International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

Examining Errors in Forensic Signature Analysis: A Statistical Perspective

Alexander Plant[1], Aziz Ben Jemia[2], Stephen Foster[3]

[1]University College of London,
[2]University of San Francisco,
[3]Penn State York

Corresponding Author:
Aziz Ben Jemia
University of San Fransisco
E-mail: benjemiaa21@gmail.com

Copyright: ©2024 Jemia A. B. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Received: 09-12-2024, Manuscript No. JQR/IJPLD/14; **Editor Assigned:** 10-12-2024, Manuscript No. JQR/IJPLD/14; **Reviewed:** 22-12-2024, Manuscript No. JQR/IJPLD/14; **Published:** 31-12-2024

ABSTRACT:

This paper investigates modern viewpoints and real-world applications of statistical analysis in the forensic examination of signatures. It explores the use of confidence interval analysis, and the determination of optimal sample sizes based on the desired confidence level. Furthermore, it discusses the subjective nature of expert opinions and elucidates approaches to incorporate and analyze these perspectives using relevant data and statistical tests. Lastly, the study examines the challenges associated with meeting the 'beyond reasonable doubt' standard of proof in real-world scenarios.

KEYWORDS: Forensic, Statistical Analysis, Ramifications

1. INTRODUCTION

The demand for forensic document examination to be quantified in a way that conclusions can be given with statistical significance has increased in recent years within Switzerland (1). In this analysis of simulated datasets, we delve into how we can reach statistically robust conclusions based on the opinions of forensic signature examiners.

We consider two types of data which can be collected from the examination. Firstly, we look at how we can apply confidence intervals to binary data from the examination process, and also how we can determine optimal sample sizes for this type of analysis. We then go on to discuss how to analyze more robust data from the examination; it is more appropriate for the expert to provide a range of confidence in how much the signature under question matches the genuine one (e.g., Likert-type scale responses). Lastly, we examine the ramifications of satisfying the 'beyond reasonable doubt' criminal justice standard of proof.

In this paper we look at existing research regarding the applications of statistics to the signature examination process, alongside research contributions which provide potentially flawed guidelines for statistical approaches to signature examination. We then move on to discuss the statistical methods which are both applicable to and appropriate for the signature examination process. After showcasing these methods, we will discuss potential next steps regarding this area of research.

Literature Review

Forensic practitioners use a series of steps to determine the strength of forensic handwriting evidence. As detailed in Kulik and Nikonets (27), cases are accepted if handwriting has enough complexity and identifiable features to be able to be examined more closely. If this is the case, examination of various aspects of the signature are compared to known writing examples (if possible), after which a conclusion is made to determine the strength of conclusion about the evidence. The Scientific Working Group for Forensic Document Examination (SWGDOC) produced a series of

terms used for qualifying the strength of evidence, ranging from identification to elimination. Evidence may also be quantified by a likelihood ratio (LR), which can help to give clarity as to the likelihood of finding the presented signature evidence given that the suspect is the author of a questioned text compared with the likelihood of such evidence if a random person besides the suspect is an author of the text (Taroni et al., 2012). Recent work has even begun tory and model distributions for this specific likelihood ratio given particular context differences in court cases. It is important to note that forensic examiner evidence is requested by the court system as a means of providing accurate and understandable interpretation for the jury. However, some court requests may create a perception which erroneously represents the forensic examiners' confidence. For example, a court may request:
"Please state, preferably in %, what is the degree of certainty of your conclusions [regarding whether a questioned signature was written by a particular known writer]. If, even under the best-case scenario there remains an unavoidable error margin for this analysis according to the state of scientific and technical knowledge, please state what that error margin is (preferably in %)." (1).

In this case, it is important to fully understand the probabilities being represented by the forensic examiner. In forensic handwriting analysis, the goal is not to determine the probability that a suspect is guilty, but rather to assess the likelihood of observing certain handwriting characteristics under two competing hypotheses:

- H1: The signature in question was written by the defendant.
- H2: The signature in question was written by someone else (e.g., a forger).

The analysis should focus on determining the probability of observing the given handwriting evidence (E) under each of these hypotheses. Specifically, the expert should aim to assess P(E|H1) — the likelihood of seeing the observed evidence if the defendant is indeed the author of the signature — and P(E|H2) — the likelihood of observing the evidence if the signature was created by someone else. The request of the court above, discussed more fully in Marquis et al. (1), was interpreted as an assessment of the probability of the defendant writing the questioned signature given the evidence—signified as P(H1|E). The goal of the current paper is to explore the proposition of the court and how it may lead to erroneous perceptions of examiner evidence.

There exist relatively few papers on the applications of statistics within the forensic signature examination process. A wide selection of papers exists discussing statistical applications but focused on the analysis of the physical properties of a single signature as opposed to statistics used to develop robust conclusions regarding the whole examination (2, 3, 4). For example, Kovari & Sharaf (2) provide insight into the efficacy of a program used to detect false signatures based on both single and multiple parameter systems. Unfortunately, this contrasts with how forensic signature testimony is often provided in court, which involves the testimony of experts who assess the veracity of signature documents.

In 2017, Marquis et al. (1) explore the application of statistical methods to derive statistically sound assertions about forensic signature evidence. Using a case study approach to estimate the level of accuracy of an expert signature witness, the researchers attempt to use a Bayesian statistical approach to create a framework for assessing the probability of being right (or wrong) in assessing a signature. Marquis et al. arrive at a probabilistic conclusion, stating, "There is an X% probability that the signature is genuine." While this approach is satisfying on the surface, there are inherent issues with the statistical framework which need to be illuminated. Bayesian estimation involves using prior odds to "update" assumptions about subsequent data. The method by which these prior odds are established—based on propositions presented by the court, which in turn rely on other evidence introduced during the case—will be discussed in detail in this article. The notion that evidence presented in court should influence the statistical analysis of signature examination is potentially concerning one—as noted by Alewjinse et al., if a sample of cases conclude with accurate signature identification in 99 of 100 cases, this should not significantly shift the base rate (prior odds) applied to the subsequent cases. Considerable differences likely emerge between cases, and setting a standard based on prior "success" may lead to erroneous conclusions.

Broadly, the utilization of Bayesian methods incorporating a prior assumption may present challenges in real-world contexts, introducing the potential for biases, which should be minimized. Prior research has implicated the importance of addressing bias in expert testimony in legal proceedings, in general (5), and in signature examination

specifically (6). More direct to Marquis et al., Morrison et al. note that the decision of what the numerator is in the Bayesian prior is often left up to subjective factors of the examiner and, although multiple scenarios may be presented to the court which collectively explain the development of an examiner's prior, it is nearly impossible to know if these cases may all have been subject to biases in some way or another. Caution is warranted when Marquis et al. notes that "frequencies can be used to inform one's degree of belief (i.e. probability) of an event" and "knowledge and the experience of the examiners" can be used, in part, to identify probabilities. In order to assess the probability requested by the court in the aforementioned scenario [P(H1|E)], prior probabilities would be needed in order to determine the likelihood ratio, which precludes the objectivity in analysis needed for a trustworthy approach to the evidence.

We will argue that the analysis of the signature, and the subsequent development of likelihoods, should be conducted independently of prior cases/scenarios impacting the assigning of likelihoods or priors. Therefore, we encourage the use of a frequentist approach in this context, as opposed to the Bayesian approach.

