2017

U-Statistics for Characterizing Forensic Sufficiency Studies

Cami Fuglsby
South Dakota State University

U-STATISTICS FOR CHARACTERIZING FORENSIC SUFFICIENCY STUDIES

BY

CAMI FUGLSBY

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Mathematics

Specialization in Statistics

South Dakota State University

2017

U-STATISTICS FOR CHARACTERIZING FORENSIC SUFFICIENCY STUDIES

Cami Fuglsby

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Mathematics with a specialization in Statistics degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Christopher Saunders, Ph.D. Date
Thesis Advisor

Kurt Cogswell, Ph.D. Date
Head, Department of Mathematics & Statistics

Dean, Graduate School Date

ACKNOWLEDGEMENTS

I would like to acknowledge my committee members: Dr. Christopher Saunders, Dr. Cedric Neumann, Dr. Semhar Michael, and Dr. Cody Wright. Specifically, I would like to thank my advisor, Dr. Christopher Saunders, for his continual support, encouragement, and for fueling my interest in forensic statistics.

I would like to also acknowledge the computational support of Sciometrics; the FBI Laboratory for supplying the set of handwritten documents; and Dr. JoAnn Buscaglia from the FBI Laboratory for her reviews and comments.

Lastly, I owe a big thank you to the Department of Mathematics and Statistics for supporting me as a graduate student. Without the department, I would not have had the opportunity to learn from incredibly knowledgeable researchers in the interesting and continually growing field of forensics.

This work is based in part on publication number 10-01 of the Laboratory Division of the FBI. The preliminary 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.

Preliminary aspects of this research summarized in this thesis was supported in part by Awards No. 2009-DN-BX-K234 and 2014-IJ-CX-K088 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 US Department of Justice.

TABLE OF CONTENTS

NOTATION **vii**

LIST OF FIGURES **ix**

LIST OF TABLES **x**

LIST OF ALGORITHMS **xi**

ABSTRACT **xii**

CHAPTERS **1**

1 Introduction & Overview **1**

1.1 Introduction 1

1.2 RMP and Sufficiency 4

1.3 Estimating the RMP and the RNMP 6

1.3.1 Properties of the RMP and RNMP 8

1.3.2 Overview of Thesis 12

2 Simulations **14**

2.1 Simulated Writing Samples 14

2.1.1 Subsampling Versus Resampling 16

2.1.2 Estimating RMP 18

2.1.3 Estimating RNMP 20

2.1.4 The Form of the Standard Error 21

2.1.5 Estimating the Standard Error 24

2.2 Applications 26

2.2.1 Determining an Appropriate Threshold Value 32
2.2.2 Estimating the RMP as a Function of Length of Writing Samples 34
2.2.3 Estimating the TMP and Standard Error 36
2.2.4 Notes-Approximation Results for Sub-sampling U-Statistics estimates of RMP 39
2.2.5 Contributions to Designing a Study of Handwriting Individuality 45

3 Conclusion **61**

REFERENCES **64**

CURRICULUM VITAE **68**

NOTATION

N**:** Represents the set of writing samples, where each sample was generated by a different writer.

D_i**:** Represents the writing sample produced from the i^{th} writer.

$D_i^{*(k)}$**:** Represents the k^{th} simulated writing sample produced from the i^{th} writer.

$s(\cdot,\ \cdot)$**:** Represents the similarity score used to compare two documents together.

τ**:** Represents the threshold value used to determine if two documents are declared a 'match' or 'non-match'.

$I(s(\cdot,\ \cdot) > \tau)$**:** The Indicator Function, where if $s(\cdot,\cdot) > \tau$, then $I = 1$, and if $s(\cdot,\cdot) \leq \tau$, then $I = 0$. Sometimes simplified to m_{ij}.

l**:** Represents a specific letter in a-z, A-Z, or $0 - 9$.

$\#(\cdot)$**:** Represents the number of unique items from $\cdot$, for example, in the set $A = \{a, a, b, c\}$, there are four items but only three unique items, and thus $\#(A) = 3$.

n_i**:** Represents the number of words in the i^{th} document.

n**:** Represents the number of words in a document or chosen from a document.

LIST OF FIGURES

1.1 An Illustration on how the RMP decreases as the amount of writing available increases, Found and Bird (2016, pg. 37). 5

2.1 A cursive and text form of the modified "London Letter" (Saunders et al. 2011a) 27

2.2 Binomial RNMP by Word Count 48

2.3 Binomial RNMP by Word Count 49

2.4 Corrected Chi-Square RNMP by Word Count 50

2.5 Corrected Chi-Square RNMP by Word Count 51

2.6 Uncorrected Chi-Square RNMP by Word Count 52

2.7 Uncorrected Chi-Square RNMP by Word Count 53

2.8 Kullbeck-Liebler RNMP by Word Count 54

2.9 Kullbeck-Liebler RNMP by Word Count 55

2.10 Fitted Simple Logistic Regression Model for Binomial RMP 56

2.11 Fitted Simple Logistic Regression Model for Uncorrected Chi-Square RMP 56

2.12 Fitted Simple Logistic Regression Model for Corrected Chi-Square RMP 57

2.13 Fitted Simple Logistic Regression Model for Kullbeck-Liebler RMP . 57

2.14 Subsampling Estimates of the Variance of the Conditional Match Probability; using the theoretical threshold for the Binomial match and the empirical threshold for the two χ^2 classifiers. 58

2.15 Fitted Simple Logistic Regression Model for Binomial TMP 59

2.16 Fitted Simple Logistic Regression Model for Uncorrected Chi-Square TMP 59

2.17 Fitted Simple Logistic Regression Model for Corrected Chi-Square TMP 60

2.18 Fitted Simple Logistic Regression Model for Kullbeck-Liebler TMP . 60

LIST OF TABLES

2.1 Fitted Simple Logistic Regression on the Models for RMP 35

2.2 Fitted Simple Logistic Regression on the Models for TMP 39

LIST OF ALGORITHMS

1 RMP 19

2 RNMP 20

3 TMP 25

4 TMP Estimate 37

ABSTRACT

U-STATISTICS FOR CHARACTERIZING FORENSIC SUFFICIENCY STUDIES

CAMI FUGLSBY

2017

One of the main metrics for deciding if a given forensic modality is useful across a broad spectrum of cases, within a given population, is the Random Match Probability (RMP), or the corresponding discriminating power. Traditionally, the RMP of a given modality is studied by comparing full 'templates' and estimating the rate at which pairs of templates 'match' in a given population. This strategy leads to a natural U-statistic of degree two. However, in questioned document examination, the RMP is studied as a function of the amount of handwriting contained in the two documents being compared; turning the U-statistic into a U-process. This work is focused on providing background on forensic sufficiency studies, RMP, and the U-processes that naturally arise in this class of problems.

CHAPTER 1

Introduction & Overview

1.1 Introduction

One potential goal of a forensic document examiner (FDE) is to determine the writer of a given document. One way to reach this goal is to compare the features of the questioned document to the features of a sample document where the writer is known. Keep in mind that, due to intra-writer variability, observing a 'perfect match' between two writing samples that were written by the same source is typically not expected, and may be indicative of a forgery.

One reason why two writing samples provided by the same individual (source) may never have the exact same handwriting characteristics is the natural variation in an individual's handwriting. Comparing writing samples is ultimately comparing writing habits between distinct individuals, which is discussed by Huber and Headrick (1999, pp. 73-74). The habits formed by an individual are the characteristics of their writing that are measured by the features or qualities. There are many factors that affect how an individual writes, a few may be who taught them at a young age, what country or region they are from, if the language they are writing in is their first language or not, etc. Throughout this paper, we will refer to an individual's entire body[1] of handwriting as their writing profile. One common belief is that an individual's writing profile is better described as a probability distribution across generated documents

[1]This includes everything an individual has ever written and could possibly write.

from that individual rather than a static characteristic of that individual, such as DNA or a fingerprint (Bulacu and Schomaker, 2007).

One method that a forensic document examiner might use is taking advantage of an automated comparison procedure that quantifies the variability in handwriting. Such a procedure would start by scanning in two writing samples, convert the samples into a set of quantitative features, and then compute a similarity score based on the quantitative features as a measure of the similarity of the writing profiles which generated the two writing samples. Then a threshold value can be introduced where two samples can be declared a 'match' if the similarity score falls above the pre-defined threshold value, and the two samples can be declared a 'non-match' if the similarity score falls below the threshold.

Using a pre-defined threshold value for a given automated comparison procedure allows for a measure of the consistency of the writing profiles generating the two samples. However, simply because the similarity score declares two samples to be a 'match' does not always mean that the same writer generated both writing samples, which could lead to an error. When two writing samples provided by different writers are declared 'match', then a false match error has occurred. Similarly, when two writing samples provided by the same writer are declared 'no-match', then a false no-match error has occurred[2]. These two errors are results of between-writer similarity and within-writer dissimilarity (Risinger and Saks, 1996), respectively, which we are characterizing.

Determining the rates of the two different errors is useful when determining a comparison procedure's ability to discriminate between writers. One way to measure the false match error rate is what we will refer to in this paper as the Random Match Probability (RMP). We define the RMP as the probability that two randomly selected

[2]These are analogous to false positives and false negatives.

writing samples from two randomly selected writers from a relevant population are declared to match by the given comparison procedure. Another way to interpret it is the rate of false match errors "averaged" over all of the relevant[3] writing samples. Similarly, we refer to the false no-match error rate as the Random Non-Match Probability (RNMP). The RNMP is defined as the probability that two randomly selected writing samples from one randomly selected writer from a relevant population are declared to not match under the given comparison procedure. Another way to interpret the RNMP is the rate of false no-match errors "averaged" over all relevant writing samples. It is important to clarify that the RMP $\neq 1-$RNMP, simply because of the different conditioning that occurs; $P\,(\text{match}|\text{different source documents})$ and $P\,(\text{non-match}|\text{same source documents})$. The RMP and RNMP depend on the comparison procedure being used, specifically the associated similarity score and threshold value used when declaring a match or non-match based on the produced similarity score. Other factors the RMP and the RNMP depend on include:

- The relevant population generating the writing samples used in the comparison. Some writing profiles are harder to distinguish between than others.
- The lengths and the content of the writing samples being compared, i.e. the number of words and the distribution of the letters used in the samples.

In order to use a given comparison procedure, it is important to know what the RMP and the RNMP that is associated with that procedure. This work will examine a class of statistics for investigating the RMP and the RNMP with a given comparison procedure. More specifically, examine how the RMP and the RNMP change as the length of the writing samples change, as this aids in determining how accurate the comparison procedure is expected to be concerning different lengths of the writing samples. Another area of interest is how the RMP and the RNMP depend on the

[3]Generally refers to the relevant population the samples were collected from, where one would want to know the commonalities and errors of that specific population.

content of the writing samples, though in this paper we do not address this. In this paper, we assume that the content of the writing samples being compared is similar to the frequency of letters as they appear in English writing.

1.2 RMP and Sufficiency

The RMP shows up 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).) Aitkin and Taroni (2004) describe the RMP in two different ways. First, as a way to measure how accurate a comparison method is when it comes to differentiating between biometric samples that come from different sources. Second, as a way to determine the strength of a declared match, as it may be evidence to support that two biometric samples were produced by a common source. For these reasons, a comparison procedure would ideally have a smaller RMP.

Found and Bird (2016, pg. 37) illustrated how, ideally, the RMP decreases as a function of the amount of writing available increases, as shown in Figure 1.1. As can be seen from the following definition, the 'Likelihood of a Chance Match' corresponds with our definition of the RMP. "If we were to choose random samples of handwriting (from different individuals) containing identical text and proceed through a stroke by stroke analysis of the concatenations, then as the complexity increases (as reflected in the number of strokes, for example), the likelihood that the samples will diverge in some way from each other would, in general, increase." (Found and Bird (2016, pg. 37))

Keep in mind that while a small RMP is ideal, a match under the given comparison procedure does not strictly mean that the two biometric samples arose from a common source, there still exists the possibility that two sources have the same or extremely

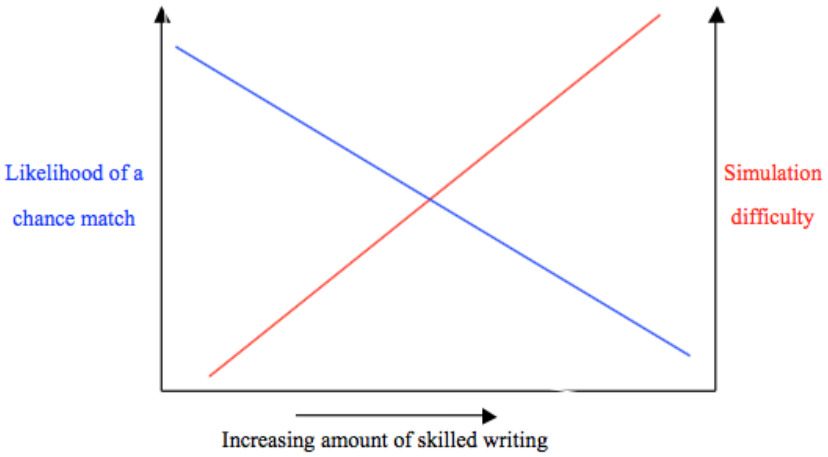


Figure 1.1: An Illustration on how the RMP decreases as the amount of writing available increases, Found and Bird (2016, pg. 37).

similar biometric profiles. As mentioned in Saks and Koehler (2008), infrequency cannot be equated to uniqueness. Balding (2005) uses the phrase "the uniqueness fallacy" to describe the fallacy in cases involving DNA evidence, where a certain set of genetic markers are declared to be unique as they are expected to occur less than once out of five billion, which is approximately the earth's population. Thus, if writing profiles are unique, we cannot say a trace writing sample is unique, we cannot equate the uniqueness of a writing profile to the uniqueness of a trace.

With this in mind, we can say that the RMP is associated with the "degree of individuality" of writing profiles from a population, as well as associated with a comparison method, which can create an upper bound on the degree of individuality. See Bolle, et al. (2004) and Saunders et al. (2011a) for a detailed discussion of this relationship. 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 the probability that there is at least one match (in the case of comparing DNA profiles) using population genetics modeling. This depends on the population size and on the number of loci compared in the typing. However, models of this type have not been developed to sufficiently characterize an individual's writing profile. This paper will propose an alternative comparison methodology to that type of modeling; it will provide information about the RMP and how it changes compared to the length of the writing samples that are compared. Note that with DNA evidence, a RMP is typically referencing the probability (or chance) that a new DNA profile matches the observed profile.

1.3 Estimating the RMP and the RNMP

One way to estimate the RMP and the RNMP of a given comparison procedure is to calculate them from a population of known writing samples.

We will be considering a collection of writing samples, where each sample was written by a different writer, with a total of N writers. Assume that the writers are a random sample from a relevant population of writers[4], and that the writing sample from the writer is characteristic of their own writing profile[5]. The writing samples themselves may include one or more documents and may have been collected at varying times

[4]We assume that the population of writers is so large that we can treat the sampled writers as independent and identically distributed (*i.i.d.*) with respect to some distribution on the relevant population of writers.

[5]In other words, each writing sample is assumed to be randomly generated from that writer's writing profile.

or in different environments. With these assumptions, the whole collection of writing samples is independent and identically distributed ($i.i.d.$).

One way to estimate the RMP involves measuring the similarity between two writing samples. This creates pairwise comparisons. Let D_i and D_j denote the two writing samples that arose from the i^{th} and j^{th} writer, respectively, where $i \neq j$. Then let $s(D_i, D_j)$ be the score that measures similarity between the two writing samples, where we assume $s(D_i, D_j) = s(D_j, D_i)$. We denote τ to be the threshold which declares two writing samples a match. If $s(D_i, D_j) > \tau$, then the two samples are declared to match. If $s(D_i, D_j) \leq \tau$, then the two samples are declared to not match. Measuring the proportion of declared matches gives an unbiased estimator of the RMP for the used comparison procedure. See 1.3.1 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 a Wald-type upper confidence bounds for the RMP.

