Scientific Approaches to Statistical Analysis and Collection of Handwriting Databases

Christopher P. Saunders
South Dakota State University

Disclaimer

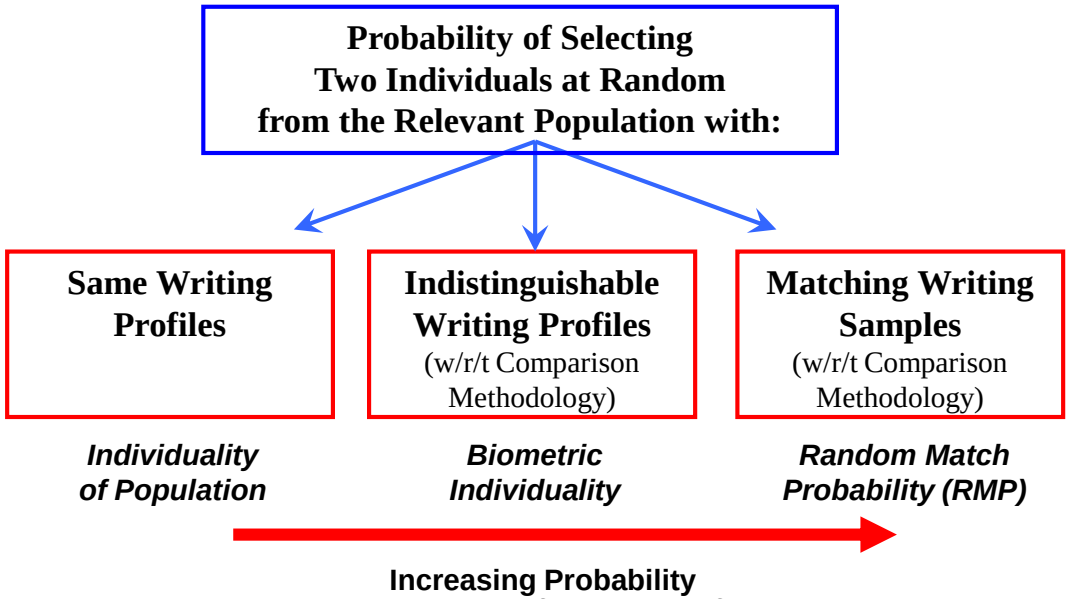
This work was supported in part under a Contract Award from the Counterterrorism and Forensic Science Research Unit of the Federal Bureau of Investigation's Laboratory Division.

- Names of commercial manufacturers are provided for information only and inclusion does not imply endorsement by the FBI.

- Points of view in this document are those of the authors and do not necessarily represent the official position of the FBI or the US Government.

Individuality

Biometric Individuality (of a **population** with respect to a **comparison methodology**): The probability that two (different) randomly selected writers from the population have indistinguishable writing profiles with respect to the comparison methodology being used.



- Fixed RNMP so we can study the RMP as a function of document size.
  - We will control the RNMP at 1% and model the RMP as a function of the size of writing samples selected from each writer's body of handwriting.

Match Probability

- *Random Match Probability* (RMP) is the chance of randomly selecting two subjects from the population and then randomly selecting a writing sample (of a given size) from each subject that is declared a match by the biometric.

- *Random Non-Match Probability* (RNMP) is the chance of randomly selecting a single subject and then sampling two documents from the selected subject's body of handwriting that are declared a non-match with respect to the biometric

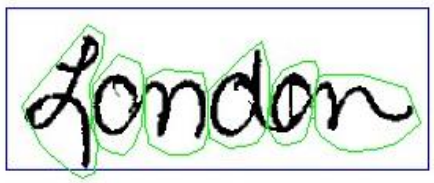
Pilot Study

- ~434 different writers
  - Approximately 10 samples (5 in print and 5 in cursive) of a modified “London Letter” paragraphs per writer
  - Collected from volunteers at the FBI, training classes, and at various conferences over a 2-year period.
  - Two of the five script paragraphs from each of 100 writers.
    - “FBI 100” data set

Data Processing

- Automated process represents each segment by a graphical isomorphism
  - Referred to as an isocode.
- Each document is reduced to the frequency of isocodes used to write each letter.
- Writing samples then consists of a set of isocode/letter pairs.
  - Each writing sample is represented as a cross-classified table of isocode by letter.

