Determining the Number of Test Fires Needed to Represent the Variability Present Within a Firearm

Eric Law

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation
Law, Eric, "Determining the Number of Test Fires Needed to Represent the Variability Present Within a Firearm" (2016). Graduate Theses, Dissertations, and Problem Reports. 6043.
https://researchrepository.wvu.edu/etd/6043
Determining the Number of Test Fires Needed to Represent the Variability Present Within a Firearm

Eric Law

Thesis submitted
to the Eberly College of Arts and Sciences
at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in
Forensic & Investigative Science

Dr. Keith Morris, Ph.D., Chair
Dr. Jacqueline Speir, Ph.D.
Dr. Afzel Noore, Ph.D.
Dr. Casey Jelsema, Ph.D.

Department of Forensic & Investigative Science

Morgantown, West Virginia
2016

Keywords: Firearms, IBIS®, breech face, firing pin
Copyright 2016 Eric Law
ABSTRACT

Determining the Number of Test Fires Needed to Represent the Variability Present Within a Firearm

Eric Law

Many studies have been performed in recent years in the field of firearm examination with the goal of providing an objective method for comparisons of fired cartridge cases. To date there has been no research to support the number of test fires needed to represent the variability present within a firearm. When a suspect firearm is submitted to a firearm examiner, test fires are performed to collect fired cartridge cases for comparison purposes. Typically, two to four test fires are performed for every firearm. The recovered cartridge cases are compared to each other to determine which characteristics from the firearm are reproducing, and then compared to any cartridge cases collected at a crime scene. The aim of this research was to determine the number of test fires firearm examiners should perform when a suspect firearm is submitted to the lab to balance cartridge case acquisition time with performance accuracy.

Each firearm in the IBIS® database at West Virginia University is represented by 100 fired cartridge case entries. This translates to about 5,000 breech face and firing pin match scores. Random samples of scores were taken separately from the breech face match score and firing pin match score lists. This subset was compared to the total match distribution of the firearm using a hybrid equivalence test to determine if the subset of similarity scores were statistically equivalent to the larger distribution of scores. The hybrid equivalence method was composed of three tests: the Kolmogorov-Smirnov two-sample test, the Mann-Whitney equivalence test, and the Fligner-Killeen test. A combination of the results of each test was utilized to determine equivalence or non-equivalence between the sampled and match distributions.

For the sampled distribution to remain above 80% equivalent to the match distribution, a minimum of 15 cartridge cases should be used to model the match distribution. Thirty cartridge cases is a conservative estimate, being that it is enough to be 100% equivalent to the actual firearm match distribution of 100 fired cartridge cases.
# Contents

1 Introduction .............................................. 1

2 Methods .................................................. 13
   2.1 Sampling Comparison Scores ......................... 13
   2.2 Equivalence Testing .................................. 17
      2.2.1 Kolmogorov-Smirnov Test for Equivalence .... 18
      2.2.2 Hybrid Equivalence Test ......................... 19

3 Results and Discussion ................................. 25
   3.1 Sampling Comparison Scores ......................... 25
   3.2 Equivalence Testing .................................. 36
      3.2.1 Kolmogorov-Smirnov Test for Equivalence .... 36
      3.2.2 Hybrid Equivalence Test ......................... 42

4 Conclusion ............................................ 57

5 Appendix A: Additional Figures ....................... 59
   5.1 Density Estimates .................................... 59
   5.2 Sampling Comparison Scores ......................... 68
   5.3 Equivalence Testing .................................. 70
   5.4 Hybrid Equivalence Test .............................. 85

6 Appendix B: R Scripts ................................ 113
   6.1 Bootstrap ............................................. 113
   6.2 Kolmogorov-Smirnov Equivalence Test ............... 115
   6.3 Equivalence Method .................................. 123

7 Appendix C: The Bootstrap ............................. 136
1. Introduction

Firearm examiners are often asked to determine whether or not a cartridge case recovered at a crime scene had been fired by a firearm found in possession of a suspect. Comparisons are performed by comparing the impressions left on the cartridge cases, specifically the firing pin and breech face impressions. A comparison microscope is used to examine the impressions, allowing for the similarities and differences to be noted. Finally, conclusions are made based on the opinion of the examiner.

A cartridge is made up of four main components: the cartridge case (casing), the primer, the propellant (powder), and the bullet (Figure 1.1 on the following page). When a cartridge is chambered in a firearm and the trigger is pulled, the firing pin, or striker in some firearms, makes contact with the primer on the base of the cartridge, creating the firing pin impression (Figure 1.2 on the next page). This primer, when struck, ignites the propellant within the cartridge creating gases which cause a rapid increase in pressure. The pressure forces the bullet out of the cartridge case and out of the barrel of the firearm. As the bullet is traveling down the barrel, the gas pressure is also pushing the cartridge case back up against the breech face of the firearm, resulting in the breech face impression (Figure 1.2 on the following page). The fired cartridge case eventually is ejected from the firearm, whether by the automatic cycling of semi-automatic or automatic firearms, or manually such as in pump action, break action and bolt action firearms, or revolvers.

Two common types of semi-automatic pistol operation are blowback operation and recoil operation. In blowback operation, the breech is not mechanically locked at the time of firing. Inertia of the bolt and recoil springs keep the breech from opening until the cartridge is fired [2]. The breech is opened by the energy from the rearward movement of the cartridge case as pressure increases by the expanding gases produced from ignition of the propellant. Recoil operated firearms have a breech which is mechanically locked by use of locking lugs, displayed in Figure 1.3 on page 3 [2]. Upon firing, the barrel link (Figure 1.3 on page 3) pivots which unlocks the breech and allows the slide to move rearward, extracting and ejecting the fired cartridge case. As the barrel link pivots and the breech unlocks, the barrel tilts downward slightly. If the firing pin is still in contact with the primer as this occurs, it may leave a
Figure 1.1: Four components making up a cartridge [1].

Figure 1.2: Firing pin impression and breech face impression on the base of a fired Winchester .40 S&W cartridge case (left) and Federal .40 S&W cartridge case (photographs by Eric Law 24-Jan-2015 and 20-Apr-2015). The breech face impression, firing pin impression, and firing pin drag mark are labeled.

drag mark on the primer. The firing pin drag mark is a class characteristic of recoil operated firearms, and is not present in blowback operated firearms.

Different breech face impression patterns are imparted by a firearm on to cartridge cases and are also considered class characteristics. These patterns include smooth, granular, circular, crosshatch, arch, and parallel [4]. The manufacturing process determines the pattern type associated with breech faces. Different cartridge loadings have different amounts and types of propellant which will burn at different rates and generate different levels of pressure within the cartridge. The difference in the amount of pressure will effect the extent to which the cartridge case is pressed up against the breech face. In addition, the type of primer can also have an effect on the resulting breech face impressions. Variations in primer hardness (malleability) due to its composition effects the extent to which impressions are imparted on
Figure 1.3: Recoil operation pistol detailing the location of the locking lugs and barrel link [3].

to cartridge cases. Impressions from a single firearm across different compositions of brass primers may show variation in reproducibility.

Automated databases exist in order to equip the firearms examiner with computerized image analysis for the identification of bullets and cartridge cases recovered in open cases, and link to a database of test fired weapons [5]. The Integrated Ballistics Identification System, or IBIS®, is one such system used to accomplish these two tasks.

The IBIS® is made up of two components: BRASSCATCHER™ for cartridge cases and BULLETPROOF™ for bullets [5]. This study is limited to only cartridge cases, and therefore only the BRASSCATCHER™ component of the IBIS®. The IBIS® station consists of image acquisition hardware, storage media for acquired images, case information input, surface signature analysis, and correlation scoring to an image database.

The Data Acquisition Station (DAS) is used to acquire images of breech face impressions, firing pin impressions, and ejector marks in 2D. A new database entry is created and the relevant information is entered, including caliber, firing pin type, manufacturer, case identifier, and date. The fired cartridge case is then placed on the stage of the microscope. The cartridge case is oriented by drag mark, if present, or breech face pattern if no drag mark is present. The cartridge case fits into a specially designed clamp that holds it in place during image acquisition. The magnification is selected, the focus is automatically adjusted, and the lighting is manually adjusted by use of a bar controlling light intensity. The DAS then automatically finds the breech face impression and firing pin impression areas of the cartridge case (although the user may intervene and manually adjust these areas) and acquires two separate images. These two images, one of the breech face impression and one of the firing pin impression, are stored on an optical disk. The firing pin and breech face impressions are
imaged separately so that images from the firing pin impression made by firearms that have firing pins which can rotate within the breech block are independently correlated without respect to firing pin orientation [5].

After image acquisition, the Signature Analysis Station (SAS), responsible for database correlations and comparisons, uses a proprietary algorithm to process the image and assign a mathematical signature to both the firing pin and breech face impressions [5]. Upon saving the images acquired from the DAS, the database automatically runs a search for each cartridge case. Specialized algorithm(s) are used to process the unique digital signatures extracted from the cartridge case images. The software then displays the results of the search, and can be ordered by firing pin-, breech face-, or ejector mark-scores. The correlation score provided for the firing pin and breech face are unit-less, and describe the amount of similarity between the entered cartridge case and the reference cartridge cases from the database. Because the firing pin and breech face impressions are made during different stages of the firing process, the scores can be used independently of each other. Larger numbers signify greater similarity as determined by the database. The examiner can then bring up the image of the unknown cartridge case along side the reference cartridge case to perform an on-screen visual comparison. The reference case identifier can be used to locate the physical reference cartridge case from the laboratory inventory and be physically compared to the unknown cartridge case using a comparison microscope.

The IBIS® correlation process involves the following steps, using both coarse and fine correlation algorithms [6]:

1. The IBIS® selects a sample of cartridge case exhibits from the database based on general class characteristics, such as caliber and firing pin shape.
2. Coarse breech face and firing pin correlation algorithms are run which are fast and reject unlikely matching candidates.
3. The breech face and firing pin coarse score lists are sorted independently.
4. The top 10% best scored cartridge case exhibits in the breech face and firing pin lists are kept for further analysis.
5. A fine correlation algorithm is run on the breech face and firing pin with the reduced sample.
6. The breech face and firing pin lists are independently sorted.
7. The top 7% of the scores of the exhibits provided by the fine correlation algorithm from both the breech face and firing pin lists are kept.
The IBIS® does not store all correlation results; only the exhibits that scored most similar with the fine correlation algorithm are kept [6].

One of the initial concerns of an automated database, such as the IBIS®, is the consistency and reliability in the results obtained when different examiners enter exhibits [5]. For cartridge cases, the focus and magnification are adjusted automatically and the lighting is electronically monitored which all help to reduce operator error. A study conducted by the Bureau of Alcohol, Tobacco, and Firearms (ATF) San Francisco Laboratory Center evaluated the performance of the IBIS® hardware and software [5]. Pairs of cartridge cases from over 200 pistols representing 9 mm Luger, .380 ACP, and .45 ACP calibers were used, and no microscopical pre-screening of the case pairs prior to selection for testing was performed. The “twin” cartridge case was found in the first position of ranked scores between 73% to 87% of the time, and 90% to 93% of the time in the top five positions. The correct “twin” cartridge case was located in the first position 100% of the time for Glock™ 9 mm caliber pistols.

Previous research has been conducted in order to evaluate inter- and intra-examiner variation in cartridge case acquisition using the IBIS®. Three main parameters can be changed within the IBIS® when entering cartridge cases: ring selection, cartridge case orientation, and lighting intensity. Inter- and intra-examiner variation for these three parameters was shown to have no effect on true positive correlation scores or rank in the list of results [7]. Furthermore, it was determined that user input is not necessary at all in cartridge case acquisition. To evaluate these three parameters, cartridge cases fired from 9 mm Luger and .40 Smith & Wesson™ caliber handguns were used and entered into IBIS® 88 times using specific combinations of ring selection, cartridge case orientation, and lighting. To evaluate the effect of ring selection, average measurements of the diameter of the breech face and firing pin were compared [7]. A significant difference was found to exist between the average measurements done automatically compared to the average measurements done manually at a confidence level of 95%. Automatic measurements were statistically significantly higher than those done manually. ANOVA testing was done to determine if variation in ring selection affected the correlation scores or true positive position. No statistically significant difference was found to exist between automatic or manual selection on scores or position at a confidence level of 99%.

A receiver operating characteristic (ROC) curve is a technique for visualizing, organizing, and selecting classifiers based on their performance [8]. ROC analysis is a generally useful performance graphing method. These curves are two dimensional graphs in which true positive rate is plotted on the y-axis and false positive rate is plotted on the x-axis. A ROC curve depicts relative trade-offs between benefits (true positives) and costs (false positives).
A discrete classifier will produce a (false positive rate, true positive rate) pair corresponding to a single point in the ROC graph \[8\]. The lower left point (0,0) in a ROC graph represents a classifier which never issues positive classification, the upper right point (1,1) represents a classifier which unconditionally issues a positive classification, and the point (0,1) represents a perfect classifier which never issues a false positive \[8\]. The diagonal line \( y = x \) represents randomly guessing a class. The area under the ROC curve numerically corresponds to the performance of a method with a value of one being desirable.

Analysis of the ROC curve for the .40 S&W caliber data by Scicchitano (2011) showed that manual selection was slightly more discriminating. Data for 9 mm Luger firearms showed similar results. There was still no statistically significant effect on correlation score or true positive position.

To evaluate the effect of cartridge case orientation, cartridge cases were entered into the IBIS\textsuperscript{R} using eleven orientations centered around 3 o’clock \[7\] (the IBIS\textsuperscript{R} manual states to orient with the firing pin drag mark at the 3 o’clock position, if present \[4\]). ANOVA testing showed no statistically significant differences between the averages of correlation scores and true positive positions of all orientations. In addition, no statistically significant differences were observed in the average score of number one rank cartridge cases in the database.

Lighting was the third and final parameter evaluated. To test the effect of lighting, cartridge cases were entered into IBIS\textsuperscript{R} using the entire range of available light measured in 10% increments from 0% to 100\% \[7\]. Resulting correlation scores were examined and compared and it was found that lighting had a Gaussian effect on the correlation scores obtained. The optimum range was found to be between 15\% and 40\% intensity. To test this range, cartridge cases were entered into the IBIS\textsuperscript{R} at intensities of 20\%, 28\%, and 38\%, with automatic lighting used as a fourth measure for comparison. When comparing true positive correlation scores and true positive positions, ANOVA testing confirmed no significant difference existed between the four groups, confirming that optimum lighting intensity existed as a range and not a single value.

Many recent challenges have been brought forward on the basis that firearm examination procedures and conclusions are subjective in nature and based on the training and experience of each individual examiner. Common criticisms include \[9\]:

- Whether tool manufacture results in sufficient individuality and the related issue of subclass characteristics;
- If tool surfaces changing over time effecting individualization;
- Lack of objective criteria;
- The need for statistics and databases because absolute identification is not possible;
- Lack of adequate validation;
- Inadequate proficiency testing.

To help address some of the above criticisms, more scientific research with objective data are needed [10]. Ten consecutively manufactured Ruger® P-series 9 mm Luger caliber autoloading pistol slides were used to test the hypothesis that firearms are capable of leaving unique marks. One complete Ruger® P-95™DC 9 mm Luger pistol was also obtained. The slide accompanying this particular pistol was not part of the consecutive machining. The Ruger® production manager detailed the machining steps used to produce the pistol breech faces, which are therefore responsible for the microscopic marks left on the breech faces.