We can use a similar process to estimate the RNMP. Instead of one writer providing one writing sample, suppose that each writer provides two writing samples. Then $s(D_{i,1}, D_{i,2})$ is the similarity score between the two writing samples provided by the same writer. Using the same τ defined above, if $s(D_{i,1}, D_{i,2}) \leq \tau$, then the two writing samples written by the same writer are declared a non-match. Calculating the proportion of non-matches when comparing two documents written by the same writer produces an estimate of the RNMP for the used comparison procedure. See 1.3.1 for more details about this estimator of the RNMP.

Though this process gives an idea of the RMP and the RNMP, it does not provide information on the relationship between the RMP, RNMP, and the sizes of the writing samples. The writing samples themselves may have been different lengths as well. Even if the writing samples in the collection were approximately the same size, we only see how the RMP and the RNMP are affected by that one size. This would mean

we would have to obtain multiple writing samples of different lengths from a single source, and do this for all N sources. Instead of collecting multiple samples, these samples can instead be "simulated" from a single collection of writing samples, where the size of the simulated samples will be less than the size of the observed writing sample.

1.3.1 Properties of the RMP and RNMP

The following results are all in Serfling (1980) and are included for completeness.

For N documents from N writers, where the the i^{th} document is written by the i^{th} writer, let $D_i : i = 1, 2, ..., N$ be *i.i.d.*. Let $s(D_i, D_j)$ be the similarity score associated with the comparison procedure which compares the documents D_i and D_j written by the i^{th} and j^{th} writers, where $i \neq j$. Let τ be the threshold used to determine if the similarity score produced a match or no-match.

One method that can be used to estimate the RMP, which we denote as θ, for a given comparison procedure is to calculate the proportion of similarity scores that are declared to match, which we denote as $m_{ij} = 1$ if D_i and D_j match, and zero if they do not.

$$\tilde{\theta} = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} m_{ij} \tag{1.1}$$

Denote

$$\theta = P(s(D_i, D_j) > \tau),$$

which means that $\theta = E(m_{ij})$. We can show that $\tilde{\theta}$ is unbiased:

Proof:

$$\begin{aligned} E(\tilde{\theta}) &= E\left(\binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} m_{ij}\right) \\ &= \binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E(m_{ij}) \\ &= \binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \theta \\ &= \binom{N}{2}^{-1}\binom{N}{2}\theta \\ &= \theta. \end{aligned}$$

■

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

$$\mathrm{Var}(\tilde{\theta}) = \frac{4(N-2)}{N(N-1)}\sigma_c^2 + \frac{2}{N\,(N-1)}\theta(1-\theta) \tag{1.2}$$

where

$$\sigma_c^2 \equiv \mathrm{Var}[E(m_{ij}|D_i)] \text{ for any } j \neq i. \tag{1.3}$$

Proof: Let $\underset{\sim}{m}$ be the $\binom{N}{2} \times 1$ vector of all of the match outcomes, and $\underset{\sim M}{1}$ represents a column vector with M rows with each element as a one. Then
$\tilde{\theta} = \binom{N}{2}^{-1} \underset{\sim\binom{N}{2}}{1^t}\underset{\sim}{m}$ and $E(\tilde{\theta}) = \theta$.

$$\begin{aligned} Var(\tilde{\theta}) &= Var\left(\binom{N}{2}^{-1}\underset{\sim\binom{N}{2}}{1^t}\underset{\sim}{m}\right) \\ &= \binom{N}{2}^{-2}\underset{\sim\binom{N}{2}}{1^t} Var\left(\underset{\sim}{m}\right)\underset{\sim\binom{N}{2}}{1} \end{aligned}$$

To understand what the variance term looks like, it helps to look at the covariance matrix.

$$Var\left(\underset{\sim}{m}\right) = \begin{Bmatrix} cov(m_{12}m_{12}) & cov(m_{12}m_{13}) & \dots & cov(m_{12}m_{N-1\ N}) \\ cov(m_{13}m_{12}) & cov(m_{13}m_{13}) & \dots & cov(m_{13}m_{N-1\ N}) \\ \vdots & \ddots & \dots & \vdots \\ cov(m_{N-1\ N}m_{12}) & cov(m_{N-1\ N}m_{13}) & \dots & cov(m_{N-1\ N}m_{N-1\ N}) \end{Bmatrix}$$

$$= \begin{Bmatrix} \theta(1-\theta) & \sigma_c^2 & \dots & 0 \\ \sigma_c^2 & \theta(1-\theta) & \dots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \theta(1-\theta) \end{Bmatrix}$$

The $\theta(1-\theta)$ is the variance of a binomial, since $m_{ij} = m_{ij}$ produces one of two outcomes, $\{0, 1\}$, and thus is a binomial random variable. The term σ_c^2 represents the covariance between two comparisons that share exactly one document in common, $cov(m_{ij}m_{ih})$. Finally, the zeros show the covariance of between two comparisons of sets of four documents, each generated by a different writer, since the documents are considered *i.i.d.*.

The representation of the form of the matrix is

$$\theta(1-\theta)\, I_{\binom{N}{2}\times\binom{N}{2}} + \sigma_c^2 \left(P_{\binom{N}{2}\times N} P^t_{N\times\binom{N}{2}} - 2I_{\binom{N}{2}\times\binom{N}{2}}\right),$$

where

$$P = \begin{Bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{Bmatrix}.$$

P is a matrix where the row represents the index of pairs of the comparisons, and the

columns represent each one of N individual documents. For the ij^{th} row, there is a one in the intersection of the i^{th} column and the j^{th} column, and zeros for every other column in that row. The matrix PP^t is used to represent each place in the covariance matrix where there is a covariance between two comparisons that share exactly one document in common. The identity matrix multiplied by two is included since PP^t will have an additional 'two' along the diagonal, which is not included in the original matrix.

We can use linear algebra to solve out the variance using the form of the matrix as follows:

$$\begin{aligned}
&\binom{N}{2}^{-2} \underset{\sim}{1}^t_{\binom{N}{2}} Var\left(\underset{\sim}{m}\right) \underset{\sim}{1}_{\binom{N}{2}} \\
&\qquad = \binom{N}{2}^{-2}\left[\theta\left(1-\theta\right)\binom{N}{2} + \sigma_c^2\left(N(N-1)^2 - 2\binom{N}{2}\right)\right] \\
&\qquad = \theta\left(1-\theta\right)\binom{N}{2}^{-1} + \sigma_c^2\left(\frac{N(N-1)^2((N-2)!)^2}{(N!)^2} - 2\binom{N}{2}^{-1}\right) \\
&\qquad = \theta\left(1-\theta\right)\binom{N}{2}^{-1} + \sigma_c^2\left(\frac{4(N-2)}{N(N-1)}\right).
\end{aligned}$$

The first step multiplied out the one-vectors, which essentially summed over the identity matrices as well as PP^t. ■

Note that σ_c^2 does not depend on i or j because $\{D_i : i = 1, 2, ..., N\}$ are assumed to be *i.i.d.*. Also, the "bar" in $E(m_{ij}|D_i)$ denotes conditional expectation. So, $E(m_{ij}|D_i)$ can be viewed as a conditional match probability - namely, the probability that a randomly selected writing sample matches a specific writing sample D_i.[6]

Based on the asymptotic distribution of a U-statistic, an approximate $100(1-\alpha)\%$

[6]Note that the first term in (2) involving σ_c^2 dominates $\text{Var}(\tilde{\theta})$ for large values of N, as the second term quickly converges to zero.

Wald-type upper confidence bound on the RMP is:

$$\tilde{\theta} + z_\alpha \sqrt{\text{Var}(\tilde{\theta})} \approx \tilde{\theta} + \frac{2 z_\alpha \sigma_c}{\sqrt{N}} \text{ for large } N \tag{1.4}$$

where $\text{Var}(\tilde{\theta})$ (or σ_c) can be replaced by a consistent estimator, such as the one that is presented in Wayman (2000) by Bickel, 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 *i.i.d.* collection of pairs of writing samples $\{(D_{i,1}, D_{i,2}) : i = 1, 2, ..., N\}$. Let $s(D_{i,1}, D_{i,2})$ denote the similarity score (associated with the comparison procedure) that compares the two writing samples $D_{i,1}$ and $D_{i,2}$ 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:

$$\hat{\gamma} = N^{-1} \sum_{i=1}^{N} I\{s(D_{i,1}, D_{i,2}) \leq \tau\}. \tag{1.5}$$

1.3.2 Overview of Thesis

Our main focus is investigating simulated writing samples that were generated from a body of genuine writing samples can be used to examine how RMP and RNMP change with respect to the size of the writing samples in the comparison. We then use this to produce the standard deviation, which is used to construct upper confidence bounds. Following that, we will examine 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 Sciometrics. Using these, we will make an example of the proposed methodology, as well as possible applications, for instance generating useful information to help in designing an empirical study focused on handwriting individuality.

CHAPTER 2

Simulations

2.1 Simulated Writing Samples

Between the limited lengths of documents and not being able to obtain a sample from every writing profile, there simply may not be enough writing samples to capture the wide range of writing profiles among a population and to assess the RMP and RNMP of a given comparison procedure, as well as their relationship to the lengths of the provided documents. It would be possible, using Monte Carlo simulation[1], to produce writing samples of any specified length if the source's writing profile is known. Bear in mind, a source's writing profile is hardly ever known. However, if reasonable models for writing profiles are available, the parameters could be estimated from the models of the writing profiles. Using various resampling methods such as the Bootstrap algorithm, a parameter could generate any number of writing samples of any size (Efron and Hastie, 2016). To date, reasonable models for writing profiles have not been developed.

Another method to generate simulated writing samples would be to subsample the words from a single writing sample for each source. This proposed methodology will allow various lengths of simulated samples to be produced, which allows for the relationship between RMP and writing sample length to be studied. We can also generate multiple simulated samples from a single writing sample, which allows for

[1]Monte Carlo simulation generates samples to estimate properties of a distribution.

the RNMP to be studied as well.

This idea of creating simulated samples from the given 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[2].

Most resampling methodologies are examples of the plug-in principle in statistics. Essentially, the plug-in principle works by estimating a property of a population using the statistic that was derived from the sample. Resampling substitutes the sample for the population and then draws samples (i.e. resamples) to imitate the process of constructing the sampling distribution.

Resampling methods still tend to rely on the same Monte Carlo techniques used when the population distribution is known. In theory, one could generate all possible simulated samples that are generated from the given samples, however, this process would be too time consuming and computer intensive to consider. Monte Carlo resampling is instead used to limit the number of simulated samples produced. There is one main difference between Monte Carlo simulation and resampling. Monte Carlo simulation assumes that the underlying distribution is known. Resampling methods assumes that the underlying distribution is not known, and thus the simulation is then based on the sampled data.

Simulated samples that are generated from the sampled data can be applied to many complicated statistical analyses. Regardless of the application though, it is important that the simulated samples imitate the distribution of the sampled data. This ensures that the properties of the simulated samples can generate valid estimators of the

[2]Note that this does not account for outside factors, i.e. writing surfaces, drug usage, or intentionally disguising their own writing or replicating another writing profile.

population characteristics of interest.

There is an important distinction to be made when subsampling from the sampled data. Sampling words without replacement produces writing samples with a similar distribution to the associated writing profile. If instead we sampled words with replacement, the simulated samples would be distributed according to a slightly different writing profile, as seen in 2.1.1. The estimators based on the latter subsampling method will not necessarily be consistent if the number of sources increases to infinity while the writing sample lengths remain short (i.e. consists of a small number of words).[3]

In the following subsections, we will detail the subsampling method we propose for estimating the RMP, RNMP, and the standard error associated with the estimator of the RMP described in 1.3.1. We continue to assume that the writing samples in the collection are *i.i.d.* and each sample is provided by one of each N writers.

2.1.1 Subsampling Versus Resampling

In this section, we will 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.

Derivation 1: Suppose the original writing sample with V words is represented as $\{W_1, W_2, ..., W_V\}$ where W_i denotes the features of the i^{th} word in the writing sample. (In our case, the features consist of the word being written and the isocodes representing the character's shapes.) Assume that this vector is an *i.i.d.* sample from a multinomial distribution with r categories and associated probability vector

[3]When we say that an estimator is consistent, it means that for a sufficiently large number of writers, it is expected that the estimator is very close to the real value for the entire population.

$\mathbf{p} = \{p_1, p_2, .., p_r\}$, which we represent as: $W_i \overset{i.i.d.}{\sim}$ Mult(1,$\mathbf{p}$), $i = 1, ..., V$. Let $Y_j = \#\{W_i \text{ in the } j^{th} \text{ category}\}$, $j = 1, 2, ..., r$. 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: $(Y_1, Y_2, ..., Y_r) \sim$ Mult(V,$\mathbf{p}$).

First, suppose the simulated writing sample $\{W_1^*, W_2^*, ..., W_n^*\}$ is generated by sampling $n \leq V$ words at random without replacement from the original writing sample $\{W_1, W_2, ..., W_V\}$. Since $\{W_1^*, W_2^*, ..., W_n^*\} \subset \{W_1, W_2, ..., W_n\}$, $W_i^* \overset{i.i.d.}{\sim}$ Mult(1,$\mathbf{p}$), $i = 1, ..., n$. So, if $Y_j^* = \#\{W_i^* \text{in the } j^{th} \text{ category}\}$, $j = 1, 2, ..., r$, $(Y_1^*, Y_2^*, ..., Y_r^*) \sim$ 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 $\{W_1^*, W_2^*, ..., W_n^*\}$ is generated by sampling $n \leq V$ words at random <u>with</u> replacement from the original writing sample $\{W_1, W_2, ..., W_V\}$. Let $\mathbf{Y}^* = (Y_1^*, Y_2^*, ..., Y_r^*)$, where $Y_j^* = \#\{W_i^* \text{ in } j^{th} \text{ category}\}$, $j = 1, 2, ..., r$. 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 Mult(1,$\hat{\mathbf{p}}$). So, conditional on the observed writing sample, $W_i^* | W_j \overset{i.i.d.}{\sim}$ Mult(1,$\hat{\mathbf{p}}$), $i = 1, ..., n$, $j = 1, ..., V$, and $\mathbf{Y}^* | \{W_1, W_2, ..., W_V\} \sim$ 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 $\sum_{j=1}^r x_j = n$,

$$P(\mathbf{Y}^* = \mathbf{x}) = \sum_y P(\mathbf{Y}^* = \mathbf{x} | \mathbf{Y} = \mathbf{y}) P(\mathbf{Y} = \mathbf{y})$$

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

Substituting the multinomial probabilities,

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

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

$$V^{-n} E\left[\Pi_{j=1}^r Y_j^{x_j}\right] = V^{-n} E\left[Y_1^n\right] > p_1$$

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. ■

2.1.2 Estimating RMP

In order to investigate how the RMP changes as a function of the lengths of writing samples, we propose Algorithm 1 to estimate the RMP when comparing two writing samples of a specified length.

Analysis of the set of data that Algorithm 1 creates will provide information on the distribution of the similarity score between two writing samples provided by two different writing sources.

Note that in Step 3 of Algorithm 1, the k represents the number of simulations that are performed for the chosen pair of writers and the given sample size n.

In Step 4 of Algorithm 1, the match is determined by the previously defined threshold value of τ. The proportion of pairs where $I\left(s(D_i^{*(k)}, D_j^{*(k)})\right) = 1$, or $\tilde{\theta}$ from (1), is an

Algorithm 1: RMP

Data: A set of N writing samples from N writers. The set of writing samples for the i^{th} writer will consist of n_i words.

