ &= \delta_1(1-\delta_1)\end{aligned} \tag{4}$$

Substituting in (3) and (4) into (2),

$$\mathrm{Var}(U_n) = N^{-1}\left\lfloor 2(n-2)(\delta_2 - \delta_1^2) + \delta_1(1-\delta_1) \right\rfloor \tag{5}$$

2. Relationships Between δ_1 and δ_2

2.1 Derived From Representation as a U-Statistics

Relationships between ξ_1 and ξ_2 given by Lee (1990) can be used to derive relationships between δ_1 and δ_2.

First, $\xi_1 \geq 0$ (Lee 1990, p. 10). That is, the correlation between $\psi(X_1, X_2)$ and $\psi(X_1, X_3)$ is non-negative, as excepted due to the common term X_1. However, as shown in (3), $\xi_1 = \delta_2 - \delta_1^2$. Thus, $\xi_1 \geq 0$ implies $\delta_1^2 \leq \delta_2$.

Second, $\xi_1 \leq \xi_2/2$ (Lee 1990, p.15). So, substituting (3) and (4), $2(\delta_2 - \delta_1^2) = 2\xi_1 \leq \xi_2 = \delta_1(1-\delta_1)$. Rewriting, $\delta_2 \leq \delta_1(1+\delta_1)/2$. Also, since δ_1 is a probability, $\delta_1 \leq 1$ which implies $(1+\delta_1)/2 \leq 1$ and thus that $\delta_1(1+\delta_1)/2 \leq \delta_1$. So, viewing the number of matches as a *U*-statistic leads to the following bounds on δ_2 in terms of δ_1.

$$\delta_1^2 \leq \delta_2 \leq \frac{\delta_1(1+\delta_1)}{2} \leq \delta_1 \tag{6}$$

These relationships between δ_1 and δ_2 lead to bounds on the variance of U_n. Using the relationship $\delta_1^2 \leq \delta_2$ in (5),

$$\begin{aligned}\mathrm{Var}(U_n) &= N^{-1}\left\lfloor 2(n-2)(\delta_2 - \delta_1^2) + \delta_1(1-\delta_1) \right\rfloor \\ &\geq N^{-1}\left[2(n-2)(\delta_1^2 - \delta_1^2) + \delta_1(1-\delta_1)\right] = \frac{\delta_1(1-\delta_1)}{N}\end{aligned}$$

Thus, the variance of U_n is bounded below by the variance of the proportion of "successes" in a binomial distribution with number of trials equal to N and success probability on each trial equal to δ_1, i.e., the variance of a sample proportion associated with treating the $N \equiv \binom{n}{2}$ pairwise comparisons as N independent trials.

Note that the variance of U_n equals the binomial variance when $\delta_1^2 = \delta_2$, i.e., when

$$P[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1] = \delta_2 = \delta_1^2 = P[\psi(X_1, X_2) = 1] \times P[\psi(X_1, X_3) = 1].$$

This shows that treating the number of matches in N pairwise comparisons as binomial underestimates the variance whenever X_1 matching one other observation, say X_2, increases its chances of matching yet another observation, say X_3.

Using the relationship $\delta_2 \le \delta_1(1+\delta_1)/2$ in (5),

$$\begin{aligned} \mathrm{Var}(U_n) &= N^{-1}\left\lfloor 2(n-2)(\delta_2 - \delta_1^2) + \delta_1(1-\delta_1) \right\rfloor \\ &\le N^{-1}\left[2(n-2)(\delta_1(1+\delta_1)/2 - \delta_1^2) + \delta_1(1-\delta_1) \right] = \frac{\delta_1(1-\delta_1)}{n/2} \end{aligned}$$

Thus, the variance of Y_n is bounded above by the variance of a binomial distribution with number of trials equal to $n/2$ and success probability on each trial equal to δ_1. This corresponds to $n/2$ independent trials, which would occur if the X's were paired (assuming *n* is even) and only the $n/2$ pairs were compared. This shows that treating the number of matches in N pairwise comparisons as $n/2$ independent trials overestimates the variance.

The difference between an upper confidence interval for the success probability in binomial trials differs significantly when the number of trials equals N vs. $n/2$. Table 1 shows the exact 95% upper confidence bounds associated with observing 0 successes in N vs. $n/2$ binomial trials. The "correct" bounds for the RMP based on $U_n = 0$ are somewhere between these two extremes.

Table 1: Upper confidence bound for the success probability when 0 successes are observed.

	95% Upper Confidence Bound	
n	Number of Trials: $N = \binom{n}{2}$	Number of Trials: $n/2$
10	6.4 E-02	4.5 E-01
50	2.4 E-03	1.1 E-01
100	6.1 E-04	5.8 E-02
500	2.4 E-05	1.2 E-02
1,000	6.0 E-06	6.0 E-03

5,000	2.4 E-07	1.2 E-03
10,000	6.0 E-08	6.0 E-04
50,000	2.4 E-09	1.2 E-04
100,000	6.0 E-10	6.0 E-05
1,000,000	6.0 E-12	6.0 E-06

Thus, inference performed on the N pairwise comparisons as if they are all independent will be liberal as the variance is underestimated. Inference performed treating the N pairwise comparisons as $n/2$ independent trials will be conservative as the variance is overestimated.

2.2 Derived From Inequalities

In some scenarios, a "tighter" upper bound on δ_2 in terms of δ_1 is possible. Consider a population Ω that can be divided into $K+1$ groups Ω_k Necessarily, $\Omega = \bigcup_{k=0}^{K} \Omega_k$.

Suppose:

a) When an object is selected at random from this population, π_k is the probability that the object belongs to group Ω_k.

b) For groups 1 through K:

   i. Any two distinct objects from the same group "match", i.e., $\psi(x_1, x_2) = 1$ if $x_1, x_2 \in \Omega_k$ for some $k \in \{1, 2, ..., K\}$, $x_1 \neq x_2$.

   ii. Any two objects from different groups "do not match", i.e., $\psi(x_1, x_2) = 0$ if $x_1 \in \Omega_{k_1}$ and $x_2 \in \Omega_{k_2}$ for some $k_1, k_2 \in \{1, 2, ..., K\}$, $k_1 \neq k_2$.

c) Any two distinct objects from Ω_0 "do not match", i.e., $\psi(x_1, x_2) = 0$ .if $x_1, x_2 \in \Omega_0$, $x_1 \neq x_2$.