Each of the following steps has the potential to leave microscopic markings on the breech face of the firearm [10]. The pistol slide blank is first shaped on a Matsuura computer numerically controlled (CNC) machine, which typically involves milling cuts that could be seen on the breech face. A gang or step broach is used to cut the final breech face profile. This involves a tooted tool being passed over the breech face in a single pass to remove material. The tool is passed over in a single direction. This step will typically leave parallel, striated marks on the breech face. This step removes all toolmarks created in the previous step. The slide surfaces are deburred by a die grinder and by hand to remove all rough edges. Next, the cartridge recess is milled which helps the extractor hold a spent cartridge case in place. The slide is then both polished and heat treated, hardening the steel making it strong enough to withstand firing. Rough edges are further removed by tumbling in a ceramic medium. After tumbling, the slides are subjected to an automated sand and bead blasting process. This blasting media may strike the breech face, unintentionally, in a completely random fashion. The firing pin aperture is then cut, leaving a tapered edge at the firing pin aperture. Finally, the slide is cleaned, blued, and inspected. Once completed, each slide is subjected to proof testing, as well as function test fires which include at least ten cartridges. This adds up to a total of eleven cartridges being fired from the pistol slide before leaving the factory.

The study by Weller et al. (2012) was limited to one type of ammunition for the purpose of focusing on similarities and differences between the breech faces [10]. Winchester® ammunition with 147 grain bullets were selected because upon testing the breech face marks were well defined and covered the whole primer area. Nine cartridges were fired from each of the consecutively manufactured slides for a total of 90 collected cartridge cases. In addition, five cartridge cases were collected from the pistol which was not part of the consecutively manufactured breech faces. Confocal microscopy was used for the image acquisition of the
collected cartridge cases. Confocal microscopy allows for the detection of three dimensional topography by use of pinhole optics. In conventional microscopy, the entire specimen is flooded evenly with light from a light source. However, a confocal microscope uses point illumination and a pinhole in an optically conjugate plane in front of the detector to eliminate out-of-focus signal (the name “confocal” stems from this configuration). Only the light reflected back from the current focal plane can focus through the pinhole and onto the detector [10]. Cross correlation functions were used for each pairwise comparison.

The toolmarks were found to be a mixture of parallel marks, from broach cutting, and random roughness (granular breech face pattern), from sand and bead blasting [10]. Forensic Sil (Loci Forensic Products, Nieuw-Vennep, the Netherlands) casts of each breech face were collected before any firing and compared. Forensic Sil is a silicone rubber developed to be a quick and easy, non-toxic imprinting material. Correspondence of parallel lines was observed between all consecutively manufactured breech faces, but not with the nonconsecutively manufactured breech face. The average score for matches from the ten consecutively made breech faces was 0.82 out of a perfect correlation of 1.0 with a standard deviation of 0.06. The average score for nonmatches from the ten consecutively made breech faces was 0.20 with a standard deviation of 0.03. No overlap was observed between the matching and nonmatching correlation scores. Five test fires were conducted from the nonconsecutively made breech face to test the presence of subclass characteristics. Subclass characteristics would be reflected as higher nonmatch scores, and no significant difference was found indicating that subclass characteristics had no mathematically detectable influence on the results.

Weller et al. (2012) was the first study to use confocal microscopy and cross correlation analysis on consecutively manufactured firearm surfaces [10]. Firearms were able to be distinguished from one another by marks left from the manufacturing process. The predominant marks were from sand and bead blasting, and there was no significant influence by subclass characteristics.

A specific area in this field in need of more research is the determination of the number of test fires needed to understand the variability within a single firearm. Riva and Champod (2014) published a paper with the goal of obtaining an objective assignment of the weight of cartridge case comparison evidence [11]. They looked at firing pin and breech face impression marks and used an automated comparison system based on 3D technology, a µScan from Nanofocus® AG. This system allowed measuring the topography of an object based on the confocal principle. In the preprocessing stage they based the detection of outliers, or points that did not accurately represent the surface of an object in the z-axis, on the measurement of the quantity of light reflected back to the detector. This was affected mainly by the slope of the surface: the more vertical the surface the more diminished the light is in returning
back to the detector. All points showing reflectance below 10% were eliminated and replaced by a value obtained by linear interpolation using neighboring points. Once outliers had been taken care of, the firing pin and breech face impressions were separated for independent comparison.

The Iterative Closest Point (ICP) algorithm was used to align the firing pin impression images during the comparison phase which minimized the distance between points in one impression and the closest points in the other impression. The ICP algorithm has become the dominant method for aligning three dimensional models based on geometry [12]. The basic ICP concept includes six stages [12]:

1. Selection of some set of points in one or both models.
2. Matching these points to samples in the other model.
3. Weighting the corresponding pairs appropriately.
4. Rejecting certain pairs based on looking at each pair individually or considering the entire set of pairs.
5. Assigning an error metric based on the point pairs.
6. Minimizing the error metric.

The breech face impression images were aligned using a different, but similar, algorithm during comparisons. This algorithm based the alignment on the finest characteristics and minimized the influence of the biggest features, such as the firing pin drag mark. Three similarity scores were generated from these comparisons which quantitatively represented the amount of morphological similarities or differences. These scores were Euclidean distance, correlation index, and the difference between the angle of normal vectors. The Euclidean distance was the ordinary distance measured between two points, the correlation index took into account the number of points superimposed, and the difference between the angle of normal vectors was the median of the difference at each point on both surfaces. These three scores were applied separately to both firing pin impression images and breech face impression images for a total of six scores per cartridge case.

For objective assignment of the weight of the evidence, a likelihood ratio was computed for each questioned cartridge case. First, they created three different distributions. Two distributions were called the within-source distributions, and represented cartridge cases fired from the same firearm. The third distribution was the between-source distribution and represented cartridge cases fired by different firearms. For computation of these distributions,
SIG Sauer® firearms of similar models were used so all firearms would have the same class characteristics. They created three distinct groups: 79 cartridge cases fired by 79 different firearms pairwise compared (between-source distribution), 60 cartridge cases fired by one particular firearm pairwise compared (within-source distribution W1), and 60 cartridge cases fired by another firearm pairwise compared (within-source distribution W2). Geco Sintox® 9 mm Luger ammunition was used for all groups.

The automatic system developed could not easily discriminate firing pin impressions left by different SIG Sauer® firearms. They noticed much better separation of their data when plotting the breech face scores meaning the breech face was more discriminating than firing pin for this study. However, when the firing pin impression was considered jointly with the breech face impression, even better separation could be seen than with either of the impressions alone. Tippett plots were used to evaluate the performance of the system developed. Likelihood ratios supporting the wrong proposition were rarely assigned, and the likelihood ratios generated were very indicative of the true state under both propositions as seen in Table 1.1.

<table>
<thead>
<tr>
<th></th>
<th>LR &lt; 1 Under $H_p$</th>
<th>Max LR</th>
<th>Min LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firearm W1</td>
<td>4.40%</td>
<td>3.92x10^{19}</td>
<td>3.5x10^{-9}</td>
</tr>
<tr>
<td>Firearm W2</td>
<td>0.26%</td>
<td>2.11x10^{22}</td>
<td>0.043</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LR &gt; 1 Under $H_d$</th>
<th>Max LR</th>
<th>Min LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firearm W1</td>
<td>0%</td>
<td>0.224</td>
<td>3.3 x 10^{-20}</td>
</tr>
<tr>
<td>Firearm W2</td>
<td>0.09%</td>
<td>13.2</td>
<td>1.4x10^{-17}</td>
</tr>
</tbody>
</table>

Table 1.1: Percentage of likelihood ratios less than one under the prosecutor’s hypothesis ($H_p$) and greater than one under the defense hypothesis ($H_d$) are shown along with the maximum and minimum likelihood ratios calculated for both scenarios. The 0% for firearm W1 does not indicate the error rate is equal to zero, but only that there were no observations for the number of comparisons performed in the study [11].

Two important considerations which need to be taken into account are that the within-source distribution must be recomputed for every single firearm (no default distribution can be used), and the fact that 60 cartridge cases were used to model both within-source distributions. It is unrealistic to expect a firearm examiner to perform 60 test fires from firearms submitted to the lab. More research is necessary in order to determine the appropriate number of test fires to be conducted in order to balance acquisition time with performance [11].
Currently, no standard exists on the number of test fires to be conducted. Hatcher (1935), revised by Lieutenant Colonel Jury and Weller under the direction of General Hatcher in 1953 entitled *Firearms Investigation, Identification, and Evidence*, is an older reference which discusses the number of test fires to be conducted before making comparisons. No specific number of test fires is referenced under the cartridge case comparisons section, however, it has been established that a small number of test bullets should be fired with four being about the maximum under normal circumstances [13]. Furthermore, when trying to duplicate a special condition, several times that number may be required. In one specific instance, 48 shots were fired in a single suspect firearm before getting the type of sheared primer desired for comparison.

*The Handbook of Firearms and Ballistics* by Heard (2008) recommends a slightly different number of test fires be performed. A minimum of four, but preferably more, rounds of ammunition of the same make and type from the crime scene should be fired [2]. It is suggested the rounds should be collected in sequence because there could be some change in the barrel due to rusting or other factors which could have some effect on the markings present. In addition, the test fires should be cross compared to understand the similarities between them and which marks can be disregarded. This is assuming that four test fires is enough to represent the variability present in that specific firearm.

In *Firearms, the Law, and Forensic Ballistics* by Warlow (2012), no actual number is given to assist the examiner in deciding how many test fires should be conducted. The only answer given is that test fires should always be conducted using several different cartridge loadings [14]. The nature of the impressions will differ from one type of loading to another due to the different levels of pressure generated during firing. Even when firing several shots using the same loading, differences will most likely be observed in how well certain marks are reproduced due to natural variations in pressure levels generated.

The Association of Firearm and Tool Mark Examiners (AFTE) is an international organization dedicated to the advancement of the field of firearm and toolmark identification. They provide training, guidelines, and standards for firearm examiners to reference and follow. AFTE recommends a minimum of two test shots be fired and recovered [15]. No other recommendations are given regarding test fires other than that the type of firearm and ammunition tested should dictate the recovery method used. No rationale is provided as to why two test fires are recommended. AFTE also recommends that test fires produced be compared to each other to determine what microscopic characteristics are reproducing for that specific firearm [15]. This would be difficult to accomplish if an examiner followed the minimum recommendation of two test fires since two cartridge cases alone would not be enough to define the reproducible characteristics within a firearm. Therefore, the purpose
of this study is to address the number of test fires needed to model the match distribution of a firearm.
2. Methods

2.1 Sampling Comparison Scores

The IBIS® compares cartridge cases in a pairwise fashion: the first cartridge case entered from a firearm would be compared to nothing, the second cartridge case would be compared to the first, the third cartridge case would be compared to the second and first, etc. After all cartridge cases were entered into the IBIS®, all of the correlations were deleted and re-run for the purpose of comparing every single cartridge case to all others in the database. All correlation data was extracted into a Microsoft® Excel® spreadsheet. Because all of the correlations had been re-run, the duplicate comparisons needed to be deleted from each data file so that only pairwise comparisons were left. An R script was created to remove duplicate comparisons prior to testing for equivalence.

There were 48 data files which were composed of 30 unique 9 mm firearms (based on serial number). Some firearms were present in multiple data files because manufactured ammunition as well as reloaded ammunition were fired and these were in their own respective data files. All firearm match distribution sizes, number of cartridges fired from each, and the type of ammunition fired are displayed in Table 2.1 on the following page. This data is as extracted from the IBIS® database. The match distribution size does not necessarily equal the number of pairwise comparisons as would be calculated from the number of fired cartridges (Equation (2.1), where $n$ is the number of fired cartridge cases). This may be due to the IBIS® missing some comparisons in the database.

$$\text{# of Comparisons} = \frac{n(n - 1)}{2} \quad (2.1)$$

An adaptive kernel density estimator was used to estimate the probability density function of the match and non match data for ten firearms. Adaptive in this sense refers to local smoothing of the density estimate to obtain a better global estimate [16]. The adaptive approach varies the size of the kernels used depending on the location of the samples to
Table 2.1: Match distribution sizes, number of cartridges fired, and ammunition information for each firearm data file.

<table>
<thead>
<tr>
<th>Serial Number</th>
<th>Make</th>
<th>Model</th>
<th>Cartridge Cases Fired</th>
<th>Match Distribution Size</th>
<th>Ammunition Type</th>
<th>Primer Brand</th>
<th>Powder Brand</th>
<th>Powder Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAN_XXD724</td>
<td>Arminius</td>
<td>m698</td>
<td>190</td>
<td>4774</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>CAN X667727</td>
<td>SCCY</td>
<td>m9CPX</td>
<td>99</td>
<td>4762</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>CYN X667727</td>
<td>SCCY</td>
<td>m9CPX</td>
<td>97</td>
<td>4429</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>CPN X667727</td>
<td>SCCY</td>
<td>m9CPX</td>
<td>96</td>
<td>3743</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>CNX X667727</td>
<td>SCCY</td>
<td>m9CPX</td>
<td>99</td>
<td>4637</td>
<td>Reload</td>
<td>Remington</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>CYN X667727</td>
<td>SCCY</td>
<td>m9CPX</td>
<td>98</td>
<td>3672</td>
<td>Blazer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDN X77862</td>
<td>Springfield</td>
<td>m6X9</td>
<td>98</td>
<td>4726</td>
<td>Reload</td>
<td>Federal</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>FEN X77862</td>
<td>Springfield</td>
<td>m6X9</td>
<td>190</td>
<td>4884</td>
<td>Reload</td>
<td>Federal</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>FNX X77862</td>
<td>Springfield</td>
<td>m6X9</td>
<td>99</td>
<td>4832</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>FPN X77862</td>
<td>Springfield</td>
<td>m6X9</td>
<td>99</td>
<td>4468</td>
<td>Reload</td>
<td>Remington</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>FYN X77862</td>
<td>Springfield</td>
<td>m6X9</td>
<td>99</td>
<td>4805</td>
<td>Reload</td>
<td>Remington</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>FNX X77862</td>
<td>Springfield</td>
<td>m6X9</td>
<td>98</td>
<td>4504</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>GAN X9L713</td>
<td>Glock</td>
<td>m19Gen4</td>
<td>99</td>
<td>4818</td>
<td>Reload</td>
<td>Federal</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>GSN X9L713</td>
<td>Glock</td>
<td>m19Gen4</td>
<td>100</td>
<td>4757</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>ONS X9L713</td>
<td>Glock</td>
<td>m19Gen4</td>
<td>99</td>
<td>4267</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>BCR X9X728</td>
<td>Hi-Point</td>
<td>m9C9</td>
<td>99</td>
<td>4809</td>
<td>Federal</td>
<td>American Eagle</td>
<td>TiteGroup</td>
<td></td>
</tr>
<tr>
<td>BRC X9X728</td>
<td>Hi-Point</td>
<td>m9C9</td>
<td>98</td>
<td>4649</td>
<td>Reload</td>
<td>Remington</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>BRC X55249</td>
<td>Hi-Point</td>
<td>m9C9</td>
<td>130</td>
<td>4756</td>
<td>Reload</td>
<td>CCI</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>BHN X55420</td>
<td>Hi-Point</td>
<td>m9C9</td>
<td>100</td>
<td>4664</td>
<td>Reload</td>
<td>CCI</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>BHN X55457</td>
<td>Hi-Point</td>
<td>m9C9</td>
<td>100</td>
<td>4771</td>
<td>Reload</td>
<td>CCI</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>KTC X9P79</td>
<td>Keltec</td>
<td>m9P79</td>
<td>101</td>
<td>4961</td>
<td>Variable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAN X4521</td>
<td>Ruger</td>
<td>m9L9</td>
<td>98</td>
<td>4577</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>BHN X4220</td>
<td>Ruger</td>
<td>m9L9</td>
<td>99</td>
<td>4714</td>
<td>Reload</td>
<td>CCI</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>BHN X32654</td>
<td>Ruger</td>
<td>m9L9</td>
<td>130</td>
<td>4924</td>
<td>Reload</td>
<td>CCI</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>BRF X44279</td>
<td>Ruger</td>
<td>m9P5</td>
<td>99</td>
<td>4873</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>BRF X14279</td>
<td>Ruger</td>
<td>m9P5</td>
<td>100</td>
<td>4872</td>
<td>Reload</td>
<td>Federal</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>BPN X14279</td>
<td>Ruger</td>
<td>m9P5</td>
<td>91</td>
<td>3661</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>RUGX X9863</td>
<td>Ruger</td>
<td>m8E9</td>
<td>99</td>
<td>4827</td>
<td>Variable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RKN X44279</td>
<td>Ruger</td>
<td>m9P5</td>
<td>99</td>
<td>4821</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>SSW X2X66</td>
<td>SigSauer</td>
<td>m25A</td>
<td>99</td>
<td>4830</td>
<td>Variable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWX X3X922</td>
<td>Smith&amp;Wesson</td>
<td>m309</td>
<td>97</td>
<td>3165</td>
<td>Variable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAN X45181</td>
<td>Taurus</td>
<td>m247G2</td>
<td>99</td>
<td>4838</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>TNN X3420</td>
<td>Taurus</td>
<td>m247G2</td>
<td>99</td>
<td>4820</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>TNN X20264</td>
<td>Taurus</td>
<td>m905</td>
<td>97</td>
<td>4595</td>
<td>Reload</td>
<td>Seller &amp; Bellot</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>TPN X55720</td>
<td>Taurus</td>
<td>m247G2</td>
<td>99</td>
<td>4351</td>
<td>Reload</td>
<td>Remington</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>TRM X85262</td>
<td>Taurus</td>
<td>m2550</td>
<td>99</td>
<td>4822</td>
<td>Variable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRM X55720</td>
<td>Taurus</td>
<td>m247G2</td>
<td>101</td>
<td>5029</td>
<td>Variable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTN X55720</td>
<td>Taurus</td>
<td>m247G2</td>
<td>99</td>
<td>4850</td>
<td>Reload</td>
<td>Federal</td>
<td>Hodgdon</td>
<td>TiteGroup</td>
</tr>
<tr>
<td>TVX X45599</td>
<td>Taurus</td>
<td>m247G2</td>
<td>99</td>
<td>4791</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>TNN X45485</td>
<td>Taurus</td>
<td>m247G2</td>
<td>99</td>
<td>4386</td>
<td>Reload</td>
<td>CCI</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>TXN X45598</td>
<td>Taurus</td>
<td>m247G2</td>
<td>98</td>
<td>4738</td>
<td>Reload</td>
<td>TaloAmmos</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>UPN X5X056</td>
<td>SigSauer</td>
<td>m2100</td>
<td>99</td>
<td>4912</td>
<td>Reload</td>
<td>Remington</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>WBN X9S99</td>
<td>Keltec</td>
<td>m9P79</td>
<td>93</td>
<td>4255</td>
<td>Reload</td>
<td>Federal</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>WNN X9S979</td>
<td>Keltec</td>
<td>m9P79</td>
<td>98</td>
<td>4670</td>
<td>Reload</td>
<td>CCI</td>
<td>Accurate</td>
<td>#2</td>
</tr>
<tr>
<td>WTN X9S979</td>
<td>Keltec</td>
<td>m9P79</td>
<td>98</td>
<td>4752</td>
<td>Reload</td>
<td>Remington</td>
<td>IMR</td>
<td>PB</td>
</tr>
<tr>
<td>WNNX X9S98</td>
<td>Keltec</td>
<td>m9P79</td>
<td>96</td>
<td>4555</td>
<td>Reload</td>
<td>Federal</td>
<td>Accurate</td>
<td>#2</td>
</tr>
</tbody>
</table>