Result: A matrix of the size $10 \times K \times \binom{N}{2}$ by 6: $\{n, k, i, j, s(D_i^{*(k)}, D_j^{*(k)}) : k = 1, 2, ..., K; i, j = 1, ..., N; i \neq j, n \in \{10, 20, 30, ..., 100\}\}$.

begin
- **for** $W_i; i \in \{1, ..., N-1\}$ **do**
  - **for** $W_j; j \in \{W_i + 1, ..., N\}$ **do**
    - **for** $n \in \{10, 20, 30, ..., 100\}$ **do**
      - **for** $k \in \{1, ..., K\}$ **do**
        - Randomly select, without replacement, n words from the n_i words in W_i, generating the pseudo-document $D_i^{*(k)}$, and n words from the n_j words in W_j, generating the pseudo-document $D_j^{*(k)}$.
        - Calculate the score, $s(D_i^{*(k)}, D_j^{*(k)})$.
      - **end**
    - **end**
  - **end**
- **end**

end

unbiased estimator of the RMP for comparing two writing samples. As the number of writers increase for fixed lengths of writing samples being compared, this estimator of the RMP is consistent; this is due to the variance decreasing as the number of writers increase to infinity. This can easily be seen by looking at the following relationship between $\tilde{\theta}$ and $\tilde{\theta}^*$, where

$$\tilde{\theta}^* = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=2}^{N} \left[K^{-1} \sum_{k=1}^{K} I\left(s\left(D_i^{*(k)}, D_j^{*(k)} \right) > \tau \right) \right].$$

$$\begin{aligned} Var\left(\tilde{\theta}^*\right) &= Var\left(E\left(\tilde{\theta}^* | D_1, ..., D_N\right)\right) + E\left(Var\left(\tilde{\theta}^* | D_1, ..., D_N\right)\right) \\ &= Var\left(\tilde{\theta}\right) + \frac{E\left(\theta_{ij}\left(1 - \theta_{ij}\right)\right)}{K} \binom{N}{2}^{-1} \\ &\implies |Var\left(\tilde{\theta}^*\right) - Var\left(\tilde{\theta}\right)| \leq \frac{1}{4K} \binom{N}{2}^{-1} \end{aligned}$$

The two estimators, $\tilde{\theta}$ and $\tilde{\theta}^*$, are unbiased estimators of the RMP, and are further explored in Section 2.2.4.

2.1.3 Estimating RNMP

We can modify Algorithm 1 to produce Algorithm 2 to study the relationship between RNMP and the length of the writing samples. The main difference is that estimating the RNMP relies on comparing two simulated writing samples produced by the same source, instead of two simulated writing samples produced by two different sources.

Algorithm 2: RNMP

Data: A set of N writing samples from N writers. The set of writing samples for the i^{th} writer will consist of n_i words.

Result: A matrix of $10 \times N \times K$ by 4 scores: $\{n, k, i, s(D_{i,1}^{*(k)}, D_{i,2}^{*(k)}) : k = 1, 2, ..., K; i = 1, ..., N\}$.

begin
- **for** $W_i; i \in \{1, ..., N\}$ **do**
  - **for** $n \in \{10, 20, 30, ..., 100\}$ **do**
    - **for** $k \in \{1, ..., K\}$ **do**
      - Randomly select, without replacement, $2n$ words from the n_i words in W_i, the first n generating the first pseudo-document $D_{W_{i,1}}^{*(k)}$, and the remaining n words generating the second pseudo-document $D_{W_{i,2}}^{*(k)}$.
      - Calculate the score, $s(D_{W_{i,1}}^{*(k)}, D_{W_{i,2}}^{*(k)})$.
    - **end**
  - **end**
- **end**

end

Analysis of the set of data that Algorithm 2 creates will provide information on the distribution of the similarity score between two writing samples provided by the same source.

In Step 3 of Algorithm 2, the match is determined by the previously defined threshold

value of τ. The proportion of pairs where $I\left(s(D^{*(k)}_{W_{i,1}}, D^{*(k)}_{W_{i,2}})\right) = 1$, or where there is no match, is an unbiased estimator of the RNMP for comparing two writing samples. As the number of writers increase for fixed lengths of writing samples being compared, this estimator of the RNMP is consistent.

As mentioned in 2.1.2, the proposed algorithm does not investigate the dependency of the RNMP with respect to the content of the writing samples being compared; the criterion used when selecting the words that create the simulated samples would have to be modified.

2.1.4 The Form of the Standard Error

Both the variance in (1.2) of the point estimator of the RMP defined in (1.1) and the associated Wald-type upper confidence bound on the RMP, $\tilde{\theta} + 2z_\alpha\sigma_c/\sqrt{N}$ are functions of the RMP as well as σ_c defined in (1.3). 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 *i.i.d.*. Under this assumption, $E(m_{ij}|D_i)$ does not depend on j and $E[E(m_{ij}|D_i)] = E(m_{ij}) = \theta$ for any $j \neq i$. So, for any $j \neq h \neq i$,

$$\begin{aligned}
\sigma_c^2 &= \text{Var}[E(m_{ij}|D_i)] \\
&= E[E(m_{ij}|D_i)E(m_{ih}|D_i)] - [E[E(m_{ij}|D_i)]]^2 \\
&= E[E(m_{ij}m_{ih}|D_i)] - \theta^2 \\
&= E(m_{ij}m_{ih}) - \theta^2 \\
&= Cov\,(m_{ij},\ m_{ih})
\end{aligned} \tag{2.1}$$

since

$$\begin{aligned} m_{ij}m_{ik} &= I\left\{s(D_i, D_j) > \tau\right\} I\left\{s(D_i, D_h) > \tau\right\} \\ &= I\left\{s(D_i, D_j) > \tau \text{ and } s(D_i, D_h) > \tau\right\}, \end{aligned}$$

the term in (6) 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 (6), this probability, which we refer to as the tri-match probability (TMP), when combined with the RMP, determines σ_c.

$$\sigma_c^2 = TMP - (RMP)^2$$

An important note is that both the TMP and the RMP, for reasonable matching algorithms, will tend to decrease as the length of the writing samples increase. This implies that by a small increase in the length of the writing sample, we have a dramatic decrease in the upper confidence bound for the RMP. Further down, we will use simple logistic regression models to characterize these relationships. The ability to characterize the behavior of this relationship is an important factor in determining the amount of writing the individuals write as well as the number of individuals to include in a study to explore handwriting individuality and sufficiency.

As discussed at the end of 2.1.5, the output from Algorithm 3 can be used to estimate the TMP for documents of a fixed length. Specifically, consider the output from Algorithm 3:

$$\left\{\left(s(D_i^{*(k)}, D_j^{*(k)}), s(D_i^{*(k)}, D_h^{*(k)}), s(D_h^{*(k)}, D_j^{*(k)})\right) : k = 1, 2, ..., K\right\}.$$

For a fixed threshold τ, this data set can be converted into a set of pairs that flag

whether each of the pairs of pseudo-documents match. Defining

$$m_{ij}^* = I\left\{s(D_i^*, D_j^*) > \tau\right\},$$

$$m_{ih}^* = I\left\{s(D_i^*, D_h^*) > \tau\right\},$$

and

$$m_{hj}^* = I\left\{s(D_h^*, D_j^*) > \tau\right\},$$

the output from Algorithm 3 can be viewed as:

$$\left\{(m_{ij}^{*(k)}, m_{ih}^{*(k)}, m_{hj}^{*(k)}) : k = 1, 2, ..., K\right\},$$

which provides information about the dependence of the TMP and σ_c on the sizes of the writing samples; the proportion of triplets for which both match, i.e.,

$$\tilde{\psi}_{hij}^*(n) \equiv K^{-1} \sum_{k=1}^{K} \frac{m_{ij}^{*(k)} m_{ih}^{*(k)} + m_{ij}^{*(k)} m_{hj}^{*(k)} + m_{ih}^{*(k)} m_{hj}^{*(k)}}{3}. \tag{2.2}$$

This will be used to construct an Incomplete U-Statistic of degree three that can be used to estimate the TMP. Under the assumption that $\nu(n) = E\left(m_{ij}^*(n),\ m_{ih}^*(n)\right)$, the results from Algorithm 3 can be used to construct a natural Incomplete U-Statistic of degree three:

$$\tilde{\nu}^*(n) = \binom{N}{3}^{-1} \sum_{h} \sum_{i} \sum_{j} \tilde{\psi}_{hij}^*. \tag{2.3}$$

This statistic in (2.3) will be a consistent estimator of ν_c as the number of writers increases for fixed sizes of writing samples. The proof will be analogous to the one used for the Incomplete U-Statistic of the RMP.

Alternatively, one can estimate σ_c using the relationship in (6) that $\sigma_c^2 = \nu_c - \theta^2$. 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 $\tilde{\theta}(n)$, then

$$\tilde{\nu}^*(n) - [\tilde{\theta}(n)]^2$$

is also a consistent estimator of σ_c by Slutsky's lemma (Serfling, 1980).

2.1.5 Estimating the Standard Error

The standard error of the estimate of the RMP (see 2.1.4) depends on using both the RMP and depends on the tri-match probability (TMP). We define the TMP as the probability of randomly selecting three writing samples from three different sources where the writing sample provided by the first source matches the writing sample from both the second source and the third source.

As the definition suggests, estimating the TMP involves comparing three simulated writing samples instead of two. The following algorithm is proposed to study the relationship between the TMP and the lengths of the writing samples being compared.

Analysis of the pairs of scores that make up this data that Algorithm 3 creates will provide information on the TMP and subsequently, the standard error of an estimate of the RMP.

In Step 3 of Algorithm 3, the match is determined by the previously defined threshold value of τ. The proportion of triplets where there is a match, is an unbiased estimator of the TMP for comparing three writing samples. As the number of writers increase for fixed lengths of writing samples being compared, this estimator of the TMP is consistent.

Algorithm 3: TMP

Data: A set of N writing samples from N writers. The set of writing samples for the i^{th} writer will consist of n_i words.

Result: A matrix of $10 \times \binom{N}{3} \times k$ by 8 paired scores: $\{n, k, h, i, j,\ s(D_i^{*(k)}, D_j^{*(k)}), s(D_i^{*(k)}, D_h^{*(k)}), s(D_h^{*(k)}, D_j^{*(k)}) : k = 1, 2, ..., K;\ n \in \{10, 20, 30, ..., 100\};\ h, i, j = 1, ..., N; h \neq i \neq j\}$.

begin
- **for** $W_h, h \in \{1, ..., N-2\}$ **do**
  - **for** $W_i, i \in \{W_h + 1, ..., N-1\}$ **do**
    - **for** $W_j, j \in \{W_i + 1, ..., N\}$ **do**
      - **for** $n \in \{10, 20, 30, ..., 100\}$ **do**
        - **for** $k \in \{1, ..., K\}$ **do**
          - Randomly select, without replacement, n words from the n_h words in W_h, generating the pseudo-document $D_h^{*(k)}$, n words from the n_i words in W_i, generating the pseudo-document $D_i^{*(k)}$, and n words from the n_j words in W_j, generating the pseudo-document $D_j^{*(k)}$.
          - Calculate the score, $s(D_h^{*(k)}, D_i^{*(k)})$
          - Calculate the score, $s(D_h^{*(k)}, D_j^{*(k)})$.
          - Calculate the score, $s(D_i^{*(k)}, D_j^{*(k)})$.
        - **end**
      - **end**
    - **end**
  - **end**
- **end**

end

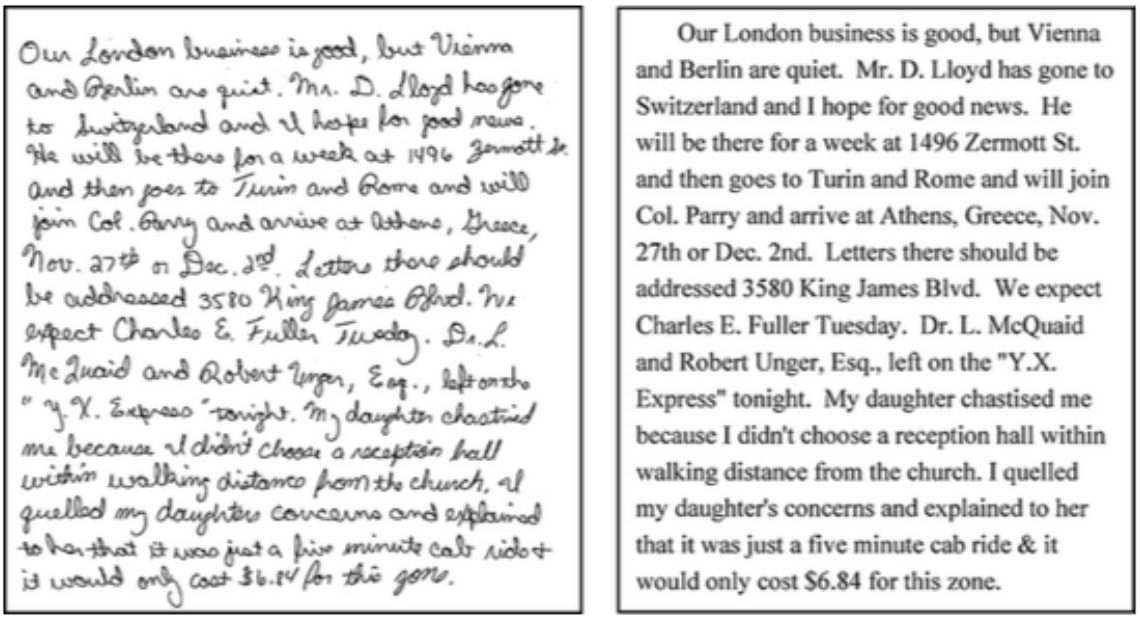
2.2 Applications

Throughout this section, we will demonstrate how the proposed algorithms mentioned in 2.1.5 can be used in a variety of applications associated with automated comparisons of writing samples. The comparison procedures we will be using are two developed by the Document Forensics Laboratory at George Mason University, and one developed at SDSU, and a collection of research writing samples collected by the FBI Laboratory and processed by Sciometrics. Note that the algorithms themselves are general - applicable to any comparison procedure and collection of writing samples.

The collection of writing samples we use was formed 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. Note that this collection is not a random sample representative of some relevant population, but instead is a convenience sample. With this in mind, the purpose of using this collection in this study is not to make a statement about the properties of any specific population, but to illustrate the algorithms proposed in 2.1.5.

Figure 2.1 displays the text and a writing sample for the modified "London Business Letter" (Osborn, 1929; Saunders et al. 2011a for the modification). For the collection of the writing samples, each volunteer was asked to provide a total of ten samples of the modified London Business Letter; five written in cursive and five written in hand printing. We will refer to this letter as the modified "London Letter". A Forensic Document Examiner made modifications to the original "London Letter" that consisted of inserting two sentences at the end in order to incorporate some occurrences of specific letter combinations (e.g. "ch", "qu", "ll"). The 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".



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 was just a five minute cab ride & it would only cost $6.84 for this zone.

Figure 2.1: A cursive and text form of the modified "London Letter" (Saunders et al. 2011a)

In order to assess the similarity between two documents using an automated process, the writing samples first had to be quantified, which we will give a brief description of how. Following manual character segmentation of each document, a proprietary automated process was used to give a representation to each segmented character. This representation is a mathematical graphic isomorphism where the internal structure can be enumerated by a code, and from now on we refer to as an isocode. The result of this process translates documents into isocode frequencies, which can be represented as a cross-classified table of letters by isocodes (Saunders et al., 2011b, Hepler et al., 2012, Davis et al., 2012).

The collection of writing samples used in this study consists of mostly cursive documents and some printed documents from which five documents each were provided by the first 100 volunteers, with a total of 500 documents after processing. The writing samples provided by a single writing source were combined for this study, this gives

one writing sample per writing source in the resulting collection.

Chi-Squared Classifier

As proposed in 2.1.5, we consider using the Chi-Squared Classifier as the similarity score, which is based on Pearson's chi-squared statistic (Saunders et al., 2011b). The development of this Chi-Square Statistic is from Davis et al., 2012.

Consider two simulated documents D_i^* and D_j^* where i and j represent the i^{th} and j^{th} writing profiles, $i \neq j$. For the l^{th} letter, consider the two vectors of counts produced from the simulated documents, n_{il} and n_{jl}, respectively. Using these two vectors, we construct a $2 \times \#(I_{obs})$, where I_{obs} is the set of observed unique isocodes of the l^{th} letter from the two documents, which we label as M. The Pearson Chi-Squared Statistic as outlined in Agresti (2012 pgs. 75-76) is as follows.

$$X_l^2 = \sum_i \sum_j \frac{(M_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

Where $\hat{\mu}_{ij} = \frac{M_{i+}M_{+j}}{n_{il}+n_{jl}}$ which is derived from the maximum likelihood estimates of the joint probabilities of each element of M. For this statistic, the degree of freedom is calculated by

$$df = 2 \times \#(I_{obs}) - 1.$$

The Corrected Chi-Square has the form

$$X_l^2 = \sum_i \sum_j \frac{(|M_{ij} - \hat{\mu}_{ij}| - 0.5)^2}{\hat{\mu}_{ij}}.$$

To produce the main "Omnibus" score to compare two documents,

1. Calculate the Chi-Square statistic for each letter and record the degrees of

freedom, which is outlined above.