Assumption (b) is equivalent to the assumption that for three randomly chosen objects $\{X_1, X_2, X_3\}$ from this population:

$$P[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1, \psi(X_2, X_3) = 0] = 0$$

In other words, if two objects $\{x_2, x_3\}$ each match a third x_1, then necessarily x_2 and x_3 match each other. Another way to state Assumption (b) is:

$$P[\psi(X_1, X_2) = 1, \psi(X_1, X_3) = 1] = P\left\lfloor X_1 \in \Omega_k, X_2 \in \Omega_k, X_3 \in \Omega_k \text{for some } k \in \{1, 2, ..., K\} \right\rfloor$$

Under these assumptions, $\delta_1 = \sum_{k=1}^{K} \pi_k^2$, $\delta_2 = \sum_{k=1}^{K} \pi_k^3$, and $\delta_2 \leq \delta_1^{3/2}$. Substituting the relationship $\delta_2 \leq \delta_1^{3/2}$ into (10),

$$\begin{aligned}\mathrm{Var}(U_n) &\le N^{-1}\left\lfloor 2(n-2)(\delta_1^{3/2}-\delta_1^2)+\delta_1(1-\delta_1)\right\rfloor \\ &= \frac{\delta_1}{N}\left[2(n-2)(\delta_1^{1/2}-\delta_1)+(1-\delta_1)\right]\end{aligned} \tag{7}$$

2. Upper Confidence Bound

When performing inference, it is useful to consider instead

$$Y_n \equiv NU_n = \sum_{i<j}\psi(X_i, X_j)\,. \tag{8}$$

instead of U_n. From (1) and (2), it follows that:

$$E(Y_n) = N\delta_1 \tag{9}$$

$$\begin{aligned}\mathrm{Var}(Y_n) &= N\left[2(n-2)(\delta_2-\delta_1)+(2n-3)\delta_1(1-\delta_1)\right] \\ &= N\left[2(n-2)(\delta_2-\delta_1^2)+\delta_1(1-\delta_1)\right]\end{aligned} \tag{10}$$

From (7), assuming the relationship $\delta_2 \le \delta_1^{3/2}$ holds,

$$\begin{aligned}\mathrm{Var}(Y_n) &\le N\delta_1\left\lfloor 2(n-2)(\delta_1^{1/2}-\delta_1)+(1-\delta_1)\right\rfloor \\ &\le N\delta_1(1-\delta_1^{1/2})\left[1+(2n-3)\delta_1^{1/2}\right] \equiv \sigma_B^2(\delta_1)\end{aligned} \tag{11}$$

This bound $\sigma_B^2(\delta_1)$ on $\mathrm{Var}(Y_n)$ (associated with the relationship $\delta_2 \le \delta_1^{3/2}$) combined with Cantelli's inequality can be used to construct an upper confidence bound for δ_1. In particular, the set $\{\delta_1 : N\delta_1 < Y_n + k(\alpha)\sigma_B(\delta_1)\}$ is a $100(1-\alpha)\%$ upper confidence bound for δ_1 where $k(\alpha) = \left(\frac{1-\alpha}{\alpha}\right)^{1/2}$. Since this is an increasing function of δ_1, the upper confidence bound for δ_1 is the maximum value of δ_1 satisfying:

$$\delta_1 \le \frac{Y_n}{N} + \frac{k(\alpha)}{\sqrt{N}}\left\{\delta_1(1-\delta_1)^{1/2}\left[1+(2n-3)\delta_1^{1/2}\right]\right\}^{1/2}$$

Agresti, A., Coull, B.A., 1998. Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions. Am Stat 52, 119–126.
Altman, D.G., Machin, D., Bryant, T.N., Gardner, M.J. (Eds.), 2000. Statistics with Confidence: Confidence Intervals and Statistical Guidelines, 2nd ed. ed. BMJ Books, Bristol.
Arvesen, J.N., 1969. Jackknifing U-Statistics. The Annals of Mathematical Statistics 40, 2076–2100.
Bain, L.J., Engelhardt, M., 1992. Introduction to Probability and Mathematical Statistics, 2nd ed. ed, The Duxbury advanced series in statistics and decision sciences. PWS-KENT Pub., Boston.
Bedrick, E.J., Aragon, J., 1989. Approximate Confidence Intervals for the Parameters of a Stationary Binary Markov Chain. Technometrics 31, 437–448.
Blyth, C.R., Still, H.A., 1983. Binomial Confidence Intervals. Journal of the American Statistical Association 78, 108–116.
Brown, L.D., Cai, T.T., DasGupta, A., 2001. Interval Estimation for a Binomial Proportion. Statistical Science 16, 101–117.
Casella, G., 2002. Statistical Inference, 2nd ed. ed. Thomson Learning, Australia ; Pacific Grove, CA.
Edwardes, M., 1994. Confidence Intervals for a Binomial Proportion. Statistics in Medicine 13, 1693–1698.
Efron, B., Stein, C., 1981. The Jackknife Estimate of Variance. The Annals of Statistics 9, 586–596.
Fleiss, J.L., Levin, B.A., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, 3rd ed. ed, Wiley series in probability and statistics. J. Wiley, Hoboken, N.J.
Gastwirth, J.L., Rubin, H., 1971. Effect of Dependence on the Level of Some One-Sample Tests. Journal of the American Statistical Association 66, 816–820.
Ghosh, B.K., 1979. A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter. Journal of the American Statistical Association 74, 894–900.
Hanley, J.A., Lippman-Hand, A., 1983. If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators. JAMA: The Journal of the American Medical Association 249, 1743–1745.
Hoeffding, W., 1948. A Class of Statistics with Asymptotically Normal Distribution. The Annals of Mathematical Statistics 19, 293–325.
Jones, D.A., 1972. Blood Samples : Probability of Discrimination. Journal of the Forensic Science Society 12, 355–359.
Jovanovic, B.D., Zalenski, R.J., 1997. Safety Evaluation and Confidence Intervals When the Number of Observed Events Is Small or Zero. Annals of Emergency Medicine 30, 301–306.
Klotz, J., 1973. Statistical Inference in Bernoulli Trials with Dependence. The Annals of Statistics 1, 373–379.
Ladd, D.W., 1975. An Algorithm for the Binomial Distribution with Dependent Trials. Journal of the American Statistical Association 70, 333–340.
Lee, A.J., 1990. U-Statistics: Theory and Practice, Statistics, textbooks and monographs. M. Dekker, New York.
Maesono, Y., 1997. Edgeworth Expansions of a Studentized U-Statistic and a Jackknife Estimator of Variance. Journal of Statistical Planning and Inference 61, 61–84.
Maesono, Y., 1998a. Asymptotic Mean Square Errors of Variance Estimators for U-Statistics and Their Edgeworth Expansions. Journal of Japan Statistical Society 28, 1–19.
Maesono, Y., 1998b. Asymptotic Comparisons of Several Variance Estimators and their Effects for Studentizations. Annals of the Institute of Statistical Mathematics 50, 451–470.
Maesono, Y., 2005. Higher-Order Comparisons of Asymptotic Confidence Intervals. Journal of Statistical Planning and Inference 133, 359–379.
Miao, W., Gastwirth, J.L., 2004. The Effect of Dependence on Confidence Intervals for a Population Proportion. The American Statistician 58, 124–130.
Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. Statistics in Medicine 17, 857–872.
Newman, T.B., 1995. If Almost Nothing Goes Wrong, Is Almost Everything All Right? Interpreting Small Numerators. JAMA: The Journal of the American Medical Association 274, 1013–1013.
Rohatgi, V.K., Saleh, A.K.M.E., 2001. An Introduction to Probability and Statistics, 2nd ed. ed, Wiley series in probability and statistics. Wiley, New York.
Ross, S.M., 2006. A First Course in Probability, 7th ed. ed. Pearson Prentice Hall, Upper Saddle River, N.J.
Saks, M.J., 2010. Forensic Identification: From a Faith-Based "science" to a Scientific Science. Forensic Science International 201, 14–17.

