

Statistical Treatment of Class Evidence: Trace Element Concentrations in Bullet Lead

Alicia L. Carriquiry, Michael Daniels, Hal S. Stern
Department of Statistics, Iowa State University and Ames Laboratory

May 4, 2000 (Table 4 corrected April 19, 2002)

1 Introduction

Forensic evidence plays an important role in the final courtroom decision concerning the guilt or innocence of a suspected criminal. The question that forensic examiners try to answer is whether two items, one found at the crime scene and one found on the suspect, have a common origin. They may also try to quantify the quality of the evidence, e.g., by specifying the probability that two items might appear to match by coincidence. Examples of the kinds of evidence that may be considered include traces of DNA, traces of blood, glass fragments, and the subject of the work reported here, trace element concentrations in bullet lead.

The application of modern scientific methods to determine whether traces of DNA found at the crime scene in the form of hair or bodily fluids match the DNA of a suspect is now commonplace, as is the use of probability and statistics to quantify the strength of the evidence. There is now a push to extend such approaches to other types of evidence such as synthetic fibers, glass, and bullets. These classes of evidence introduce complications not present with biological specimens; the population distribution of biological specimens, such as DNA or blood type, are well known and largely invariant over time whereas for other types of evidence, such as bullets, the population distribution is largely unknown. To determine such population (or reference) distributions for trace element concentrations in bullet lead, which will be our focus, or other non-biological evidence types, one must try to exploit features specific to the manufacturing process of the product.

We consider two approaches to the problem of determining if bullet fragments found at a crime scene and bullets found with a suspect appear to have a common origin and assessing the significance of such evidence. The first approach, which we refer to as the significance test/coincidence probability approach, consists of two steps: first, determining whether the two items being compared are indistinguishable or ‘match’ via a statistical significance test, and second, determining the importance that can be assigned to the declared match. The second step typically involves data collection and data analysis to determine the probability that a match would occur by chance or coincidence. Many statistical procedures have

been developed for the first step. Most, like the common t-test are known as parametric approaches because they focus on specific parameters (the mean) and are derived under assumptions about the distribution of measurements. In this work, we introduce a non-parametric or empirical alternative. One disadvantage of the significance test approach is its reliance on a binary decision rule at the first step. Although it can be appropriate for discrete or categorical data (like A/B/O/AB blood type) to say that two items match, it is more natural with continuous data (like trace element concentrations) to assess the degree to which two objects match rather than attempt a single match/not-match dichotomy.

The second approach to assessing the quality of trace evidence uses the likelihood ratio to provide a quantitative measure of the quality of the evidence. The likelihood ratio approach can be motivated by Bayes' Theorem. The odds form of Bayes' Theorem holds that the posterior odds for an event (in this case, that the evidence resulted from contact between the suspect and the crime scene) are the product of the prior odds and the likelihood ratio. The likelihood ratio compares the chances of obtaining measurements like those obtained for the bullet fragment at the crime scene and for the suspect's bullets under the two competing hypotheses (guilty versus not guilty or contact versus no contact). From this perspective the likelihood ratio supplies a one-number summary of the degree of match and the significance thereof. The main disadvantage of this approach is that a large number of assumptions are often required to estimate a likelihood ratio. We will develop the likelihood ratio approach for bullet fragment evidence.

The results of our analysis suggest some difficulty in reliably measuring the quality of bullet lead evidence. The likelihood ratio approach is developed, but only for a special case that is likely to be rare. There are some computational difficulties in extending the likelihood ratio approach to more realistic scenarios. The empirical test that is developed appears to have good properties. There is however still no reliable measure of the probability of a coincidental match for the test. Our results serve primarily to highlight the importance of the manufacturing process in assessing bullet evidence. The data made available to us have been collected after the manufacturing process is complete, from bullets purchased at stores or found during the course of investigations. Our results clearly demonstrate that there would be a benefit to data collected from the manufacturer prior to the packaging of bullets into boxes.

The next section provides some details on the bullet manufacturing process, introducing ideas and terminology that motivate the analyses that follow. In Section 3 we describe two data sets compiled by the Federal Bureau of Investigation (FBI) which we have used in developing and testing our methods. An informal preliminary analysis of the bullet data is presented in Section 4. Some of the things learned there helped guide subsequent work on the likelihood ratio approach (Section 5) and the empirical testing approach (Section 6). For the most part we attempt to provide general descriptions of the work. A summary and some final comments are given in Section 7. Technical details that support the general descriptions are provided in Section 8.