2. Then, sum all of the statistics that were calculated for each letter, and separately sum up all of the degrees of freedom that were calculated for each letter.

$$X_T^2 = \sum_l X_l^2$$
$$DoF_T = \sum_l DoF_l$$

3. With the Omnibus Chi-Square Statistic and its corresponding degrees of freedom, 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.

$$P(X_{DoF}^2 \geq X_T^2 | DoF = DoF_T)$$

Where X_{DoF}^2 is the Chi-square random variable associated with the calculated Degree of Freedom.

Kullbeck-Liebler Distance

The development of the Kullbeck-Liebler Distance is presented from Hepler et al. (2012) for clarity.

Consider two simulated documents D_i^* and D_j^* where i and j represent the i^{th} and j^{th} writing profiles, $i \neq j$. For the l^{th} letter, consider the two vectors of counts produced from the sim documents, n_{il} and n_{jl}, respectively. Using these two vectors, we construct a $2 \times \#(I_{obs})$, where I_{obs} is the set of observed unique isocodes of the l^{th} letter from the two documents. We define

$$\nu_{il_k} = \frac{n_{il_k} + I_{obs}^{-1}}{n_{il_+} + 1}$$

and

$$\nu_{jl_k} = \frac{n_{jl_k} + I_{obs}^{-1}}{n_{jl_+} + 1},$$

where $n_{il_+} = \sum_{k=1}^{I_{obs}} n_{il_k}$, $n_{jl_+} = \sum_{k=1}^{I_{obs}} n_{jl_k}$, and $k = 1, ..., I_{obs}$ represents the number of unique isocodes used for the letter l from either document. We define the dissimilarity score for the l^{th} letter as

$$\delta(n_{il}, n_{jl}) \equiv \sum_{k=1}^{I_{obs}} \nu_{jl_k} ln(\frac{\nu_{jl_k}}{\nu_{il_k}}).$$

Note that when $I_{obs} = 1$, the dissimilarity score is zero. To calculate the dissimilarity score between two documents, i.e. using all letters $l = 1, ..., L$, we first want to define a set of weights such that the sum of the weights equals one,

$$\lambda_l \propto \begin{cases} \frac{1}{\sqrt{\frac{1}{n_{il_+}}}\sqrt{\frac{1}{n_{jl_+}}}} & min(n_{il_+}, n_{jl_+}) \geq 1 \\ 0 & otherwise. \end{cases}$$

This ensures a letter has a weight only when it occurs in both documents. To calculate the dissimilarity score of the two documents,

$$\delta(n_i, n_j) = \sum_{l=1}^{L} \lambda_l \delta(n_{il}, n_{jl}).$$

Multinomial

Let $D_1|\#(l) \sim Multi(\#(l), \underset{\sim}{\theta_1})$, where $\underset{\sim}{\theta_1} = (p_{1,1}, p_{1,2}, ..., p_{1,I})^T$, $p_{1,j} \sim F$, $j = 1, ..., I$ where I is the total number of isocodes known. F is the distribution function defined on a simplex of $\mathbb{R}^j$. Let C_{lj} represent the counts of the j^{th} isocode for the l^{th} letter of the questioned document. θ_{lm} is the proportion of the j^{th} isocode for the l^{th} letter.

Then,

$$\begin{aligned} f(C_l|\theta_l) &= \binom{C_{l+}}{C_{l1}C_{l2}...C_{lI}} \Pi_{j=1}^{I} \theta_{lj}^{C_{lj}} \\ &= C_{1+}! \Pi_{j \in \{C_{lj}>0\}} \frac{\theta_{lj}^{C_{lj}}}{C_{lj}!} \Pi_{j \in \{C_{lj}=0\}} \frac{\theta_{lj}^{C_{lj}}}{1}. \end{aligned} \tag{2.4}$$

If we let $j^* = j$ if $C_{u_{lj}} > 0$ for each l, then

$$f(C_{u_l}|\theta_l) = \binom{C_{u_{l+}}}{C_{u_{l1^*}}C_{u_{l2^*}}...C_{u_{lJ^*}}} \Pi_{j^*=1}^{J^*} \theta_{lj^*}^{C_{u_{lj^*}}}. \tag{2.5}$$

Note that $\sum_{j^*=1}^{J^*} \theta_{lj^*} < 1$.

This implies that we can only consider the observed categories plus one, and thus reduces the dimensionality.

Binomial-Poisson

This score is referred to as the Binomial Method throughout this paper.

This matching score is based on a poisson process. We will consider each isocode/letter pairing and assume the two documents being compared are the same length. With this set in place, we test to see if the two sets of the poisson process are equivalent by testing if the counts of each isocode/letter pairing are equal across the two documents. If we condition on the total number of counts in a isocode/letter pairing, we are essentially testing

$$H_0 : N_1 \sim Binom\left(p = 0.5, N_1 + N_2\right). \tag{2.6}$$

If we use the exact level α p-value (Casella and Berger, 1990, pg. 368) for this test, we will have one Uniform random variable for each observed isocode/letter pairing

between the two documents. We then combine the p-values using Fisher's Method (Fisher 1948).

$$\sum_{j,l}^{I_{obs},L_{obs}} -2ln(p_{j,l}) \sim \chi^2_{2\times I_{obs}L_{obs}} \tag{2.7}$$

We use a probability inverse transform to produce a theoretical Uniform similarity score, as long as the two documents were written by the same writer.

2.2.1 Determining an Appropriate Threshold Value

The RMP and the RNMP can be used to select a suitable threshold to use with a comparison procedure for declaring a match between two writing samples. One method for selecting a threshold value is to choose the threshold so that the rate of false match errors equals the rate of false no-match errors. The resulting rate is called the equal error rate (EER) and is a standard method used to compare the "matching" accuracy across comparison procedures, mainly comparison procedures designed for biometric authentication systems. Another method for selecting a threshold value is to give a pre-specified rate of no-match errors, e.g. 1%. This method is frequently used in forensic settings where the consequences of the two types of errors are not equal, with the false match error commonly considered to have heavier consequences.

With respect to the proposed algorithms, we will consider determining the threshold value that will produce a RNMP of 1%. If the individual characters chosen for the algorithm are considered a random sample from an individual's writing profile, then the similarity score calculated with the Chi-Squared Classifier is related to an approximate p-value. If we assume independence across the characters, (theoretically) the similarity score, when applied to two randomly selected writing samples from the same source, will have approximately a uniform distribution, independent of the length and content that make up the two writing samples in the comparison. This

suggests that the 1% RNMP threshold for the Chi-Squared Classifier should be 0.01, with the assumption that the characters are independent.

Assuming the characters are independent is questionable, though. Which would mean the 1% RNMP threshold might not be 0.01, and might change with the length and content of the writing samples. Since the two samples are from the same source, we will use Algorithm 2 and simulate writing samples to study empirically the dependence on the length (though not necessarily the dependence on the content) by estimating the RNMP for a variety of lengths of writing samples. The algorithms are applied to words, and thus the dependence between characters within a word is preserved.

We applied Algorithm 2 with 10 different lengths of the simulated writing samples, starting at $n = 10$, and increasing by increments of 10, until $n = 100$. Each length chosen was ran through the algorithm nine times, producing nine sets of 10 scores for each comparison.

Figures 2.2-2.9 are plots of the empirical cumulative distribution function (ECDF)[4] of the nine similarity scores produced using Algorithm 2 for each of the 10 lengths of writing samples. The 45-degree line that is overlaid on a few of the plots is the cumulative distribution function (CDF) for the uniform distribution.

The behavior of the EDCF in Figures 2.4-2.9 are similar for a given method regardless of the 10 different lengths of writing samples. Comparing the EDCF to the CDF for a uniform distribution, there is very little similarity between the two, as the EDCF fluctuates around the CDF. This suggests that for all 10 lengths of writing samples, the similarity scores for within-writer comparisons tend to be more concentrated toward smaller values than would be expected if the similarity scores followed a uniform distribution.

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

Notice in nearly every plot in Figures 2.4-2.9, the ECDF is near or below the uniform CDF line when the score values are less than 0.01; and the EDCF appears to be moving closer to the CDF (as the score values increase towards 0.01) as the length of the writing samples increase in terms of word counts. Even though the uniform approximation is not accurate across the similarity score values, this observation suggests that the 1% RNMP threshold is close to 0.01, or maybe even slightly larger than 0.01. With this in mind, 0.01 appears to be a reasonable choice (though conservative, particularly for smaller lengths of writing samples) for the threshold value 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%. A threshold value of 0.01 is conservative, particularly for shorter lengths of writing samples, and using a conservative value for the RNMP threshold will result in overestimating the RMP associated with the actual 1% RNMP threshold, and so the RNMP threshold should be bigger. However, this appears to be less of an issue as the length of writing samples being compared increases.

2.2.2 Estimating the RMP as a Function of Length of Writing Samples

We can investigate the RMP associated with the Chi-Squared Classifier and a threshold of $\tau = 0.01$ by applying one of our proposed algorithms.[5]

We applied Algorithm 1 with 10 different lengths of the simulated writing samples, starting at $n = 10$ and increasing by increments of 10, ending when $n = 100$.[6] Each length chosen was ran through the algorithm 9 times, producing 9 sets of 10 scores

[5]This threshold corresponds to a rate of no-match errors of at most 1%, which was suggested by the results in the previous sub-section.

[6]It is not possible to accurately estimate very small RMP with $K = 9$, and so we considered common lengths of writing samples to be at most 100.

Model	Intercept	Slope
Binomial Theoretical Threshold	2.985	-0.058
Binomial Empirical Threshold	2.985	-0.058
Uncorrected Chi-Square Theoretical	2.876	-0.109
Uncorrected Chi-Square Empirical	2.548	-0.111
Corrected Chi-Square Theoretical	3.526	-0.093
Corrected Chi-Square Empirical	2.577	-0.107
Kullbeck-Liebler Empirical	1.405	-0.017

Table 2.1: Fitted Simple Logistic Regression on the Models for RMP

per comparison.

As illustrated in Figures 2.10-2.13, the proportion of the scores that exceeded 0.01 was calculated for each set of scores. This graph illustrates the dependency of the RMP on the length of the writing samples being compared, where the RMP is approaching zero as the lengths of the writing samples increase. In order to produce a reasonable estimate of the trend, we fitted a logistic regression model cubic in length of writing sample. Table 1 shows the resulting fit, and is plotted as a solid line in Figures 2.10-2.13. Based on the fitted logistic curve, the RMP associated with the Uncorrected Chi-Squared Classifier with threshold 0.01 is less than 10% when comparing lengths of writing samples to be at least 50 words long, and less than 1% when comparing lengths of writing samples to be at least 65 words long.

The logistic parameters are estimated by solving the following equation

$$\hat{\beta} = argmin_{\beta} \sum_{n} \left(\hat{\theta}^{*}(n) - \theta_{\beta}(n) \right)^{2},$$

recalling that

$$ln \left(\frac{\theta_{\beta}(n)}{1 - \theta_{\beta}(n)} \right) = \beta_0 + \beta_1 n.$$

2.2.3 Estimating the TMP and Standard Error

One approach to constructing upper confidence bounds on the RMP would be to use a Wald-type upper confidence bound, which is described in 1.3.1. A Wald-type upper confidence bound is an approximate confidence set based on the asymptotic normality of a parameter estimate. For a given parameter estimate $\tilde{\theta}$, it typically has the form $\tilde{\theta} + \mathcal{Z}_{1-\alpha} \times SE$, where $\mathcal{Z}_{1-\alpha}$ is the z-value associated with the chosen α-level, and SE stands for the standard error (or any consistent estimator of the standard error). For example, for a 95% upper confidence bound, the standard error would be multiplied by 1.645.

As mentioned above, in order to calculate this type of upper confidence bound, the standard error of the estimated RMP needs to be known. However, the standard error of the RMP estimator is related to the TMP, which can be estimated using Algorithm 3.

Given enough time, we would apply Algorithm 3 with 10 different lengths[7] of the simulated writing samples, a threshold of $\tau = 0.01$, would start at $n = 10$ and increasing by increments of 10, ending when $n = 100$. Each length chosen would be ran through the algorithm 9 times, producing 9 sets of 10 scores per comparison.

The application of Algorithm 3 using the same values of K and choices of writing sample length is computationally complex, as it would take upwards of three months to run. We propose a new algorithm to randomly sample approximately 600 triplicates from the entire set of possible triplicates[8].

As Figures 2.15-2.18 illustrates, the proportion of pairs for each set of score pairs where both scores exceeded 0.01 was calculated. Figures 2.15-2.18 shows the dependency of

[7] The values of K and the choices of writing sample length are the same as the RMP application in 2.2.2.

[8] 600 was chosen because that is approximately how many calculations that could be performed in two days using available computational resources.

Algorithm 4: TMP Estimate

Data: A set of N writing samples from N writers. The set of writing samples for the i^{th} writer will consist of n_i words.
The set of each triplicate formed by the N writers,
$T = \{\{W_1, W_2, W_3\}, \{W_1, W_2, W_4\}, ..., \{W_{N-2}, W_{N-1}, W_N\}\}$, where T has $\binom{N}{3}$ elements.
$T_M = \{\{W_{1,1^\star}, W_{1,2^\star}, W_{1,3^\star}\}, \{W_{2,1^\star}, W_{2,2^\star}, W_{2,3^\star}\}, ..., \{W_{M,1^\star}, W_{M,2^\star}, W_{M,3^\star}\}\}$, a random sample from T of M triplicates.
Result: A matrix of $10 \times K \times M$ by 8 paired scores:
$\{n, k, \{m, 1^\star\}, \{m, 2^\star\}, \{m, 3^\star\},\ s(D^{*(k)}_{m,1^\star}, D^{*(k)}_{m,2^\star}),\ s(D^{*(k)}_{m,1^\star}, D^{*(k)}_{m,3^\star}),$
$s(D^{*(k)}_{m,2^\star}, D^{*(k)}_{m,3^\star}) :\ k = 1, 2, ..., K;\ \text{for } n \in \{10, 20, 30, ..., 100\};\ m = 1, ..., M\}$.

```
begin
    for {W_{m,1*}, W_{m,2*}, W_{m,3*}}, m = 1, ..., M do
        for k ∈ {1, ..., K} do
            for n ∈ {10, 20, 30, ..., 100} do
                Randomly select, without replacement, n words from the n_{m,1*}
                 words in W_{m,1*}, generating the pseudo-document D^{*(k)}_{m,1*}, n
                 words from the n_{m,2*} words in W_{m,2*}, generating the
                 pseudo-document D^{*(k)}_{m,2*}, and n words from the n_{m,3*} words in
                 W_{m,3*}, generating the pseudo-document D^{*(k)}_{m,3*}.
                Calculate the score, s(D^{*(k)}_{m,1*}, D^{*(k)}_{m,2*})
                Calculate the score, s(D^{*(k)}_{m,1*}, D^{*(k)}_{m,3*}).
            end
        end
    end
end
```

the TMP on the length of the writing samples being compared; the TMP approaching zero as the length of the writing samples grow larger. Following in the path of the RMP, we again plotted a logistic regression model, cubic in length of writing samples, to provide a reasonable fit and a smooth curve to represent the relationship between the TMP and the length of the writing samples. Table 2 summarizes the results of the logistic curve, and is plotted as the solid line in Figures 2.15-2.18.

Figures 2.15-2.18 also include a dotted line for comparison. This is the logistic model fit to the estimated RMP from 2.2.2. If we compare the two curves, observe that the estimated RMP and the estimated TMP act similar for shorter lengths of writing samples. Also note that the estimated TMP drops down more rapidly as the length of the writing samples increase.

Figure 2.14 illustrates the relationship between the estimated variance of the conditional match probabilities[9] of the estimated RMP and the number of writers across the different lengths of writing samples. Using (2) given in 1.3.1[10], these estimates combine the two logistic curves shown in Figures 2.15-2.18, as well as an additional logistic model. The logistic model for the RMP in Figures 2.10-2.13 is fit to the proportion of pairs from the simulated samples that match when comparing the first and the second simulated samples in the output from Algorithm 3. The estimates shown in Figures 2.15-2.18 use the same type of logistic model, also cubic in length of writing sample, and is fit to the proportion of pairs from the simulated samples that match when comparing the first and the third simulated samples in the output from Algorithm 3. Even though the standard error is affected by both the length of the writing samples and the number of writers represented, it is more sensitive to the lengths than to the number of writers.