The list of breech face and firing pin match score correlation data generated by the IBIS® for each individual cartridge case comparison, random samples of raw scores were taken with replacement ranging from two scores up to 200. The empirical cumulative distribution functions (CDFs) were calculated using the `ecdf` function within the R `stats` package [18].

A class of statistical tests called cumulative distribution function (CDF) comparison tests were utilized to compare the sampled scores to the match distribution. Included in this class are the Kolmogorov-Smirnov test, Cramer-von Mises test, and Anderson-Darling test which are all used to determine if two samples were drawn from the same population [19]. The test

avoid oversmoothing where lots of observables are present and undersmoothing where less observables are present (typically at the tails of the distribution) [17].

From the list of breech face and firing pin match score correlation data generated by the IBIS® for each individual cartridge case comparison, random samples of raw scores were taken with replacement ranging from two scores up to 200. The empirical cumulative distribution functions (CDFs) were calculated using the `ecdf` function within the R `stats` package [18].

A class of statistical tests called cumulative distribution function (CDF) comparison tests were utilized to compare the sampled scores to the match distribution. Included in this class are the Kolmogorov-Smirnov test, Cramer-von Mises test, and Anderson-Darling test which are all used to determine if two samples were drawn from the same population [19]. The test
statistic and p-value for each were saved to a CSV file at each number of scores sampled. For the Kolmogorov-Smirnov test, the critical value at 95% confidence was calculated as shown in Equation (2.4) [19]. If the test statistic for a given sampling was less than the critical value, then the null hypothesis that the two samples have identical cumulative distribution functions was not rejected.

For the Kolmogorov-Smirnov test, the two sample, two sided version of the test was applied. The null and alternate hypotheses are [19]:

\[ H_0 : F(x) = G(x) \text{ for all } x \text{ from } -\infty \text{ to } +\infty \]
\[ H_1 : F(x) \neq G(x) \text{ for at least one value of } x \]

The test statistic \( D \) is defined as the maximum vertical deviation between the empirical distribution functions of two samples as shown in Equation (2.3) [19].

\[ D = \sup_x |S_1(x) - S_2(x)| \] (2.3)

The null hypothesis \( H_0 \) is rejected at the level of significance \( \alpha \) if the test statistic exceeds the \( 1-\alpha \) quantile as calculated using Equation (2.4) at the 95% confidence level, where \( m \) and \( n \) are the sizes of the two samples [19].

\[ W_{.95} \approx 1.36 \sqrt{\frac{m + n}{mn}} \] (2.4)

The Kolmogorov-Smirnov test was implemented using \texttt{ks.test} available within the R \texttt{stats} package [18]. Using \texttt{ks.test}, the sampled data were compared to the match data to determine if any significant differences were present.

The Cramer-von Mises test for two samples was also used to test if two samples were drawn from the same population [19]. This test seems to make more effective use of the data because unlike the Kolmogorov-Smirnov test which locates only the maximum vertical deviation between two CDFs, the Cramer-von Mises tests calculates the sum of the squared differences at all sample points. The hypotheses are the same as for the Kolmogorov-Smirnov
test in Equation (2.2) on page 15 [19]. The test statistic is the sum of the squared differences between the CDFs multiplied by a normalization factor accounting for both sample sizes [19]:

\[ T = \frac{mn}{(m + n)^2} \left\{ \sum_{i=1}^{n} [S_1(X_i) - S_2(X_i)]^2 + \sum_{j=1}^{m} [S_1(Y_j) - S_2(Y_j)]^2 \right\} \]  (2.5)

For the Cramer-von Mises test the 95% confidence critical value was located in a table [19]. The test statistics were compared to the critical value, and the null hypothesis that the two samples have identical distribution functions was not rejected if the test statistic was less than the critical value. Likewise, for the Kolmogorov-Smirnov and Cramer-von Mises tests, if the p-value was greater than 0.05, the null hypothesis was not rejected.

The Cramer-von Mises test was performed using `cvm.test` within the R `dgof` package [20].

The Anderson-Darling test was developed as another cumulative distribution function comparison test. It has two advantages over the Kolmogorov-Smirnov test [21]: 
a) it is especially sensitive at the tails of the distributions, and b) there is evidence it is better at detecting very small differences, especially with large sample sizes. The two sample Anderson-Darling test statistic is calculated as follows [21]:

\[ AD = \frac{1}{mn} \sum_{i=1}^{n+m} (N_iZ_{(n+m-ni)})^2 \frac{1}{iZ_{(n+m-i)}} \]  (2.6)

In Equation (2.6) \( Z_{(n+m)} \) represents the combined and ordered samples \( X_{(n)} \) and \( Y_{(m)} \) of size \( n \) and \( m \) and \( N_i \) represents the number of observations in \( X_{(n)} \) that are equal to or smaller than the \( i \)th observation in \( Z_{(n+m)} \) [21].

The Anderson-Darling test was performed using `ad.test` within the R `kSamples` package [22].

One of the most valuable features of the Kolmogorov-Smirnov two sided test statistic is that it may be used to form a confidence band for the true unknown distribution function [19]. A random sample can be drawn, and from that sample an upper and lower limit that contain the unknown parameter between them with probability \( 1-\alpha \) may be calculated. In other words, a random sample can be drawn from some population whose distribution function is completely unknown, and the upper and lower limits could be placed on the graph and the statement could be made that the unknown distribution function lies entirely within those bounds with probability \( 1-\alpha \).
A random sample of scores was taken from the firearm match distribution, and the two sided Kolmogorov-Smirnov critical value was calculated at 95% confidence using Equation (2.4) on page 15. The critical value was plotted both above (upper bound) and below (lower bound) the CDF of the random sample of scores. This created the confidence band in which the unknown distribution function was completely contained at 95% confidence.

2.2 Equivalence Testing

The most obvious difference between classical hypothesis testing and equivalence testing is that the null and alternate hypotheses are switched [23]. In tests of equivalence the null hypothesis states there is a difference in the samples, while the alternate hypothesis states the samples are equivalent. Using a traditional comparative test to establish equivalence frequently leads to incorrect conclusions because the burden of proof is on the wrong hypothesis (that there is a difference) [24]. In this instance, a nonsignificant result only implies equivalency cannot be ruled out; traditional hypothesis tests cannot prove that a state of no difference between two groups exists [24]. The general form of an equivalence test is shown in Equation (2.7).

\[ H_0 : \mu_1 \neq \mu_2 \]
\[ H_1 : \mu_1 = \mu_2 \]  

(2.7)

Determination of the equivalence margin, \( \Delta \), is the most important step in testing for equivalence [24]. This value defines the equivalence region and determines the result of the test. This \( \Delta \) value will define how close is enough to be considered equivalent. In Equation (2.8), \( H_0 \) is defined as nonequivalent because the difference in means of the two groups is greater than the equivalence margin \( \Delta \) and \( H_1 \) is defined as equivalent because the difference in means of the two groups is less than the equivalence margin.

\[ H_0 : |\mu_1 - \mu_2| > \Delta \]
\[ H_1 : |\mu_1 - \mu_2| < \Delta \]  

(2.8)
The most commonly used test for equivalence is the two one-sided test (TOST) [24]. Restating H₀,

\[ H₀: \mu₁ - \mu₂ < -\Delta \text{ or } \mu₁ - \mu₂ > \Delta \] (2.9)

To declare the two means as equivalent, both criteria in Equation (2.9) must be rejected. The above equations use mean when discussing equivalence testing, but mean may be substituted for dispersion (variance) or whatever parameter is of interest.

### 2.2.1 Kolmogorov-Smirnov Test for Equivalence

The Kolmogorov-Smirnov test has a one sided version. The null and alternative hypotheses are stated as follows in the one sided case [19],

\[ H₀ : F(x) \geq G(x) \]
\[ H₁ : F(x) < G(x) \text{ for at least one value of } x \] (2.10)

The converse of Equation (2.10) is also true for the opposite one sided test. Naturally, because there is a one sided version of the Kolmogorov-Smirnov test, a two one sided test for equivalence makes sense [25].

\[ D^+ = |\max_t (F_Y(t) - F_X(t))| \]
\[ D^- = |\min_t (F_Y(t) - F_X(t))| \] (2.11)

The null and alternative hypothesis may be stated,

\[ H₀^+ : F_Y(t) \leq F_X(t) \]
\[ H₁^+ : F_Y(t) > F_X(t) \text{ for at least one value of } t \] (2.12)

\[ H₀^- : F_Y(t) \geq F_X(t) \]
\[ H₁^- : F_Y(t) < F_X(t) \text{ for at least one value of } t \]
These hypotheses may then be translated, within a general equivalence interval \( \Delta \),

\[
H_0^- : |\min_t (F_Y(t) - F_X(t))| \geq \Delta \\
H_A^- : |\min_t (F_Y(t) - F_X(t))| < \Delta
\]  

(2.13)

The two one sided “negativist” null hypotheses to test for equivalence take the same form because both \( D^+ \) and \( D^- \) are always non-negative.

\[
H_{01}^- : D^+ \geq \Delta \\
H_{02}^- : D^- \geq \Delta
\]  

(2.14)

The test statistics are \( D_1^+ \) and \( D_2^- \) which correspond to \( H_{01}^- \) and \( H_{02}^- \), respectively, are

\[
D_1^+ = \Delta - D^+ = \Delta - |\max_t (F_Y(t) - F_X(t))| \\
D_2^- = \Delta - D^- = \Delta - |\min_t (F_Y(t) - F_X(t))|
\]  

(2.15)

If \( D_1 \geq \Delta \) and \( D_2 \geq \Delta \), then \( H_{01}^- \) and \( H_{02}^- \) are rejected. Otherwise, fail to reject \( H_{01}^- \) and \( H_{02}^- \). If both \( H_{01}^- \) and \( H_{02}^- \) are rejected, then \( H_0^- \) is rejected, establishing evidence of distribution equivalence.

To test the performance of the equivalence test version of the Kolmogorov-Smirnov test, random normal distributions were created with the same mean and standard deviation. The Kolmogorov-Smirnov equivalence test was used to determine if it could detect that the distributions were in fact the same. The value chosen for \( \Delta \) was 0.1. This was a large value for the equivalence margin because a CDF ranges from 0 to 1. A 0.1 equivalence margin represents a 10% difference between two CDFs.

### 2.2.2 Hybrid Equivalence Test

A hybrid approach to equivalence testing was implemented after trying the equivalence test version of the Kolmogorov-Smirnov test. This method utilized three statistical tests: the Mann-Whitney equivalence test for distribution location, the Fligner-Killeen test for dispersion equivalence, and the classic two-sided Kolmogorov-Smirnov test (not the Kolmogorov-Smirnov equivalence test presented in Section 2.2.1 on page 18). Each test had to result in equivalence to conclude that the sampled distribution was equivalent to the match distribution. The two-sided Kolmogorov-Smirnov test was not used to imply that the two
distributions were equivalent, but it was a level of protection against different distributions which resulted in equivalence from the Mann-Whitney and Fligner-Killeen tests.

In testing for differences in location of two distributions the two sample t-test is the most powerful test to use with the assumption that the two populations are normally distributed \[19\]. The assumption of normality can be hard to verify, and non-normal distributions may result in very little power when using the t-test compared to the Mann-Whitney test, which is especially true when one or both samples contain outliers \[19\]. The asymptotic relative efficiency of the Mann-Whitney test compared with the t-test under perfect normal distributions that are identical except for their means is 0.955 \[19\]. With any departures from normality, the asymptotic relative efficiency increases above 1. The power obtained by the Mann-Whitney test for equivalence shows little sensitivity to changes in form of the baseline distribution \[26\].