2 Bullet Manufacturing Process

At the present time there are four major U.S. manufacturers of bullets: Cascade Cartridge, Federal, Remington, and Winchester. Though there is some variation in the manufacturing process across companies, the basic process is fairly consistent. It was described for us by Charles Peter, Scientist at the FBI laboratory in Washington, DC.

Bullets are produced from lead alloy obtained from local smelters. Manufacturers set specifications or guidelines for the alloy, e.g., a target range for the concentration of a particular element. One potential difficulty is that these need not be fixed over time. At the factory, the lead alloy from the smelter is melted down and mixed in large vats. Bullets are produced from the raw material in these vats and then subsequently packaged in boxes with each box containing fifty bullets. Approximately 300,000 bullets may be produced from the raw material in a single vat. The storage of bullets within the manufacturing plant and the packaging process are such that bullets from several different vats can end up in the same box, though the degree to which this happens varies from manufacturer to manufacturer.

We expect that bullets produced from the same vat of raw material are more alike than bullets produced from different vats. This is a key observation that guides the attempt to determine whether two bullets were manufactured at the same time. One difficulty in the analyses that follow is the absence of controlled data collection from the manufacturers that would allow this assumption to be examined more carefully. If samples of bullets known to have been produced at the same time could be obtained, it would be possible to quantify the variation in trace element concentrations among bullets from the same vat, and for bullets from different vats.

3 Data

Building sensible methods for assessing bullet evidence requires data. Two distinct data sources are used in the remainder of this report. The primary data source on which we have relied is an FBI laboratory study carried out by E. R. Peele, D. G. Havekost, C. A. Peters, J. P. Riley, R. C. Halberstam, and R. D. Koons (the results of which were published in the Proceedings of the 1991 International Symposium on the Forensic Aspects of Trace Evidence). Four full boxes (50 bullets each) of .38 caliber cartridges loaded with 158 grain, round nose bullets were obtained from each the four major U. S. manufacturers. Two of the boxes in each group were packaged on the same date. The lead tip of each bullet was quartered and concentrations of five trace elements were measured on three of the quarters: copper, arsenic, bismuth, silver, and antimony.

A second data set made available to us is the existing FBI bullet database containing trace element concentrations for over 13,000 bullets collected during the course of investigations and other lab studies over the past ten years. The set includes bullets from U.S. and international manufacturers. It also includes a variety of calibers.

The lab data set was provided first and used in a variety of analyses. The larger database plays only a small role in the analyses that follow. The larger database is discussed further in Section 4.4 below.

4 Preliminary Analysis

We begin by examining the data from the 800 bullet FBI laboratory study to explore the nature of trace element data. The exploratory analysis also focuses on the possibility of identifying groups of bullets that originated from the same vat or raw material during the manufacturing process.

4.1 Summary statistics

Table 1 gives the mean concentration of each of five trace elements in the 200 bullets from each manufacturer. This table provides support for the notion that trace element concentrations can be used to assess bullet evidence. Bullets manufactured by Cascade and Federal have extremely high average concentrations of Antimony whereas Remington and Winchester bullets are quite low on that element. Figure 1 contains a number of scatterplots showing the average measurement (over the three measurements) for each bullet on two of the elements; bullets from different manufacturers are identified by different symbols. Figure 1 suggests that having measurements on just two elements is often enough to identify the manufacturer of a bullet, at least within the context of this study.

Manufacturer	Sb	Cu	As	Bi	Ag
Cascade	26836	262	233	128	37.6
Federal	27437	278	1381	16	65.5
Remington	7289	400	105	169	37.0
Winchester	4605	238	36	115	40.3

Table 1: Mean concentration of each of five trace elements for 200 bullets from each of four manufacturers.