Second, it is important to recall that simulation of handwriting may be difficult to detect, and prior information being applied to likelihoods in the Bayesian approach may be built around expectations stemming from poorly skilled/low quality forgeries in previous cases. As described comprehensively in Morris (28), there are circumstances where a signature is being effectively copied via tracing, so no new features from the simulator are woven into the signature's makeup (p. 155). Morris notes that these circumstances make it "virtually impossible" to identify the simulator, which could lead to erroneous conclusions if such similarity is based on prior information suggesting that variation is expected or required between any two signatures. In other words, expert forgeries are increasingly difficult to detect, raising the potential for erroneous claims of similarity. While it is difficult to estimate the frequency of expert forgeries, one must consider that most forgers are likely to strive for high-quality forgeries, and that this could be more achievable in certain experienced criminal circles.

Third, the methods described by Marquis et al. rely on binary decision-making (Hp: the questioned signature was written by the defendant vs. Hd: the questioned signature was simulated by an unknown writer), developing likelihood estimates based on patterns of features which do (or do not) match reference signature. However, as Stern, H.S. and colleagues (7) note, the complexity of signatures varies, suggesting that a more nuanced approach might be preferable. This would involve the examiner assessing the similarity of the signature on a scale, thereby producing ordinal data. While this approach has been supported by recent work showing examiner confidence seems to be an important indicator of true positive and true negative detection rates (8).

Adjacent to our point on ordinal data, we will also discuss the importance of confidence intervals in the statistical evidence provided to courts. As discussed by Morrison et al. (9), the courts call for "error margins" alongside statistical evidence for forensic signature testimony. While Morrison et al. do not provide explicit solutions, we propose that confidence intervals, used extensively in quantitative analyses, may provide the necessary nuance to contextualize forensic signature testimony more appropriately.

Finally, we will be addressing the application of statistical evidence to the "beyond a reasonable doubt" claim requested by many court systems. In forensic contexts, random match probability (RMP) is often used as a standard for forensic evidence, which essentially assesses the probability of a match being detected based purely on coincidence—in the context of areas like DNA evidence, RMP is sometimes set as high as 1 in 1 billion (10). Although DNA evidence tends to have a much smaller RMP given the methods being used, forensic signature analysis sometimes has an affiliated RMP of 1 in 100, meaning only 1 of 100 signatures would be expected to match if the signature came from a different source. This presents a final conundrum—while forensic evidence can have a significant impact on perceptions of guilt (11). Much research has suggested that forensic handwriting evidence struggles to meet the strict thresholds of conservative RMPs, inviting the possibility for verbally framing signature evidence to appear more credible (12). The current research will finally discuss the simulation data evidence in the context of RMPs, "beyond a reasonable doubt", and the application of forensic signature evidence in the court system.

2. METHODS

In this section, we explore suitable methods of statistical analysis for forensic signature examination. As outlined in the introduction, we will examine statistical approaches suitable for two distinct types of data. Firstly, we look at the applications of confidence intervals to binary data, and how sample sizes affect this method of analysis.

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

Subsequently, we delve into the examination of appropriate statistical analysis methods for ordinal data. This type of data contains more information than binary data, therefore can yield more robust conclusions.

Two types of data

Up to this point, we have introduced two types of data that can be gathered during the examination process (binary and ordinal). Here, we elaborate on the significance of this distinction and clarify how the process involves two distinct forms of data.

In a court proceeding, as a means of assessing forensic signature evidence, an examiner will be asked whether a signature under question is genuine. They will be given several genuine signatures and, using what they learn from these signatures, they will then opine on whether the questioned signature is legitimate or not.

Binary Data

If an examiner is asked to determine whether a signature is legitimate, they will have to compare the genuine signatures in the sample to the signature under question. First, it must be noted that 100% "matches" are not possible in this context aside from simulated signatures, as no signature is 100% identical to another (this is why "absolute conclusions" are warned against by best practices in this area). It is for this reason most signature research has tried to assess signature features which are generally similar across instances (13). In the concluding decision made by a practitioner, a choice can be made to indicate whether or not there is evidence in support of identification to varying degrees—the ENFHEX's Best Practice Manual uses the terminology "full similarity" to indicate high probability of for the H1 hypothesis that the defendant wrote the signature of interest. It should be noted that 1-to-1 comparisons are not made, but a set of features and their overlap with a set of reference signatures are used to lead to the final decision about likelihoods. This is one inherent flaw to binary decisions, like placing a conclusion in "full similarity" or not, which may not capture the variability between genuine signatures encountered when an examiner is identifying features in a signature evaluation. Regardless, to make the process a data driven one with binary approaches, the examiner can count how many signatures in the sample are meaningfully similar to the signature under question ("given this comparison, the signature is indicated as genuine"); the more signatures that are consistent with the signature in question, the better the findings correspond to the examiner's expectations regarding a potential identification. This produces a sample, for example of 10 1's and 5 0's, meaning that 10 comparisons indicate the signature is genuine, and 5 comparisons indicate the signature is not genuine. This is analogous to the process discussed in Marquis, R, et al. (1).

Ordinal Data

If you ask a binary question, does this signature match this signature', you will get a binary answer, which often lacks information and can (potentially) lead to lower confidence when it comes to analyzing this data with statistical methods. Considering recent work that suggests even experts have somewhat high rates of error (14), binary data may be a method which helps to perpetuate this error into the court's decisions. What if we instead request the examiner to assess the degree of resemblance between the questioned signature and an authentic one, employing a rating system ranging from 1 to 5? The breakdown of the rating system could follow a scale as outlined below:

1 = The expert believes, to the best of their ability, that the signature is clearly not genuine 2 = Unlikely genuine signature
3 = Potentially genuine signature
4 = Likely genuine signature
5 = The expert believes, to the best of their ability, that the signature is indeed genuine.

This is known as a Likert scale, which will give us a sample of ordinal data. Such data is more information dense than mere binary data. Although it is challenging to locate references to this type of examination in current signature examination literature, the concept of forensic signature examiners utilizing the Likert scale to assess the complexity of signatures has been previously utilized (13), and Likert scales are used extensively in academic research on confidence in judgments about signatures (15). In this case, the examiner uses an ordinal scale from 1 to 5 to express the degree of similarity between the questioned signature and reference signatures. This scale does not represent the overall probability that the defendant authored the signature but reflects the strength of the evidence for or against

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

authorship. These ratings could then be used to help calculate the LR, which compares how likely the evidence is under two competing hypotheses: H1 (the defendant wrote the signature) and H2 (someone else wrote the signature). The court will then combine the LR with prior information to determine the overall probability of authorship. This indicates that examiners may be capable of applying this method during the examination as described here, thus providing the appropriate level of understanding for the court. Overall, a few issues are present in the current literature. First, there is a lack of consideration for between-signature variability coming from the same source—while binary data does appear to inherently fall victim to this issue, a simulation which shows the outcomes of using binary data for signature analysis has not yet been conducted. Second, although ordinal data does appear to be more aptly address variability in signatures, simulation data on this possibility is also not found in the literature. The current research seeks to fill these gaps in literature.

Methods of analysis and sample size implications for binary data

One way to exemplify potential issues with binary and ordinal data is to simulate datasets, an approach used in some previous studies on forensic analyses (16)—the benefits of simulated data is that they can allow for artificial toggling of important parameters one might encounter in real- world circumstances. In this section we will focus specifically on the analysis of the binary data and the effects of differing sample sizes. We construct confidence intervals for proportions to establish a range within which we have confidence that the signature in question is what would be considered full similarity. Confidence intervals are built based on responses of expert signature examiners—although each examiner's confidence level is not assessed, the sample size determines the width of the interval. We go on to discuss potential calculations for optimal sample sizes.