[9] As is clear from the form of the standard error of the U-Statistic, the variance of the conditional match probability is proportional to the standard error of the estimated RMP.

[10] This is not suggesting that the best estimator of the standard error of this data is the combination of fitted logistic models for the RMP and TMP. The estimator discussed here is to show one approach to estimating the standard error. See 2.1.4 for a more detailed discussion of this issue.

Model	Intercept	Slope
Binomial Theoretical Threshold	2.957	-0.078
Binomial Empirical Threshold	2.743	-0.078
Uncorrected Chi-Square Theoretical Threshold	3.189	-0.157
Uncorrected Chi-Square Empirical Threshold	2.649	-0.155
Corrected Chi-Square Theoretical Threshold	3.983	-0.132
Corrected Chi-Square Empirical Threshold	2.767	-0.155
Kullbeck-Liebler Empirical Threshold	0.821	-0.0182

Table 2.2: Fitted Simple Logistic Regression on the Models for TMP

The variance of the conditional match probability plots in Figure 2.14 suggest that for a short length of writing samples (i.e., 100 words), the standard error of the estimated RMP is going to be essentially zero. Knowing this, using shorter lengths of writing samples will not be as useful when trying to precisely bound a very small RMP. Doubling the length of the writing samples to 200 does not provide much in reducing the standard error.

2.2.4 Notes-Approximation Results for Sub-sampling U-Statistics estimates of RMP

In this section we will discuss some of the statistical properties of the estimates of the RMP from the previous sections. The main goal is to demonstrate that the limiting form of the sub-sampling estimates, for a specified word count, are U-statistics of degree two (or three in the case of the estimates of the variance of the conditional match probability). We will use a slightly different set of notation to facilitate the discussion of the theoretical properties.

Assume we have N vectors of counts denoted as D_i, $i = 1,\ 2, \ldots, N$. We will assume that $D_i = \sum_{j=1}^{n_i} D_{ij},\ j = 1, 2, 3, \ldots, n_i$, where for fixed i,

$$D_{ij} \overset{i.i.d.}{\sim} Multi(1,\ p_i).$$

This immediately implies that, conditional on p_i and n_i,

$$D_i \sim \ Multi(n_i,\ p_i).$$

We will assume that $p_i \overset{i.i.d.}{\sim} F, \quad i = 1, 2, \ldots, N$, where F is a distribution function defined on a simplex of $\mathbb{R}^M$.

Let

$$m_{ij} = m\left(D_i,\ D_j\right) = I\left(s\left(D_i,\ D_j\right) > \tau\right),$$

where $s\left(\cdot,\ \cdot\right)$ is a function that maps two vectors of counts to $[0,\ \infty)$ as a similarity score. Then the random match probability conditional on the length of the two documents being compared is

$$\theta\left(n_i,\ n_j\right) = Em_{ij}.$$

We are specifically interested in estimating the random match probability between two documents of common length n of interest, with n_i being the actual number of words in the i^{th} document; if $n \leq n_i$ for $i = 1,\ 2, \ldots, N$, then $\theta\left(n\right) = Em_{ij}$.

A sub-sampled document of size n from a writing sample, say D_i^*, is generated by sampling n words without replacement from the original n_i words that make up D_i. Denote the k^{th} sub-sampled document from D_i as the ${D_i}^{*(k)}$.

Lemma 2.1:
Let D_i and ${D_i}^{k*}$ be defined as above. Then,

i) Conditional on D_i; ${D_i}^{*(k)}$ are i.i.d. multivariate Hyper-geometric random variables, for $k = 1, 2, 3, \ldots$

ii) ${D_i}^{*(k)} \overset{ind.}{\sim} Multi(n,\ p_i),\ i = 1,\ 2, \ldots, N.$

Proof: Part i) By definition of the hyper-geometric random variables.

Part ii) For a multinomial random vector, we can take advantage of the fact that, for f being probability mass functions associated with multinomial random vector X,

$$f(X) = f(X_1)f(X_2|X_1)f(X_3|X_2X_1)\dots f(X_I|X_{I-1}\dots X_1)$$

to allow us to only focus on binomial and hypergeometric random variables. Let X and Y be random variables such that

$$X \sim Binomial(M, P)$$

and

$$Y|X = x \sim HyperGeometric(x, J - x, m).$$

Let $B_j \overset{iid.}{\sim} Bernoulli(p)$. Then X can be written as

$$X = \sum_{j=1}^{J} B_j,$$

furthermore we can decompose X as

$$\begin{aligned} X &= \sum_{j=1}^{m} B_j + \sum_{j=m+1}^{J} B_j \\ &\equiv Y + Y^c. \end{aligned}$$

Next we will look at all possible permutations of the B_j's and consider a probability distribution on the permutations such that each permutation is equally likely to be observed. Let Q be a random vector taking values on $\mathcal{Q}$; the set of permutations of the objects $\{1,\ 2, \dots, J\}$. Now for $Q = q$, let $j^\star$ be the $j^{\star th}$ entry in the vector q, we

can then write

$$\begin{aligned} X|_{Q=q} &= \sum_{j^\star=1}^{J} B_{j^\star} \\ &= \sum_{j^\star=1}^{m} B_j + \sum_{j^\star=m+1}^{J} B_{j^\star} \\ &\equiv Y_q + Y_q^c. \end{aligned}$$

Since the B_j's are *i.i.d.*, then there is no dependence on the selected permutation. This directly implies that Y and Y_q have the same distribution. Therefore, $Y \sim Binomial(m, p)$ and $Y^c \sim Binomial(N-m, p)$, which in turn implies that ${D_i}^{*(k)} \overset{i.i.d.}{\sim} Multi(n_i,\ p_i),\ i = 1,\ 2, \ldots, N.$ ■

An implication of Lemma 1. is that

$$\theta(n) = Em_{ij}^{*(k)} = Em\left(D_i^{*(k)},\ D_j^{*(k)}\right).$$

Consider two writing profiles for which we have observed writing samples from, denote these profiles as p_i and p_j, let $T_i \sim Multi(n,\ p_i)$ and $T_j \sim Multi(n,\ p_j)$.

Then define

$$\theta_{ij}(n) = E\left(m\left(T_i,\ T_j\right)|\, p_i,\, p_j\right)$$

and note that

$$E\left(\theta_{ij}(n)\right) = \theta(n).$$

For two given writers, say i and j, $\theta_{ij}(n)$ is the probability that we observe a match between two of these writers' documents that are composed of n words.

Next consider a document, D, made up of n_D distinct words. There are exactly

$$R = \binom{n_D}{n}$$

distinct subsets of words of size n; denote the set of all possible subsets as

$$\left\{D^{(r)}\right\}_{r=1}^{R}.$$

Consider D_i and D_j, as well as their corresponding subsets of words of size n :

$$R_i = \binom{n_i}{n}$$

$$\left\{D_i^{*(r)}\right\}_{r=1}^{R_i},$$

$$R_j = \binom{n_j}{n}$$

and

$$\left\{D_j^{*(r)}\right\}_{r=1}^{R_j}.$$

A possible estimate of $\theta_{ij}(n)$ is

$$\widetilde{\theta}_{ij}(n) = {R_i}^{-1}{R_j}^{-1}\sum_{r=1}^{R_i}\sum_{r'=1}^{R_j} m\left({D_i}^{*(r)},\ {D_j}^{*(r')}\right).$$

Next we will define a U-statistic that estimates θ under the above assumptions:

$$\tilde{\theta}_N(n) = \binom{N}{2}^{-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\widetilde{\theta}_{ij}(n).$$

Lemma 2.2:

Let $\{D_i\}_{i=1}^{N}$ *satisfy the above assumptions. Let* $\left\{D_i^{*(r)}\right\}_{r=1}^{R_i}$ *be defined as above. Let* $\theta(n)$ *and* $\tilde{\theta}_N(n)$ *be defined as above.*

Then, for fixed n*,*

$$\sqrt{N}\left(\tilde{\theta}_N(n) - \theta(n)\right) \rightsquigarrow N(0,\ 4\xi_1),\ as\ N \to \infty,$$

for $\xi_1 = \text{var}\left(\widetilde{\theta}_i\right) > 0$ *and* $\widetilde{\theta}_i(n) = E\left(\widetilde{\theta}_{ij}(n)\middle| D_i\right)$*, for* $i = 1, 2, ..., N$.

Proof: As mentioned above, $\tilde{\theta}_N(n)$ is a U-statistic of degree two. The kernel of $\tilde{\theta}_N(n)$ is $\tilde{\theta}_{ij}(n)$ which is a bounded symmetric function with expectation θ. As long as $\text{var}\left(\tilde{\theta}_i\right) > 0$, the result follows from Theorem A of Section 5.5.1 of Serfling (1980). ■

As in 2.1.2, $D_1^{*(k)}$ and $D_2^{*(k)}$ be sub-sampled documents from the k^{th} draw and

$$\tilde{\theta}^* = \binom{N}{2}^{-1} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[K^{-1} \sum_{k=1}^{K} I\left(s\left(D_i^{*(k)}, D_j^{*(k)}\right) > \tau\right)\right].$$

[Note that the distribution of $K\tilde{\theta}^*(n)$ given $\{D_i\}_{i=1}^{N}$ is binomial with probability of success $U_N(n)$ and K trials.] The variance of $\tilde{\theta}^*(n)$ is then

$$\begin{aligned} Var\left(\tilde{\theta}^*\right) &= Var\left(E\left(\tilde{\theta}^*|D_1, ..., D_N\right)\right) + E\left(Var\left(\tilde{\theta}^*|D_1, ..., D_N\right)\right) \\ &= Var\left(\tilde{\theta}\right) + \frac{E\left(\theta_{ij}\left(1 - \theta_{ij}\right)\right)}{K}\binom{N}{2}^{-1} \\ &\Longrightarrow \left|Var\left(\tilde{\theta}^*\right) - Var\left(\tilde{\theta}\right)\right| \leq \frac{1}{4K}\binom{N}{2}^{-1} \end{aligned}$$

Note that as K, the number of iterations in the subsampling algorithm increases, the

variance of $\tilde{\theta}^*(n)$ approaches the variance of $\tilde{\theta}(n)$.

An additional aspect of this development, which I plan on working on for my Ph.D., is that we have developed a class of U-statistics which are indexed by the length of the documents considered in the kernel. This generalizes the U-statistics that we have focused on in this research project into a class known as U-processes. This representation is a major goal of this specific research project.

2.2.5 Contributions to Designing a Study of Handwriting Individuality

One critical part of planning a study is to select the number of observations. One procedure is to designate a desired margin of error associated with estimating a parameter of interest and select the number of observations which produce the given 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. Discussed in detail in Bolle et al. (2004) and Saunders et al. (2011a), and mentioned in 1.2, an upper bound on the RMP is also an upper bound on the infrequency of matching writing profiles. Thus, for an empirical study of the individuality of handwriting within a specific population, one might consider selecting the number of writers to produce a desired upper bound on the RMP (assuming the "true" value for the RMP is very close to zero).

Obtaining a small upper bound when there are zero observed matches is ideal. The smallest possible upper bound on the RMP (when writing samples are compared pairwise) occurs when there are no observed matches from a collection of writing samples taken from a large number of writers.

Using a Wald-type upper bound (1.3.1) with an estimated standard error based on simulated samples, as described in 2.2.3, requires a large number of writers to produce writing samples in order to be very precise. Though, typically, a large set of writing samples is not available at the planning stages of a study. One alternative would be to use one of the proposed estimators of the standard error that is based entirely on an observed set of writing samples without subsequent subsampling. However, most of these proposed estimators, such as those proposed 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) show that adding two "successes" and two "failures" to the sample gives coverage probabilities close to the nominal confidence levels of an adjusted Wald interval. In order to investigate the coverage probability of a Wald-type upper confidence bound on the RMP, we have performed a small simulation study with a similar adjustment of adding one match and one no match[11] to Wayman's (2000) estimate of standard error. Our preliminary investigations suggest that this adjustment produces coverage probabilities close to the nominal confidence levels.

When adding one match and one no match to the sample when there are no observed matches, the formula for the 95% upper confidence bound using Wayman's (2000) estimate of the standard error simplifies to $4.65/[(N+1)(N+2)]$, where N is the number of writers. This shows that the more writers involved in the study, the upper bound for when there are no observed matches is smaller, and will result in a conservative interval in the sense that the coverage probability has increased. For example, in a sample of 963 writers with no observed matches would yield a 95% upper confidence bound on the RMP of 5 in one million, and a sample 2,154 writers

[11]To be more specific, the procedure consists of adding one sample that matches exactly one observed sample, and adding one sample that does not match any of the observed samples. This ensures there is exactly one match out of all of the comparisons.

with no observed matches would yield a 95% upper confidence bound on the RMP of about one in 1 million.

Keep in mind that the examples mentioned above assumes no observed matches when the writing samples are compared pairwise. And, given that the 'true' RMP is not zero, for a fixed length of writing samples and fixed RNMP threshold, the probability of observing a match increases as more writers are introduced and being compared. However, the length of the writing samples affects the RMP and in turn affects the probability of observing no matches, as shown in Figures 2.10-2.13. Therefore, in order to obtain a small upper bound, the length of the writing samples as well as the number of writers must be considered.

The modeling techniques developed and proposed from this work will allow researchers to accurately model the necessary match probabilities needed to design individuality or sufficiency studies as in Saunders et al. (2011).