Data Processing



Letter = *n*
Isocode = 4;112

Accumulate across characters

Frequency Distribution of Letter/Isocode Usage in a Single Writing Sample

	Isocode 1	**Isocode 2**	**. . . .**	**Isocode *M***	
1					1
. . . .					**. . . .**
9					1
A					1
. . . .					**. . . .**
Z					1
a					35
. . . .					**. . . .**
z					2

Sub-sampling Algorithm: RNMP

RNMP sub-sampling algorithm :

1. Randomly select one writer.

2. For the selected writer, construct two "random" writing samples by selecting, without replacement, a pre-specified number of characters from that individual's collection of documents: *n*1 being the number of characters making up the writing sample from the first writing sample, and *n*2 being the number of characters making up the second writing sample.

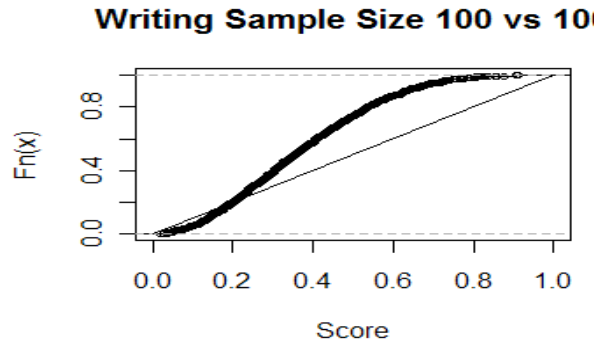
3. Compare the two "random" writing samples using the chi-squared similarity score.

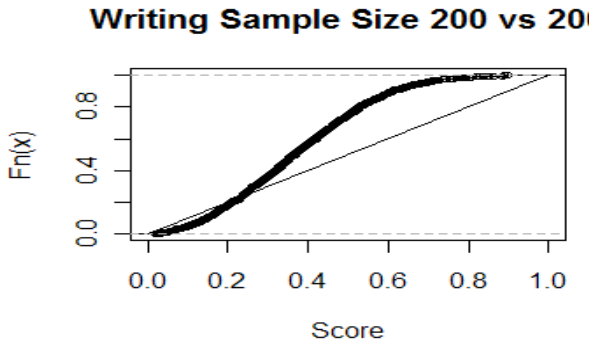
Application of the re-sampling algorithm many, many times over a variety of writing sample sizes results in a set of "data" of the form:

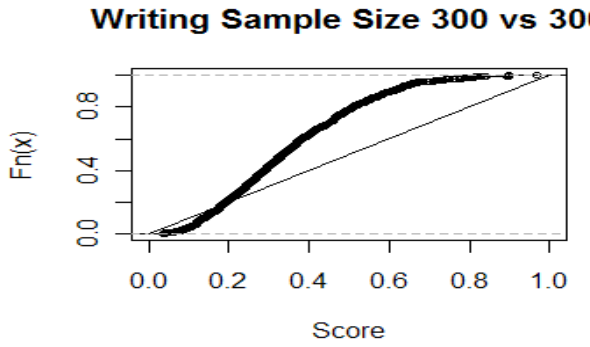
(*n*1, *n*2, *x*) where *x* = chi-squared similarity score.

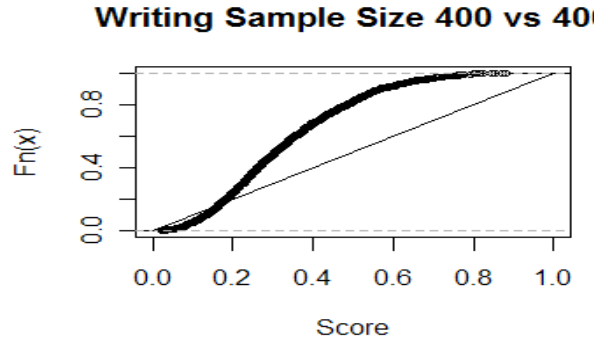
Sub-Sampled Within-Writer Similarity Scores