One final issue that can be examined with these data concerns the measurement variability. The three measurements of each single bullet provide data for estimating the standard deviation of the measurement process. Table 2 presents these standard deviations by manufacturer for each trace element. There is considerable variability in the standard deviations – some are quite large. Table 3 repeats the standard deviation calculations after taking the logarithm of each measurement. Note that taking the logarithm of the measurements makes

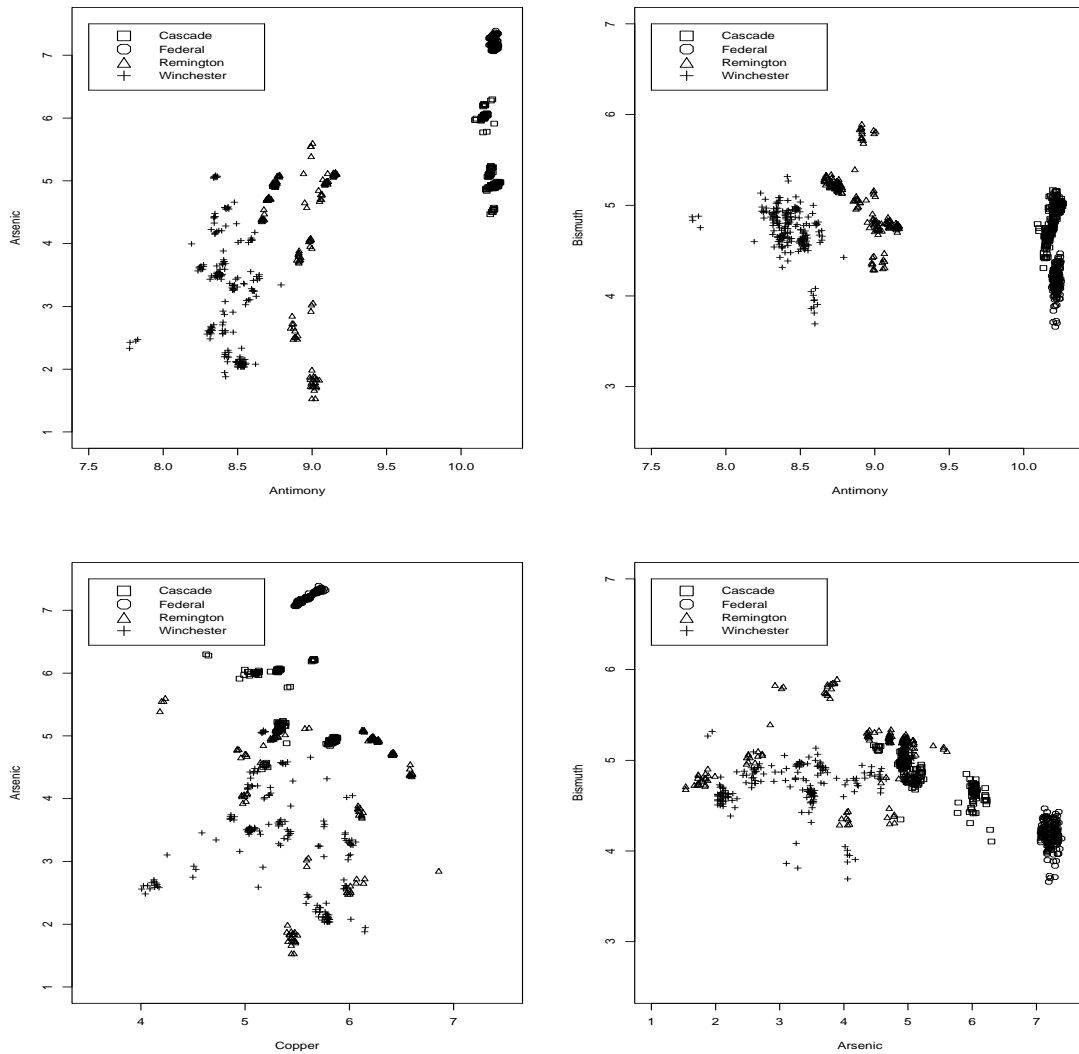


Figure 1: Scatterplots showing mean of the three measurements (on logarithmic scale) for each bullet on two elements. Bullets from different manufacturers are plotted with different symbols as indicated in the legend.

the measurement variability more consistent across manufacturers and elements. This is fairly common for physical measurements of this type. For the analyses that follow we use the logarithms of the measurements in our analyses.

Manufacturer	Sb	Cu	As	Bi	Ag
Cascade	427	6.5	5.4	4.8	0.7
Federal	317	5.2	21.5	0.4	4.2
Remington	60	5.3	2.4	5.2	0.6
Winchester	46	4.2	1.5	4.8	0.7

Table 2: Estimated standard deviation of measurement errors for 200 bullets from each of four manufacturers.

Manufacturer	Sb	Cu	As	Bi	Ag
Cascade	0.016	0.025	0.028	0.038	0.019
Federal	0.012	0.019	0.016	0.023	0.067
Remington	0.008	0.016	0.038	0.034	0.018
Winchester	0.010	0.021	0.057	0.044	0.019

Table 3: Estimated standard deviation of measurement errors after taking logarithms of measurements for 200 bullets from each of four manufacturers.