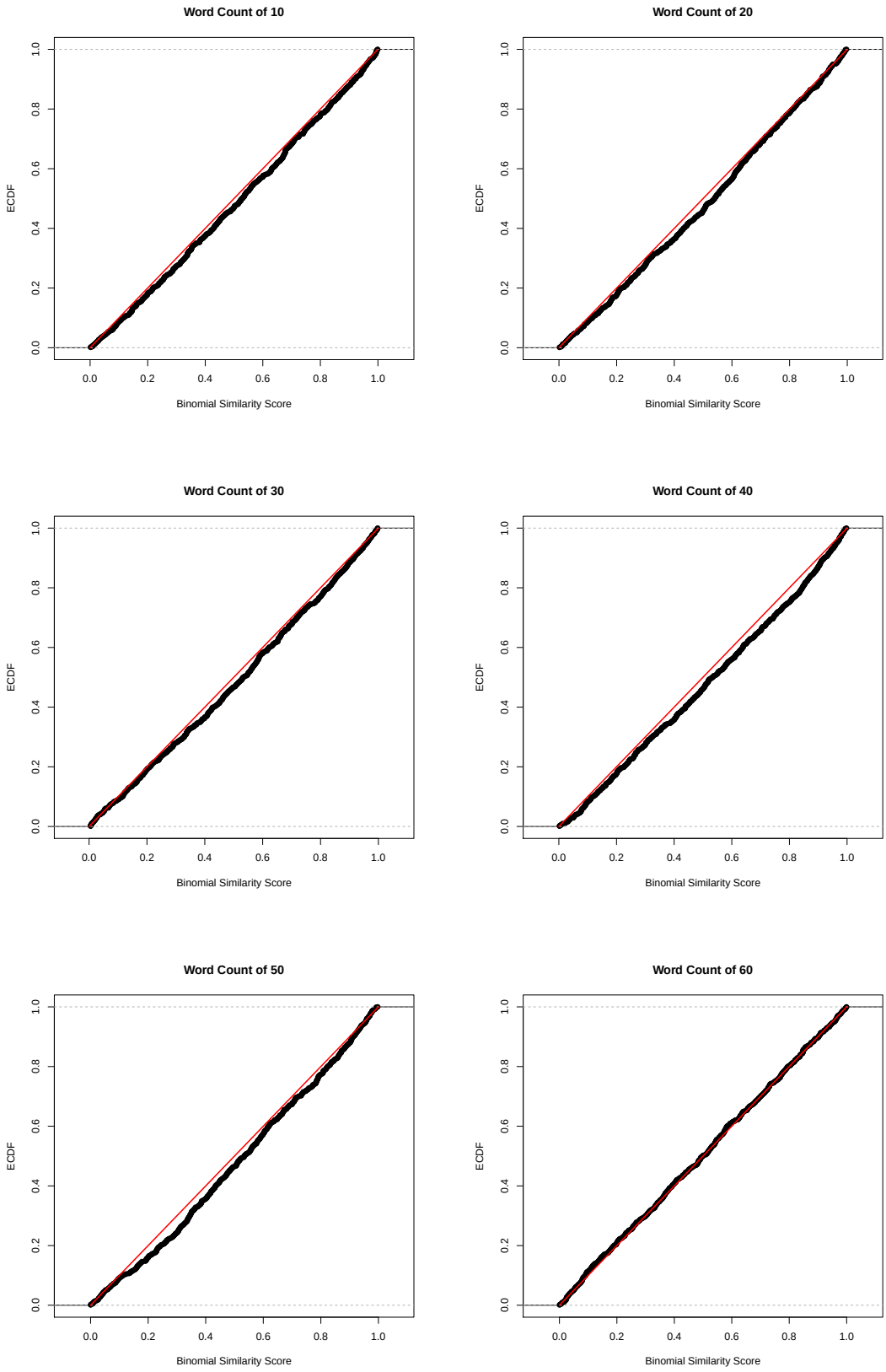


Figure 2.2: Binomial RNMP by Word Count

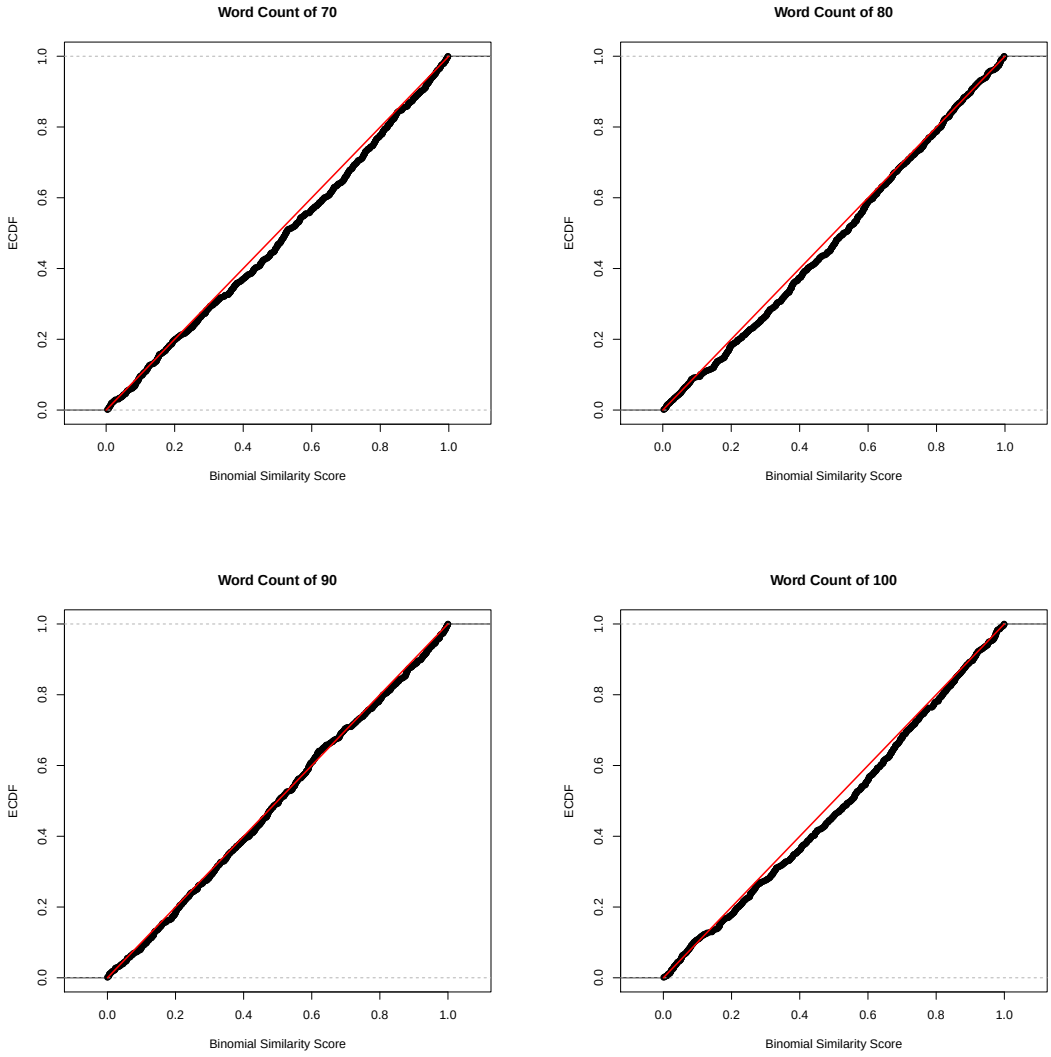


Figure 2.3: Binomial RNMP by Word Count

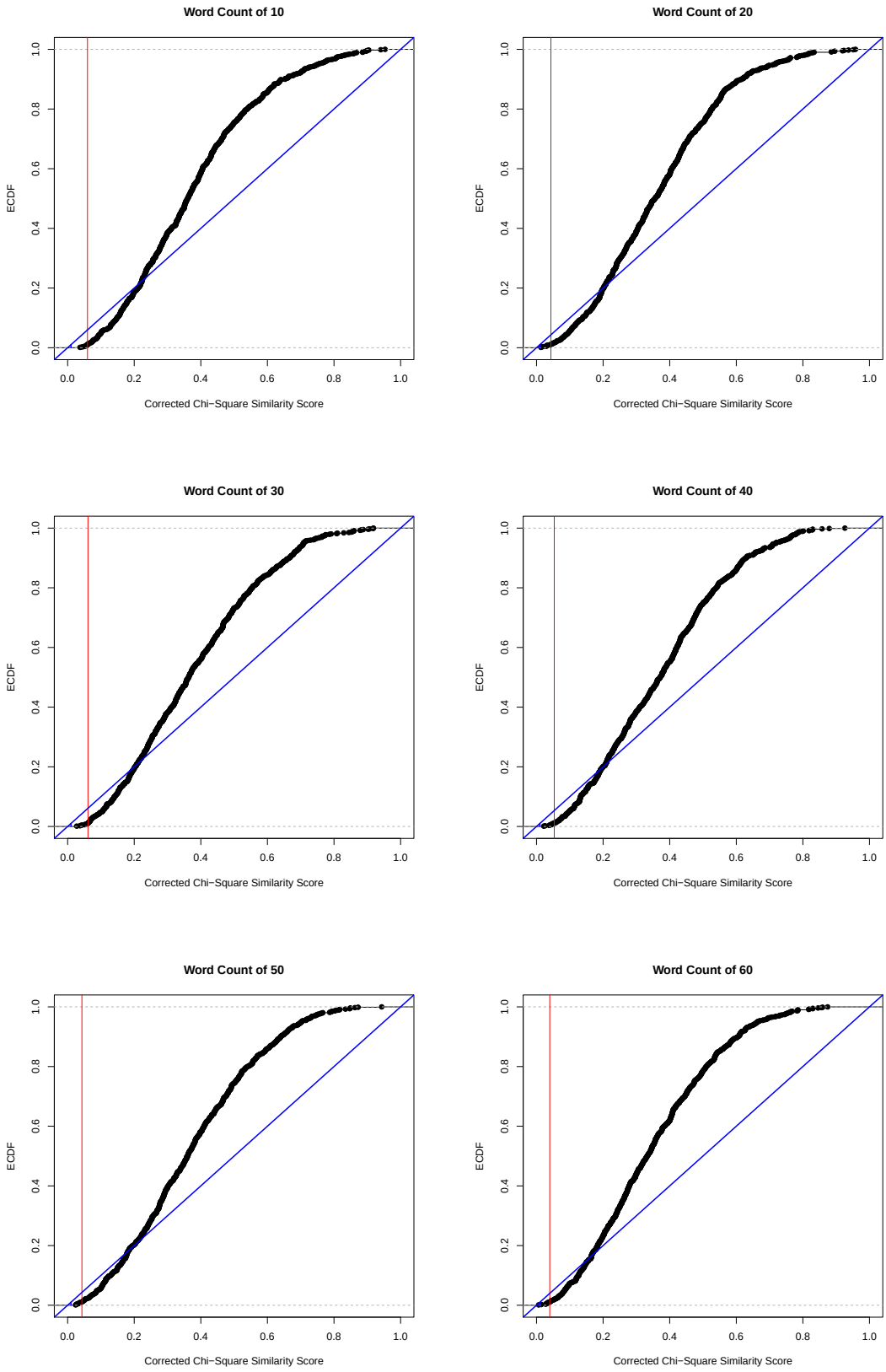


Figure 2.4: Corrected Chi-Square RNMP by Word Count

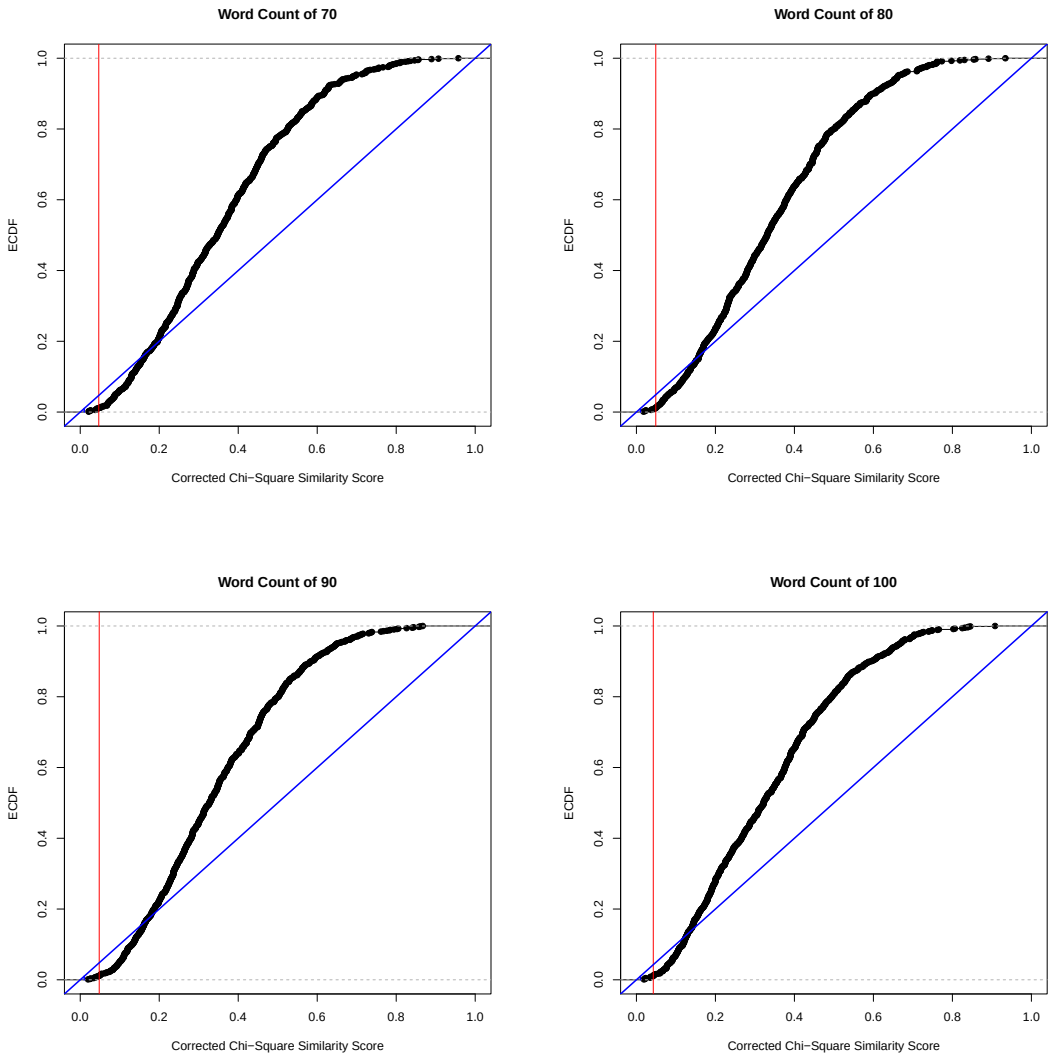


Figure 2.5: Corrected Chi-Square RNMP by Word Count

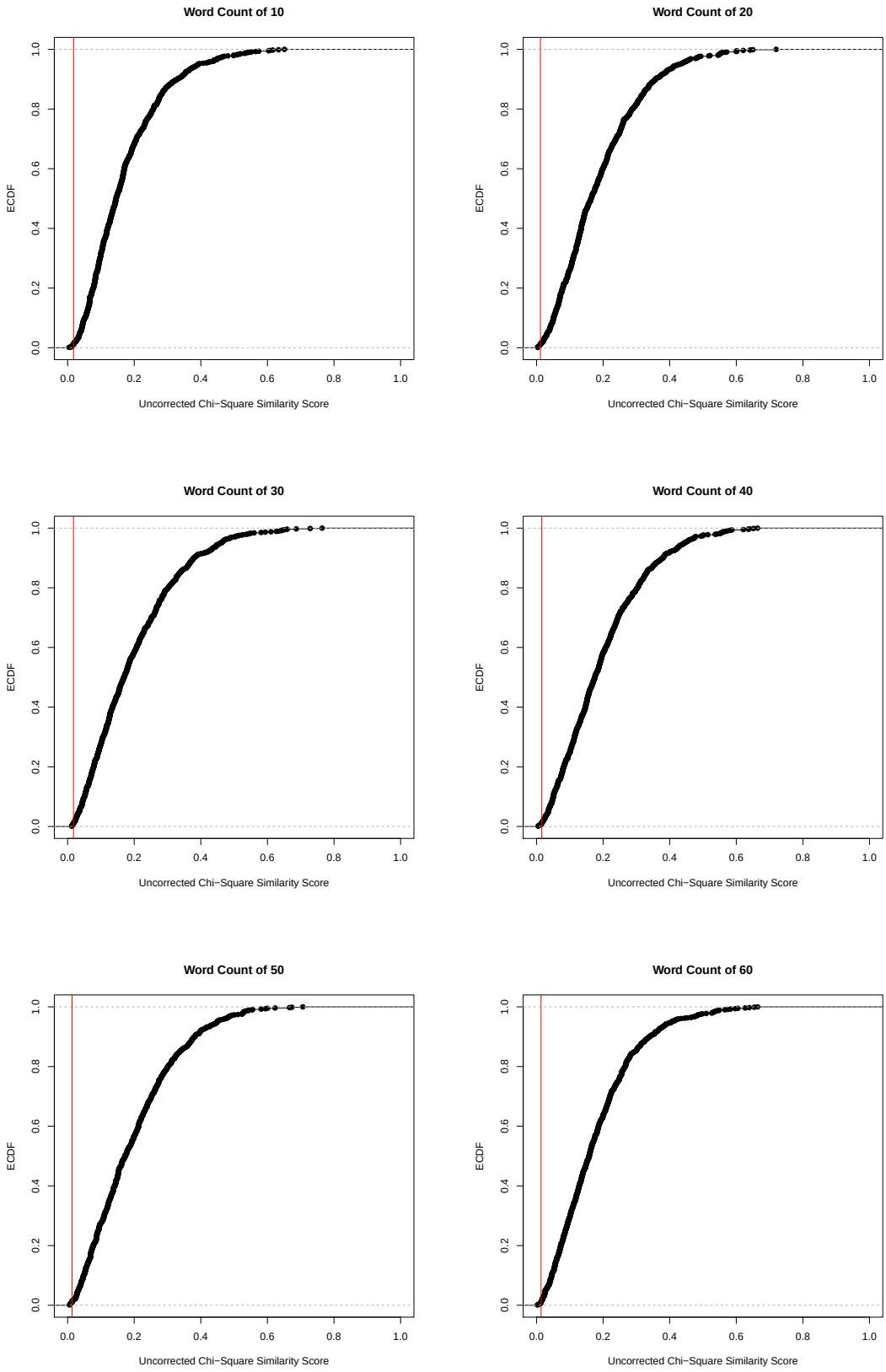


Figure 2.6: Uncorrected Chi-Square RNMP by Word Count

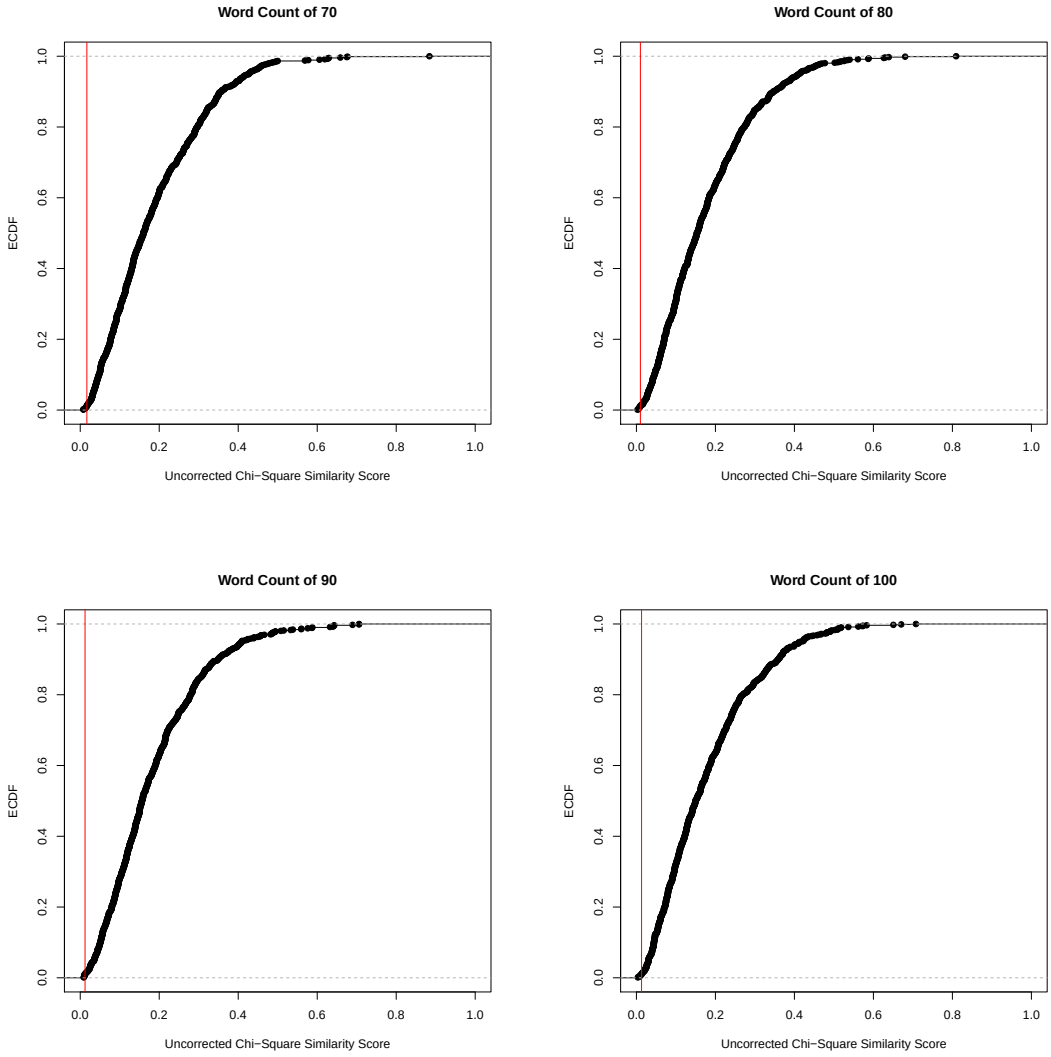


Figure 2.7: Uncorrected Chi-Square RNMP by Word Count