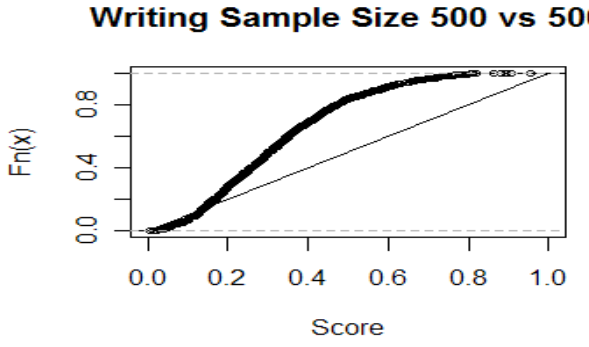
(K=1000 for each document size)

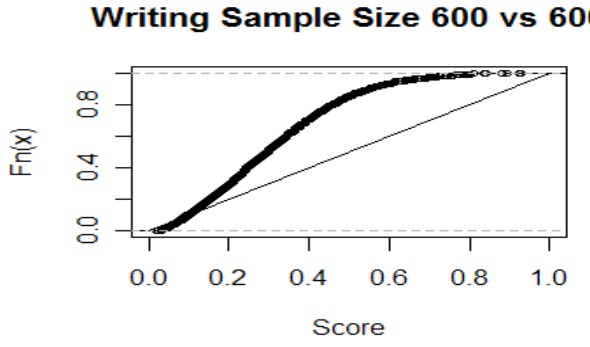


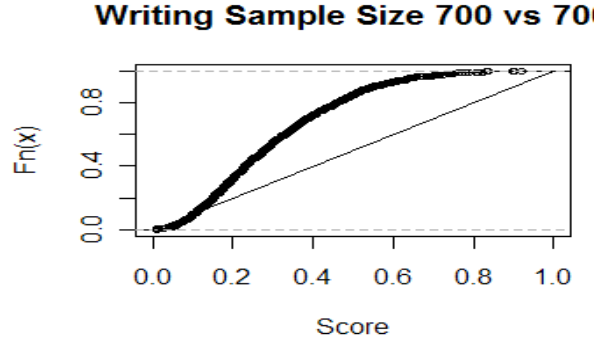


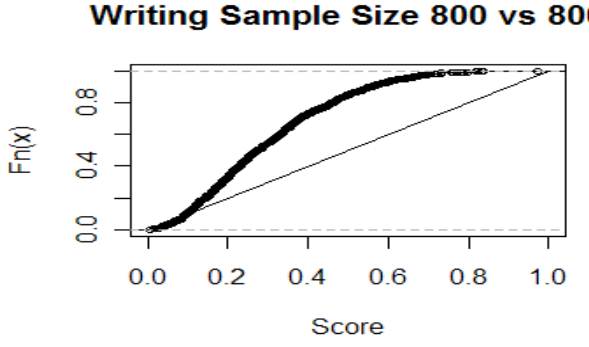


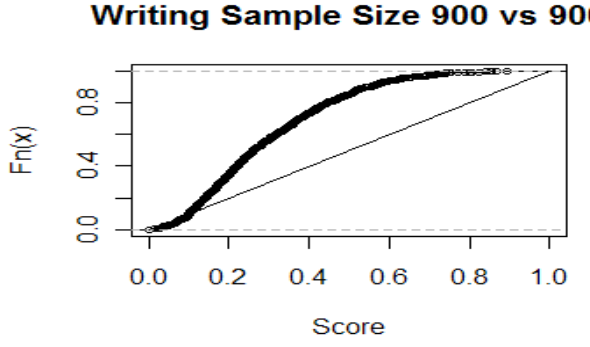












Sub-sampling Algorithm: RMP

RMP sub-sampling algorithm :