Schucany, W.R., Bankson, D.M., 1989. Small Sample Variance Estimators for U-Statistics. Australian Journal of Statistics 31, 417–426.
Sen, P.K., 1960. On Some Convergence Properties of U-Statistics. Calcutta Statistical Association Bulletin 10, 1–18.
Serfling, R.J., 1968. The Wilcoxon Two-sample Statistic on Strongly Mixing processes. The Annals of Mathematical Statistics 39, 1202–1209.
Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.
Smalldon, K.W., Moffat, A.C., 1973. The Calculation of Discriminating Power for a Series of Correlated Attributes. Journal of the Forensic Science Society 13, 291–295.
StatXact 8 User Manual, 2007. . Cytel Inc.
Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., Aitken, C., 2010. Data Analysis in Forensic Science: A Bayesian Decision Perspective, Statistics in Practice. John Wiley and Sons, Chichester.
Tobi, H., van den Berg, P.B., de Jong-van den Berg, L.T., 2005. Small Proportions: What to Report for Confidence Intervals? Pharmacoepidemiology and Drug Safety 14, 239–247.
Vollset, S.E., 1993. Confidence intervals for a binomial proportion. Statistics in Medicine 12, 809–824.
Wayman, J.L., 2000. Confidence Interval and Test Size Estimation for Biometric Data, in: Wayman, J.L. (Ed.), National Biometric Center Collected Works: 1997-2000. National Biometric Test Center, San Jose, CA, pp. 89–99.
Wolff, S., Gastwirth, J., Rubin, H., 1967. The Effect of Autoregressive Dependence on a Nonparametric Test (corresp.). IEEE Trans. Inform. Theory 13, 311–313.

Appendix 3

ROC Curves for Statistical Methods of Evaluating Evidence: Common Performance Measures Based On Similarity Scores

R. B. Patterson, J. J. Miller, and C. P. Saunders†

George Mason University, Fairfax, USA

Summary. We demonstrate the benefits of receiver operating characteristic (ROC) curves for measuring the performance of four statistical methods applied to forensic data. The statistical methods evaluate whether multivariate trace evidence found at a crime scene and on a suspect arise from the same source. Each method produces a numerical value that indicates the degree of association between two pieces of evidence. We treat such a number as a similarity score which provides a univariate measure of association between two observations. The number of false positives and false negatives at nominal thresholds on similarity scores for each method make interpreting performance ambiguous. Instead, ROC curves created with the similarity scores depict the full range of error rates achievable with each method. Moreover, they show the differentiation of similarity scores on common axes of true positive and false positive rates and so do not depend on the scales of the similarity scores. ROC curves provide an objective and comprehensive methodology for comparing the performance of the distinct statistical methods. We analyzed the performance of the statistical methods under several scenarios with data consisting of elemental composition measurements of glass fragments. Overall, the methods perform similarly and very accurately, with values often near 0.99 for the area under the ROC curve.

Keywords: Receiver operating characteristic (ROC) curves; Evaluation of evidence; Forensic science; Likelihood ratio; Test statistic; Multivariate data

1. Introduction

In forensics, many methods of evaluating evidence deal with whether two observations come from the same source. Such methods often produce a wide range of numerical values as output, which may support the hypothesis of a common source or the hypothesis of a different source. Yet, in measuring the performance of these methods, most often single values serve as fixed cutoffs on the output from the methods. For instance, a test statistic may have a cutoff at the significance level of 0.05. A likelihood ratio method may use the number one as the cutoff. These fixed values lead to single sets of errors (i.e., false positives and false negatives) as measures of performance when assessing the procedures. Comparing sets of just two numbers may make interpreting performance ambiguous. However, instead of measuring performance by one pair of error rates, we could apply techniques that consider all possible cutoffs and thus the full range of performance. Examples of such techniques include Tippett plots, detection error tradeoff (DET) curves, and receiver operating characteristic (ROC) curves. In this article, we will demonstrate the use of ROC curves with forensics data and detail their benefits.

Numerous fields use ROC curves to evaluate the performance of classification and prediction methods (Swets et al., 2000). Zhou et al. (2002), Pepe (2004), Lasko et al. (2005), and Zou et al. (2007) discussed their application in medicine. The general introduction by Fawcett (2006) mentioned their role in data mining and machine learning. In this article we demonstrate the utility of ROC curves in forensics, where we seek to measure the performance of methods that evaluate evidence.

We introduce ROC curves in general by following the development presented in the recent book by Krzanowski and Hand (2009) before we detail their use in forensics. In many fields, a classification method assigns observations to one of two classes. For example, in medicine a diagnostic test may predict whether a patient is healthy or sick. In astronomy, the problem may be to detect whether an observed object is an asteroid. In machine learning, an algorithm may classify a web search result as relevant or not. In these

†*Address for correspondence:* R. Bradley Patterson, Department of Statistics, George Mason University, 4400 University Drive, MSC 4A7, Fairfax, VA, 22030, USA.
E-mail: rpatter4@gmu.edu

examples, each observation belongs in truth to one of two populations, which we label positive and negative for generality.

A classification method often maps an observation to a univariate score and then classifies the observation based on the score. For instance, the method may classify new observations as positive or negative depending on whether their scores are above a certain value. We can measure the performance of the classifier by studying the separation of the distributions of scores for the positive and negative populations.

We could measure a classification method's performance by choosing a single threshold on scores. Then based on whether an observation has a score above or below the threshold, we could classify it as positive or negative. However, we may assign the wrong class and thereby cause an error. With multiple observations, we can compute the rates of these errors. Then, the *false positive rate* refers to the fraction of times that the classification method incorrectly assigns observations from the negative class to the positive class. Similarly, the *false negative rate* refers to the fraction of times that the classification method incorrectly assigns observations from the positive class to the negative class. The *true negative rate* is the complement of the false positive rate, and the *true positive rate* is the complement of the false negative rate. A single threshold on scores yields one set of rates, but we could vary the threshold to find all possible rates.

ROC curves depict the full range of error rates for a classification method. They plot the true positive rate against the false positive rate for all threshold values. Furthermore, because the relative ordering of scores for the positive and negative populations determines the ROC curve, we may use ROC curves to compare classification methods that generate scores on different scales. ROC curves indicate the differentiation of the scores for the positive and negative populations. Summary measures from ROC curves also provide insight into performance. For example, the area under the curve (AUC) corresponds to the Wilcoxon rank-sum test for the distributions of scores for the positive and negative populations. It gives the probability that a randomly chosen observation from the positive class will have a higher score than a randomly chosen negative observation. When dealing with numerous ROC curves, we may use their AUCs as average measures of performance to facilitate their interpretation.

In forensics, we evaluate pieces of evidence to weigh whether they come from the same source. We may form two hypotheses about a pair of observations, one of which we find at a crime scene and the other on a suspect. Suppose that the prosecution hypothesizes that the observations come from the same source while the defense hypothesizes that they come from different sources. (More specifically, the defense may hypothesize that the observation found on a suspect arises from a source randomly selected from a reasonable alternative population.) Many methods in forensics produce a numerical value that indicates the degree of association between two pieces of evidence. We may treat such a number as a similarity score, comparable to the scores discussed previously for classification methods in general. A similarity score is a univariate measure of association between two observations. High similarity scores support the hypothesis that the pair of observations belong to the same source. Low similarity scores support the hypothesis that the pair of observations come from different sources. A method's performance depends on its capability of supporting the hypothesis that corresponds to the truth.

Unlike many other fields, forensics does not technically have two populations as in the preceding examples. However, the methods of evaluating evidence do generate two distributions of similarity scores, one for pairs from the same source and another for pairs from different sources. The separation of those distributions is critical. We propose measuring that separation with ROC curves. We can still label the distributions as positive and negative. In the context of forensics, a *false positive* corresponds to misleading evidence in favor of the prosecution, and a *false negative* corresponds to misleading evidence in favor of the defense.