The probability of obtaining an observation from the first distribution which exceeds some single observation drawn from the second distribution can be represented by Equation (2.16) \[26\]. This is the mean difference scaled by the standard deviation. If the mean difference is zero, then the equation is equal to a probability of 0.5 on the normal CDF. The testing problem is shown in Equation (2.17) \[26\]. A distribution free solution to this testing problem can be based on the large sample distribution of the estimator \(W_+\) of \(\pi_+\) \[26\]. This estimator is given as the Mann-Whitney two sample test statistic shown in Equation (2.18). The equivalence bounds were set to 0.5 ± 0.2 for the range of probabilities on the CDF in determining equivalence. This led to a 0.3 to 0.7 equivalence margin.

\[
\pi_+ = \Phi \left( \frac{\xi - \eta}{\sqrt{2} \sigma} \right) \tag{2.16}
\]

\[
H : \pi_+ \leq 1/2 - \epsilon'_1 \text{ or } \pi_+ \geq 1/2 - \epsilon'_2
\]

\[
K : 1/2 - \epsilon'_1 < \pi_+ < 1/2 + \epsilon'_2 \tag{2.17}
\]

\[
W_+ = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{(0,\infty)}(X_i - Y_j) \tag{2.18}
\]

The F-test for homogeneity of variance is very sensitive to the assumption of normality, and for this reason is not safe to use unless it is known that the populations are normal \[19\]. The Fligner-Killeen rank-based procedure for two-sample homogeneity of variance testing
and estimation, however, is optimal under normality, and unlike the traditional F-test for variance, possesses robustness in validity and power for non-normal situations [27]. A large Monte Carlo study of dispersion tests over a wide variety of situations was performed by Conover, Johnson, and Johnson (1981). The traditional F-test for homogeneity of variance did poorly, and in certain situations the empirical α levels exceeded 0.80 when the nominal α level was 0.05 [28]. A test with a statistic similar to that of Fligner-Killeen performed much better, but had the tendency to be liberal in some situations, whereas the Fligner-Killeen test was found to be valid over all situations tested [27] [28].

The Fligner-Killeen test in itself is not an equivalence test. The \texttt{fk.test} function within the R package \texttt{npsm} calculated the test statistic and p-value, as well as a confidence interval with upper and lower bounds [27] [29]. If two distributions have the same variance, the ratio of variances of each respective distribution will be one because the variances would be equal. An equivalence test for dispersion would be set up to include a range of values, around one, which would be considered close enough to one to say no difference in dispersion was detected. The critical bounds for the equivalence margin were chosen to be (0.66, 1.50) which allowed for a 33% difference in dispersion. The interval (0.66, 1.50) is on the scale of standard deviation because even though the Fligner-Killeen test was comparing the variances, the results it provided were in terms of standard deviation. If the resulting 95% confidence interval from the \texttt{fk.test} function in R was included entirely within the equivalence limits, then it was determined that the two distributions were equivalent in dispersion.

To test the power of the equivalence testing method, simulations were run using distributions with the same mean and standard deviation. Statistical power is defined as the probability that a false null hypothesis will be rejected [30]. Power is inversely related to the probability of making a Type II error, or false negative [30]. If the power of a test is set at 0.50, or 50%, then there is a 50/50 chance of rejecting a null hypothesis that is false. If power is set to a higher level, 0.90 for example, then the chance of detecting an effect is improved, however, the sample size for a power of this magnitude may be too high for practical purposes [30].

The appropriate level of power for a test was suggested by Cohen (1988) to be set at 0.80 [30]. This would lead to an 80% probability of detecting a real effect and a 20% probability of making a Type II error. This creates a good balance between Type I and Type II risk. Most scientists view Type I errors to be more serious than Type II errors [30]. Cohen proposed that Type I errors should be treated four times more seriously than Type II errors [30]. If alpha significance (Type I error) is set to 0.05, then beta significance (Type II error) should be set to 0.20. If a 5% chance of making a Type I error can be tolerated, then a 20% chance of making a Type II error should be tolerated [30].
The R code for the Mann-Whitney equivalence test was available online [26]. This code, however, was extremely inefficient. It included multiple nested for loops which tend to be slow when done in R, and to do a simulation test with 10,000 iterations it would have taken 27 days to finish. The code was posted to the website StackExchange to get input from others on how to improve the speed [32]. StackExchange user “Gabe” responded with a solution and the corresponding R code. The nested for loops were coded in the language Fortran within RStudio, and then written as an R function. This allowed R to call on Fortran to run the nested for loops. With this implemented, the time it took to run 10,000 iterations of the code was down to between two and three hours, compared to 27 days with the nested for loops written in R. The Fortran code is shown as part of the R Script in Appendix B: R Scripts (Section 6.3 on page 123).

To allow R to compile the Fortran code the following method had to be used:

1. The packages inline and Rcpp were downloaded and installed through RStudio [33] [34] [35];
2. Rtools was downloaded and installed to the computer C drive;
3. A new path had to be set in the beginning of the R code to direct the code to the Fortran compiler with Rtools (Appendix B: R Scripts (Section 6.3 on page 123).

Performance of the equivalence method was tested as follows:

1. Random normal distributions were generated that had the same mean and standard deviations, and were set up to simulate the data from the firearms. One generated distribution had 60 values, and 50 values down to ten in increments of five and were compared to a second generated distribution of 5,000 values;
2. The two-sample Kolmogorov-Smirnov test, Mann-Whitney test for equivalence, and Fligner-Killeen test were run on each iteration;
3. The above procedure was repeated 10,000 times for each smaller distribution size.

Each of the 48 data files were tested for equivalence separately to ensure that ammunition fired was the same between cartridge cases for each comparison. The number of scores sampled from the match distribution of each firearm began at 60 and then decreased from 50 down to ten in increments of five, with one thousand iterations being run at each sample level. Samples were taken from the match distribution without replacement, and once a sample was done for each iteration, those sampled scores were removed from the match distribution to simulate independence between the data sets. The R script for firearm equivalence testing
output a separate file for both breech face and firing pin for each firearm data file. Plots
detailing the number of scores equivalent to the match distribution at each number sampled,
as well as the number of unique cartridge cases for both the breech face and firing pin files,
were created for each data file.

The following method was implemented for each data file:

1. The data file was subset to include only matches;
2. \( n \) random samples were taken from the list of breech face and firing pin match scores
   without replacement;
3. The \( n \) random samples were removed from the list of match scores for both breech face
   and firing pin;
4. The distribution of sampled scores was compared to the match distribution using
   the two-sample Kolmogorov-Smirnov test, Mann-Whitney test for equivalence, and
   Fligner-Killeen test;
5. The above procedure was repeated 1,000 times for each of the number of scores sampled
   which was 60 scores, and 50 scores to ten in increments of five;
6. The results from the tests were plotted as percent equivalent out of the 1,000 iterations
   performed at each sample level.

Each test was designed to output a value of zero if the distributions being compared were
equivalent (no significant difference for the Kolmogorov-Smirnov test), or a one if they were
non-equivalent (significant difference for the Kolmogorov-Smirnov test). The sum of these
values was taken to combine the results of the three tests. If the sum of all three tests for a
given iteration was zero, then the sampled and match distribution were said to be equivalent.
If the sum of all three tests was greater than zero for a given iteration, then the sampled
and match distribution were said to be non-equivalent.

The data was also sampled for three firearms (CAN X66727, TAN X45401, and FXN
X77862) at the cartridge case level prior to testing for equivalence. Two, three, four, five,
eight, and ten cartridge cases were sampled, along with all of their respective pairwise com-
parisons, removed from the list of all cartridge cases fired by that firearm, and then compared
using the equivalence method performing 1,000 iterations at each number sampled for a total
of 6,000 iterations per firearm.

Furthermore, to better represent a realistic number of comparisons that would be per-
formed by a firearm examiner, after sampling at the cartridge case level all non-pairwise
comparisons were removed leaving only inter-comparison scores between the sampled cartridge cases. For example, ten sampled cartridge cases would be made up of 45 unique pairwise comparison scores (as calculated using Equation (2.1) on page 13). The number of cartridge cases sampled were 10, 15, 20, 25, and 30, and 1,000 iterations were performed at each number sampled.
3. **Results and Discussion**

3.1 **Sampling Comparison Scores**

Kernel density estimates are included to show that the match and nonmatch densities are not normal, and vary in shape as well as separation between the matches and nonmatches. In Figure 3.1 on page 28, a Taurus™ 24/7 G2 firearm, there is good separation between the matches and nonmatches for both breech face and firing pin scores. However, in Figure 3.2 on page 29, another Taurus™ 24/7 G2 firearm, only little separation between breech face and firing pin matches and nonmatches can be seen. Refer to Appendix A: Additional Figures (Section 5.1 on page 59) for kernel density estimates of additional firearms.

Comparison scores were sampled from the actual match distribution and compared using the Kolmogorov-Smirnov test, Cramer-von Mises test, and Anderson-Darling test. The number of scores sampled were 150, 100 to 10 in increments of 10, and 10 to 2 in increments of 1. Refer to Appendix A: Additional Figures (Section 5.2 on page 68) for data tables with test statistics and p-values for each of the three tests at different.

In Figure 3.3 on page 30 the variability within the p-value for the Kolmogorov-Smirnov test is shown. The Cramer-von Mises test and Anderson-Darling tests resulted in variability similar to the Kolmogorov-Smirnov test. The red line indicates the 95% confidence threshold, with a p-value falling below the line being significantly different from the match distribution and a p-value falling above the line being not significantly different. Thirty random samples were taken at each number of scores sampled.

The results of the Cramer-von Mises two sample test is shown in Figure 3.4 on page 31, with green areas indicating failure to reject the null hypothesis (no statistically significant difference found) and red areas representing rejecting the null hypothesis (evidence to suggest a difference exists). Two scores were sampled with one score plotted on the x-axis and the other score on the y-axis. The intersection point is the green or red color indicative of the p-value from the Cramer-von Mises test.

The Kolmogorov-Smirnov critical value at 95% confidence increases as the sample size decreases. The critical values are represented for breech face and firing pin in Figure 3.5 on
page 32 and Figure 3.6 on page 33, respectively, by the red lines. Each blue dot is the test statistic calculated using Equation (2.3) on page 15. A lot more variability is present the smaller the sample size becomes. Any point above the 95% confidence level is significantly different from the parent distribution, and only a few instances of this exist.

The Kolmogorov-Smirnov confidence band plots for breech face and firing pin scores are shown in Figure 3.7 on page 34 and Figure 3.8 on page 35, respectively. The red lines in the plot area show the upper and lower bounds for the number of scores sampled. Two scores do not give any real indication of where the firearm match CDF lies. As the number of scores increases up to 50, a much better representation is presented as given by the tighter bounds on where the firearm match CDF lies.

In a hypothesis test there are two propositions: the null hypothesis and the alternative hypothesis. The null hypothesis is assumed to be true until the data provide sufficient evidence to support the alternative hypothesis. The degree of statistical evidence needed to support the alternative hypothesis is the confidence level and is one minus the alpha value (Type 1 error rate), which occurs when you incorrectly reject the null hypothesis. The alpha value is frequently 0.05 and corresponds to a 95% confidence level. This leads to accepting a 5% chance of rejecting the null hypothesis even if it is true. If the evidence fails to reject the null hypothesis, it does not mean the null hypothesis is true because a hypothesis test does not determine which hypothesis is true or even which is most likely. A hypothesis test only assesses whether there is available evidence to reject the null hypothesis [36].

The Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests are all under the same null and alternative hypotheses, where the alternative is that there is a difference between the two empirical cumulative distribution functions being compared. To show that two distributions are the same, the null hypothesis that the two empirical cumulative distribution functions are equal must not be rejected. This rationale is flawed because failing to reject the null hypothesis is not the same as “accepting” the null hypothesis to be true. The test results in this instance only state that there is not enough evidence to support the alternative hypothesis that the empirical cumulative distribution functions are different, and therefore does not give evidence to support the null hypothesis that the empirical cumulative distribution functions are equal.

When sampling different numbers of scores from the match distribution and using the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests to compare, the resulting test statistics and p-values almost always led to failing to reject the null hypothesis. However, this did not mean that the distributions were equivalent. As shown in Figure 3.9 on page 37, the plot with two samples is visually different from the match distribution, however, the Kolmogorov-Smirnov test indicated they were not significantly different. This is because
the sample size was so small that the critical value covered almost the entire range of the cumulative distribution plot, as represented by the red 95% confidence bound in Figure 3.5 on page 32.

Further support for two scores being different than the match distribution based on visual examination was provided by the Kolmogorov-Smirnov confidence band plots. Two sampled scores gave no real indication of where the match CDF would lie as the whole range of the plot area is included within the bounds. Once 50 scores were sampled, a much better representation of the location and shape of the actual match CDF was observed. This provided more evidence that a classic hypothesis test was not going to lead to a correct interpretation of the results.
Figure 3.1: Taurus™ 24/7 G2 (X45401) match density estimate (blue solid line) and non-match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 3.2: Taurus™ 24/7 G2 (X55720) match density estimate (blue solid line) and non-match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 3.3: Variability in p-value within sampling levels for the Kolmogorov-Smirnov two sample test. The red line is at $\alpha=0.05$ (95% confidence level). Thirty iterations were performed at each number of scores sampled. Any point above the red line had no significant difference.
Figure 3.4: Cramer-von Mises two sample test results of two sampled scores compared to the match distribution from firearm X45401. One of the two scores is represented on the x-axis and the other score is represented on the y-axis.
Figure 3.5: Kolmogorov-Smirnov test statistics of the number of scores sampled shown on the x-axis based on 10 iterations. The red line is the 95% confidence level based on the critical value calculated using Equation (2.4) on page 15. Any point under the red line had no significant difference.
Figure 3.6: Kolmogorov-Smirnov test statistics of the number of scores sampled shown on the x-axis based on 10 iterations. The red line is the 95% confidence level based on the critical value calculated using Equation (2.4) on page 15. Any point under the red line had no significant difference.
Figure 3.7: Kolmogorov-Smirnov confidence band plots at 95% confidence for breech face scores.
Figure 3.8: Kolmogorov-Smirnov confidence band plots at 95% confidence for firing pin scores.
3.2 Equivalence Testing

3.2.1 Kolmogorov-Smirnov Test for Equivalence

Kolmogorov-Smirnov one sided tests were used, and the test statistic for both one sided versions of the test were coded in R and can be found in Appendix B: R Scripts (Section 6.2 on page 115). The code was adapted from the R `ks.test` source code for the purpose of providing only the one and two sided test statistics for the Kolmogorov-Smirnov test. The compareMax function calculated the maximum vertical deviation that the sampled CDF lies above the match CDF. Conversely, the compareMin function calculated the maximum vertical deviation that the sampled CDF lies below the match CDF. The compareTwo function calculated the two sided test statistic, which was equal to the larger of both one sided test statistics.

Using the proposed equivalence test version of the Kolmogorov-Smirnov test from Section 2.2.1 on page 18, the results for comparing two distributions, whether they had the same mean and standard deviation or not, always rejected equivalence.

The plots created by the code in Appendix B: R Scripts (Section 6.2 on page 115) display the match CDFs in red and the sampled CDFs in blue. The maximum deviations from both the compareMax and compareMin functions are represented by the black vertical line drawn between the two CDFs. Examples of the Kolmogorov-Smirnov test statistic are represented in Figure 3.9 on the next page and Figure 3.10 on page 38 for breech face scores, and in Figure 3.11 on page 39 and Figure 3.12 on page 40 for firing pin scores. Refer to Appendix A: Additional Figures (Section 5.3 on page 70) for additional figures with comparison between different numbers of sampled scores.
Figure 3.9: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 2 breech face comparison scores. Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 3.10: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 100 breech face comparison scores. Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 3.11: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 2 firing pin comparison scores. Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 3.12: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 100 firing pin comparison scores. Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
The Kolmogorov-Smirnov test for equivalence was promising in theory, however, after implementation it did not perform as was expected. When two random normal distributions were created, with the same mean and standard deviation, the Kolmogorov-Smirnov equivalence test always failed to reject the null hypothesis. No evidence of distribution equivalence was found in spite of the fact that the two distributions were equivalent.

The same results were obtained when creating two random normal distributions with different means and standard deviations. No evidence of equivalence was ever found. This was expected as the distributions were different, but was tested to ensure that the implementation of the Kolmogorov-Smirnov test for equivalence was not providing the opposite result.