1. Randomly select two writers without replacement.

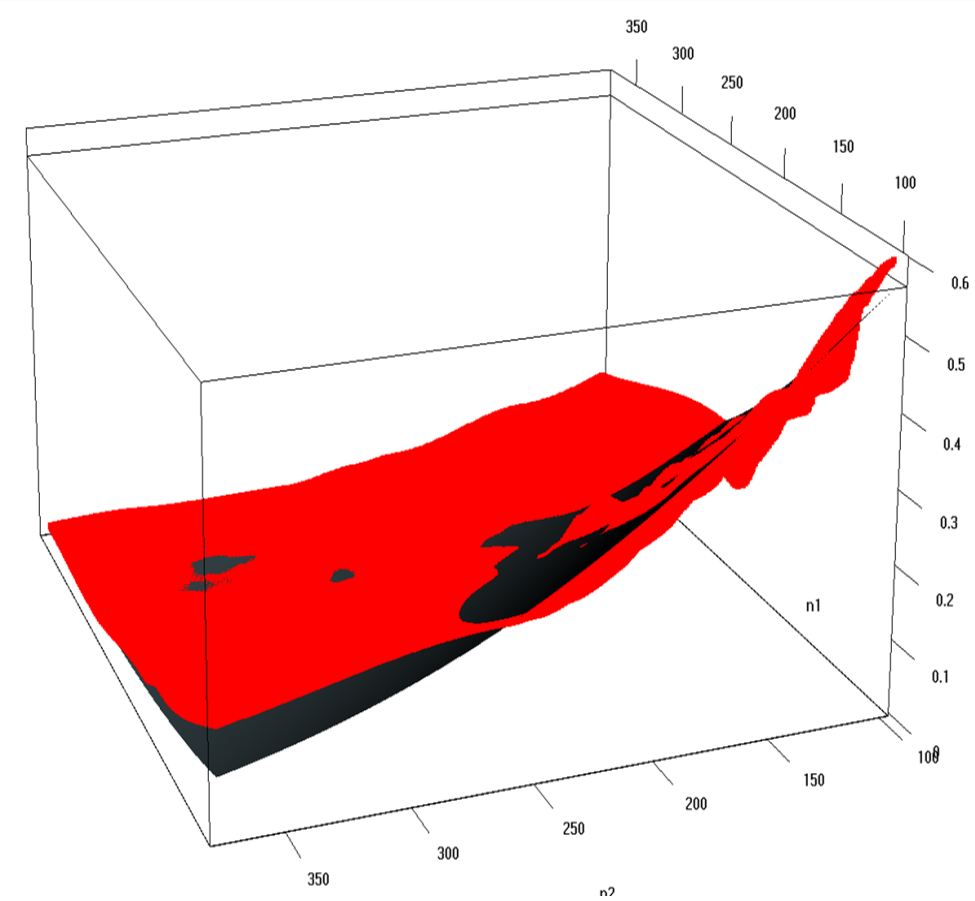
2. For each selected writer, construct a "random" writing sample by selecting, without replacement, a pre-specified number of characters from that individual's collection of documents: *n*1 being the number of characters making up the writing sample from the first selected writer, and *n*2 being the number of characters making up the writing sample from the second selected writer.

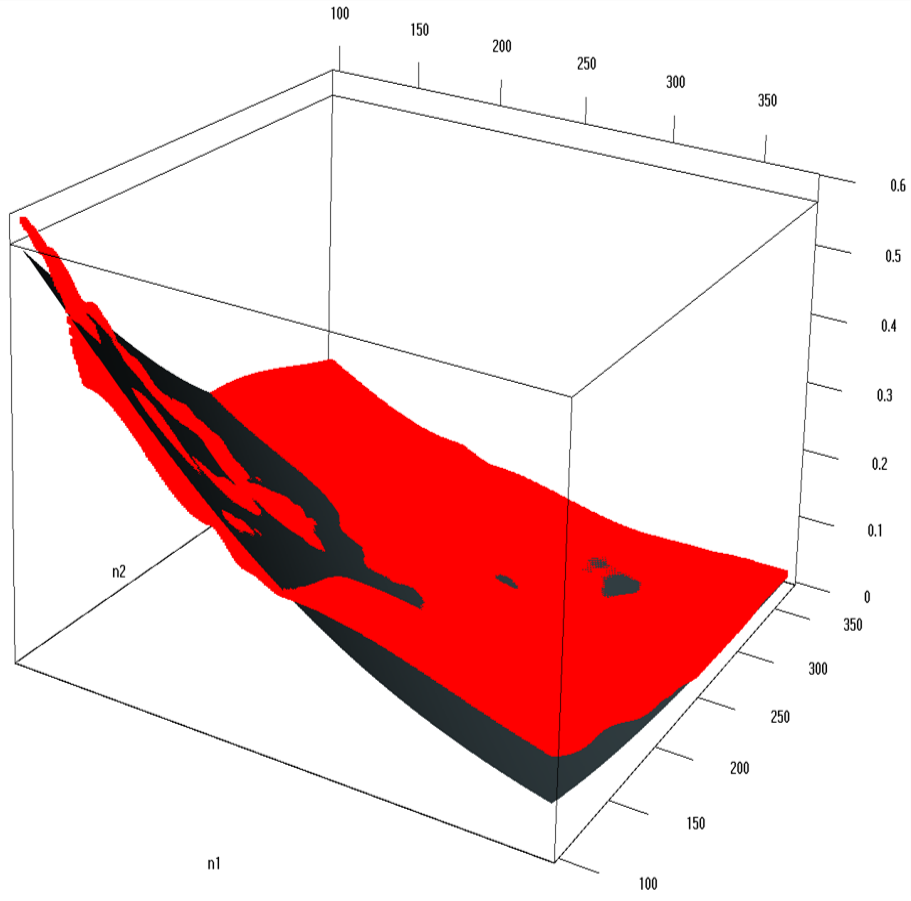
3. Compare the two "random" writing samples using the chi-squared similarity score and record whether or not a match has occurred.