4.2 Differentiating among manufacturers

For this data set the manufacturer of each bullet is known. The problem of building a rule from a training sample to identify the manufacturer of a bullet from a vector of trace element concentration measurements is an example of what is known within the field of statistics as a classification problem. There are a number of statistical approaches to such problems including both formal methods (discriminant analysis, classification trees) and informal methods (graphical techniques). In the present case any of these approaches can easily discriminate amongst the four manufacturers for the 800 bullets in the laboratory study.

The statistical package Splus allows one to build classification trees which classify objects (bullets) based on simple binary decision rules. In this case the classification tree is easy to describe: if a bullet has antimony concentration less than 5713.34 then it is a Winchester bullet, if the antimony concentration is greater than 5713.34 and less than 16965.8 then

it is a Remington bullet, if the antimony concentration is greater than 16965.8 then we must consult a second element, with high levels of arsenic (greater than 856.835) indicating a Federal bullet and low arsenic indicating a Cascade bullet. This can be represented graphically as in Figure 2. Using these simple rules all but one of the 800 bullets is classified correctly. A purely graphical approach can also be used; in that case classifications are made by referring to plots like those in Figure 1.

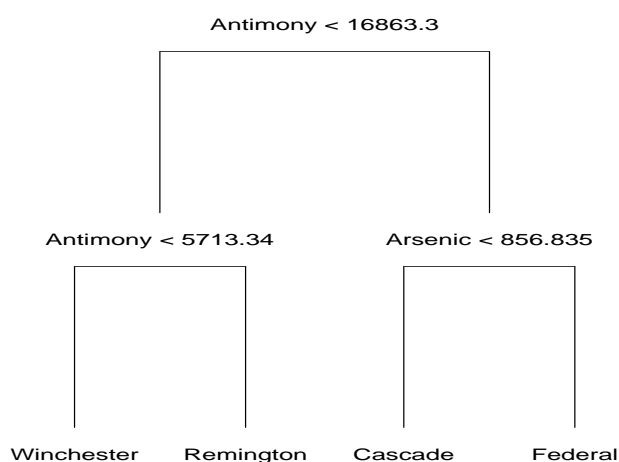


Figure 2: Classification tree for identifying the manufacturer of a bullet.

As this part of the problem is relatively straightforward for the laboratory data, we put it aside for the moment and assume that bullet manufacturer can be easily identified. This seems clearly to be true if we can measure enough elements; for the lab data two elements are generally enough. It should be noted however that automatically applying the classification tree from above to the large FBI bullet database was not successful - for one thing the measurements appear to be recorded in different units or on different scales.

4.3 Groups within manufacturers

Recall that our plan to use quantitative measures to assess bullet evidence is based on the supposition that bullets manufactured from the same raw material (i.e., in the same vat) are likely to have more similar trace element concentration measurements than bullets manufactured from different raw material at the same manufacturer. Identifying groups of

bullets manufactured from the same raw material, these are called compositional groups in the FBI study, is an example of what is known as a clustering problem in statistics because the number (or even existence) of such groups is not known for sure. We explored a large number of approaches to this problem including traditional clustering methods (known as hierarchical clustering) where a measure of the distance between two clusters is defined and then applied repeatedly to merge nearby clusters, graphical methods based on viewing representations of the data in one-, two-, or three-dimensions, and a method designed to mimic the manual clustering algorithm that was used in the FBI's laboratory study. The method that seemed to work best is known as model-based clustering and we next describe that method.

Model-based clustering (as described in an article by J. Banfield and A. Raftery in the journal *Biometrics* in 1993) hypothesizes that each cluster (group) can be modeled via a Gaussian or normal distribution with a specific covariance structure; we note that the distribution of the trace element concentrations is better approximated by a normal distribution once we have taken the logarithm of the measurements as suggested earlier. The normal distribution takes the vectors of trace element concentrations to be centered around a typical value for each cluster. The variability of the measurements for bullets within a cluster can be allowed to vary for each cluster or be restricted to be constant for all clusters from a given manufacturer. A formal measure (based on the normal density function) is proposed by the developers of model-based clustering for estimating the number of groups present in a data set.