Confidence intervals

The type of confidence interval used is the Clopper–Pearson interval. This calculation method is particularly suitable for dealing with smaller sample sizes, which is important given the restricted number of signatures that might be available for sampling (17). This method uses the relationship between the binomial distribution and the beta distribution to generate the interval, using the following formula:

$$\left[B\left(\frac{\alpha}{2}; X, n-X+1\right), B\left(1-\frac{\alpha}{2}; X+1, n-X\right)\right]$$

Where $\alpha\alpha$ is our confidence level, n is sample size, X is the number of signatures which match. With BB (.: $\alpha\alpha\alpha\alpha, \beta\beta\beta\beta$), the beta distribution quantile function (18).

We will investigate the effects of sample size on confidence interval width and upper/lower limits. We look at confidence intervals based on four sample sizes being 5, 12, 30, 100. While there are no set standards for sample sizes in this area, there are times when sample sizes are quite small (n = 5), although many experts recommend at least n = 12, per common practice. Sample sizes of 30 and 100 are less likely to be encountered (although still found in practice) and can help to illustrate the changes in confidence intervals at relatively large sample sizes. We also consider the number of signatures matches within each sample size, to gain a clear understanding of how the confidence intervals change as these two variables – sample size and proportion of signature matches – vary.

Minimum Sample Size

We look at methods of calculating a minimum sample size to be used within the confidence interval analysis. But first, let's discuss the accuracy of the examiners and their opinions and how this relates to sample size calculations.

In their 2018 paper, Stern, H.S., and colleagues (7) explore the diverse complexities involved in examining signatures, which can range from indecipherable signatures to ones that can be easily replicated. This variability means that even highly skilled examiners may face challenges in accurately determining whether a signature indicates a full similarity (e.g., relative similarity across many signature features), supporting the H1 hypothesis that the defendant wrote a questioned signature. Such difficulties introduce the possibility of errors in the signature analysis

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

process. This thought lends itself well to the optimal sample size calculation, as we can account for potential mistakes/errors.

$$n \approx \frac{z_{\alpha/2}^2 * \hat{p} * (1 - \hat{p})}{d^2}$$

The above equation gives us the minimum sample size needed for estimating the population proportion. Looking at the equation, $z_{\alpha/2}^2$ for the statistical significance we require in the confidence nterval, we are using the commonly adopted 95% significance level. The $pppppp$ is the expected population proportion. Taking into account our previous discussion, even if the signature was authentic, we must consider a margin of error introduced by the examiners (or erroneous signatures). We will examine a range of error allowances from 1% to 10%. Finally, the term $dddd2$ accounts for the desired margin of error in the confidence interval. While we test a range from 1% to 8%, it is commonly acknowledged at the 95% significance level that a margin of error between 4% and 8% is acceptable within most statistical analysis (19). Nevertheless, given the nature of this data and the potential impact of the statistical conclusions derived from it, we conduct testing with a margin of error as low as 1%, which would equate to only a 1% chance of error given that the defendant wrote the signature in question – although there is no probabilistic threshold for evidence to be admitted in court, the standard of proof required in criminal proceedings in Switzerland and most jurisdictions is described as "beyond a reasonable doubt" and is likely perceived as quite low error probabilities for a decision that has been made on evidence.

Methods Of Analysis for Ordinal Data

An examiner may want to assess the degree of similarity between signatures using an ordinal scale, similar to the scale offered by the ENFHEX's Best Practice Manual. As mentioned in Stern, H.S, et al. (7), signatures vary in complexity, meaning that the confidence the examiner has that a signature is genuine may not be binary (100% confidence in full similarity, or 0% confidence in full similarity). Instead, they may prefer to rate how likely the signature features are similar on a scale of 1 to 5 (as described in the data section above). We now have a sample of ordinal data ranging from 1 to 5 which can be analyzed using different statistical methods compared to binary data.

Confidence Intervals

To calculate confidence intervals for ordinal data, we will be using a bootstrapping method. This is a common method employed when building confidence intervals using ordinal data, as it provides robust estimates without having to make too many assumptions about the data. The primary assumption that is unlikely to be satisfied by ordinal data is its normal distribution. In fact, we would anticipate the data to be heavily skewed towards the upper end (close to 5) when a signature is fully similar. Through bootstrapping, we can construct confidence intervals without assuming a specific underlying distribution (20).

Studying the impact of sample size on the confidence intervals constructed from ordinal data is more uncertain compared to binary data, as the variability within the sample significantly influences the width of the confidence interval. We will examine confidence intervals for various sample sizes and sample compositions.

Additionally, just as we examined the margin of error for the binary confidence intervals, we will also explore the impact of sample size on the breadth of the ordinal confidence intervals. We will demonstrate the noticeable reductions in the breadth of the ordinal confidence intervals as the sample size increases. While the binary confidence intervals are constructed around a proportion, allowing us to express the margin of error as a percentage, the ordinal

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

confidence intervals, centered around a mean between 1 and 5, will be evaluated based on the absolute width of the interval.

3. RESULTS

Here we see the results of all methods of analysis described above, starting with binary data then moving on to ordinal data.

Analysis of binary data

As mentioned earlier, we will examine confidence intervals for various sample sizes (5, 12, 30, and a range of signatures matches within each sample.