Application of the re-sampling algorithm many, many times over a variety of writing sample sizes results in a set of "data" of the form:

(*n*1, *n*2, *x*) where *x* = 1 if the two writing samples match; 0 if the two writing samples do not match.

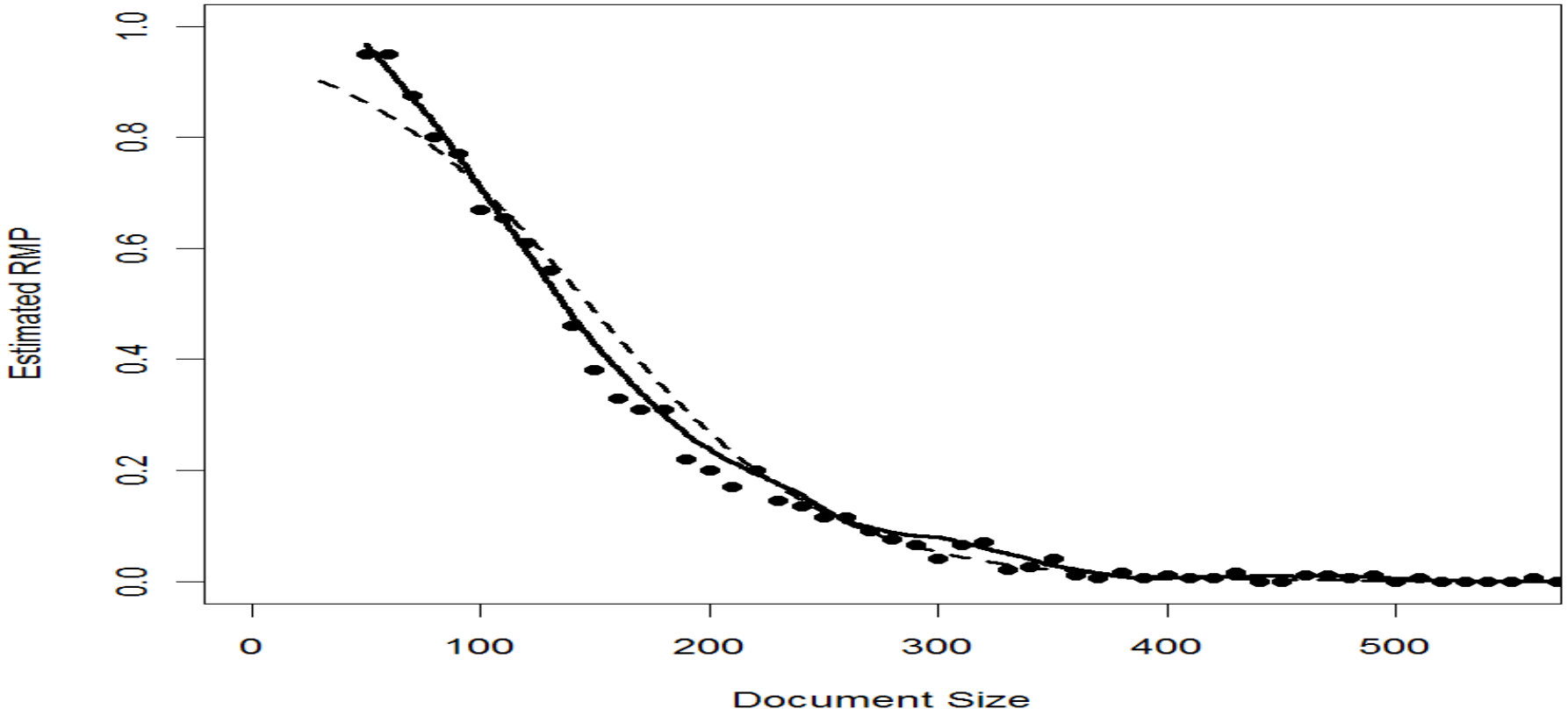
RMP Modeling





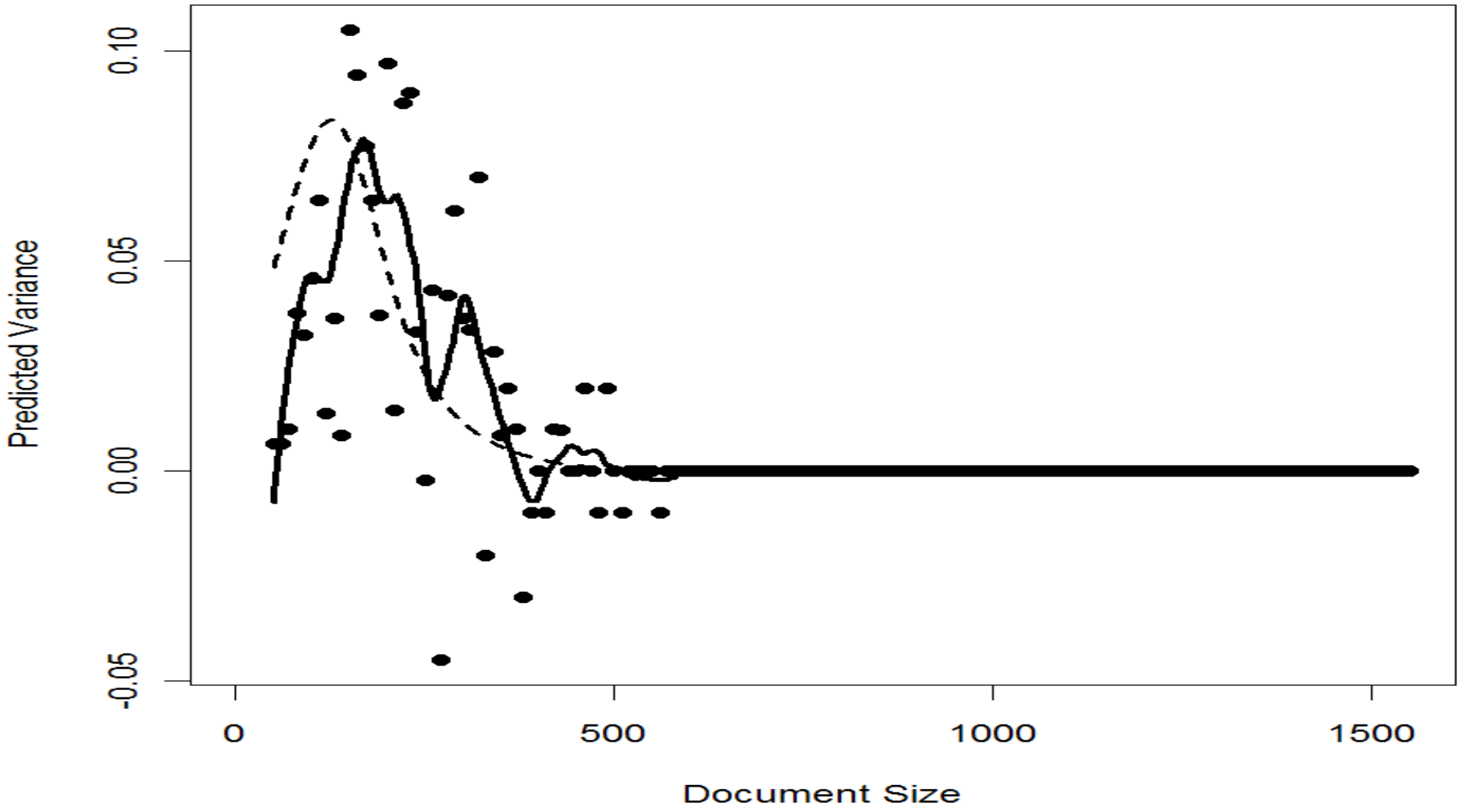
	Coefficient	Standard Error
(Intercept)	-2.28675	0.07749
*n*1	0.01075	0.00059
*n*2	0.00998	0.00059