ROC curves offer several benefits to forensics. To start, they capture the full range of error rates achievable with a method. They also depict the relative separation of the distributions of similarity scores from a given method. This then allows for comparisons of methods that produce scores on different scales. Additionally, an important characteristic for a method of evaluating pairs of evidence is the probability that a randomly selected pair from the same source would have a higher similarity score than a randomly selected pair from different sources, which the AUC can estimate. We will discuss these benefits and more details of ROC curves relevant to forensics in Section 2.2.2.

To show the value of ROC curves in forensics, we applied them to measuring the performance of methods of evaluating trace evidence in the form of glass fragments. The methods, based on test statistics and likelihood ratios, came from an article by Aitken and Lucy (2004). Test statistics and likelihood ratios both provide measures of association between two samples. So we interpreted those values as similarity scores,

with which we created ROC curves for the same data as the original article. The ROC curves provided measurements of the full performance of the methods across all thresholds as well as an even basis for comparison. All of the methods performed very well.

2. Data and Methodology

2.1. Data

We use the glass data published online with the article by Aitken and Lucy (2004) at

`http://www.blackwellpublishing.com/rss/`

The complete data set includes measurements of elemental composition for each of five (n) fragments from 62 (m) window panes, giving a total of 310 ($N = mn$) observations. Each observation of a fragment includes measurements of four elements (Si, K, Ca, and Fe), which we transform as in the original paper to three (p) variables: log(Ca/K), log(Ca/Si), and log(Ca/Fe). An additional variable available for grouping the data is the type of window, of which there were three. Sixteen panes came from the first type, 16 from the second, and 30 from the third.

2.2. Methodology

To demonstrate the usefulness of ROC curves in forensics, we chose specific methods of evaluating evidence, but many others could benefit from ROC curves as well. After describing the chosen methods, we detail the application of ROC curves in forensics and their relation to other common measures of performance.

2.2.1. Procedures of evaluating evidence and similarity scores

We selected methods of evaluating evidence from the paper by Aitken and Lucy (2004) for use with ROC curves. The authors studied methods of evaluating trace evidence in the form of glass fragments. The forensic question was whether glass fragments recovered from a suspect had the same source as glass fragments found at a crime scene. We selected the following four methods of evaluating evidence as reported in Aitken and Lucy (2004):

(a) *multiple t-statistics*, based on the largest absolute value of multiple t-statistics;
(b) *T^2-statistic*, based on the value of Hotelling's T^2-statistic;
(c) *normal-based LR*, based on a likelihood ratio where multivariate normal probability densities represent both within-group and between-group variability;
(d) *density-based LR*, based on a likelihood ratio where within-group variability is assessed using a multivariate normal probability density function and between-group variability by using a multivariate kernel density estimate.

The above methods all map two samples to a number, which we interpret as a similarity score. The negative values from the multiple t-statistics and T^2-statistic methods and the raw values from the LR methods constitute similarity scores. We do not claim that a given value of a similarity score should have a specific interpretation. The reader may decide what meaning to attach.

2.2.2. Details of ROC curves

ROC curves have appeared in previous studies of the performance of methods in forensics. We refer the reader to examples by Whittaker et al. (1998); Phillips et al. (2001); Gonzalez-Rodriguez et al. (2005); Martin-de-las-Heras and Tafur (2009); and Tuceryan et al. (2011). As forensic methods assign more degrees of certainty (i.e., more categories or ranks) or continuous values to the evaluation of evidence, ROC curves become even more useful. Below we reiterate some of the benefits of ROC curves to forensics and introduce new ones.

For assessing the performance of methods of evaluating evidence, ROC curves offer several advantages. First, ROC curves are independent of the scale, calibration, or normalization of similarity scores generated

by a given method. The curves depend on only the order of similarity scores from the positive and negative distributions. Thus, we may fairly compare methods of evaluating evidence that produce similarity scores on different scales. Second, ROC curves depict the differentiation of similarity scores from the positive and negative distributions and hence a method's inherent capability of separating true positives and true negatives. Curves that approach the upper-left corner more closely indicate more separation between the distributions. Third, ROC curves present a complete picture of possible error rates achievable with a method. Each point on an ROC curve gives the true positive rate and false positive rate for a particular threshold on the similarity scores. So instead of choosing an arbitrary threshold on similarity scores, we could pick a threshold based on the error rates. In all, ROC curves offer complete and objective comparisons of the performance of methods of evaluating evidence.

Other possible techniques of measuring performance include the detection error tradeoff (DET) curve and Tippett plot. In biometrics, the DET curve aids in assessing a comparison method that does not possess a pre-specified threshold. The DET curve is similar to the ROC, but plots both error rates on normal deviate scales (Martin et al., 1997). Tippett plots have appeared in forensics, particularly for studying the performance of likelihood ratio methods (Aitken and Taroni, 2004). We describe Tippett plots in terms of our existing presentation of negative and positive distributions of similarity scores for pairs of observations from different sources and the same sources. We may form cumulative distribution functions (CDFs) of these distributions along the univariate axis of similarity scores. For a specific method of evaluating evidence, a Tippett plot shows one minus the CDF for each of the two distributions versus the similarity score. So a Tippett plot consists of two curves that give the fractions of the negative and positive distributions above the similarity score. By contrast, an ROC curve shows one minus the CDF of the positive distribution plotted against one minus the CDF of the negative distribution. While a Tippett plot clearly depends on similarity scores' scale, an ROC curve is independent of their scale. Thus, ROC curves make comparing methods that produce similarity scores on different scales easier. After choosing error rates at which to operate with a given method, we may identify the corresponding threshold on similarity scores by examining the Tippett plot or by finding the value of the parameter underlying the ROC curve.

Several aspects of ROC curves provide further insight into methods of evaluating evidence. An important value for measuring performance of methods in forensics and biometrics is the equal error rate (EER). The error rate achieved at the threshold where the false positive rate matches the false negative rate defines the EER. On an ROC plot, the EER occurs where a line running from the point (0,1) to (1,0) intersects the curve. As we will demonstrate in Section 3, the EER often does not match the error rates at nominal cutoffs. The AUC gives two valuable indications of the performance of a forensic method. First, the empirical AUC estimates the probability that a randomly selected pair from the same source would have a higher similarity score than a randomly selected pair from different sources. As Hanley and McNeil (1982) reported, this estimate is related to the nonparametric Wilcoxon rank-sum test of equal distributions. Second, the empirical AUC gives the mean true positive rate averaged uniformly across the false positive rate (Krzanowski and Hand, 2009). This average true positive rate is clearly distinct from the average error rate calculated as the mean of the false positive and false negative rates at a single threshold. By taking the average true positive rate over all possible false positive rates, we obtain a broader average measure of performance with the empirical AUC.

2.2.3. Analysis

We applied the four procedures mentioned above to several subsets of the data with different allotments of fragments to the control and recovered samples. To measure the performance of the procedures, we created ROC curves and computed the AUC values. We describe the relevant details of the calculations below.

To start, we introduce some notation. Let $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3})'$ be the vector of observations, where x_{ijk} is the k^{th} variable of the j^{th} fragment from the i^{th} window. Then the mean of the i^{th} window is: $\bar{\mathbf{x}}_i = \frac{1}{5}\sum_{j=1}^{5} \mathbf{x}_{ij}$.