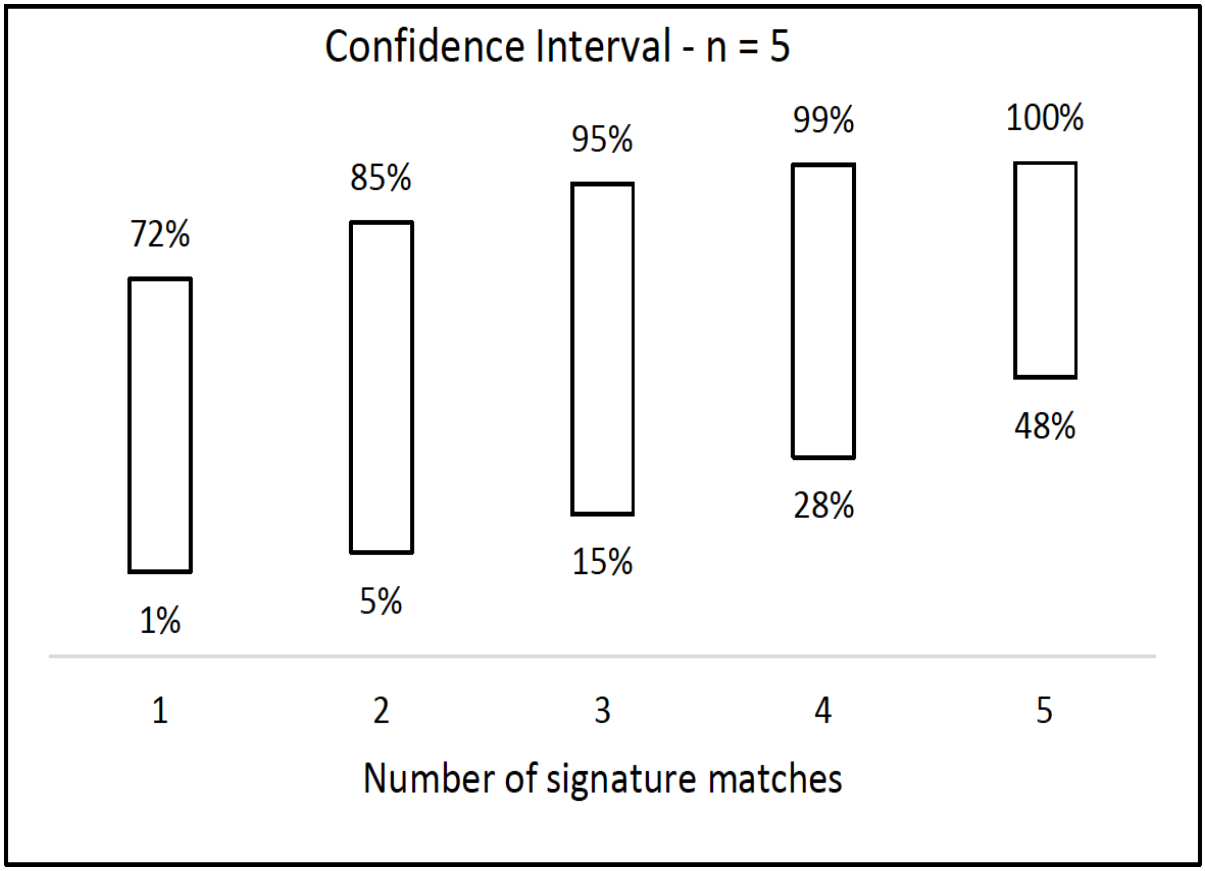

Figure 1: confidence intervals for sample size of 5

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

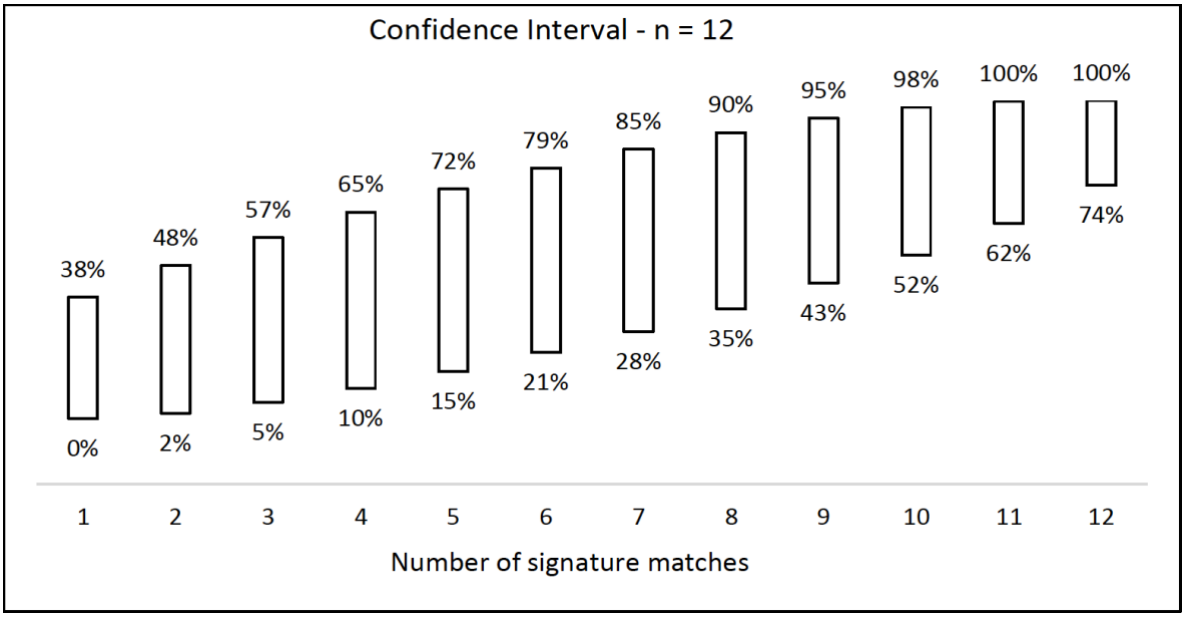

Figure 2: Confidence intervals for sample size of 12

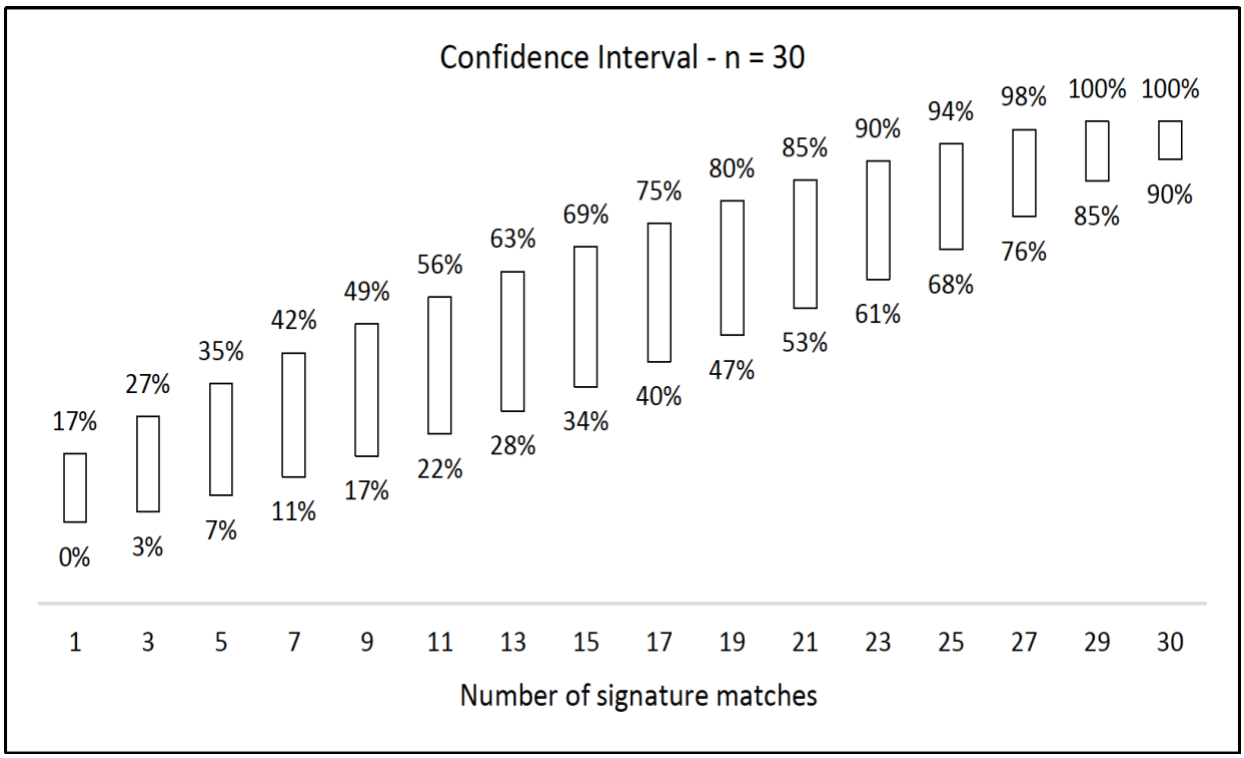

Figure 3: Confidence intervals for sample size of 30

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

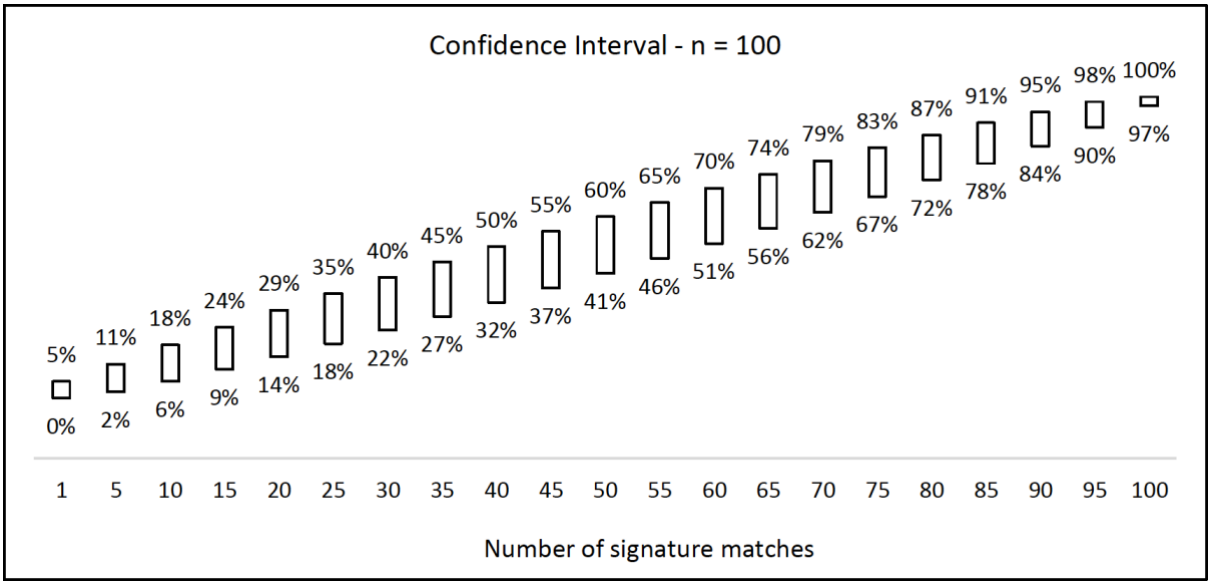

Figure 4: Confidence intervals for sample size of 100

After examining the confidence intervals mentioned, two clear observations stand out. Firstly, the intervals are noticeably narrower and show higher percentages when there are more signature matches. The main point to note is the difference in the confidence intervals as the sample size increases. For example, when all 12 out of 12 signatures are considered strong evidence for a match, the lower bound of the confidence interval is 74%, indicating a 74% chance that the evidence is strong for the signature being genuine. However, when all 30 out of 30 signatures match in this sense, the lower bound of the confidence interval increases to 90%, signifying much greater confidence. Therefore, even though all signatures are strong evidence matches in both samples, confidence is notably higher with a larger sample size. This supports the claim that larger sample sizes can create a higher "lower threshold" for confidence.

Margin of error for binary data

Here we see how sample sizes affect confidence interval widths. The confidence interval width is twice the margin of error, and the graph below distinctly illustrates the correlation between sample size, margin of error, and the proportion of signature matches. This relationship demonstrates how confidence intervals are influenced by sample sizes and the proportion of signature matches.

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

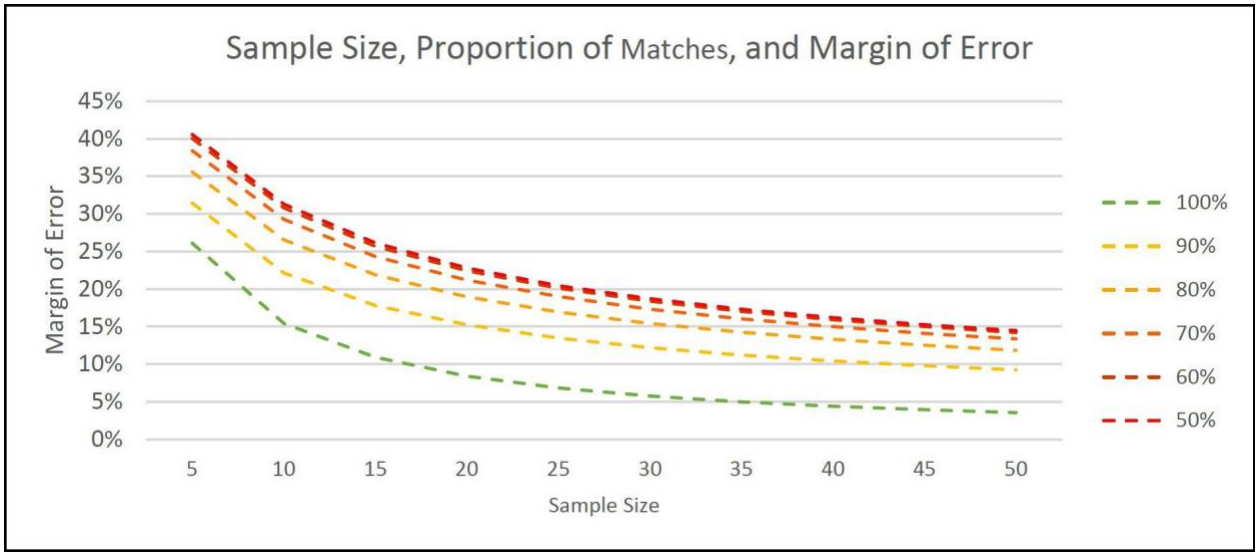

Figure 5: Sample sizes and margin of error

The visualization presented in Figure 5 allows us to understand how different elements affect the margin of error in our confidence interval. It illustrates the impact of sample size and the proportion of signature matches on the margin of error, highlighting the relationship between these elements and the precision of the interval estimates for the population proportion.

The horizontal axis illustrates the sample size, demonstrating a clear trend of decreasing margin of error as the sample size increases. Although the rate of reduction diminishes, consistent reductions in margin of error are evident with larger sample sizes. The five lines representing the proportion of signature matches in the sample show an intuitive relationship: as the proportion of matching signatures increases, the margin of error decreases. This emphasizes two important points: firstly, a larger sample size leads to a lower margin of error. Secondly, it reveals that the margin of error will remain high when the proportion of signature matches is low, indicating a higher likelihood of an illegitimate signature.