The decision factor for this equivalence test is if $D_1 \geq \Delta$ and $D_2 \geq \Delta$, then $H_{01}$ and $H_{02}$ are rejected. If $\Delta$ is set at 0.1, then the maximum vertical deviation, $D$, will never be greater than or equal to $\Delta$. This is true for any value chosen for $\Delta$ based on the test statistics in Equation (2.15) on page 19 because $\Delta$ is outside of the absolute value. This implementation of an equivalence test version of the Kolmogorov-Smirnov test would never reject both null hypotheses to establish evidence of distribution equivalence.

The Kolmogorov-Smirnov equivalence test did not work, however, it would be ideal because it would be sensitive to location, dispersion, and shape of the distributions being compared.
3.2.2 Hybrid Equivalence Test

The power curve in Figure 3.13 details the power of each of the three individual tests (Kolmogorov-Smirnov, Mann-Whitney test for equivalence, and the Fligner-Killeen test) as well as the overall method combining the three tests. Ten thousand iterations were run at each sample level and compared to a distribution of 5,000 samples. The method is at 80% power at a sample size of around 40.

For the Mann-Whitney equivalence bounds, $0.5 \pm 0.2$ was chosen. This bound was used because comparing two normal distributions each with a sample size of 36 resulted in 82% power, while the $t$-test resulted in 86% power [26]. When performing simulations with random equivalent normal distributions of sizes to simulate the firearm data, the Mann-Whitney test performed as expected in terms of power (Figure 3.13).

The equivalence bounds for the Fligner-Killeen test were chosen to be $(0.66, 1.50)$, allowing for a 33% difference in the variances between two distributions. The range $(0.75, 1.33)$, allowing for a 25% difference in variance, was tested in a simulation using random equivalent normal distributions, but resulted in poor power under these optimal conditions. For this reason, the equivalence margin was increased to incorporate a slightly larger difference in variance that resulted in the power seen in Figure 3.13.

![Equivalence Method Power Curve](image)

*Figure 3.13: Power curve for the equivalence test method as well as each of the three individual tests.*
The `fk.test` function in R was providing an error message when the firearm data was being run through the equivalence test script. Some iterations would give an error message while others would run perfectly. The error was in relation to `drop.test` which was a function called on by `fk.test` in computing a p-value. To correct for this, the source code was adjusted to skip the `drop.test` function altogether, and output an “NA” for the p-value. For this implementation, only the confidence interval was utilized, so a p-value of “NA” did not effect the results. The `fk.test` function was then recompiled, avoiding `drop.test`, therefore fixing the error being received. The new function was named `fk.ind` and can be seen as part of the equivalence method R code in Appendix B: R Scripts (Section 6.3 on page 123).

Equivalence method plots such as in Figure 3.14 on the next page, show the percent of equivalent distributions on the y-axis at each of the sample levels on the x-axis. The secondary x-axes are the mean number of unique cartridge cases composing the number of scores sampled for breech face and firing pin results. The corresponding standard deviations are shown below the means of the two groups.

In all equivalence results plots, four letters before the serial number in the title indicates that a combination of manufactured and reloaded ammunition was present in the data file. In these files, manufactured ammunition is compared against reloaded ammunition and vice versa. Ten scores was the smallest number sampled from the firearm match distributions. At ten scores, all firearms were under 30% equivalent for both breech face and firing pin results, with many being at 0% equivalent (except for a Springfield™ XD9 in Figure 5.27 on page 87 which is at about 35% equivalent at ten sampled scores).

The same SCCY™ CPX firearm (serial number X66727) is displayed in Figure 3.14 on the next page, Figure 3.15 on page 45, and Figure 3.16 on page 46, and the cartridge cases contributing to each of these plots were reloads with different combinations of primer and powder manufactures. The cartridge cases in Figure 3.14 on the following page performed better in breech face than firing pin using TulAmmo brand primers with IMR PB powder. The cartridge cases in Figure 3.15 on page 45 had Federal brand primers with Hodgdon TiteGroup powder and the cartridge cases in Figure 3.16 on page 46 had Sellier and Bellot brand primers with Hodgdon TiteGroup powder, and these both performed better in firing pin than breech face. The number of scores sampled to remain at about 80% equivalent varies between the three plots, however, the same general shape can be seen.

The plots shown in Figure 3.17 on page 46 and Figure 3.18 on page 47 are both Taurus™ 24/7 G2 firearms (with different serial numbers) and performed well with 80% equivalent being reached at just over 30 scores for Figure 3.17 on page 46 and right at 30 scores for Figure 3.18 on page 47. Less of a difference was observed between breech face and firing pin
Figure 3.14: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

compared to the SCCY™ CPX firearms in Figure 3.14, Figure 3.15 on the following page, and Figure 3.16 on page 46.

A Hi-Point™ C9, shown in Figure 3.19 on page 47, was one of the best performers with 80% equivalent being reached at about 25 scores sampled. The breech face and firing pin follow the same trend with little space between the two throughout the range of scores sampled.

The plot shown in Figure 3.20 on page 48 is the same firearm as in Figure 3.14 and is based on sampling at the cartridge case level instead of sampling random comparison scores. The percentage of equivalent distributions stayed relatively constant sampling two cartridges cases compared to ten. The breech face data approaches 50% equivalent, and is at its maximum, at three sampled cartridge cases. The firing pin data never exceeds 40% equivalent, but reaches it at four sampled cartridge cases. A similar trend with the percent equivalent staying fairly constant from two cartridge cases sampled up to ten can be seen for two additional firearms in Figure 3.21 on page 48 and Figure 3.22 on page 49.

The results when sampling at the cartridge case level for each individual test based on the number of total iterations out of the 6,000 run that were determined to be non-equivalent are displayed in Table 3.1 on page 49 for each of the three firearms. The Kolmogorov-Smirnov test resulted in non-equivalence for between 50% and 76% of the scores from the
Figure 3.15: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

three firearms tested. The Mann-Whitney and Fligner-Killeen tests for equivalence rejected equivalence at a much smaller rate, between 1% and 10%.

The Kolmogorov-Smirnov test was removed from the sampled cartridge case data to determine how well the firearm would perform when just using the Mann-Whitney and Fligner-Killeen equivalence tests. The resulting plots are shown in Figure 3.23 on page 50, Figure 3.24 on page 50, and Figure 3.25 on page 51 and are very different from the plots shown in Figure 3.20 on page 48, Figure 3.21 on page 48, and Figure 3.22 on page 49 where the Kolmogorov-Smirnov test is included. These plots detail the number of cartridge cases sampled to be compared to the match distribution of the firearm. The secondary x-axes show the mean number of comparison scores sampled with the corresponding standard deviations below for both breech face and firing pin. At ten sampled cartridge cases, the three firearms were each at or just below 100% equivalent.
Figure 3.16: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 3.17: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 3.18: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 3.19: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 3.20: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases.

Figure 3.21: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases.
Figure 3.22: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases.

Table 3.1: Percent of sampled distributions not equivalent to the firearm match distribution when sampling cartridge cases broken down by test.

<table>
<thead>
<tr>
<th>Test</th>
<th>BF</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAN X66727</td>
<td>0.55</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>FXN X77862</td>
<td>0.56</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>TAN X45401</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Figure 3.23: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases without using the Kolmogorov-Smirnov test.

Figure 3.24: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases without using the Kolmogorov-Smirnov test.
Figure 3.25: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases without using the Kolmogorov-Smirnov test.
The equivalence test plots display the number of comparison scores needed to represent the match distribution of the firearm with the current match distribution size in the IBIS® database. The plots do not describe anything about the quality of the underlying match distribution itself. Some comparisons were not present in the data files from the IBIS®. Examining the breech face and firing pin impressions of the cartridge cases that the IBIS® missed when pairwise comparing them may lead to more insight into performance of the IBIS®.

The number of cartridge cases for both breech face and firing pin scores are displayed on the secondary x-axes. This is the number of unique cartridge cases from the database that the number of sampled comparison scores are composed of. This number of cartridge cases does not provide the number of test fires needed in a strict sense. For example, 45 comparison scores could be obtained by performing ten test fires and pairwise comparing them, however, Figure 3.14 on page 44 shows that 35 cartridge cases are needed for 45 comparison scores. Pairwise comparing thirty five cartridge cases would lead to 595 comparison scores, more than the 45 scores shown in Figure 3.14 on page 44.

The equivalence method plots are based on randomly sampling comparison scores. Assuming 45 scores are desired, then ten cartridge cases could be fired and pairwise compared to obtain those 45 scores. However, the structure of pairwise comparing ten cartridge cases to obtain the 45 scores desired is different than a random sampling of 45 scores from a list of 100 pairwise compared cartridge cases.

In addition to sampling comparison scores, cartridge cases were sampled with all of their respective pairwise comparisons. Practically, sampling cartridge cases makes more sense than comparing individual comparison scores because a firearm examiner does a number of test fires and pairwise compares them. When the Kolmogorov-Smirnov test was included in the sampled cartridge case results, the percent equivalent stayed relatively constant from two to ten sampled cartridge cases (Figure 3.20 on page 48). Sampling cartridge cases and all of their respective pairwise comparisons led to a large number of scores being compared to the match distribution of the firearm. At this large sample size, the Kolmogorov-Smirnov test had too much power, meaning any very small deviation between the distributions would result in the Kolmogorov-Smirnov test rejecting the null hypothesis that there were no differences present. For this reason the Kolmogorov-Smirnov test results were removed leading to the plot shown in Figure 3.23 on page 50. Without the Kolmogorov-Smirnov test, the percent equivalent drastically improved.

These plots are describing the number of cartridge cases needed to represent the 100 cartridge cases which currently make up most of the firearms in the IBIS® database. This number of cartridge cases is indicative of the number needed to represent the match dis-
tribution of the firearm, under the assumption that 100 cartridge cases is enough, as the distribution of the ten cartridge cases is equivalent to the distribution of 100.

The issue addressed above with the plots based on sampling comparison scores is also present in the plots based on sampling cartridge cases. In Figure 3.20 on page 48, for example, ten cartridge cases is composed of 483 comparison scores for both breech face and firing pin. This is due to the database size, as there are a large number of samples already present. Pairwise comparing ten cartridge cases to an existing database of 100 will absolutely lead to more comparison scores than pairwise comparing ten cartridge cases among only each other as would be the case with an unknown firearm not currently present in the database.

To simulate the addition of an unknown firearm, only pairwise comparison scores were kept from the list of sampled cartridge cases. The results are displayed in Figure 3.26 on the next page, Figure 3.27 on the following page, and Figure 3.28 on page 55. The number of cartridge cases is on the x-axis, with the number of comparison scores on the secondary x-axes. The number of comparison scores may deviate slightly from the exact number of expected pairwise comparisons because the IBIS® missed some comparisons when running its correlation algorithm. For example, 10 cartridge cases should equal 45 pairwise comparison scores, but the firearm FXN (Figure 3.27 on the following page) only had 42 scores on average at 10 sampled cartridge cases over the 1,000 iterations.

The percent equivalent decreased as the number of sampled cartridge cases increased due to the Komogorov-Smirnov test increasing in power as the sample size increased. For this reason the Kolmogorov-Smirnov test was removed and the resulting plots are shown in Figure 3.29 on page 55, Figure 3.30 on page 56 and Figure 3.31 on page 56. At 30 sampled cartridge cases for all three firearms the percent equivalent was at 100%. Twenty sampled cartridge cases provided at least 90% power and 15 sampled cartridge cases provided at least 80% power for each firearm. At less than 15 sampled cartridge cases the power dropped below 80%, increasing the probability of false negatives.
Figure 3.26: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases.

Figure 3.27: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases.
Figure 3.28: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases.

Figure 3.29: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases without using the Kolmogorov-Smirnov test.
Figure 3.30: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases without using the Kolmogorov-Smirnov test.

Figure 3.31: Percent of sampled distributions equivalent to the firearm match distribution when sampling cartridge cases without using the Kolmogorov-Smirnov test.
4. Conclusion

From the data presented in this study, two cartridge cases are not enough to represent the match distribution of a firearm. Only one comparison score can be measured when comparing two cartridge cases from an unknown firearm, and a match distribution could not be estimated with only one value.

In the study done by Riva and Champod (2014), the two firearm match distributions created were each composed of 1,770 comparison scores from pairwise comparing 60 fired cartridge cases [11]. Based on a database of 100 pairwise compared cartridge cases for a given firearm, 1,770 comparison scores and 60 fired cartridge cases are both more than enough to be representative of that population.

From the present simulations, to remain above 80% equivalent (providing a good balance between Type I and Type II errors), a minimum of 15 cartridge cases should be used to model the match distribution of an unknown firearm. Thirty cartridge cases, half of what was used in the study done by Riva and Champod (2014), is a conservative estimate, being that it is enough to be 100% equivalent to the actual firearm match distribution of 100 fired cartridge cases. This is based on sampling cartridge cases with only their inter-comparison scores and not using the Kolmogorov-Smirnov test (see Figures 3.29 to 3.31 on pages 55 to 56).

In a case where a shooting occurred, a forensic scientist is going to compare a cartridge case recovered at a crime scene to test fires from a firearm found in possession of a suspect. A power curve, similar to the one in Figure 3.29 on page 55, could be developed for that make and model of firearm (if it is currently in the database of the agency) based on the type of ammunition, such as manufactured or reloaded, if known, prior to test firing. This power curve would allow the forensic scientist to perform an assessment of the correct number of test fires necessary for that make and model of firearm and ammunition combination. If the make and model of firearm of interest is not present in the database, then the examiner must perform the number of test fires necessary for the worst performing firearm known.

The most important area for further research would be to expand on this study by including other makes and models of 9 mm firearms, as well as additional calibers, specifically by sampling cartridge cases with only inter-comparison scores. This would allow for inter- and intra-caliber comparisons of the number of cartridge cases necessary to model match
distributions. Further generalization of results may also be possible to determine if other firearms and calibers would follow the same trends as seen in Figures 3.29 to 3.31 on pages 55-56.

It would also be helpful to develop an understanding of why the IBIS® fails to recover some known matches when running its correlation algorithms. Because of its proprietary algorithms it is not possible to observe how the IBIS® functions. By examining features between known match cartridge cases the IBIS® does not recover, more insight into the operation of the correlation algorithms may be possible.
5. Appendix A: Additional Figures

5.1 Density Estimates
Figure 5.1: Ruger® LC9 (X32446) match density estimate (blue solid line) and non match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 5.2: Ruger® LC9 (X33654) match density estimate (blue solid line) and non match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 5.3: Ruger® LC9 (X43521) match density estimate (blue solid line) and non match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 5.4: Taurus™ 24/7 G2 (X45398) match density estimate (blue solid line) and non-match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 5.5: Springfield™ XD9 (X77862) match density estimate (blue solid line) and non-match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 5.6: Hi-Point™ C9 (X80728) match density estimate (blue solid line) and non match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 5.7: Smith and Wesson\textsuperscript{TM} SD9 (XA9892) match density estimate (blue solid line) and non match density estimate (red dashed line) for breech face and firing pin correlation scores.
Figure 5.8: Arcus D98 (XXX724) match density estimate (blue solid line) and non match density estimate (red dashed line) for breech face and firing pin correlation scores.
### 5.2 Sampling Comparison Scores

Table 5.1: Comparison of the Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises test statistics and p-values for the listed number of breech face scores sampled for a given iteration.