In investigating the performance of the selected methods, we also considered their dependence on assumptions of normality and the number of fragments in the background data. All four methods assume a normal distribution for the within-group variation of fragments from the same window. The normal-based LR method also assumes a normal distribution for the between-group variation of the means of all the windows. Aitken and Lucy (2004) reported that the latter assumption may not be reasonable for this set of

data. We explored the impact of these assumptions by splitting the data into four different sets $\{\mathbf{x}_{ij}\}$, which consisted of windows from: all types ($m = 62$), only the first ($m = 16$), only the second ($m = 16$), and only the third ($m = 30$). The sets of data had a constant number ($n = 5$) of fragments per window, but a variable number (m) of windows and total number ($N = nm$) of fragments. To examine the degree of normality of the between- and within-group variation for each of these sets, we constructed Q-Q plots, presented in Table 5 of Appendix A. The distribution of between-group variation seems more normal for the sets of individual types of windows (especially those of type one) than for the set of all windows. For the within-group variation, we found type one windows exhibit the most similarity to the normal distribution and type two windows the least. The four methods of evaluating evidence also all use background data to estimate the within-group covariance matrices, and the LR procedures do so as well for the between-group covariance matrices. When analyzing a given set of data, we restricted the background data to observations from only the specific set. Then for each comparison between two groups of fragments, we excluded the two groups' observations from the background data before estimating the covariance matrices. The specific equations and results for the within- and between-group covariance matrices appear in Appendix B for the entire data.

Within a given set of the data, each window has five fragments, which we can treat as control or recovered fragments. As an example, we could let fragments one and four be the control fragments and fragments two, three, and five be the recovered fragments. Treating window a as the control and window b as the recovered is distinct from treating window b as the control and window a as the recovered for $a \neq b$. If we continue with the example above, in the first case, the comparison is between the data $\{\mathbf{x}_{a1}, \mathbf{x}_{a4}\}$ and $\{\mathbf{x}_{b2}, \mathbf{x}_{b3}, \mathbf{x}_{b5}\}$, while in the second the comparison is between $\{\mathbf{x}_{b1}, \mathbf{x}_{b4}\}$ and $\{\mathbf{x}_{a2}, \mathbf{x}_{a3}, \mathbf{x}_{a5}\}$. Additionally, because the four methods under study all treat the control and recovered data symmetrically, the scenario in which fragments one and four are the control and fragments two, three, and five the recovered will contain the same comparisons as the scenario in which fragments two, three, and five are the *control* and fragments one and four the *recovered*. We consider the following allotments of the five fragments from each window into the control and recovered data: two and two, two and three, and one and four. Within each of these scenarios, several permutations of the fragments assigned to the control and recovered data exist. For each such permutation of fragments, m^2 comparisons exist, with m between the same group and $m \times (m-1)$ between different groups.

To assess the performance of the different methods, we report the EER and the AUC of the ROC curve, computed with the `R` package `ROCR` by Sing et al. (2005). We converted the AUC and EER values into percentages by multiplying by 100. To calculate the EER, we sort the scores to use as thresholds. We then determine the curves for the rate of false negatives and the rate of false positives versus the threshold on similarity scores as follows. We treat comparisons between fragments from the same window with a score greater than or equal to the threshold as false negatives and comparisons between fragments from different windows with a score less than or equal to the threshold as false positives. The EER is then the maximum of the two rate curves at the point where they have the smallest absolute difference.

3. Results and discussion

In this section, we present results for each of the four sets of data described above. For a given scenario, we calculated the EER and AUC for each relevant permutation of fragments assigned to the control and recovered data. We then averaged the results across the permutations for the scenario as one summary. To obtain smoother values, we also combined the similarity scores across the permutations for the scenario and then calculated the EER and AUC from the pooled collection.

To contrast ROC curves with single thresholds, we also calculated the percentages of false positives and false negatives for each permutation of a scenario at nominal cutoffs of the similarity scores. For the multiple t-statistics method and the T^2-statistic method, we chose the cutoff at the 5% significance level. In both LR methods we used the number one as the cutoff. We report the percentages of each type of error averaged across the permutations of a scenario for each set of the data.

3.1. Windows of all types

We initially use all of the data so that $m = 62$ and $N = 310$. Table 1 gives the average percentages of false negatives and false positives for each method. In Table 1, the normal-based LR method has the lowest

Table 1. Average percentages of errors using all of the data.

Scenario	Errors	Multiple t-statistics	T^2-statistic	Normal-based LR	Density-based LR
2 vs 2	false neg.	7.53	6.56	0.65	1.18
	false pos.	2.41	2.36	3.74	3.55
2 vs 3	false neg.	7.90	6.77	0.48	0.97
	false pos.	2.23	2.19	3.41	3.27
1 vs 4	false neg.	8.71	7.10	0.00	0.32
	false pos.	2.71	2.65	4.14	3.90

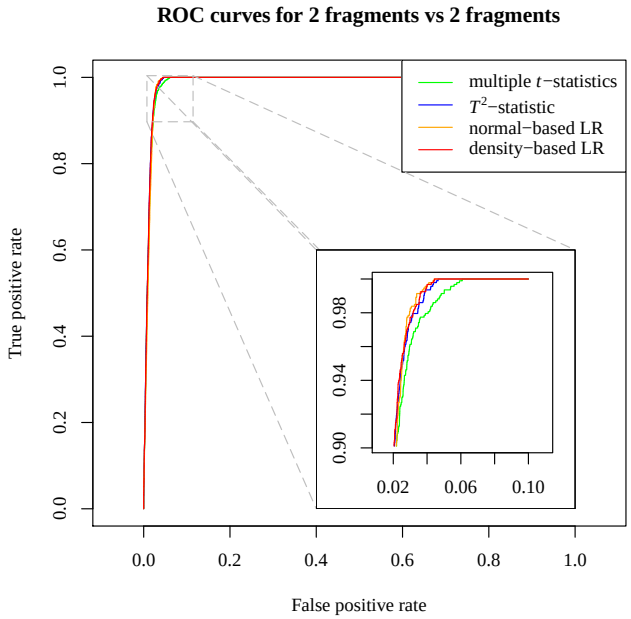


Fig. 1. ROC curves for two versus two fragments from windows of all types.

percentage of false negatives, but the T^2-statistic method has the lowest percentage of false positives. Thus, results at single, nominal thresholds on similarity scores make measuring the performances of these methods challenging.

Figures 1, 2, and 3 show the ROC curves for the pooled similarity scores. By depicting all possible error rates for the methods, the ROC curves make comparisons of performance easier and allow for the choice of threshold based on error rates. The nearly overlapping ROC curves for the T^2-statistic, normal-based LR, and density-based LR methods indicate that they perform almost equally. The corresponding EER and AUC for the ROC curves appear in Table 2. These results confirm that all of the methods perform very well. The averaged and pooled AUC values are all at least 98.8 and within less than one fifth of a percentage point across the methods for a given scenario's row. Different rankings of the methods by the EER and AUC results within a scenario occur due to the practically indistinguishable ROC curves. We emphasize primarily that all of the methods exhibit very high performance. Also, while allotting two fragments to one set of observations and three to the other leads to the best results, allotting four fragments to one set and one to the other produces the worst results.