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

Minimum Sample Size

Table 1: Minimum sample sizes for confidence interval of proportion

	Examiner Error									
MOE	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91	0.9
8%	6	12	17	23	29	34	39	44	49	54
7%	8	15	23	30	37	44	51	58	64	71
6%	11	21	31	41	51	60	69	79	87	96
5%	15	30	45	59	73	87	100	113	126	138
4%	24	47	70	92	114	135	156	177	197	216
3%	42	84	124	164	203	241	278	314	350	384
2%	95	188	279	369	456	542	625	707	787	864
1%	380	753	1,118	1,475	1,825	2,167	2,501	2,827	3,146	3,457

Using 95% significance level

Minimum Sample Sizes

Bringing all the discussed elements together, we can determine the minimum sample sizes of signatures required to ensure robust statistical confidence in our conclusions. Referring to Table 1, we find a range of optimal sample sizes based on two selected variables. For instance, if we aim for a low margin of error of 4% and assume a 95% accuracy rate for signature conclusions, the minimum sample size required is 114. However, as our desire for an extremely low margin of error leads to impractically large sample sizes (court officials may only be able to find a dozen signatures at most), this serves as a prompt to consider the use of an ordinal rating method, given its information-dense nature.

Analysis Of Ordinal Data

The testing and comparing confidence intervals constructed from binary data were straightforward due to the simplicity of the sample. However, with the sample now containing ordinal data, the examples become somewhat more intricate. Ordinal data may, in practice, be closer to the perception of confidence one actually uses for assessing if features are similar in a comparison. We look at a few example samples varying in both composition and sample size:

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

Table 2: Bootstrapped Confidence Intervals on Ordinal Data

Sample Composition						Sample Size	Confidence Interval	
Rating	1	2	3	4	5	n	Lower Bound	Upper Bound
Sample 1				1	4	5	4.40	5.00
Sample 2			1	1	3	5	3.60	5.00
Sample 3	1	1			3	5	2.00	5.00
Sample 4				2	10	12	4.58	5.00
Sample 5			1	1	10	12	4.33	5.00
Sample 6	1	1			10	12	3.58	5.00
Sample 7				6	24	30	4.63	4.93
Sample 8			3	3	24	30	4.47	4.90
Sample 9	3	3			24	30	3.77	4.77
Sample 10				20	80	100	4.72	4.88
Sample 11			10	10	80	100	4.57	4.82
Sample 12	10	10			80	100	4.02	4.57

Above we can see the confidence intervals built upon the different example samples of ordinal data. We have developed nine sample examples to understand the effects of both sample size and sample composition on the confidence interval. The four groupings of samples all have similar composition, but with different sample sizes.