<table>
<thead>
<tr>
<th>Sampling</th>
<th>KS Test Statistic</th>
<th>KS P-value</th>
<th>AD Test Statistic</th>
<th>AD P-value</th>
<th>CvM Test Statistic</th>
<th>CvM P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>0.05</td>
<td>0.79</td>
<td>0.50</td>
<td>0.79</td>
<td>0.08</td>
<td>0.78</td>
</tr>
<tr>
<td>100</td>
<td>0.06</td>
<td>0.92</td>
<td>0.40</td>
<td>0.83</td>
<td>0.04</td>
<td>0.94</td>
</tr>
<tr>
<td>90</td>
<td>0.08</td>
<td>0.57</td>
<td>1.20</td>
<td>0.27</td>
<td>0.12</td>
<td>0.51</td>
</tr>
<tr>
<td>80</td>
<td>0.06</td>
<td>0.90</td>
<td>0.37</td>
<td>0.89</td>
<td>0.05</td>
<td>0.87</td>
</tr>
<tr>
<td>70</td>
<td>0.15</td>
<td>0.08</td>
<td>1.66</td>
<td>0.14</td>
<td>0.35</td>
<td>0.14</td>
</tr>
<tr>
<td>60</td>
<td>0.10</td>
<td>0.53</td>
<td>0.67</td>
<td>0.53</td>
<td>0.11</td>
<td>0.52</td>
</tr>
<tr>
<td>50</td>
<td>0.14</td>
<td>0.29</td>
<td>2.04</td>
<td>0.09</td>
<td>0.37</td>
<td>0.22</td>
</tr>
<tr>
<td>40</td>
<td>0.08</td>
<td>0.97</td>
<td>0.25</td>
<td>0.99</td>
<td>0.03</td>
<td>0.97</td>
</tr>
<tr>
<td>30</td>
<td>0.13</td>
<td>0.73</td>
<td>0.54</td>
<td>0.72</td>
<td>0.11</td>
<td>0.59</td>
</tr>
<tr>
<td>20</td>
<td>0.12</td>
<td>0.95</td>
<td>0.39</td>
<td>0.82</td>
<td>0.06</td>
<td>0.86</td>
</tr>
<tr>
<td>10</td>
<td>0.13</td>
<td>1.00</td>
<td>0.22</td>
<td>0.98</td>
<td>0.03</td>
<td>0.99</td>
</tr>
<tr>
<td>9</td>
<td>0.20</td>
<td>0.87</td>
<td>0.35</td>
<td>0.87</td>
<td>0.06</td>
<td>0.85</td>
</tr>
<tr>
<td>8</td>
<td>0.37</td>
<td>0.22</td>
<td>1.68</td>
<td>0.16</td>
<td>0.29</td>
<td>0.17</td>
</tr>
<tr>
<td>7</td>
<td>0.33</td>
<td>0.43</td>
<td>0.92</td>
<td>0.38</td>
<td>0.13</td>
<td>0.51</td>
</tr>
<tr>
<td>6</td>
<td>0.26</td>
<td>0.91</td>
<td>1.73</td>
<td>0.15</td>
<td>0.13</td>
<td>0.51</td>
</tr>
<tr>
<td>5</td>
<td>0.36</td>
<td>0.54</td>
<td>1.06</td>
<td>0.37</td>
<td>0.20</td>
<td>0.31</td>
</tr>
<tr>
<td>4</td>
<td>0.46</td>
<td>0.37</td>
<td>1.42</td>
<td>0.17</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>3</td>
<td>0.62</td>
<td>0.21</td>
<td>1.27</td>
<td>0.26</td>
<td>0.28</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>0.35</td>
<td>0.97</td>
<td>0.32</td>
<td>0.93</td>
<td>0.06</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Table 5.2: Comparison of the Kolmogorov-Smirnov, Anderson-Darling, and Cramer-von Mises test statistics and p-values for the listed number of firing pin scores sampled for a given iteration.

<table>
<thead>
<tr>
<th>Sampling</th>
<th>KS Test Statistic</th>
<th>KS P-value</th>
<th>AD Test Statistic</th>
<th>AD P-value</th>
<th>CvM Test Statistic</th>
<th>CvM P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>0.04</td>
<td>0.99</td>
<td>0.25</td>
<td>0.97</td>
<td>0.03</td>
<td>0.98</td>
</tr>
<tr>
<td>100</td>
<td>0.07</td>
<td>0.66</td>
<td>0.48</td>
<td>0.78</td>
<td>0.07</td>
<td>0.74</td>
</tr>
<tr>
<td>90</td>
<td>0.05</td>
<td>0.99</td>
<td>0.18</td>
<td>1.00</td>
<td>0.02</td>
<td>1.00</td>
</tr>
<tr>
<td>80</td>
<td>0.08</td>
<td>0.67</td>
<td>0.59</td>
<td>0.74</td>
<td>0.08</td>
<td>0.71</td>
</tr>
<tr>
<td>70</td>
<td>0.08</td>
<td>0.72</td>
<td>0.39</td>
<td>0.87</td>
<td>0.04</td>
<td>0.93</td>
</tr>
<tr>
<td>60</td>
<td>0.11</td>
<td>0.44</td>
<td>1.23</td>
<td>0.27</td>
<td>0.19</td>
<td>0.27</td>
</tr>
<tr>
<td>50</td>
<td>0.11</td>
<td>0.53</td>
<td>1.33</td>
<td>0.31</td>
<td>0.24</td>
<td>0.20</td>
</tr>
<tr>
<td>40</td>
<td>0.16</td>
<td>0.27</td>
<td>0.90</td>
<td>0.40</td>
<td>0.17</td>
<td>0.33</td>
</tr>
<tr>
<td>30</td>
<td>0.20</td>
<td>0.17</td>
<td>2.36</td>
<td>0.07</td>
<td>0.44</td>
<td>0.04</td>
</tr>
<tr>
<td>20</td>
<td>0.17</td>
<td>0.60</td>
<td>0.72</td>
<td>0.52</td>
<td>0.13</td>
<td>0.45</td>
</tr>
<tr>
<td>10</td>
<td>0.39</td>
<td>0.09</td>
<td>1.40</td>
<td>0.19</td>
<td>0.30</td>
<td>0.12</td>
</tr>
<tr>
<td>9</td>
<td>0.24</td>
<td>0.68</td>
<td>0.45</td>
<td>0.85</td>
<td>0.08</td>
<td>0.71</td>
</tr>
<tr>
<td>8</td>
<td>0.18</td>
<td>0.96</td>
<td>0.48</td>
<td>0.83</td>
<td>0.04</td>
<td>0.90</td>
</tr>
<tr>
<td>7</td>
<td>0.52</td>
<td>0.05</td>
<td>2.26</td>
<td>0.04</td>
<td>0.49</td>
<td>0.04</td>
</tr>
<tr>
<td>6</td>
<td>0.41</td>
<td>0.26</td>
<td>1.19</td>
<td>0.25</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>5</td>
<td>0.20</td>
<td>0.99</td>
<td>0.63</td>
<td>0.62</td>
<td>0.04</td>
<td>0.96</td>
</tr>
<tr>
<td>4</td>
<td>0.41</td>
<td>0.51</td>
<td>0.65</td>
<td>0.64</td>
<td>0.13</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>0.53</td>
<td>0.37</td>
<td>1.14</td>
<td>0.32</td>
<td>0.20</td>
<td>0.26</td>
</tr>
<tr>
<td>2</td>
<td>0.66</td>
<td>0.34</td>
<td>1.05</td>
<td>0.35</td>
<td>0.22</td>
<td>0.23</td>
</tr>
</tbody>
</table>
5.3 Equivalence Testing
Figure 5.9: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 3 breech face comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.10: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 5 breech face comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.11: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 8 breech face comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.12: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 10 breech face comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.13: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 30 breech face comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.14: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 50 breech face comparison scores for a Taurus\textsuperscript{TM} 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.15: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 200 breech face comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.16: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 3 firing pin comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.17: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 5 firing pin comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.18: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 8 firing pin comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.19: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 10 firing pin comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.20: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 30 firing pin comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.21: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 50 firing pin comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
Figure 5.22: The maximum vertical deviation (black line) between the match distribution in red and sampled CDF in blue based on 200 firing pin comparison scores for a Taurus™ 24/7 G2 (SN X45401). Top plot is the distance the sampled CDF lies above the match CDF, and the bottom plot is the distance the sampled CDF lies below the match CDF.
5.4 Hybrid Equivalence Test

Figure 5.23: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.24: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.25: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.26: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.27: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.28: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.29: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.30: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.31: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.32: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.33: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.34: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.35: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.36: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.37: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.38: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.39: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.40: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.41: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.42: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.43: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.44: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.45: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.46: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.47: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.48: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.49: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.50: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.51: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.52: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.53: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.54: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.55: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.56: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.57: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.58: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.59: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.60: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.61: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.62: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.

Figure 5.63: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
Figure 5.64: Percent of sampled distributions equivalent to the match distribution of the firearm with two additional x-axes showing the number of cartridge cases present at the corresponding number of sampled scores for both breech face and firing pin.
6. Appendix B: R Scripts

6.1 Bootstrap

```r
setwd("C:/Users/Eric/OneDrive/Law_Thesis_Research/Taurus_24_7_G2 ↪ X45401/Boot_Plots_New")

load("C:/Users/Eric/OneDrive/Law_Thesis_Research/NineMM_Data/ ↪ CompleteNineMM.RData")
#load("C:/Users/lawen_000/Desktop/NineMM_Data/CompleteNineMM.RData ↪ " )

# Subset one firearm from NineMM data set by serial number
Gun <- Firearms[which(Firearms$IdentifierGun_Sample==’X45401’),]
#write.csv(Gun, file = ”X45401.csv”)

# Randomly sample two cartridge cases by exhibit number
#?set.seed()
set.seed(5)
Sample <- Gun[sample(1:nrow(Gun), 2, replace=FALSE),]
Sample$ExhibitNumber_Sample
#Sample$CaseID_Sample

# Subset randomly sampled exhibits
x <- Gun[which(Gun$ExhibitNumber_Sample==’121’),]
y <- Gun[which(Gun$ExhibitNumber_Sample==’169’),]

# Merge randomly sample cartridge cases into one data frame
```

113
\begin{verbatim}
Gun1 <- rbind(x, y)

# Pull out exhibits with Match=Yes
Gun_Match <- Gun1[which(Gun1$Match=='Yes'),]

# function which will bootstrap the standard error of the mean for Breech Face Match Scores
BF <- subset(Gun_Match, select=BF, row.names=FALSE)
row.names(BF) <- NULL
data <- BF$BF

b.mean <- function(data, num) {
  resamples <- lapply(1:5000, function(i) sample(data, replace=T))
  r.mean <- sapply(resamples, mean)
  std.err <- sqrt(var(r.mean))
  list(std.err=std.err, resamples=resamples, means=r.mean)
}

# saving the results of the function b.mean in the object b1
b1 <- b.mean(data, 5000)

# displaying the first of the 5000 bootstrap samples
b1$resamples[1]

# displaying the standard error
b1$std.err

Ecdf(b1$means)

# adaptive kernel density estimate for breech face bootstrap
x <- seq(min(b1$means), max(b1$means), 0.5)
BF_boot_dens5000 <- akj(sort(b1$means),x)
\end{verbatim}
#creating and saving breech face bootstrap histogram with adaptive density estimate overlay
png(filename="X45401_Breech_Face_Score_Bootstrap_2_Cartridge_Cases_5000_Repetitions.png", width=1000, height=800)
par(mai=c(1.02,1,1,0.42))

plot(x, BF.boot_dens5000$dens, type="l", main="Breech Face Score Bootstrap 2 Cartridge Cases 5000 Repetitions", xlab="Mean Breech Face Correlation Score", ylab="Density", cex.lab=1.5, cex.axis=1.5, cex.main=1.5, cex.sub=1.5)

dev.off()

#qq plot to test normality
png(filename="X45401_QQ_Plot_Breech_Face_Score_Bootstrap_2_Cartridge_Cases_5000_Repetitions.png", width=1000, height=800)
par(mai=c(1.02,1,1,0.42))
qqnorm(b1$means, cex.lab=1.5, cex.axis=1.5, cex.main=1.5)
qqline(b1$means, col="red", lwd=1)

dev.off()

6.2 Kolmogorov-Smirnov Equivalence Test

compareTwo <- function(x1, y1) {
  n1 <- length(x1); m1 <- length(y1)
  w1 <- c(x1, y1)
  o1 <- order(w1)
  z1 <- cumsum(ifelse(o1 <= n1, m1, -n1))
  i1 <- which.max(abs(z1))
  w1[o1[i1]]
}

compareMax <- function(x2, y2) {

}
\[ n2 \leftarrow \text{length}(x2); \quad m2 \leftarrow \text{length}(y2) \]
\[ w2 \leftarrow c(x2, y2) \]
\[ o2 \leftarrow \text{order}(w2) \]
\[ z2 \leftarrow \text{cumsum}(\text{if else}(o2 \leq n2, m2, -n2)) \]
\[ i2 \leftarrow \text{which.max}(z2) \]
\[ w2[o2[i2]] \]

\text{compareMin} \leftarrow \text{function}(x3, y3) \{ \\
\quad n3 \leftarrow \text{length}(x3); \quad m3 \leftarrow \text{length}(y3) \]
\quad w3 \leftarrow c(x3, y3) \\
\quad o3 \leftarrow \text{order}(w3) \]
\quad z3 \leftarrow \text{cumsum}(\text{if else}(o3 \leq n3, m3, -n3)) \]
\quad i3 \leftarrow \text{which.min}(z3) \\
\quad w3[o3[i3]] \}

# = = = = = = = = = = = = = = = = = = = = = = = = = =
setwd("C:/Users/Eric/OneDrive/Law_Thesis_Research/KS_TOST") 
load("C:/Users/Eric/OneDrive/Law_Thesis_Research/NineMM_Data/\rightarrow CompleteNineMM.RData") 
# = = = = = = = = = = = = = = = = = = = = = = = = = =
ConTwo \leftarrow 1.36 \# \text{used to calculate KS critical value} -- 1.36 \text{ for 95\% confidence}, 1.63 \text{ for 99\% confidence}
ConOne \leftarrow 1.22 \# 1.22 \text{ for 95\% confidence one sided test}
Delta \leftarrow 0.1

Iter \leftarrow 100
Seeder \leftarrow 15
Sample_Level \leftarrow c(200, 30, 10, 8, 5, 3, 2) 
#\text{length}(\text{Sample Level})

Output_BF \leftarrow c() 
Output_FP \leftarrow c() 

Gun_Num \leftarrow c("X45401")
for (k in 1:length(Gun_Num)) {

  Gun <- Firearms[which(Firearms$IdentifierGun_Sample==Gun_Num[k])
                 ↦ .]
  Gun_MDistribution <- Gun[which(Gun$Match=='Yes'),]

  BF_Match <- subset(Gun_MDistribution, select=c(CaseID_Sample, BF ↦ ), row.names=FALSE)
  FP_Match <- subset(Gun_MDistribution, select=c(CaseID_Sample, FP ↦ ), row.names=FALSE)

for (i in 1:length(Sample_Level)) {

  for (j in 1:Iter) {

    BF_Subset <- BF_Match[sample(1:nrow(BF_Match), Sample_Level[ ↦ i], replace=TRUE),]
    #BFks <- ks.test(BF_Subset$BF, BF_Match$BF)

    uTwo <- compareTwo(BF_Subset$BF, BF_Match$BF)  #use two sided
           ↦ compare function
    uMax <- compareMax(BF_Subset$BF, BF_Match$BF)  #use one sided
           ↦ max compare function
    uMin <- compareMin(BF_Subset$BF, BF_Match$BF)  #use one sided
           ↦ min compare function
    e.x <- ecdf(BF_Subset$BF)  #Calculate empirical CDF of subset
    e.y <- ecdf(BF_Match$BF)  #Calculate empirical CDF of match
           ↦ scores

    TS_TwoSided <- abs(e.x(uTwo) - e.y(uTwo))  #provides two
                ↦ sided KS test statistic
    TS_OneSidedMax <- abs(e.x(uMax) - e.y(uMax))  #provides one
                      ↦ sided greater KS test statistic
    TS_OneSidedMin <- abs(e.x(uMin) - e.y(uMin))  #provides one
                      ↦ sided less KS test statistic