In regard to choosing a threshold, the ROC curves show that three of the methods can achieve comparable error rates. Furthermore, a quick visual inspection of the curves reveals the attainable error rates. The choice of error rates may depend on many factors, and doing a cost-benefit analysis may suggest rates optimal for a specific application. We demonstrate just two possibilities as examples with the ROC curves

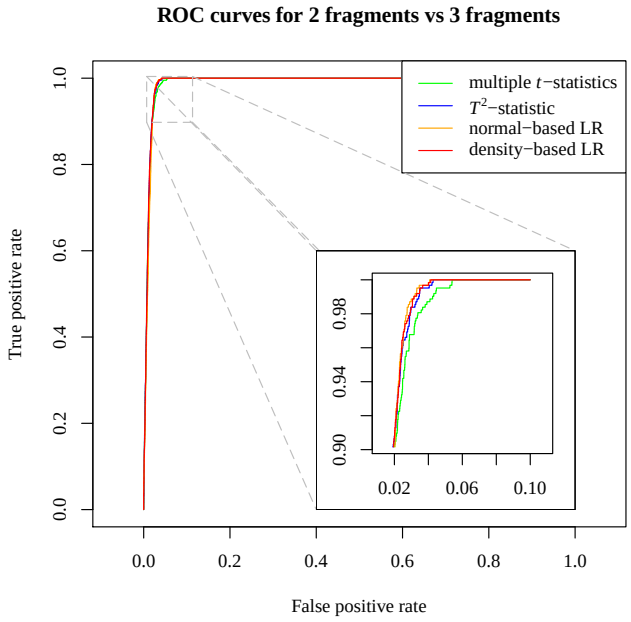


Fig. 2. ROC curves for two versus three fragments from windows of all types.

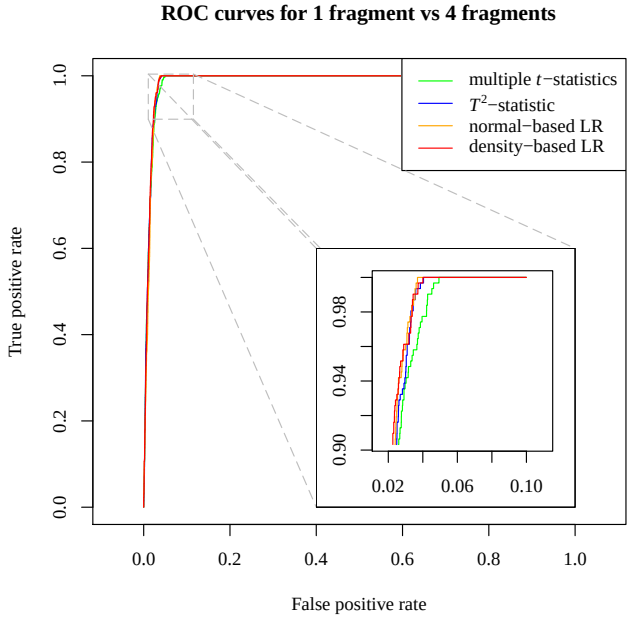


Fig. 3. ROC curves for one versus four fragments from windows of all types.

Table 2. Results from ROC curves for all of the data.

Scenario	Value	Multiple t-statistics		T^2-statistic		Normal-based LR		Density-based LR	
		EER	AUC	EER	AUC	EER	AUC	EER	AUC
2 vs 2	Averaged	3.42	98.94	3.19	99.01	3.17	98.91	3.13	99.02
	Pooled	3.23	98.94	2.90	99.01	2.80	98.91	2.90	99.02
2 vs 3	Averaged	3.28	99.03	3.14	99.08	2.95	98.98	3.04	99.09
	Pooled	3.23	99.03	2.81	99.08	2.61	98.98	2.62	99.09
1 vs 4	Averaged	3.53	98.84	3.25	98.90	3.27	98.80	3.28	98.92
	Pooled	3.69	98.83	3.23	98.90	3.23	98.81	3.23	98.92

Table 3. Average percentages of errors for windows of individual types.

Window type	Scenario	Errors	Multiple t-statistics	T^2-statistic	Normal-based LR	Density-based LR
1	2 vs 2	false neg.	6.25	5.42	2.92	2.92
		false pos.	3.39	3.33	4.78	4.81
	2 vs 3	false neg.	5.62	6.25	2.50	2.50
		false pos.	2.75	2.63	4.58	4.54
	1 vs 4	false neg.	8.75	7.50	2.50	2.50
		false pos.	3.83	3.83	5.00	5.00
2	2 vs 2	false neg.	6.25	5.42	2.92	2.92
		false pos.	20.92	20.31	20.03	19.33
	2 vs 3	false neg.	6.88	3.75	3.12	3.12
		false pos.	18.88	18.67	18.88	18.42
	1 vs 4	false neg.	5.00	3.75	3.75	3.75
		false pos.	24.17	22.17	20.83	19.83
3	2 vs 2	false neg.	9.56	5.56	0.67	1.33
		false pos.	4.54	4.57	5.87	5.81
	2 vs 3	false neg.	8.67	7.33	0.33	0.67
		false pos.	4.39	4.29	5.48	5.53
	1 vs 4	false neg.	8.00	6.67	1.33	2.67
		false pos.	5.15	4.97	6.25	6.23

for two versus three fragments. If the acceptable false negative rate were 5%, then we could achieve false positive rates of approximately 2.3% with the T^2-statistic, normal-based LR, and density-based LR methods. The corresponding thresholds on similarity scores for each would be, respectively, 8.96, 117.33, and 44.71. Alternatively, we could seek equal values of the false negative and false positive rates. Then the equal error rates and thresholds on similarity scores would respectively be 3.23% and 3.29 for the multiple t-statistics method, 2.81% and 12.44 for the T^2-statistic method, 2.61% and 38.32 for the normal-based LR method, and 2.62% and 17.98 for the density-based LR method.

3.2. Windows of individual types

In this section we use only windows of individual types. Table 3 gives the average percentages of false negatives and false positives at the nominal cutoffs for each method. In general, the methods based on likelihood ratios have lower percentages of false negatives, and those based on test statistics have lower percentages of false positives. Again, the single thresholds lead to a difficult interpretation of performance for the methods.

While we omit figures of the ROC curves to save space, we list the averaged and pooled EER and AUC

Table 4. Results from ROC curves for windows of individual types.

Window type	Scenario	Value	Multiple t-statistics		T^2-statistic		Normal-based LR		Density-based LR	
			EER	AUC	EER	AUC	EER	AUC	EER	AUC
1	2 vs 2	Averaged	5.00	99.01	5.33	99.11	5.78	98.68	5.75	98.77
		Pooled	4.17	99.03	4.17	99.09	4.17	98.65	4.17	98.75
	2 vs 3	Averaged	4.00	99.26	4.83	99.28	5.17	98.86	5.12	98.96
		Pooled	3.96	99.23	3.75	99.27	3.75	98.84	3.75	98.96
	1 vs 4	Averaged	5.17	98.74	5.42	98.86	6.25	98.35	5.58	98.41
		Pooled	5.00	98.75	5.00	98.83	5.00	98.42	5.00	98.46
2	2 vs 2	Averaged	14.89	91.33	15.53	91.94	14.86	91.87	14.78	91.75
		Pooled	15.42	91.37	14.36	92.01	14.17	91.97	14.25	91.79
	2 vs 3	Averaged	14.79	92.09	14.37	92.67	14.37	92.40	14.21	92.29
		Pooled	14.38	92.16	14.37	92.71	13.75	92.44	13.38	92.24
	1 vs 4	Averaged	19.08	89.84	16.42	90.90	16.42	91.16	15.75	91.01
		Pooled	17.92	89.86	17.50	91.04	16.25	91.29	15.58	91.20
3	2 vs 2	Averaged	5.39	97.72	4.98	97.75	4.70	97.76	4.93	97.68
		Pooled	5.13	97.72	4.74	97.74	4.46	97.76	4.67	97.67
	2 vs 3	Averaged	5.29	97.85	4.94	97.89	4.47	97.88	4.72	97.81
		Pooled	5.00	97.85	4.67	97.88	4.34	97.88	4.67	97.82
	1 vs 4	Averaged	6.28	97.53	6.32	97.59	6.30	97.70	6.67	97.59
		Pooled	5.52	97.53	5.33	97.58	5.33	97.70	5.33	97.58