Looking at sample 7 and sample 8 for instance, we can see how the composition changes the confidence interval. In sample 7 the expert has rated 80% of the signatures at level 5 and 20% of the signatures at level 4 (please see example scale in the methodology), giving a confidence interval of 4.63 to 4.93, whereas the expert in sample 8 has rated 10% at level 3, 10% at level 4, and 80% at level 5, intuitively this confidence interval is wider, with a lower bound of 4.47 and upper of 4.9. What is evident here is how, within the same sample size, two experts with slightly differing opinions produce distinct confidence intervals, consequently leading to different conclusions.

Now, let's examine samples with identical composition (i.e., same ratios of ratings) but varying sample sizes. Sample 7 and Sample 10 both have the same composition of ratings, differing only in their sample size, of n=30 and n=100, respectively. We can see the confidence interval for sample 7 is 4.63 to 4.93, whereas the confidence interval for sample 10 is 4.72 to 4.88. Here are two key observations. Firstly, as anticipated, the confidence interval is narrower for sample 10 due to the larger sample size, indicating greater confidence in the result. Secondly, the confidence interval for sample 10 is concentrated at the higher end of the scale, rather than narrowing evenly as the sample size increases. This is attributed to the sample containing data at the higher end of the scale, resulting in a more accurate confidence interval.

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

Margin of error for ordinal data

Here we see the effects of sample size on the ordinal data, clearly showing the benefits of a larger sample of signatures.

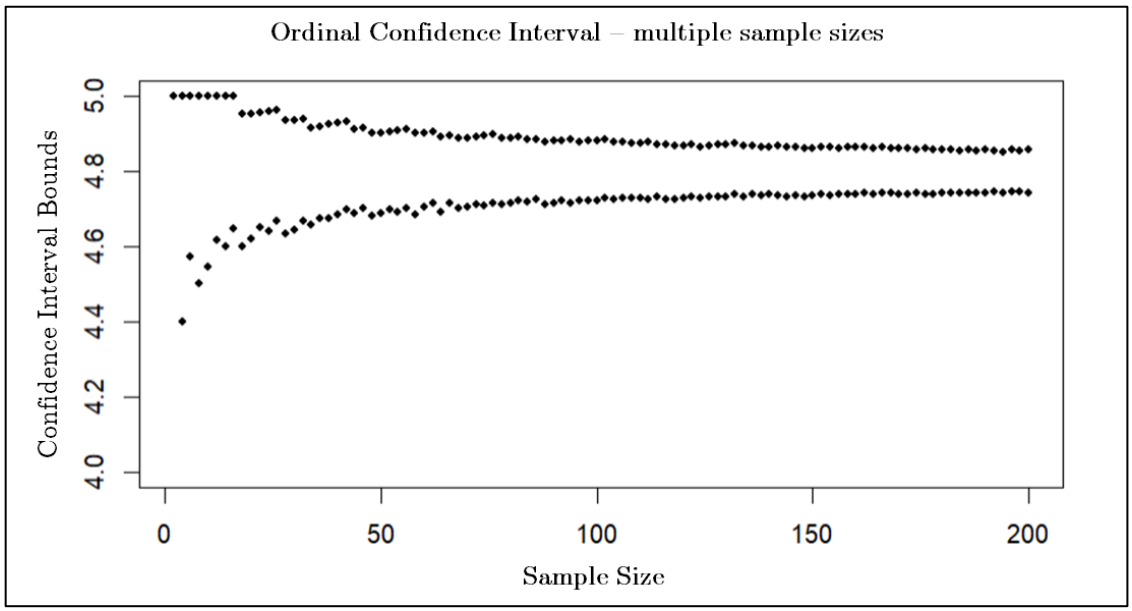

Figure 6: Ordinal confidence interval (20% rating of 4, 80% rating of 5)

Looking above at Figure 6, we aim to show the effects of larger sample sizes on the width of the ordinal confidence interval. We demonstrate a clear reduction in the margin of error as the sample size increases, with the upper and lower bounds narrowing around the mean of the expert's opinion. If the sample is essentially opposite to this, with 80% of the ratings being a 1, implying "Expert believes, to the best of their ability, that the signature is clearly not genuine", then the confidence interval will be equally narrow but positioned at the lower end of the scale, indicating stronger evidence for a forged or non-genuine signature.

As the composition of the sample becomes more varied, encompassing multiple ratings from 1 to 5, the width of the confidence interval expands greatly, indicating that the expert's opinion is not strongly consolidated around a certain rating. Our conclusion from this confidence interval would be along the lines of 'the expert's opinion isn't sufficiently strong to indicate the authenticity of the signature', which essentially reflects what the expert is implying through their varied ratings. These various conclusions are each helpful in providing context to signature evidence in a court of law.

Binary and ordinal comparison

Here we will look at the confidence intervals built on comparable binary and ordinal samples to effectively emphasize the differences between these two data types.

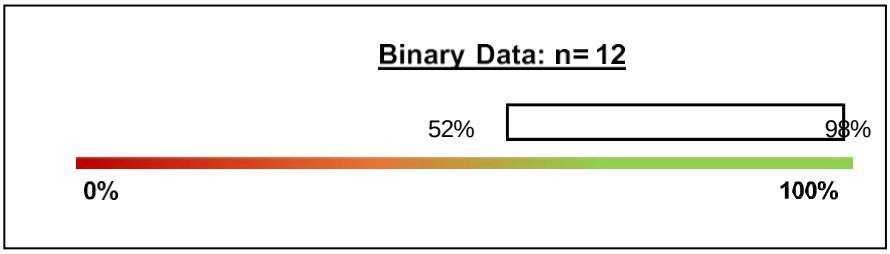

Figure 7: Binary data (n=12, 10 match, 2 unmatched)

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

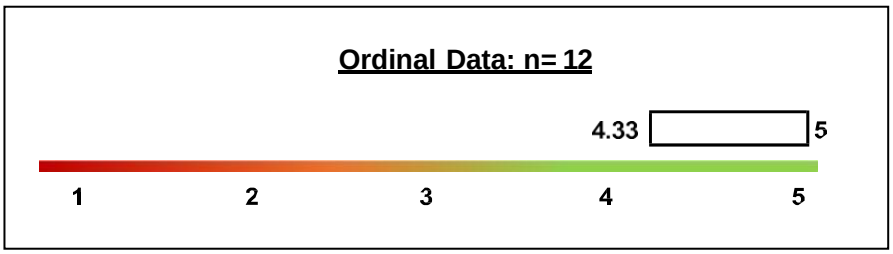

Figure 8: Ordinal data (n=12, see sample 5 in Table 2)

Upon examining Figure 7, it is evident that a sample of 12 signatures rated on a binary scale resulted in a wide confidence interval, primarily due to the small sample size. Given that the data is binary, it lacks substantial information regarding the expert opinion of the examiner. Consequently, drawing conclusions from this confidence interval would be challenging due to the significant disparity between the upper and lower bounds (the margin of error).