    D_minus <- Delta-TS_OneSidedMin

  }
}

D_minus <- Delta-TS_OneSidedMin
\[ D_{\text{plus}} \leftarrow \Delta \text{Delta} - \text{TS\_OneSidedMax} \]

\[
\text{BFks\_TwoCV} \leftarrow \text{ConTwo} \ast \text{sqrt} \left( \left( \text{length} (\text{BF\_Subset$BF$}) + \text{length} (\text{BF\_Match$BF$}) \right) / \left( \text{length} (\text{BF\_Subset$BF$}) \ast \text{length} (\text{BF\_Match$BF$}) \right) \right) \\
\text{#two sided critical value} \\
\text{BFks\_OneCV} \leftarrow \text{ConOne} \ast \text{sqrt} \left( \left( \text{length} (\text{BF\_Subset$BF$}) + \text{length} (\text{BF\_Match$BF$}) \right) / \left( \text{length} (\text{BF\_Subset$BF$}) \ast \text{length} (\text{BF\_Match$BF$}) \right) \right) \\
\text{#one sided critical value} \\
\text{BFks\_CC} \leftarrow \text{length} (\text{unique} (\text{BF\_Subset$CaseID\_Sample$})) \ \#\text{number of unique cartridge cases} \\
\]

\[
\text{Decision\_min} \leftarrow \text{if else} (D_{\text{minus}} > \Delta , ' \text{Reject}' , ' \text{Fail to Reject}' \) \ \#\text{Decision min} \\
\text{Decision\_max} \leftarrow \text{if else} (D_{\text{plus}} > \Delta , ' \text{Reject}' , ' \text{Fail to Reject}' \) \ \#\text{Decision max} \\
\]

\[
\text{Outer} \leftarrow \text{c} (\text{Sample\_Level[i]}, \text{BFks\_CC, j, TS\_TwoSided, TS\_OneSidedMax, TS\_OneSidedMin, BFks\_TwoCV, BFks\_OneCV}) \\
\text{cat} (\text{Gun\_Num[k]}, \text{Sample\_Level[i]}, j, "\text{n}") \\
\text{Output\_BF} \leftarrow \text{rbind} (\text{Output\_BF}, \text{Outer}) \\
\]

\[
\text{fname} \leftarrow \text{paste} (\text{Gun\_Num[k]}, \text{Sample\_Level[i]}, " \text{Breech\_Face\_TwoSided\_KS\_Test\_ECDF.png", sep="\""}) \\
\text{png} (\text{fname}, \text{width}=1200, \text{height}=800) \\
\text{par} (\text{mai}=\text{c} (1.2, 1.3, 1.2, 0.42)) \ \#\text{bottom, left, top and right} \\
\text{plot} (\text{e\_x, col="Blue"}, \text{main=paste (Gun\_Num[k], "\text{Breech\_Face\_TwoSided\_KS\_Test\_ECDF} with" ,Sample\_Level[i], "Scores Sampled", sep="\""}, \text{xlab="", ylab="", cex.main=2, cex=2, xlim=\text{c} (0,180)}) \\
\text{plot} (\text{e\_y, add=TRUE, col="Red"}) \\
\text{lines} (\text{c(uTwo, uTwo)}, \text{c (0,1), col="Gray"}) \\
\text{lines} (\text{c(uTwo, uTwo)}, \text{c (e\_x(uTwo), e\_y(uTwo))}, \text{lwd=2}) \\
\text{title} (\text{xlab="Correlation Score", ylab = "Probability", line = 4, cex.lab=2}) \\
\text{dev.off} ()
\]
fname <- paste(Gun_Num[k], Sample_Level[i], "Breech Face One → Sided KS Test Max ECDF.png", sep=" ")
png(fname, width=1200, height=800)
par(mai=c(1.2, 1.3, 1.2, 0.42)) # bottom, left, top and right
plot(e.x, col="Blue", main=paste(Gun_Num[k], " Breech Face One → Sided KS Test Max ECDF with Sample Level[i], " Scores Sampled", sep=" "), xlab="", ylab="", cex.main = 2, cex.axis=2, xlim=c(0,180))
plot(e.y, add=TRUE, col="Red")
lines(c(uMax, uMax), c(0,1), col="Gray")
lines(c(uMin, uMin), c(e.x(uMax), e.y(uMax)), lwd=2)
title(xlab=" Correlation Score", ylab=" Probability", line = 4, cex.lab=2)
dev.off()

fname <- paste(Gun_Num[k], Sample_Level[i], "Breech Face One → Sided KS Test Min ECDF.png", sep=" ")
png(fname, width=1200, height=800)
par(mai=c(1.2, 1.3, 1.2, 0.42)) # bottom, left, top and right
plot(e.x, col="Blue", main=paste(Gun_Num[k], " Breech Face One → Sided KS Test Min ECDF with Sample Level[i], " Scores Sampled", sep=" "), xlab="", ylab="", cex.main = 2, cex.axis=2, xlim=c(0,180))
plot(e.y, add=TRUE, col="Red")
lines(c(uMin, uMin), c(0,1), col="Gray")
lines(c(uMin, uMin), c(e.x(uMin), e.y(uMin)), lwd=2)
title(xlab=" Correlation Score", ylab=" Probability", line = 4, cex.lab=2)
dev.off()

Outer <- c()
#
FP_Subset <- FP_Match[sample(1:nrow(FP_Match), Sample_Level[i], replace=TRUE),]
#FPks <- ks.test(FP_Subset$FP, FP_Match$FP)
uTwo <- compareTwo(FP_Subset$FP, FP_Match$FP)  # use two sided
  ↪ compare function
uMax <- compareMax(FP_Subset$FP, FP_Match$FP)   # use one sided
  ↪ max compare function
uMin <- compareMin(FP_Subset$FP, FP_Match$FP)   # use one sided
  ↪ min compare function
e.x <- ecdf(FP_Subset$FP)  # Calculate empirical CDF of subset
e.y <- ecdf(FP_Match$FP)    # Calculate empirical CDF of match
  ↪ scores
TS_TwoSided <- abs(e.x(uTwo) - e.y(uTwo))  # provides two
  ↪ sided KS test statistic
TS_OneSidedMax <- abs(e.x(uMax) - e.y(uMax))  # provides one
  ↪ sided greater KS test statistic
TS_OneSidedMin <- abs(e.x(uMin) - e.y(uMin))  # provides one
  ↪ sided less KS test statistic

D_minus <- Delta - TS_OneSidedMin
D_plus <- Delta - TS_OneSidedMax

FPks_TwoCV <- ConTwo * sqrt((length(FP_Subset$FP) + length(FP_Match$FP)) / (length(FP_Subset$FP) * length(FP_Match$FP)))
  ↪ # two sided critical value
FPks_OneCV <- ConOne * sqrt((length(FP_Subset$FP) + length(FP_Match$FP)) / (length(FP_Subset$FP) * length(FP_Match$FP)))
  ↪ # one sided critical value
FPks_CC <- length(unique(FP_Subset$CaseID_Sample))  # number
  ↪ of unique cartridge cases

Decision.min <- ifelse(D_minus > Delta, 'Reject', 'Fail to reject')  # Decision min

Decision.max <- ifelse(D_plus > Delta, 'Reject', 'Fail to reject')  # Decision max
Outer <- c(Sample_Level[i], FPks_CC, j, TS_TwoSided, TS_
    OneSidedMax, TS_OneSidedMin, FPks_TwoCV, FPks_OneCV)
cat(Gun_Num[k], Sample_Level[i], j, "\n")
Output_FP <- rbind(Output_FP, Outer)

fname <- paste(Gun_Num[k], Sample_Level[i], " Firing Pin TwoSided KS Test ECDF. png", sep=" ")
png(fname, width=1200, height=800)
par(mai=c(1.2,1.3,1.2,0.42))
plot(e.x, col="Blue", main=paste(Gun_Num[k], " Firing Pin TwoSided KS Test ECDF with", Sample_Level[i], " Scores", sep=" "), xlab="", ylab="", cex.main=2, cex.axis=2, xlim=c(0,180))
plot(e.y, add=TRUE, col="Red")
lines(c(uTwo, uTwo), c(0,1), col="Gray")
lines(c(uTwo, uTwo), c(e.x(uTwo), e.y(uTwo)), lwd=2)
title(xlab="Correlation Score", ylab = "Probability", line =4, cex.lab=2)
dev.off()

fname <- paste(Gun_Num[k], Sample_Level[i], " Firing Pin OneSided KS Test Max ECDF. png", sep=" ")
png(fname, width=1200, height=800)
par(mai=c(1.2,1.3,1.2,0.42))
plot(e.x, col="Blue", main=paste(Gun_Num[k], " Firing Pin OneSided KS Test Max ECDF with", Sample_Level[i], " Scores", sep=" "), xlab="", ylab="", cex.main=2, cex.axis=2, xlim=c(0,180))
plot(e.y, add=TRUE, col="Red")
lines(c(uMax, uMax), c(0,1), col="Gray")
lines(c(uMax, uMax), c(e.x(uMax), e.y(uMax)), lwd=2)
title(xlab="Correlation Score", ylab = "Probability", line =4, cex.lab=2)
dev.off()}
fname <- paste(Gun_Num[k], Sample_Level[i], ”Firing Pin One”, “Sided KS Test Min ECDF.png”, sep=” ”)

png(fname, width=1200, height=800)

par(mai=c(1.2, 1.3, 1.2, 0.42)) #bottom, left, top and right

plot(e.x, col=”Blue”, main=paste(Gun_Num[k], ”Firing Pin One”, “Sided KS Test Min ECDF with”, Sample_Level[i], ”Scores”, “Sampled”, sep=” ”), xlab=”, ylab=”, cex.main=2, cex.axis=2, ylim=c(0,180))

plot(e.y, add=TRUE, col=”Red”)

lines(c(uMin, uMin), c(0,1), col=”Gray”)

lines(c(uMin, uMin), c(e.x(uMin), e.y(uMin)), lwd=2)

title(xlab=”Correlation Score”, ylab=”Probability”, line=4, cex.lab=2)

dev.off()

Outer <- c()

Seeder <- Seeder+2

}

colnames(Output_BF) <- c(”Sampling”, ”Unique CC”, ”Iteration”, ”KS_TwoSided Test statistic”, ”KS_OneSidedMax Test statistic”, ”KS_OneSidedMin Test statistic”, ”KS_TwoSided Critical value”, ”KS_OneSided Critical value”)

colnames(Output_FP) <- c(”Sampling”, ”Unique CC”, ”Iteration”, ”KS_TwoSided Test statistic”, ”KS_OneSidedMax Test statistic”, ”KS_OneSidedMin Test statistic”, ”KS_TwoSided Critical value”, ”KS_OneSided Critical value”)

setwd(”C:/Users/Eric/OneDrive/Law_Thesis_Research/KS_TOST”)

write.csv(Output_BF, file=paste(Gun_Num[k], ”BF Sampling Levels”, “KS TOST.csv”, sep=” ”), row.names=FALSE)

write.csv(Output_FP, file=paste(Gun_Num[k], ”FP Sampling Levels”, “KS TOST.csv”, sep=” ”), row.names=FALSE)
6.3 Equivalence Method

```r
library("npsm", lib.loc="~/R/win-library/3.3")
library("inline", lib.loc="~/R/win-library/3.3")
library("Rcpp", lib.loc="~/R/win-library/3.3")

rtools <- "C:\\Rtools\\bin"
gcc <- "C:\\Rtools\\gcc-4.6.3\\bin"
path <- strsplit(Sys.getenv("PATH"), ";")[[1]]
new_path <- c(rtools, gcc, path)
new_path <- new_path[!duplicated(tolower(new_path))]
Sys.setenv(PATH = paste(new_path, collapse = ";"))

# file.dir <- "C:/Users/Eric/OneDrive/Law_Thesis_Research/Data_Files/
# 🄵️/CleanedFiles"
save.dir <- "C:/Users/Eric/OneDrive/Law_Thesis_Research/Firearm昶/
# 🄵️/Equivalence_Test_Results_CC/New_Unique_Comparison_Data"

#Firearms <- read.csv("C:/Users/Eric/OneDrive/Law_Thesis_Research/
# 🄵️/Data_Files/CleanedFiles/FirearmsCombined.csv")
#load("C:/Users/Eric/OneDrive/Law_Thesis_Research/CleanedFiles2/
# 🄵️/FirearmsCombined.RData")
#save(Firearms, file="C:/Users/Eric/OneDrive/Law_Thesis_Research/
# 🄵️/CleanedFiles2/FirearmsCombinedNew.RData")

Con <- 1.36 # used to calculate KS critical value — 1.36 for 95%
# confidence, 1.63 for 99% confidence
Iter <- 1000
Seeder <- 3
Sample_Level <- c(10, 15, 20, 25, 30)
```
OutputCC_new <- c()
OutputCC <- c()
Output_BF <- c()
Output_FP <- c()

setwd(file.dir)
Gun_Num <- c("TAN_X45401", "CAN_X66727", "FXN_X77862")

# All 9mm Serial Numbers Below

#"AAN_XXX724", "AR98_XXX724", "CAN_X66727", "CBN_X97569", "CCN_X66727
✏", "CEN_X97570", "CFN_X97568", "CGN_X97571", "CPN_X66727", "CXN_
✏ X66727", "CZN_X66727",
#"FAN_X17849", "FDN_X77862", "FEN_X17802", "FJN_X77862", "FMN_X77862
✏", "FVN_X17841", "FXN_X77862", "GAN_XLB713", "GGN_XLB713", "GN
✏ XLB713", "GSN_XLB713",
#"HBN_XX9554", "HCN_X80728", "HFN_XX9554", "HICA_XX9554", "HIPO_X80728
✏", "HJN_X80728", "HKN_X55429", "HSN_XX9554", "HTN_XX9554", "HVN_
✏ X55420", "HWN_X55457",
#"KTC_XSBP59", "KTC_SUB_XEF603", "RAN_X43521", "RBN_X32446", "RCN_
✏ X33654", "RFN_X44279", "RRN_X44279", "RPN_X44279", "RUG9_X69363
✏", "RVN_X44279",
#"SCCY_X66727", "SFXD_XX77862", "SIG9_X82066", "SWS9_XA9892", "TAN_
✏ X45401", "TGN_X55720", "TNN_X20246", "TPN_X55720", "TRM9_X54042
✏", "TRT9_X55720",
#"TTN_X55720", "TVN_X45399", "TWN_X45405", "TXN_X45398", "UPN_XX5056
✏", "WAN_XEF603", "WBN_XSHQ79", "WCN_XSJN79", "WDN_XSJPO8", "WFN_
✏ XSBP59", "WPN_XEF603",
#"WNN_XEF603", "WNN_XAZV54", "WNN_XAZV54", "WXN_XSHQ08"

# Fortran code for first loop
loop1code <- 
integer(1:124)
real*8 tmp
do i = 1, n
  do j1 = 1, n - 1
    do j2 = j1 + 1, n
      tmp = x(i) - max(y(j1), y(j2))
      if (tmp > 0.) pihxyy = pihxyy + 1
    end do
  end do
end do

# Compile the code and turn loop into a function
loop1fun <- cfunction(sig = signature(x="numeric", y="numeric",
                                     pihxyy="integer", m="integer", n="integer"),
                        dim=c("(m)", "(n)"", "", "", ""), loop1code, language="F95")

# Fortran code for second loop
loop2code <- "
  integer i1, i2
  real*8 tmp
  do i1 = 1, m-1
    do i2 = i1 + 1, m
      do j = 1, n
        tmp = min(x(i1), x(i2)) - y(j)
        if (tmp > 0.) pihxyy = pihxyy + 1
      end do
    end do
  end do
"}