in Table 4 for each method. Recall that the AUC gives an average measure of performance. We review the results by type of window. For windows of type one, all of the methods have very high accuracy with the lowest AUC value above 98.3, and the difference among the methods for a given scenario is less than half of a percentage point. The performance of all methods suffers with windows of type two, which seem to agree least with the assumptions of between- and within-group normality. However, the different methods still have similar results; the AUC values differ by less than two percentage points across the methods. For type three windows, the results improve. The lowest AUC value is above 97.5, and the range of values across the methods for a given scenario is less than half of a percentage point. Yet despite having almost twice the amount of background data as type one windows, type three windows have slightly lower performance.

We also note the different possible thresholds on the similarity scores suggested by the ROC curves for the scenario of two versus three fragments from windows of type one. If we sought equal false positive and false negative rates, we could choose the thresholds as follows. We could achieve the EER of 3.96% for the multiple t-statistics method with a threshold of 2.95. By setting the thresholds on the similarity scores for the T^2-statistic, normal-based LR, and density-based LR methods at 11.63, 3.77, and 4.13, respectively, we could achieve the common EER of 3.75%.

4. Conclusion

We have demonstrated the use of ROC curves for measuring the performance of methods of evaluating forensic evidence. We noted that the false positive and false negative rates at nominal thresholds on output from the methods made assessing their performance unclear. Treating the output from the methods as similarity scores allowed us to analyze the methods with ROC curves. The ROC curves showed the methods' capability of discriminating between true positives and true negatives more completely. They also allowed for different thresholds on similarity scores for achieving error rates with each method.

The particular results for the publicly available glass data analyzed with the methods studied by Aitken and Lucy (2004) support additional conclusions. In general, all of the methods perform extremely well, and they can very accurately separate the glass fragments in this data set by window. Also, the normal-based

LR method does not appear to suffer appreciably from a lack of between-group normality. Although the full set of all window types exhibits departure from the assumption of between-group normality, the results for all methods with this set are almost as good as the results with the set of type one windows, which show the best agreement with the assumption of between-group normality. The higher performance with the set of type one windows over the set of type three windows suggests the larger number of windows in the background data may not overly influence the methods either. Instead, the performance with the different sets of data seems to correlate better with the normality of the within-group variability. The Q-Q plots in Table 5 of Appendix A suggest that the within-group variability of type one windows is the most normal and that of type two windows the least. Thus, the dependence on a normal distribution for within-group variability may affect all methods most.

As mentioned in the introduction, the accuracy of all four methods is very high. The smallest AUC value is still greater than 89.8, and most of the AUC values are near 99. Furthermore, the distinction among the four methods is small, with differences in AUC values often less than one percentage point. Indeed, the ROC curves for the T^2-statistic, normal-based LR, and density-based LR methods almost overlap. Thus, they can achieve nearly equivalent error rates by choosing appropriate thresholds. A researcher may prefer one method over another for philosophical reasons or computational necessity, but all methods perform almost equally well in terms of ROC curves.

Applying ROC curves to different statistical methods of evaluating forensic evidence or to the same methods covered here but with different data would offer some very interesting future research. Also, fitting parametric ROC curves to the empirical ones would provide efficient estimates with which to make additional measures of performance.

Acknowledgements

The research detailed in this article was supported in part by Award No. 2009-DN-BX-K234 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the Department of Justice.

The authors thank Dr. C. G. G. Aitken at the University of Edinburgh and Dr. D. Lucy at Lancaster University for sharing `R` code to compute the likelihood ratios.

A. Between-group and within-group distributions

The normal-based LR method assumes a multivariate normal distribution of the window means, but the density-based LR method does not. Aitken and Lucy (2004) noted the departure from normality for the window means from all of the data. To assess normality for the window means for each set, we examined Q-Q plots of Mahalanobis distances squared versus quantiles of χ^2_3, the chi-squared distribution with three degrees of freedom. The squared Mahalanobis distances for each set are:

$$\text{Mahalanobis } D^2 = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \hat{S}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}),$$

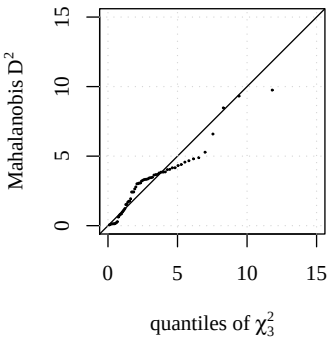
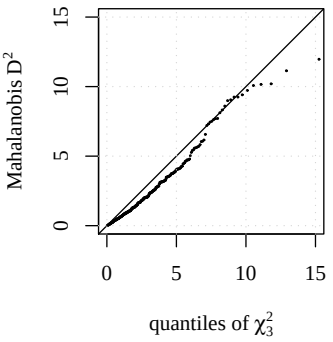
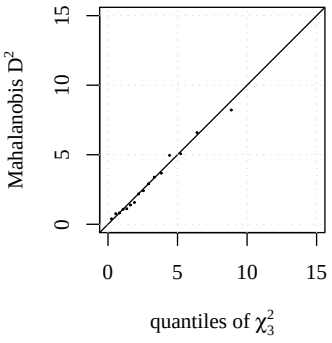
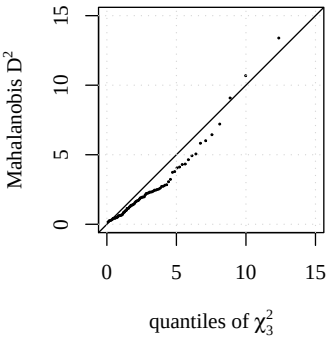
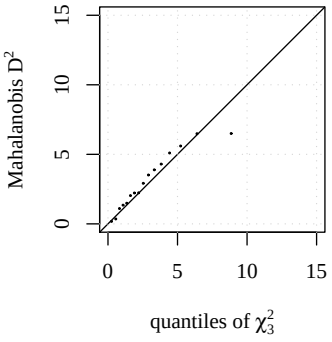
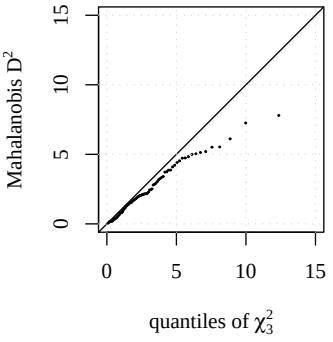
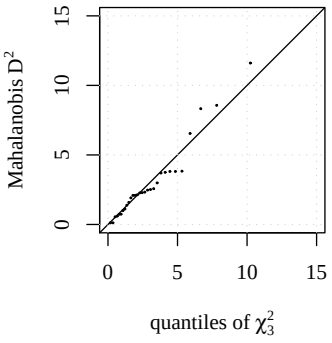
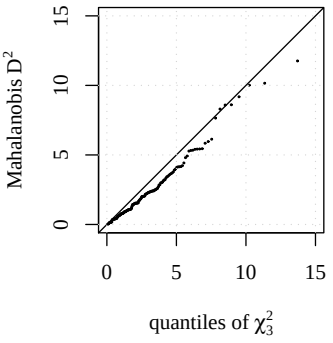
where i ranges over the groups in the set, $\bar{\mathbf{x}}$ is the mean of the set, and $\hat{S}$ is the estimated covariance matrix. Table 5 suggests that the window means from only a single type assume a more normal distribution. We also created Q-Q plots of the within-group variation, presented in the far right column of Table 5. The within-group variability appears most normal for windows of type one and least normal for windows of type two.