Figure 8 depicts a confidence interval constructed from a comparable sample of the same size. In this case, the collected data is ordinal, providing more detailed information about the expert's opinion. Consequently, this leads to a confidence interval with a lower margin of error, resulting in a narrower interval. As a result, drawing a more confident conclusion regarding the authenticity of the signature is easier with the confidence interval built on ordinal data, as opposed to the one derived from binary data.

Certainly, it is important to consider that if the ordinal sample of the same size comprised numerous ratings on the lower end of the scale (1 or 2), the resulting confidence interval would indeed be wider. This is understandable, as the expert's confidence in certain signatures strongly indicating lack of genuineness would be evident. The advantage of utilizing the ordinal scale lies in the expert's ability to exercise discretion, enabling them to adjust their level of belief in genuineness for each signature comparison.

Beyond Reasonable Doubt

Finally, let's assess the feasibility of meeting the requirements of the criminal justice system for achieving 'beyond reasonable doubt' standard of proof. It is generally accepted that the standard of proof needed to attain this level of certainty can be quantified at 99% (6), implying that the statistical findings from signature examination must indicate at least a 99% probability of the signature being authentic. Although the need to empirically assess confidence in signature matches (21), there has been justified discussion as to the misleading nature of presenting forensic signature evidence in this manner to the courts (22).

As demonstrated earlier, for binary data, hundreds to thousands of signatures are required to achieve a 1% margin of error. However, this is often impractical in the real world due to resource constraints or a limited number of signatures available from the person under investigation.

Recognizing this limitation, we explore methods to reduce sample sizes by utilizing more informative ordinal data.To attain the 'beyond reasonable doubt' level of certainty with ordinal data, the confidence interval would need to have a lower bound no less than 4.95 (equivalent to 99%). Here, we present samples that meet this criterion, examining both sample size and composition.

International Journal for Public Policy, Law and Development
Volume 1, Issue 1, Year 2024

Table 3: Confidence Intervals 'Beyond Reasonable Doubt'.

	Sample Composition					Sample Size	Confidence Interval	
Rating	1	2	3	4	5	n	Lower Bound	Upper Bound
Sample 1	1	0	0	0	239	240	4.95	5
Sample 2	0	1	0	0	179	180	4.95	5
Sample 3	0	0	1	0	119	120	4.95	5
Sample 4	0	0	0	1	59	60	4.95	5
Sample 5	2	0	0	0	398	400	4.95	5
Sample 6	0	2	0	0	298	300	4.95	5
Sample 7	0	0	2	0	198	200	4.95	5
Sample 8	0	0	0	2	98	100	4.95	5

Looking at Table 3 above, it is evident that the sample sizes needed to achieve a 1% margin of error are lower compared to those required when dealing with binary data. However, these numbers are still rather unrealistic to achieve in real-world scenarios: although smaller than those needed with binary data, the sample sizes are still excessively large. The primary concern arises from the fact that if there is even just one signature that the examiner rates as a 4 instead of a 5, then the confidence interval is already at the threshold of proving genuineness beyond a reasonable doubt, even with a sample size of 60.

In the event of a signature comparison rated at 3 by the examiner, indicating relative uncertainty, an additional 119 comparison rated at 5, where the examiner is as confident as possible in the signature's authenticity, would be required to achieve a 99% level of certainty that the signature is genuine. This underscores the misalignment between the expectations of the justice system and what is practically achievable through the application of statistics.

4. DISCUSSION

In our analysis of simulated data, we have outlined an initial approach to employing statistical methods in forensic signature examination. This encompasses the use of statistics to analyse the opinions of signature examination experts, as well as the potential for signature examiners to gather diverse data pertaining to their opinions. We started by exploring the application of confidence intervals for proportions in analysing binary data obtained during the examination process, which is the type of data examined in Marquis et al. (1), which uses Bayesian statistical methods on signature examination data. We have shown the effects of sample sizes, and how smaller sample sizes produce very unreliable conclusions even when all signatures in a sample are indicated as genuine by the expert. Then finally we have shown methods of calculating minimal sample sizes using best practice statistical methods. However, as mentioned, the limitations of binary data when the sample size is small is something which should be addressed, as in real world scenarios the number of signatures in a sample may be limited, which is why we go on to discuss the potential to use ordinal data.

We discuss how the signature examiner during the comparison of two signatures may prefer to rate the potential genuineness on a scale of 1 to 5, rather than producing a binary "yes" or "no" type of answer to the question 'is this signature genuine?'. As mentioned in Stern et al. (7), the complexity of signatures varies, meaning that the opinion of the signature examiner might not be a black and white one; instead, they may aim to convey, based on their professional judgment, the potential genuineness of the signature on a scale ranging from 1 to 5. This type of data

contains more information compared to binary data; it enables us to derive meaningful conclusions from smaller samples, unlike binary data, which may not yield robust conclusions with a similar sample size. The increased information content is precisely why a smaller sample size suffices. One might also consider if these effects may differ depending on the nature of the signatures, whether they are electronic or handwritten. However, recent work has shown that there is little discrepancy in expert judgments of signatures between the two mediums (23), suggesting that the methods found herein can be applied to both forms of signature examination which are found in the court system.

Concluding our analysis, we elaborate on the robustness, appropriateness, and applicability of the statistical methods in the signature examination process. Nonetheless, this does not imply that these methods should be disregarded. Instead, we advocate for the integration of the criminal justice system's requirements within the framework of statistical methods.

It is our recommendation that firstly the use of frequentist statistical methods as described here be employed within any statistical analysis of signature examination data, given the potential complications of Bayesian methods; and secondly that signature examination data should be collected using an ordinal scale, enabling examiners to assess signatures based on their subjective yet expert opinion, rather than mandating binary responses. In response to the following request:

"Please state, preferably in %, what is the degree of certainty of your conclusions [regarding whether a questioned signature was written by a particular known writer]. If, even under the best-case scenario there remains an unavoidable error margin for this analysis according to the state of scientific and technical knowledge, please state what that error margin is (preferably in %)." (1).

We have proposed methods to develop statistically significant intervals of confidence from which we can draw conclusions regarding the authenticity of a signature. All statistical analyses conducted in the study were performed at a significant level of 95%. The methods employed in our study are based on a frequentist statistical approach, eliminating the need to introduce subjective priors into the analysis, as was the case within Marquis et al. (1), a practice criticized both by our study and by Morrison and colleagues (9). Based on our current analysis, it appears clear that the "degree of certainty" requested by the courts is often either unattainable via empirical approaches or presented in a manner which misrepresents the nature of the evidence. This creates a difficult circumstance for forensic signature evidence to be presented as the court's expectations currently stand. On a similar note, it is clear that the forensic analysis procedure and examiner approach, in and of themselves, are not flawed. Instead, we propose that the court may need a better understanding of the forensic examiner approach to be able to ask more appropriate questions which match the evidence able to be provided by the examination. If the court request requires prior probabilities to make a judgment about the forensic evidence (e.g., details about the health of the defendant as noted in ENFSI's BPM in Section 7.3.3), then it is possible that the court request should be amended to a more reasonable state.

5. RECOMMENDATIONS

Our study establishes groundwork for employing statistics to draw conclusions from the examination process, as well as serving as a platform for further research within the field. We have delineated the theoretical implications of sample sizes and the benefits of ordinal data collection. A next step would be to use real world data, potentially through primary research, considering the probable scarcity of real-world ordinal data at this stage, to fully compare the methods described above, and to fully exemplify the potential advantages of the use of ordinal data in this process. Another approach may be to use higher-powered Monte Carlo simulations to estimate if other approaches can be helpful. While binary and ordinal data approaches are likely the most common measurement to be employed, one might consider if other measures can impact accuracy, like the order of false signatures in a "deck" of signatures (these effects have been found in other forensic contexts like lineup identifications; 24). Future research should consider if having examiners respond to a few additional questions might provide more helpful context for the court to make sense of signature examination evidence.

