J19 Developing a Frequency of Occurrence Proportion-Based Database in Forensic Science: A Template Using the Handwriting Database

Thomas W. Vastrick, BS, Apopka, FL 32703*

Learning Overview: The goal of this presentation is to provide attendees with a template for how to create a frequency of occurrence database for their location and for their discipline through the experiences and lessons learned developing a handwriting database over the past ten years. Attendees will learn about each step used by one of the founders and leaders of the template project.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by knowledge transfer of the experiences gained through the entire development process of one such database. As such, the worldwide forensic science community will have the benefit of the successes and failures in methodologies utilized and the reasons behind the results.

In 2010, a team of forensic document examiners and statisticians undertook to develop, from scratch, a database of frequency of occurrence proportions. This project was funded by the National Institute of Justice, United States Department of Justice with a research grant. The results were formally presented at the 2015 American Academy of Forensic Sciences Annual Scientific Meeting in Orlando, FL. Subsequent to that, team members headed two further studies that broadened the database to include numerals. To date, there are more than 800 features in the database.

Because there had been few similar studies ever conducted, the team needed to establish several protocols. These protocols, by design, were based on statistical standards as it was determined that best practices dictated that this was to be a statistically driven study concerning forensic science rather than a forensic science study using statistics. The protocols incorporated published information and procedures utilized in other areas. The areas of procedures that were addressed included population sampling, quality control, data selection, system development, reporting, potential for abuse, and integrity of the data.

The statisticians employed a multi-strata system of population sampling based on factors that had been published concerning intrinsic and extrinsic factors that affect handwriting. The team reviewed each factor and determined the best method to address the factors individually. Next, the team used recent census reports to establish the strata goals, set at 80% of the census numbers. The other 20% were permitted to be collected randomly in order to take advantage of randomness and to further address factors that may not have been considered. The population sampling method will be addressed in great detail in this presentation and implications to developing sampling protocols for regions that may have significantly less homogeneity will be discussed.

Another aspect of population sampling was the physical collecting of specimens. Again, standard methodologies needed to be followed and the method with which the team executed this aspect of the study will be discussed in detail.

Data selection was a complex process that required a significant amount of time and processes. First, features were initially selected by subject matter experts based on being deemed objective in nature, as opposed to subjective. The reason was to establish reproducibility in data entry from the many data entry personnel. The method of testing involved use of an Attribute Agreement Analysis (AAA). This presentation will go into the published AAA standards and both the successful and unsuccessful methods used in its implementation.

The system developed for the input and use of the data was complex and unique. Efforts were made to create a system that could be expandable and flexible so that future additions would be easy and use in actual casework could be applied on a national level of targeted subgroups. This presentation will discuss this in detail and demonstrate the various software programs that were developed.

Since population sampling is not precise to the entire population, but rather an estimate, it is important to know how to report results in the most accurate manners possible. This presentation will discuss how the handwriting database deals with this issue.

Finally, procedures were created to minimize the potential for any bias by the separation of duties and the restricted access to data by the forensic specialists. Conversely, the entire team contributed to ideas of potential misuse and abuse of the process and agreed that it was important to address these issues upfront. These issues will also be discussed in the presentation.

Frequency, Statistics, Database

*Presenting Author