RMP Modeling: Equal Document Sizes



	Coefficient	Standard Error
(Intercept)	-2.784783	0.1129635
Document Size	0.018923	0.0006051

The Modeled Variance



Properties of the Estimators

1. Consistent as the number of Writing Samples and the number of simulated documents tends to infinity.
   - *The Writing Sample Size can remain fixed!!*

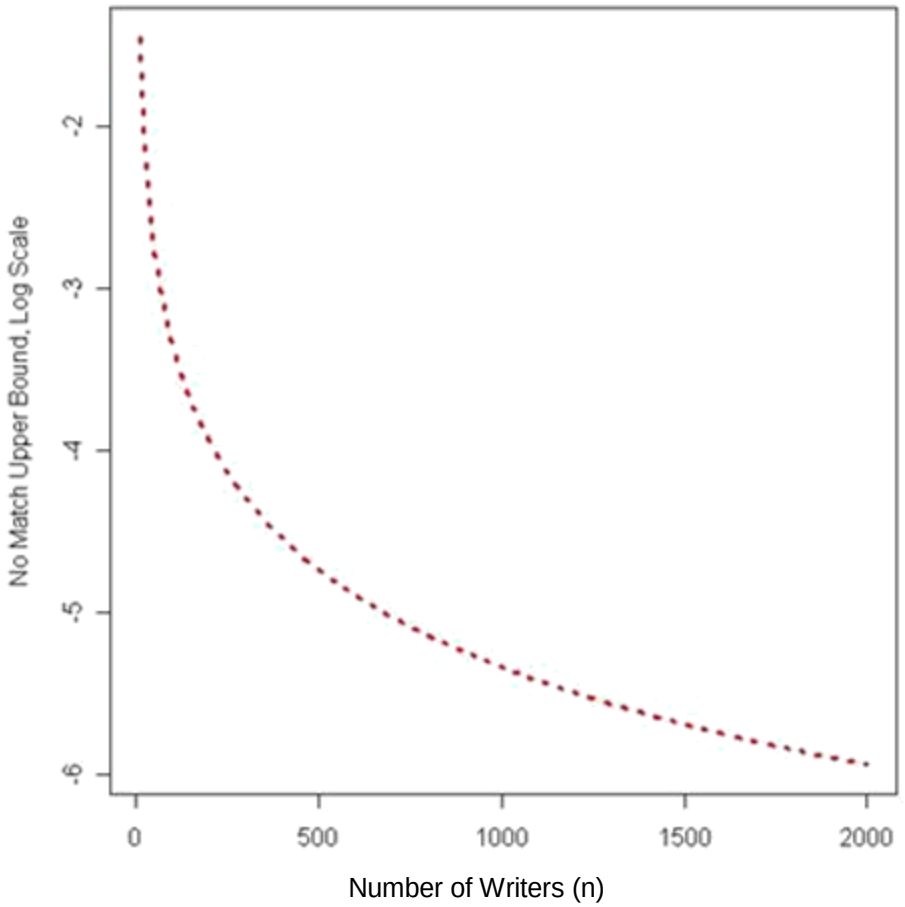
2. Asymptotically Normal Estimators.

3. Unbiased for the *RMP* and $E\left(p_i^2 \right)$.

The Design of an Individuality Study

- The sub-sampling models provide guidance on the relationship between the size of a writing sample collected and the RMP.
- Basic probability inequalities can give an idea on the behavior of upper confidence bounds on the RMP.
  - Given combination of writing sample size and number of sampled writers.
- The ideal setting is when we have a sample of documents from a large number of people and observe no matches when the collected documents are combined.

Confidence Bounds



95% Upper Confidence Bounds for the RMP when no observed matches are observed with *n* writers.