# Compile the code and turn loop into a function
loop2fun <- cfunction(sig = signature(x="numeric", y="numeric",
                                      pihxyy="integer", m="integer", n="integer"),
                        dim=c("(m)", "(n)"", "", "", ""), loop2code, language="F95")

loop3fun <- function(x, y) sum(outer(x, y, function(x, y) trunc(0.5*(
                          sign(x-y)+1))))

# Function for Two Sided KS Test
compareTwo <- function(x1, y1) {
  n1 <- length(x1); m1 <- length(y1)
  w1 <- c(x1, y1)
  o1 <- order(w1)
  z1 <- cumsum(ifelse(o1 <= n1, m1, -n1))
  i1 <- which.max(abs(z1))
  w1[o1[i1]]
}

## Start with the code of fk.test(), declared as a function now
  ↪ named fk.int()

fk.ind <- function(x, y, alternative = c("two.sided", "less", "
  ↪ greater"), conf.level = 0.95){
  fkscores = new("scores", phi = function(u) {
      (qnorm((u + 1)/2))^2 - 1
  }, Dphi = function(u) {
      qnorm((u + 1)/2)/dnorm(qnorm((u + 1)/2))
  })
  zed = c(abs(x - median(x)), abs(y - median(y)))
  n1 = length(x)
  n2 = length(y)
  n = n1 + n2
  cvec = c(rep(0, n1), rep(1, n2))
  rz = rank(zed)/(n + 1)
  v = getScores(fkscores, rz)
  num = sum(cvec * v)
  sigphi = sqrt(((n1 * n2)/n) * var(v))
  ts = num/sigphi
  see = alternative[1]
  if (see == "two.sided") {
    pval = 2 * (1 - pnorm(abs(ts)))
  }
  if (see == "greater") {
    pval = 1 - pnorm(ts)
  }
  if (see == "less") {

pval = pnorm(ts)
}
x = x - median(x)
x = x[x != 0]
x = log(x)
y = y - median(y)
y = y[y != 0]
y = log(y)
ns = length(x)
ns = ns + n2s

f = abs(qt(((1 - conf.level)/2), ns - 2))
c = c(rep(0, n1s), rep(1, n2s))
zed = c(x, y)
fitz = rfit(zed ~ c, scores = f)

tauhat <- fitz$t
n <- length(fitz$y)
ppl <- fitz$qrx1$rank
est <- fitz$coef
ses <- sqrt(diag(vcov(fitz)))
tstat <- est/ves
pval <- 2 * pt(abs(tstat), n - pp1)
coef <- cbind(est, ses, tstat, pval)

# dt <- drop.test(fitz)
# R2 <- (dt$df1/dt$df2 * dt$F)/(1 + dt$df1/dt$df2 * dt$F)
# ans <- list(coefficients = coef, dropstat = dt$F, droppval = dt

sumf <- list(coefficients = coef)
# sumf$call <- fitz$call
class(sumf) <- "summary.rfit"
# sumf

############################################################

delta = coef(sumf)[2, 1]
se = coef(sumf)[2, 2]
lb = delta - tc * se
ub = delta + tc * se
eta = exp(delta)

eta = exp(delta)

ci = c(exp(lb), exp(ub))
res <- list(estimate = eta, conf.int = ci, statistic = ts,
    p.value = pval, conf.level = conf.level)
res$call <- match.call()
class(res) <- list("rank.test")
res

f3 <- function(x, y){

    alpha <- 0.05
    m <- length(x)
    n <- length(y)
    eps1_ <- 0.2 # 0.1382 default
eps2_ <- 0.2  #0.2602 default

eqctr <- 0.5 + (eps2_ - eps1_) / 2
eqleng <- eps1_ + eps2_

wxy <- 0
pihxxxy <- 0
pihxyy <- 0

wxy <- loop3fun(x, y)  #uses function loop3fun defined above
                   \rightarrow instead of for loop

# Call new function from compiled code instead
pihxyy <- loop1fun(x, y, pihxxxy, m, n)${\text{pihxyy}}$

# Call new compiled function for second loop
pihxxxy <- loop2fun(x, y, pihxxxy, m, n)${\text{pihxxxy}}$

wxy <- wxy / (m*n)
pihxxxy <- abs(pihxxxy^2 / (m*(m-1)*n))
pihxyy <- abs(pihxyy^2 / (n*(n-1)*m))
sigmah <- sqrt(abs((wxy-(m+n-1)*wxy^2+(m-1)*pihxxxy+(n-1)*pihxyy
                  \rightarrow ) / (m*n))))

crit <- sqrt(qchisq(alpha, 1, (eqleng/2/sigmah)**2))

if (abs((wxy-eqctr)/sigmah) >= crit) rej <- 1
if (abs((wxy-eqctr)/sigmah) < crit) rej <- 0

if (is.na(sigmah) || is.na(crit)) rej <- 1

MWD_Decision <- rej

#
```r
for(k in 1:length(Gun_Num)){
  setwd(file.dir)
  datafile <- paste(as.character(Gun_Num[k]), " .csv", sep=" ")
  Firearm <- read.csv(datafile)
  Gun_MDistribution <- Firearm[which(Firearm$Match=='Yes'),]
  Gun_MDistribution <- subset(Gun_MDistribution, select=c(CaseID_sample, ExhibitNumber_Sample, CaseID_DB, ExhibitNumber_DB, BF, FP), row.names=FALSE)

  for(i in 1:length(Sample_Level)){
    for(j in 1:Iter){
      set.seed(Seeder)
      tester <- Gun_MDistribution
      testerCC <- tester
      SampledCC <- sample(unique(tester$CaseID_Sample), Sample_Leve[l[i]], replace=FALSE)

      for(b in 1:length(SampledCC)){
        out <- grep(SampledCC[b], tester$CaseID_Sample)
        tester <- tester[-out,]
        testerCC <- testerCC[out,]
        OutputCC <- rbind(OutputCC, testerCC)  #sampled CCs
      }
    }
  }
}
```
```r
#delete from OutputCC everything that does not inlcude CaseID_Sample in CaseID_DB column
test_var <- unique(OutputCC$ExhibitNumber_Sample)
for(d in 1:length(test_var)){
  out_new <- OutputCC[which(OutputCC$ExhibitNumber_DB == test_var[d]),]
  OutputCC_new <- rbind(OutputCC_new, out_new)
}

BF_Match <- tester$BF
BF_Subset <- OutputCC_new$BF

FP_Match <- tester$FP
FP_Subset <- OutputCC_new$FP

OutputCC <- c()
OutputCC_new <- c()

# Mann–Whitney Test for Equivalence
MW_DecisionBF <- f3(BF_Subset, BF_Match)

# Fligner Killeen for Dispersion Equivalence
```
zBF <- fk.ind(BF_Subset, BF_Match, alternative = c("two.
   ↩ sided"), conf.level = 0.95)
zBF.conf <- zBF$conf.int

FK_DecisionBF <- as.numeric(ifelse(zBF.conf[1] > .66 & zBF.
   conf[2] < 1.5, '0', '1'))

# KS Two Sample Test
Con <- 1.36 # used to calculate KS critical value — 1.36 for
   ↩ 95% confidence, 1.63 for 99% confidence
KS_Critical.value <- Con*sqrt(((length(BF_Subset)+length(BF.
   Match))/(length(BF_Subset)*length(BF_Match)))) # two
   ↩ sided critical value
u <- compareTwo(BF_Subset, BF_Match) # use two sided compare
   ↩ function
e.x <- ecdf(BF_Subset) # Calculate empirical CDF of subset
   ↩ e.y <- ecdf(BF_Match) # Calculate empirical CDF of match
   ↩ scores
KS_Test.statistic <- abs(e.x(u) - e.y(u)) # provides two
   ↩ sided KS test statistic

KS_DecisionBF <- ifelse(KS_Test.statistic > KS_Critical.
   ↩ value, '1', '0') # Decision — 0 means no significant
   ↩ difference

ResultsBF <- cbind(as.numeric(KS_DecisionBF), as.numeric(MW.
   ↩ DecisionBF), as.numeric(FK_DecisionBF))
SumBF <- rowSums(ResultsBF)

Outer <- c(Sample_Level[i], j, length(BF_Match), length(BF.
   Subset), KS_DecisionBF, MW_DecisionBF, FK_DecisionBF,
   SumBF)
cat(Gun.Num[k], Sample_Level[i], j, SumBF, "\n")
Output_BF <- rbind(Output_BF, Outer)
Outer <- c()

#--------------------------------------
#Mann–Whitney Test for Equivalence

MW_DecisionFP <- f3(FP_Subset, FP_Match)

#--------------------------------------
#Fligner Killeen for Dispersion Equivalence

zFP <- fk.ind(FP_Subset, FP_Match, alternative = c("two.
   ↦ sided"), conf.level = 0.95)
zFP.conf <- zFP$conf.int

FK_DecisionFP <- as.numeric(ifelse(zFP.conf[1] > .66 & zFP.
   ↦ conf[2] < 1.5, '0', '1'))

#--------------------------------------
#KS Two Sample Test

Con <- 1.36 #used to calculate KS critical value—–1.36 for
   ↦ 95% confidence, 1.63 for 99% confidence
KS_Critical.value <- Con*sqrt((length(FP_Subset)+length(FP_ 
   ↦ Match))/(length(FP_Subset)*length(FP_Match))) #two
   ↦ sided critical value
u <- compareTwo(FP_Subset, FP_Match) #use two sided compare
   ↦ function
e.x <- ecdf(FP_Subset) #Calculate empirical CDF of subset
e.y <- ecdf(FP_Match) #Calculate empirical CDF of match
   ↦ scores
KS_Test.statistic <- abs(e.x(u) - e.y(u)) #provides two
   ↦ sided KS test statistic

KS_DecisionFP <- ifelse(KS_Test.statistic > KS_Critical.
   ↦ value, '1', '0') #Decision—–0 means no significant
   ↦ difference
ResultsFP <- cbind(as.numeric(KS_DecisionFP), as.numeric(MW_DecisionFP), as.numeric(FK_DecisionFP))

SumFP <- rowSums(ResultsFP)

Outer <- c(Sample_Level[i], j, length(BF_Match), length(BF_Subset), KS_DecisionFP, MW_DecisionFP, FK_DecisionFP, SumFP)
cat(Gun_Num[k], Sample_Level[i], j, SumFP, "\n")

Output_FP <- rbind(Output_FP, Outer)

Outer <- c()

Seeder <- Seeder+3

}


setwd(save.dir)
write.csv(Output_BF, file=paste(Gun_Num[k], Sample_Level[i], "BF Firearm Equivalence Test Results Sampled CC.csv", sep=" "), row.names=FALSE)
write.csv(Output_FP, file=paste(Gun_Num[k], Sample_Level[i], "FP ↦ Firearms Equivalence Test Results Sampled CC.csv", sep=" "), row.names=FALSE)

Output_BF <- c()
Output_FP <- c()
Output_CC <- c()
7. Appendix C: The Bootstrap

All 9 mm correlation data from the IBIS® database was extracted into a Microsoft® Excel® spreadsheet. RStudio was used to subset a given firearm by its serial number and then subset further to include only the cartridge case identifier string, firing pin correlation scores, breech face correlation scores, and whether it was a match or not a match. Two cartridge cases that have been fired by the same firearm (identified by make, model, and serial number) are said to match. Two cartridge cases fired by two different firearms (identified by make, model, and serial number) are said to be a non-match.

A method called the bootstrap was used. The overall idea behind the bootstrap was to treat the original sample as a stand-in for the population [37]. The bootstrap is a statistical method which takes an original sample of values and resamples from it repeatedly with replacement a total of \( n \) times, where \( n \) is the size of the original sample, and calculates the desired statistic for each iteration [37]. Because the bootstrap samples with replacement, some values from the original sample may be selected more than once while others may not be selected at all. In this instance the number of test fires was the sample and the correlation data from the 100 fired cartridge cases in the IBIS® was the population.

The bootstrap package called boot, a library available within R, was used and resampled \( n \) times from the initial distribution of size \( n \) [38][39]. The output was a histogram of the bootstrap distribution alongside a QQ plot to show normality. This package did not allow use of the bootstrap data outside of the package’s built-in functions. To overcome this limitation, the bootstrap was coded in RStudio manually, avoiding the boot package, which allowed full access to all bootstrap data.

The bootstrap was performed at the cartridge case level. Two cartridge cases were sampled from the match distribution of one firearm with all respective correlation scores. The number of iterations of the bootstrap were chosen to be 100, 150, 250, 500, 750, 1000, and 5000. After bootstrapping the sampled cartridge cases density estimates were used to plot the bootstrap data. The Kolmogorov-Smirnov test was used to compare the bootstrap distribution to the firearm match distribution.

As the number of bootstrap iterations was increased, the distribution became more normal as would be expected based on the law of large numbers. In Figure 7.1 on page 138
the breech face and firing pin score bootstrap density estimates are displayed based on 5000 repetitions. QQ plots are next to each of these kernel density estimates to show normality. As the number of iterations increased, the smaller the standard deviation became. Refer to Figures 7.2 to 7.7 on pages 139-144 for additional figures with 100, 150, 250, 500, 750, and 1000 iterations.

When using the Kolmogorov-Smirnov test to compare the bootstrapped sample to the firearm match data, the p-value was extremely small, usually in the magnitude of $10^{-16}$ indicating the Kolmogorov-Smirnov test was finding significant differences. Increasing the number of bootstrap iterations had no real impact on increasing the p-value because it only made the bootstrap distribution more normal, therefore making it less like the actual firearm match distribution. For these reasons only one firearm was used with the bootstrap method, and it was not explored any further.

The bootstrap method was initially tested, but it soon became evident that it was not the correct approach. As the number of iterations increased the bootstrap distribution became closer and closer to modeling a perfect normal distribution and the standard deviation continued to decrease. The kernel density estimates of the firearm match distributions did not model a normal distribution as well as the bootstrap distributions. When using the Kolmogorov-Smirnov test to compare the empirical cumulative distribution functions, tremendously small p-values resulted. This was expected based on visual comparison of the distributions. For example, the bootstrap distributions for breech face and firing pin scores for a Taurus™ 24/7 G2 firearm in Figure 7.1 on the following page are different visually from the match densities for the same firearm in Figure 3.1 on page 28.
Figure 7.1: Bootstrap with 5000 repetitions for a Taurus™ 24/7 G2 (SN X45401) for breech face scores (top) and firing pin scores (bottom) with corresponding QQ plots.
Figure 7.2: Bootstrap with 100 repetitions for a Taurus$^{\text{TM}}$ 24/7 G2 (SN X45401) for breech face scores (top) and firing pin scores (bottom) with corresponding QQ plots.
Figure 7.3: Bootstrap with 150 repetitions for a Taurus\textsuperscript{TM} 24/7 G2 (SN X45401) for breech face scores (top) and firing pin scores (bottom) with corresponding QQ plots.
Figure 7.4: Bootstrap with 250 repetitions for a Taurus\textsuperscript{TM} 24/7 G2 (SN X45401) for breech face scores (top) and firing pin scores (bottom) with corresponding QQ plots.
Figure 7.5: Bootstrap with 500 repetitions for a Taurus\textsuperscript{TM} 24/7 G2 (SN X45401) for breech face scores (top) and firing pin scores (bottom) with corresponding QQ plots.
Figure 7.6: Bootstrap with 750 repetitions for a Taurus\textsuperscript{TM} 24/7 G2 (SN X45401) for breech face scores (top) and firing pin scores (bottom) with corresponding QQ plots.
Figure 7.7: Bootstrap with 1000 repetitions for a Taurus\textsuperscript{TM} 24/7 G2 (SN X45401) for breech face scores (top) and firing pin scores (bottom) with corresponding QQ plots.
References