Finally, although beyond the scope of the current article, one might consider if human examiners may be unwarranted given the mismatch between what is asked for by the courts (beyond a reasonable doubt) and what examiners are able to provide given practical sample sizes. Recent work has begun investigating the use of neural

network computers to assess forensic documents, with accuracy rates at 97% or higher across contexts (25). Other approaches, like Dynamic Time Warping, are novel and complex, but show promise in automating signature verification (26). Future work should consider exploring the integration of modern options into the signature examination toolbox.

In response to the quest for achieving a confidence level 'beyond reasonable doubt,' we deem this objective unattainable within real-world contexts—the manner in which signature examination currently takes place cannot reconcile the variability in signatures and judgments within appropriate statistical frameworks. Consequently, we advocate for acknowledging an expert opinion as such, without striving for an absolute, indisputable outcome, or pushing examiners to complete more comprehensive assessments to create a deeper knowledge of the evidence being provided to the court.

Conflict of Interest Statement

The authors have no competing interests to declare

References

1. Marquis, R., Liv Cadola, Williams David Mazzella and Hicks, T. (2017). How to answer the question of error margin in forensic signature examination with a Bayesian approach? Nowa Kodyfikacja Prawa Karnego, 45, pp.9–14. doi: https://doi.org/10.19195/2084- 5065.45.2.
2. Kovari B, Charaf H. Statistical analysis of signature features with respect to applicability in off-line signature verification. In 14th WSEAS Int. Conf. on Computers 2010 Jul 23 (Vol. 2, pp. 473-478).
3. McKeague IW. A statistical model for signature verification. Journal of the American Statistical Association. 2005 Mar 1;100(469):231-41. https://doi.org/10.1198/016214504000000827
4. Srinivasan H, Srihari SN, Beal M. Signature verification using kolmogorov-smirnov statistic. InProc. International Graphonomics Society Conference (IGS) 2005 Jun (pp. 152-156).
5. Dror IE. Human expert performance in forensic decision making: seven different sources of bias. Australian Journal of Forensic Sciences. 2017 Sep 3;49(5):541-7. https://doi.org/10.1080/00450618.2017.1281348
6. Alewijnse LC, Mattijssen EJ, Stoel RD. Minimizing bias in forensic handwriting examinations. Journal of Forensic Document Examination. 2015 Dec 31;25:17-26. https://doi.org/10.31974/jfde25-17-26
7. Stern HS, Angel M, Cavanaugh M, Zhu S, Lai EL. Assessing the complexity of handwritten signatures. Law, Probability and Risk. 2018 Jun 1;17(2):123-32. https://doi.org/10.1093/lpr/mgy007
8. Hicklin RA, Eisenhart L, Richetelli N, Miller MD, Belcastro P, Burkes TM, Parks CL, Smith MA, Buscaglia J, Peters EM, Perlman RS. Accuracy and reliability of forensic handwriting comparisons. Proceedings of the National Academy of Sciences. 2022 Aug 9;119(32):e2119944119. https://doi.org/10.1073/pnas.2119944119
9. Morrison, G.S., Ballentyne, K. and Geoghegan, P.H. (2018). A response to Marquis et al. (2017) What is the error margin of your signature analysis? Forensic Science International, 287, pp.e11–e12. doi: https://doi.org/10.1016/j.forsciint.2018.03.009.
10. Kaasa SO, Peterson T, Morris EK, Thompson WC. Statistical inference and forensic evidence: Evaluating a bullet lead match. Law and Human Behavior. 2007 Oct; 31:433- 47. https://doi.org/10.1007/s10979-006-9074-4
11. Ling S, Kaplan J, Berryessa CM. The importance of forensic evidence for decisions on criminal guilt. Science & justice. 2021 Mar 1;61(2):142-9. https://doi.org/10.1016/j.scijus.2020.11.004
12. Mnookin JL. Scripting expertise: The history of handwriting identification evidence and the judicial construction of reliability. Virginia Law Review. 2001 Dec 1:1723-845. https://doi.org/10.2307/1073905
13. Guest RM. The repeatability of signatures. In Ninth international workshop on frontiers in handwriting recognition 2004 Oct 26 (pp. 492-497). IEEE. https://doi.org/ 10.1109/IWFHR.2004.103
14. Martire KA, Growns B, Navarro DJ. What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. Psychonomic bulletin & review. 2018 Dec;25:2346-55. https://doi.org/10.3758/s13423-018-1448-3
15. Elliott SJ, Hunt AR. The challenges of forgeries and perception of dynamic signature verification. In Proceedings of the 6th International Conference on Recent Advances in Soft Computing (RASC 2006) 2006 (pp. 455-459).
16. Rabasse C, Guest RM, Fairhurst MC. A new method for the synthesis of signature data with natural variability. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008 Apr 25;38(3):691-9. https://doi.org/10.1109/TSMCB.2008.918575

17. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934 Dec 1;26(4):404-13. https://doi.org/10.2307/2331986
18. Helwig, N. (2020). Inference for proportions. Retrieved May 28, 2021 from: http://users.stat.umn.edu/~helwig/notes/ProportionTests.pdf
19. Voxco (2021) 'Discovering the margin of error in survey data, Voxco, 23 March. https://www.voxco.com/blog/margin-of-error/#%3A%7E%3Atext%3DThe%20most%20commonly%20acceptable%20margin%2C%2C%20population%20size%2C%20and%20percentage
20. Diciccio TJ, Romano JP. A review of bootstrap confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1988 Jul; 50(3):338-54. https://doi.org/10.1111/j.2517-6161.1989.tb01442.xMo
21. rrison GS, Enzinger E. What should a forensic practitioner's likelihood ratio be?. Science & Justice. 2016 Sep 1;56(5):374-9. https://doi.org/10.1016/j.scijus.2016.05.007
22. Morrison GS. What should a forensic practitioner's likelihood ratio be? II. Science & Justice. 2017 Nov 1;57(6):472-6. https://doi.org/10.1016/j.scijus.2017.08.004
23. Heckeroth J, Boywitt CD. Examining authenticity: an initial exploration of the suitability of handwritten electronic signatures. Forensic science international. 2017 Jun 1; 275:144-54. https://doi.org/10.1016/j.forsciint.2017.02.019
24. Carlson CA, Gronlund SD, Clark SE. Lineup composition, suspect position, and the sequential lineup advantage. Journal of Experimental Psychology: Applied. 2008 Jun;14(2):118. https://doi.org/10.1037/1076-898x.14.2.118
25. Dhieb T, Njah S, Boubaker H, Ouarda W, Ayed MB, Alimi AM. Towards a novel biometric system for forensic document examination. Computers & Security. 2020 Oct 1;97:101973. https://doi.org/10.1016/j.cose.2020.101973
26. Mazzolini D, Mignone P, Pavan P, Vessio G. An easy-to-explain decision support framework for forensic analysis of dynamic signatures. Forensic Science International: Digital Investigation. 2021 Sep 1; 38:301216. https://doi.org/10.1016/j.fsidi.2021.301216.
27. Kulik, S., & Nikonets, D. (2016, July). Forensic handwriting examination and human factors: improving the practice through automation and expert training. In 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC) (pp. 221-226). IEEE.
28. Morris, R. Forensic Handwriting Identification. Fundamental Concepts and Principles. 2021. Academic Press