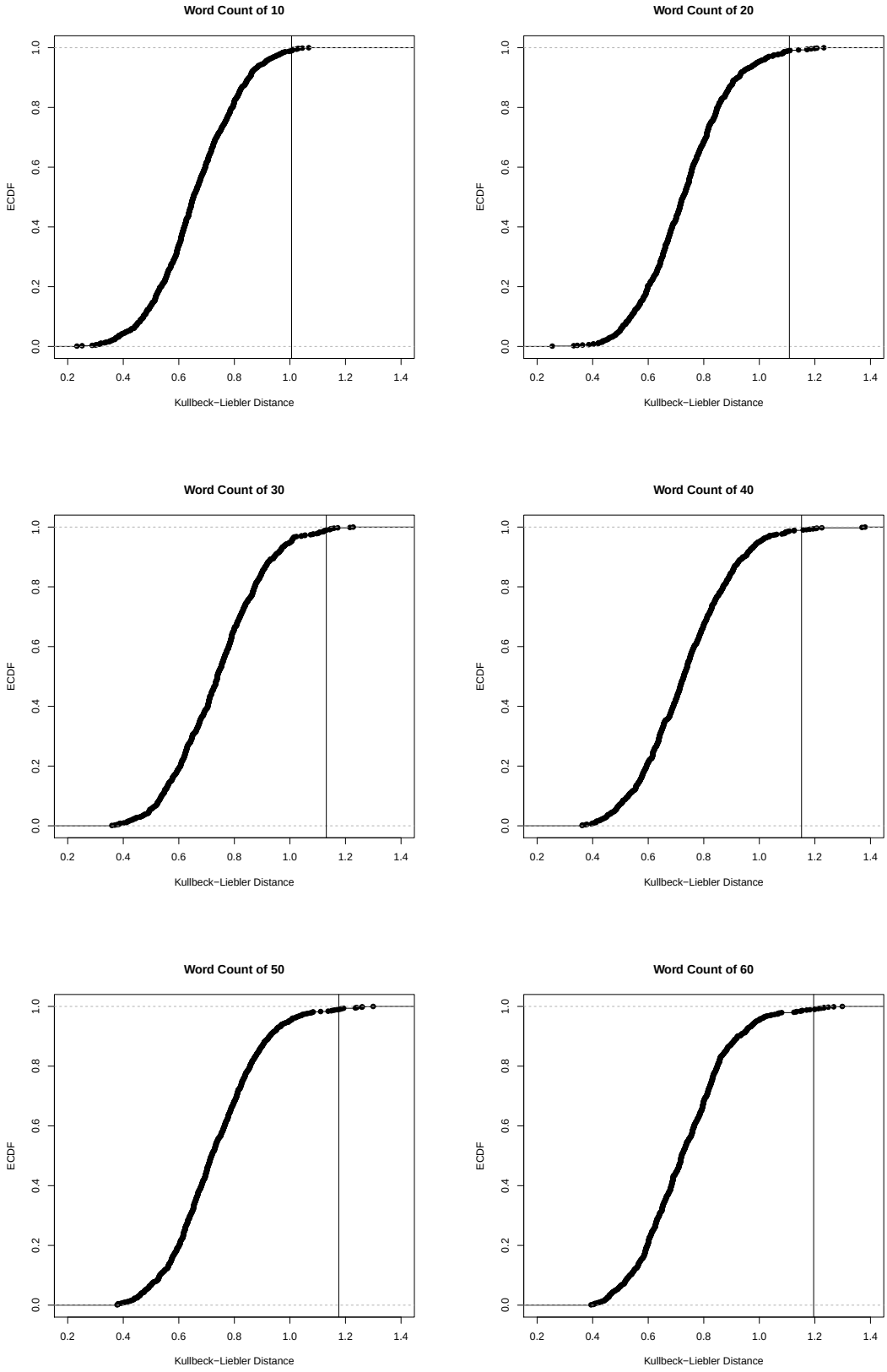


Figure 2.8: Kullbeck-Liebler RNMP by Word Count

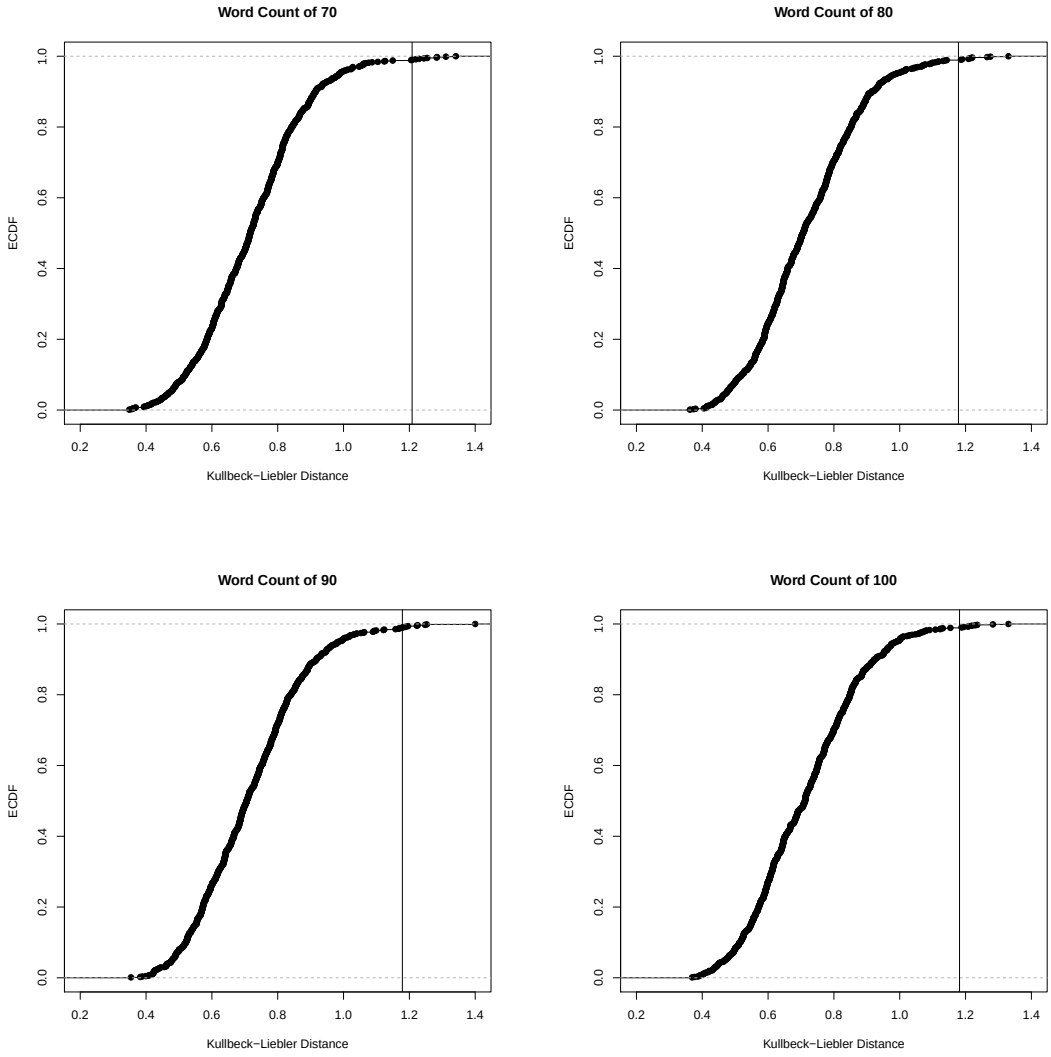


Figure 2.9: Kullbeck-Liebler RNMP by Word Count

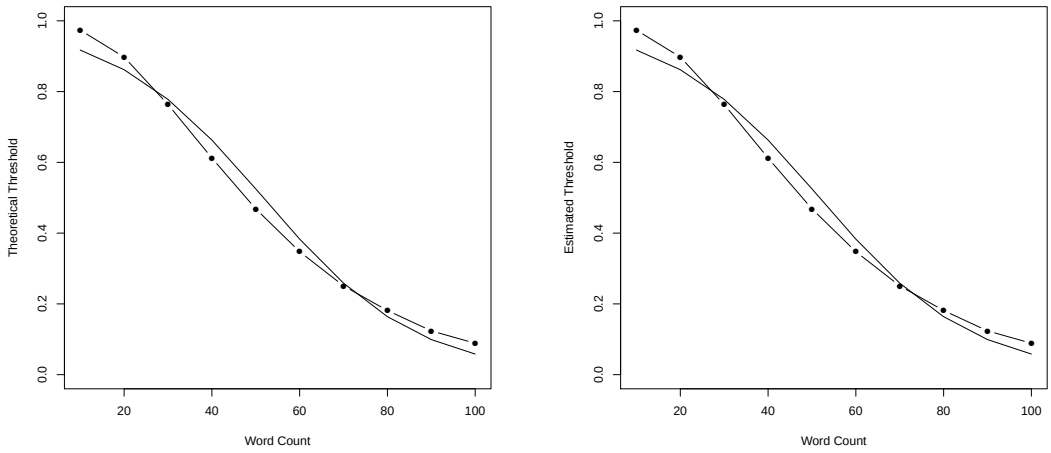


Figure 2.10: Fitted Simple Logistic Regression Model for Binomial RMP

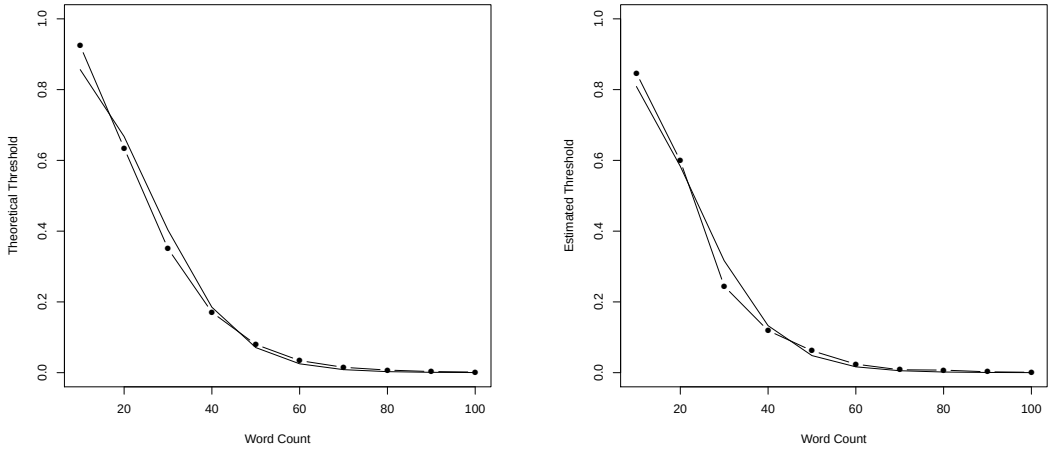


Figure 2.11: Fitted Simple Logistic Regression Model for Uncorrected Chi-Square RMP

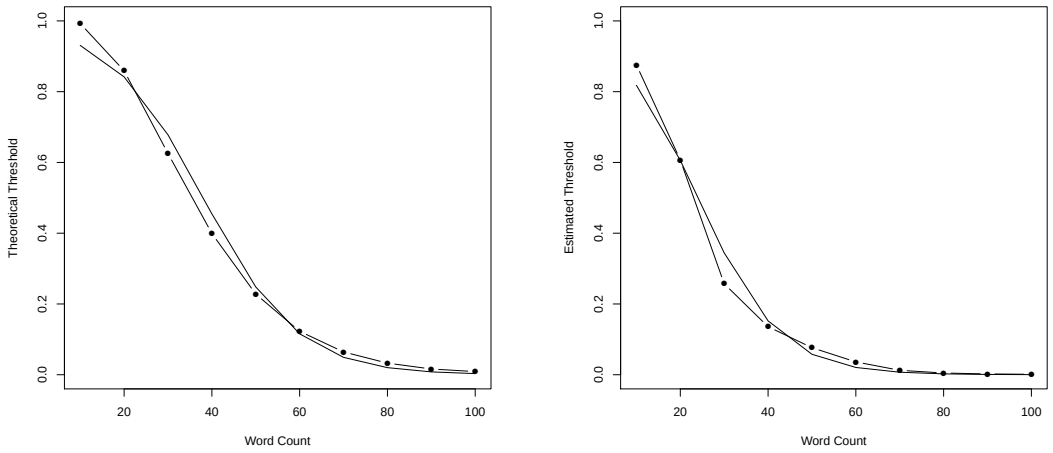


Figure 2.12: Fitted Simple Logistic Regression Model for Corrected Chi-Square RMP