B. Covariance matrices

We present the equations used to estimate the within- and between-group covariance matrices below. Letting

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)',$$

Table 5. Q-Q plots of Mahalanobis distances squared versus quantiles of χ^2_3 for between- and within-group variability for the sets of all windows and individual types.

Window type(s)	Between-group	Within-group
All		
One		
Two		
Three		

the estimate of the within-group covariance matrix U is

$$\hat{U} = W/(N-m).$$

We estimate the between-group covariance matrix C by

$$\hat{C} = \frac{B}{m-1} - \frac{W}{n(N-m)},$$

where

$$B = \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$$

and $\bar{\mathbf{x}}$ is the grand mean.

We calculated the within- and between-group matrices for each subset of the data analyzed. The estimated within- and between-group covariance matrices for all groups are:

$$\hat{U} = \begin{pmatrix} 1.68 \times 10^{-2} & 2.66 \times 10^{-5} & 2.21 \times 10^{-4} \\ 2.66 \times 10^{-5} & 6.53 \times 10^{-5} & 7.40 \times 10^{-6} \\ 2.21 \times 10^{-4} & 7.40 \times 10^{-6} & 1.33 \times 10^{-3} \end{pmatrix}$$

$$\hat{C} = \begin{pmatrix} 7.06 \times 10^{-1} & 9.88 \times 10^{-2} & -4.63 \times 10^{-2} \\ 9.88 \times 10^{-2} & 6.21 \times 10^{-2} & -6.96 \times 10^{-3} \\ -4.63 \times 10^{-2} & -6.96 \times 10^{-3} & 1.01 \times 10^{-1} \end{pmatrix}.$$

For the set of only type one windows, the estimated within- and between-group covariance matrices are:

$$\hat{U} = \begin{pmatrix} 1.57 \times 10^{-2} & 1.72 \times 10^{-4} & 3.50 \times 10^{-4} \\ 1.72 \times 10^{-4} & 5.57 \times 10^{-5} & -5.96 \times 10^{-6} \\ 3.50 \times 10^{-4} & -5.96 \times 10^{-6} & 5.04 \times 10^{-4} \end{pmatrix}$$

$$\hat{C} = \begin{pmatrix} 1.06 \times 10^{-1} & -8.33 \times 10^{-3} & 7.62 \times 10^{-2} \\ -8.33 \times 10^{-3} & 2.62 \times 10^{-3} & -6.9 \times 10^{-3} \\ 7.62 \times 10^{-2} & -6.9 \times 10^{-3} & 8.25 \times 10^{-2} \end{pmatrix}.$$

For the set of only type two windows, the estimated within- and between-group covariance matrices are:

$$\hat{U} = \begin{pmatrix} 3.78 \times 10^{-2} & -2.74 \times 10^{-5} & 7.46 \times 10^{-4} \\ -2.74 \times 10^{-5} & 6.61 \times 10^{-5} & 3.63 \times 10^{-5} \\ 7.46 \times 10^{-4} & 3.63 \times 10^{-5} & 8.64 \times 10^{-4} \end{pmatrix}$$

$$\hat{C} = \begin{pmatrix} 1.77 \times 10^{-1} & 6.75 \times 10^{-4} & 1.90 \times 10^{-3} \\ 6.75 \times 10^{-4} & 7.10 \times 10^{-4} & -1.89 \times 10^{-3} \\ 1.90 \times 10^{-3} & -1.89 \times 10^{-3} & 6.47 \times 10^{-3} \end{pmatrix}.$$

For the set of only type three windows, the estimated within- and between-group covariance matrices are:

$$\hat{U} = \begin{pmatrix} 6.15 \times 10^{-3} & -2.19 \times 10^{-5} & -1.28 \times 10^{-4} \\ -2.19 \times 10^{-5} & 7.00 \times 10^{-5} & -8.80 \times 10^{-7} \\ -1.28 \times 10^{-4} & -8.80 \times 10^{-7} & 2.02 \times 10^{-3} \end{pmatrix}$$

$$\hat{C} = \begin{pmatrix} 5.49 \times 10^{-1} & 4.80 \times 10^{-2} & 2.57 \times 10^{-2} \\ 4.80 \times 10^{-2} & 8.55 \times 10^{-3} & -1.19 \times 10^{-2} \\ 2.57 \times 10^{-2} & -1.19 \times 10^{-2} & 1.18 \times 10^{-1} \end{pmatrix}.$$

References

Aitken, C. G. G. and D. Lucy (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society Series C 53*(1), 109–122.

Aitken, C. G. G. and F. Taroni (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester: John Wiley and Sons.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters 27*(8), 861–874.

Gonzalez-Rodriguez, J., J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia (2005). Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Science International 155*(2-3), 126–140.

Hanley, J. A. and B. J. McNeil (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*(1), 29–36.

Krzanowski, W. J. and D. J. Hand (2009). *ROC Curves for Continuous Data* (1 ed.). Boca Raton: Chapman & Hall.

Lasko, T. A., J. G. Bhagwat, K. H. Zou, and L. Ohno-Machado (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics 38*(5), 404 – 415.

Martin, A., G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki (1997). The DET curve in assessment of detection task performance. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, Volume 4, pp. 1895–1898.

Martin-de-las-Heras, S. and D. Tafur (2009). Comparison of simulated human dermal bitemarks possessing three-dimensional attributes to suspected biters using a proprietary three-dimensional comparison. *Forensic Science International 190*(1-3), 33–37.

Pepe, M. S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction* (1 ed.). Oxford: Oxford University Press.

Phillips, V. L., M. J. Saks, and J. L. Peterson (2001). The application of signal detection theory to decision-making in forensic science. *Journal of Forensic Sciences 46*(2), 294–308.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005). ROCR: visualizing classifier performance in R. *Bioinformatics 21*(20), 3940–3941.

Swets, J. A., R. M. Dawes, and J. Monahan (2000). Better decisions through science. *Scientific American 283*(4), 82–87.

Tuceryan, M., F. Li, H. L. Blitzer, E. T. Parks, and J. A. Platt (2011). A framework for estimating probability of a match in forensic bite mark identification. *Journal of Forensic Sciences 56*, S83–S89.

Whittaker, D. K., M. R. Brickley, and L. Evans (1998). A comparison of the ability of experts and non-experts to differentiate between adult and child human bite marks using receiver operating characteristic (ROC) analysis. *Forensic Science International 92*(1), 11–20.

Zhou, X., D. K. McClish, and N. A. Obuchowski (2002). *Statistical Methods in Diagnostic Medicine* (1 ed.). Wiley-Interscience.

Zou, K. H., A. J. O'Malley, and L. Mauri (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. *Circulation 115*(5), 654–657.