Writing-Sample Sizes Needed for Specified Number of Writers

- Based on a one sided version of Chebyshev's inequality.
  - Sometimes called Cantilli's inequality.
  - The probability of observing no matches when comparing writing samples pairwise from each of *n* writers is at least 50%, 80%, and 95% for the following writing sample sizes

	Probability of No Matches		
Number of Writers (*N*)	50%	80%	95%
50	751	828	916
100	869	945	1032
200	985	1062	1147
500	1137	1213	1298
700	1193	1268	1353
1000	1251	1326	1411
2000	1364	1439	1523

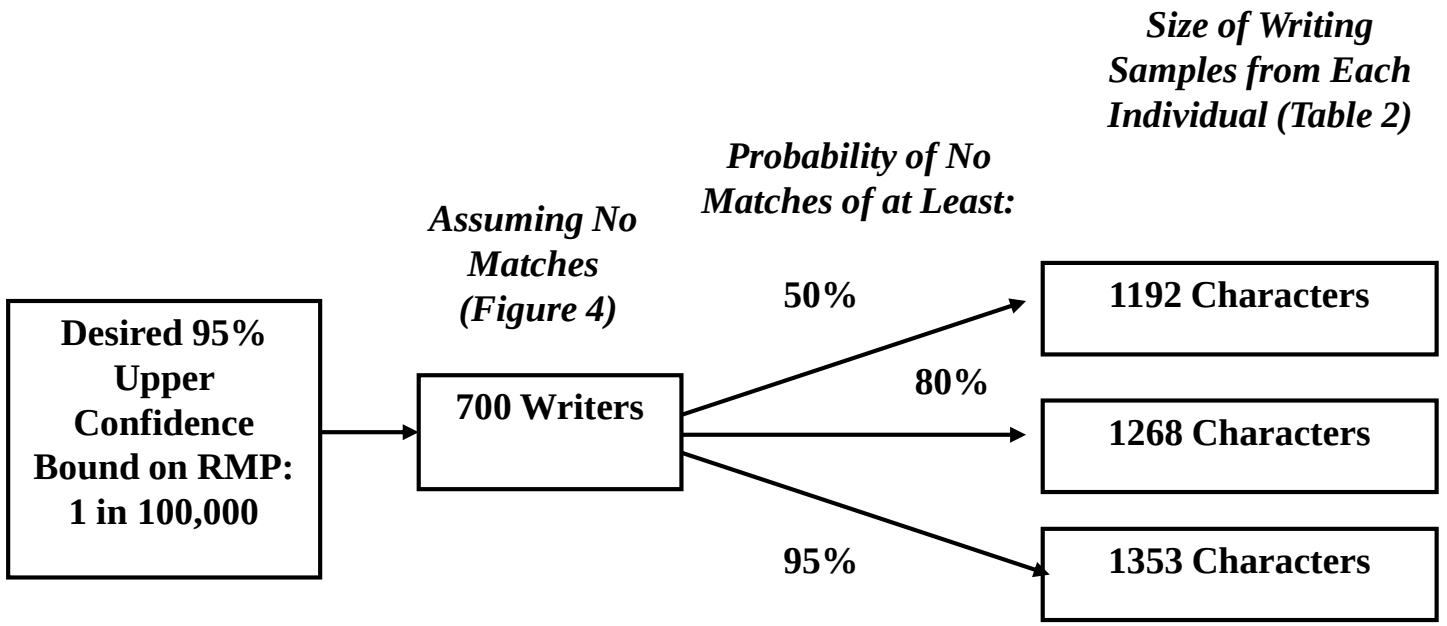
The Determination of Writing Sample Size

The chance of observing a match in the $n(n-1)/2$ pairwise comparisons is a function of the writing sample size.

For example, say the desired upper confidence bound on the RMP is 1 in 100,000.

1. Then the smallest number of writers we could use to achieve this bound is 700.
2. To have at least an 80% chance of achieving no matches in the 244650 cross- comparisons:
3. We would need to have each person submit a writing sample of about 1268 characters

An Example


Size of Writing Samples from Each Individual (Table 2)
Probability of No Matches of at Least:
Assuming No Matches (Figure 4)
Desired 95% Upper Confidence Bound on RMP: 1 in 100,000
700 Writers
50%
80%
95%
1192 Characters
1268 Characters
1353 Characters

Acknowledgments

- Linda Davis, JoAnn Buscaglia, John Miller, Danica Ommen, Don Gantz
- The IC Post Doctorial Fellowship Program
- Sciometrics and Mark Walch