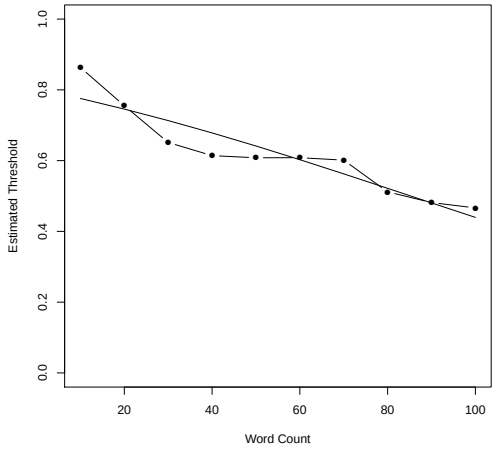


Figure 2.13: Fitted Simple Logistic Regression Model for Kullbeck-Liebler RMP

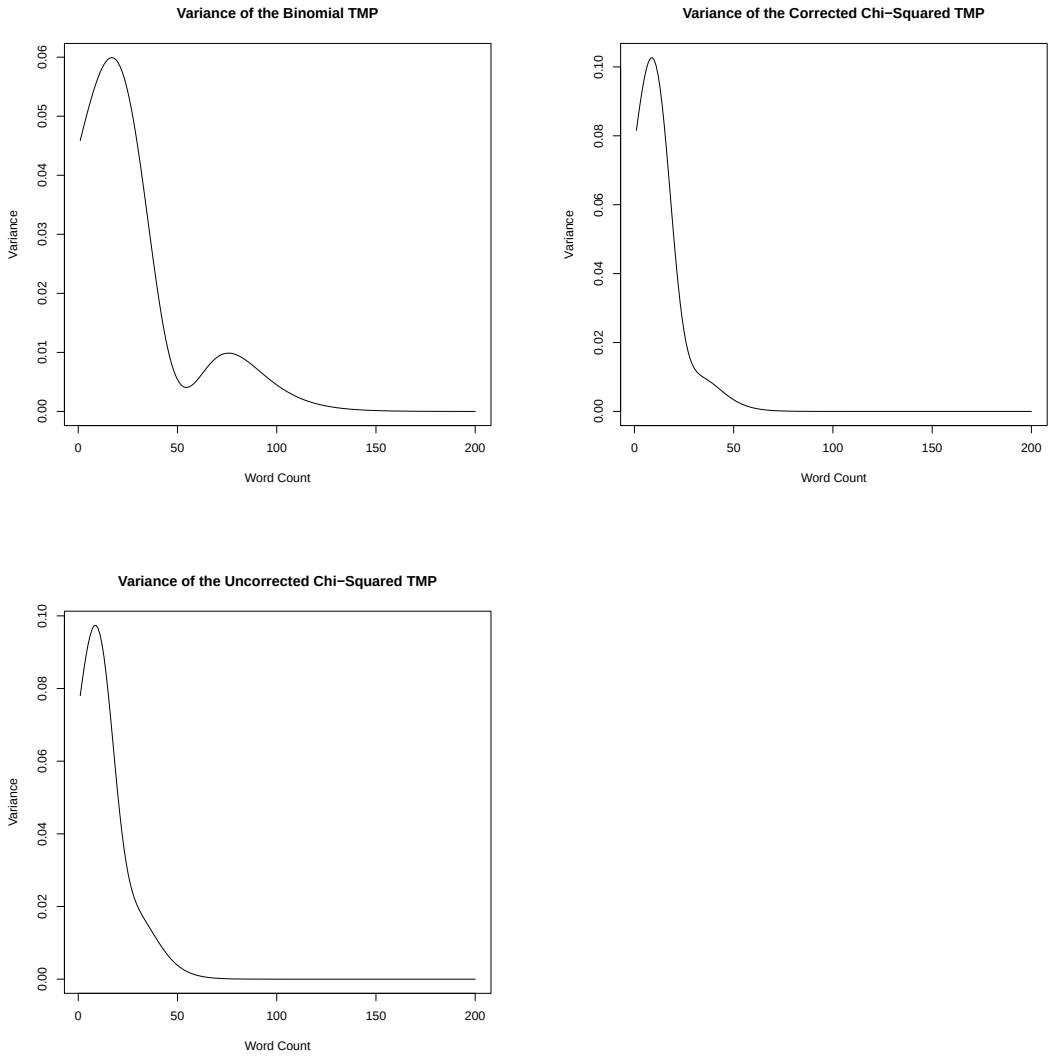


Figure 2.14: Subsampling Estimates of the Variance of the Conditional Match Probability; using the theoretical threshold for the Binomial match and the empirical threshold for the two χ^2 classifiers.

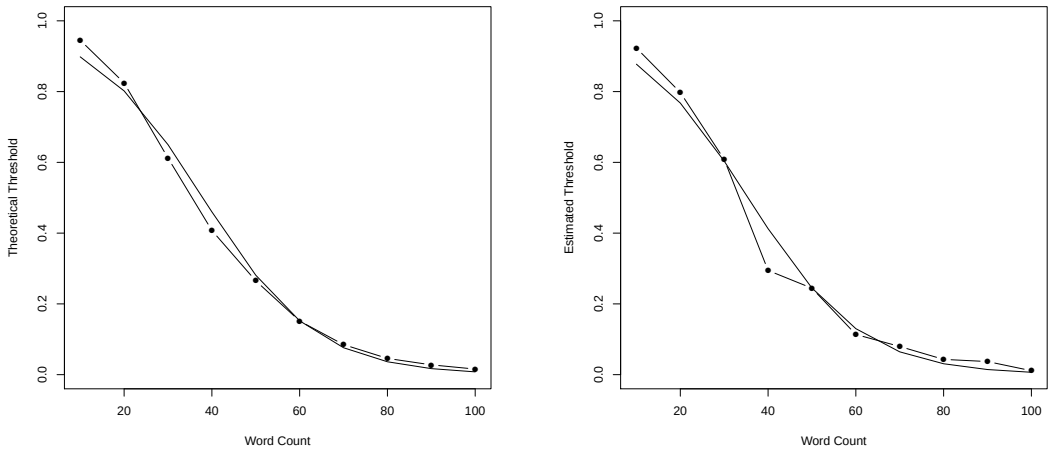


Figure 2.15: Fitted Simple Logistic Regression Model for Binomial TMP

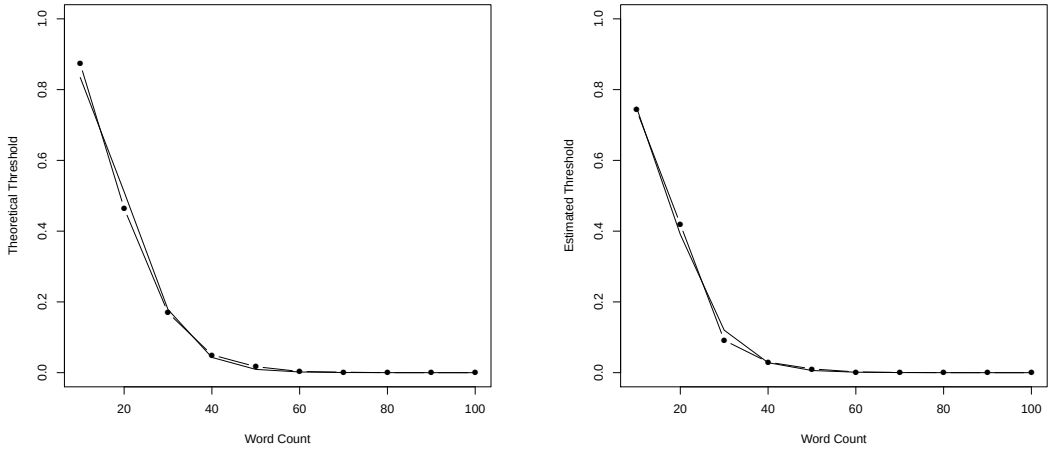


Figure 2.16: Fitted Simple Logistic Regression Model for Uncorrected Chi-Square TMP

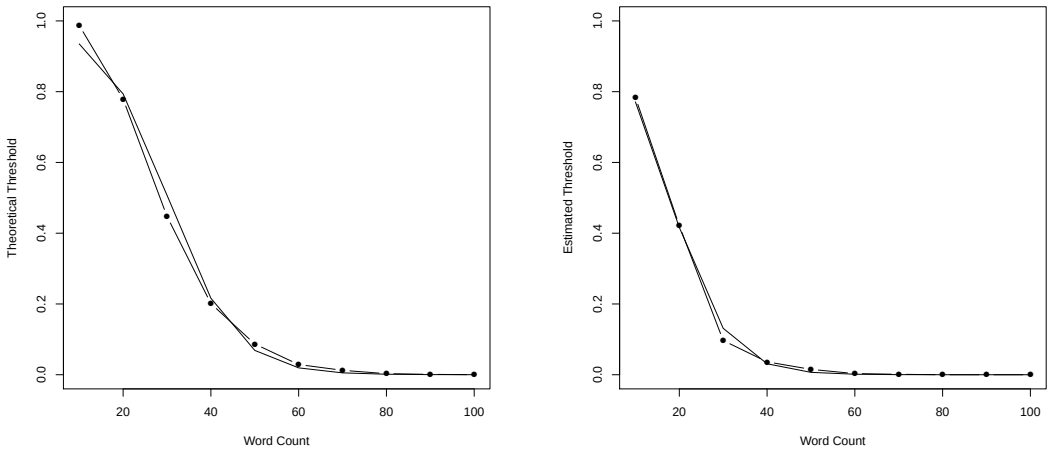


Figure 2.17: Fitted Simple Logistic Regression Model for Corrected Chi-Square TMP

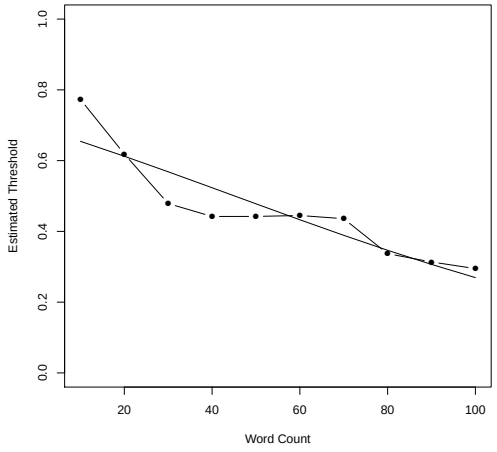


Figure 2.18: Fitted Simple Logistic Regression Model for Kullbeck-Liebler TMP

CHAPTER 3

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 missions of forensic science.*

This statement suggests that investigations into 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, no comparable modeling exists in handwriting analysis. In this paper, we have discussed one alternative for modeling the relationship between the RMP and the RNMP associated with a comparison procedure and applied to a collection of writing samples. The proposed method involves investigating the RMP and the RNMP using simulated writing samples. We have proposed algorithms for subsampling from writing samples in a collection where the subsamples are used to consistently estimate the RMP and the RNMP as a function of the lengths of the writing samples being compared. We have determined that the consistency of the subsampling estimators is only dependent on the number of writers, not on the length of the writing samples. We also have proposed a subsampling algorithm that can be used to estimate

the standard error associated with an estimator of the RMP based on all pairwise comparisons of the writing samples in a collection.

All of these algorithms have been stated in terms of a common length 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 words to generate the simulated writing samples.

Throughout this paper we introduced this subsampling methodology, which was the main objective, and have also applied the subsampling-based algorithms to a collection of writing samples. One example we have shown is how the information about the RMP can be used when organizing an empirical study of handwriting individuality within a relevant population. Caution must be taken concerning the actual values of the resulting estimates and the recommendations regarding an empirical study of handwriting individuality presented in this paper, as the collection of writing samples used to provide examples is relatively small with samples from only 100 individuals. Also, the collection of writing samples is a convenience sample and not necessarily representative of a specific population. Having a small collection limits the ability to accurately estimate very small match probabilities.

Although this paper is focused on match probabilities, the algorithms this paper proposed have potential applications in other forensic disciplines. Match probabilities are

utilized in studies of individuality and to validate the use of specific forensic techniques for individualization. However, it should be noted that they may not be the relevant measures for use in court (Stoney, 1984). A recent focus in forensic disciplines is to use the likelihood ratio, which can be used in cases such as handwriting and in the DNA practice of reporting profile frequencies. If a likelihood ratio is used, the probability that is estimated for the denominator is related to match probabilities. Currently, we are in the process of investigating the use of subsampling techniques proposed in this paper to estimate a likelihood ratio for handwriting identification.

BIBLIOGRAPHY

AGRESTI, A. (2012). Categorical Data Analysis, third edition. New York, NY: Wiley.

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.

CASELLA, G., BERGER, R, (1990). Statistical Inference. Belmont, California.

Duxbury Press.

DAVIS, L., SAUNDERS, C., HEPLER, A., BUSCAGLIA, J. (2012) Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios. Forensic science international 216(1): 146-157.

EFRON, B. AND HASTIE, T. (2016). Computer Age Statistical Inference. New York, NY, Cambridge University Press.

FISHER, R.A. (1948). Questions and answers #14. The American Statistician. 2 (5): 30-31.

FOUND, B. AND BIRD, C. (2016). The Modular Forensic Handwriting Method. Journal of Forensic Document Examination 26: 7-75.

HEPLER, A., SAUNDERS, C., DAVIS, L., BUSCAGLIA, J. (2012) Score-based likelihood ratios for handwriting evidence. Forensic science international 219(1): 129-140.

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

MILLER, J. J., PATTERSON, R. B., GANTZ, D. T., SAUNDERS, C. P., WALCH, M. A., BUSCAGLIA, J. A. (2017). A Set of Handwriting Features for Use in Automated Writer Identification. Journal of Forensic Sciences 62(3): 722-734.

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.

CURRICULUM VITAE

Education

Accepted 2017	*Ph.D. Student* *Computational Science and Statistics* South Dakota State University
2017	*M.S. in Mathematics* *emphasis in Statistics* South Dakota State University
2015	*B.S. in Mathematics* South Dakota State University

Professional Experience

05/2017 – *present*	*Graduate Research Assistant* South Dakota State University
08/2015 – *present*	*Graduate Teaching Assistant* South Dakota State University

Publications

In Preparation
Cami Fuglsby, Christopher P. Saunders, Linda J. Davis, John J. Miller, and JoAnn Buscaglia.
"Incomplete U-Statistics and Processes for the Estimation of Match Probabilities in Questioned Document Analysis"

Poster Presentations

08/2017	*"U-Processes for Characterizing Forensic Sufficiency Studies"* Cami Fuglsby, JoAnn Buscaglia, Christopher P. Saunders Joint Statistical Meetings; Baltimore, ML
Accepted 09/2017	*"U-Processes for Characterizing Forensic Sufficiency Studies"* Cami Fuglsby, Christopher P. Saunders, JoAnn Buscaglia International Conference on Forensic Inference and Statistics; Minneapolis, MN Travel support provided by: Stephen E. Fienberg CSAFE Young Investigator Award

Leadership and Service

National Service	Currently supporting the Expert Working Group on Human Factors in Handwriting Examinations Interpretation and Technology subgroup for the National Institute of Standards and Technology.
	Supporting the Questioned Document subgroup for the Organization of Scientific Area Committees for Forensic Science.
Refereeing	Journal of Forensic Sciences (2017)
Departmental Organization	Served as the president of the SDSU Math Club (2014-2015)
Volunteering	Northeastern SD Chapter of Math Counts (2014, 2016)
	Ready-SET-Go! Camp at South Dakota State University (2016)

Professional Memberships

Institute of Mathematical Statistics

Awards and Honors

2017	Stephen E. Fienberg CSAFE Young Investigator Travel Award - [$1500]

Glossary

Between-Writer: Refers to the comparison of two documents that were generated by two individual writers.

Character: Refers to the letter in a specific place in a document.

Isocode: A representation of each segmented character, represented by a mathematical graphic isomorphism where the internal structure can be enumerated by a code. Isocodes are cretaed by a proprietary automated process.

Length: The length of a document is measured by the number of characters or words written in the document. Throughout this paper we will refer to the length as the number of words in that document.

Letter: Refers to a specific letter of the English Alphabet, a number between 0-9, or a commonly used symbol, i.e. $ or &. Note that this distinguishes between capital and lowercase versions of the Alphabet, which implies for the Alphabet alone, there are 52 unique letters.

Within-Writer: Refers to the comparison of two documents that were generated by the same writer.