We used the model-based clustering to find sets of bullets within the 200 bullets for each manufacturer that appear to have come from a single batch of raw material. More details are provided in the technical appendix of Section 8.1. Tables 4-7 gives the number of clusters found for each manufacturer, the size of the clusters, and identifies the boxes from which the bullets in each cluster came.

GROUP	# OF BULLETS	SOURCE OF BULLETS
1	22	22 from box 1
2	9	9 from box 1
3	19	19 from box 1
4	87	44 from box 2, 43 from box 3
5	9	5 from box 2, 4 from box 3
6	5	1 from box 2, 3 from box 3, 1 from box 4
7	40	40 from box 4
8	9	9 from box 4

Table 4: Model-based clustering results for Cascade.

GROUP	# OF BULLETS	SOURCE OF BULLETS
1	49	18 from box 1, 13 from box 2, 18 from box 4
2	95	32 from box 1, 37 from box 2, 26 from box 4
3	6	6 from box 4
4	50	50 from box 3

Table 5: Model-based clustering results for Federal.

GROUP	# OF BULLETS	SOURCE OF BULLETS
1	63	30 from box 1, 33 from box 2
2	19	10 from box 1, 9 from box 2
3	10	4 from box 1, 1 from box 2, 5 from box 3
4	13	6 from box 1, 7 from box 2
5	31	31 from box 3
6	13	13 from box 3
7	11	1 from box 3, 10 from box 4
8	22	22 from box 4
9	18	18 from box 4

Table 6: Model-based clustering results for Remington.

GROUP	# OF BULLETS	SOURCE OF BULLETS
1	1	1 from box 1
2	25	25 from box 1
3	1	1 from box 1
4	1	1 from box 1
5	7	7 from box 1
6	1	1 from box 1
7	1	1 from box 1
8	2	2 from box 1
9	2	2 from box 1
10	4	4 from box 1
11	1	1 from box 1
12	1	1 from box 1
13	2	2 from box 1
14	1	1 from box 1
15	13	9 from box 2, 4 from box 4
16	8	4 from box 2, 4 from box 4
17	14	8 from box 2, 6 from box 4
18	13	10 from box 2, 3 from box 4
19	8	4 from box 2, 4 from box 4
20	1	1 from box 2
21	7	5 from box 2, 2 from box 4
22	2	2 from box 2
23	1	1 from box 2
24	3	1 from box 2, 2 from box 4
25	1	1 from box 2
26	9	3 from box 2, 6 from box 4
27	1	1 from box 2
28	37	13 from box 4, 24 from box 5
29	8	2 from box 4, 6 from box 5
30	1	1 from box 4
31	1	1 from box 4
32	1	1 from box 4
33	8	8 from box 5
34	6	6 from box 5
35	3	3 from box 5
36	4	1 from box 4, 3 from box 5

Table 7: Model-based clustering results for Winchester.

4.4 The FBI database

We relied primarily on the 800-bullet FBI lab data set to develop our methods. After developing the methods we examined the large FBI database in some detail. Several things made it difficult to extrapolate the things we had learned from the lab data to the larger database. First, the database includes data from many manufacturers; for some only a limited amount of data is available. Second, the measurements of the elements were much less consistent over time than was expected based on the lab study. In the lab study the four boxes with packaging dates varying over more than a decade yielded good evidence that a manufacturer’s specifications would be fairly consistent. The larger database sheds doubt on whether this is true. Bullets were entered into the database over a long time period and may represent bullets and may represent bullets manufactured over an even longer time period. There are two consequences of our difficulty working with the larger database. First, we have not used these data often in developing the methods that follow. Second, if there are not in fact consistent manufacturer patterns over long periods of time, then periodic samples from the manufacturers would be required to make the various methods for assessing evidence practical.

5 The Likelihood Ratio Approach

The likelihood ratio approach to assessing the evidence from a crime scene and a suspect provides a single quantitative measure of the probative value of the evidence. In this section we review the concept of the likelihood ratio and apply it to the case of trace element data from bullet evidence. The main results are presented in this section without technical details; the details are provided in the technical appendix (Section 8).

5.1 Introduction

To begin, let G denote the hypothesis that the bullet fragment (or fragments) from the crime scene and the bullets found with the suspect have a common source (i.e., come from a single source box (or boxes)). We abbreviate this as G for “Guilty” recognizing of course that a common source does not necessarily imply the suspect is guilty. It is common to take \bar{G} to be the complementary or opposite event, that the bullet fragments and the bullets do not have a common source. Take E to represent all of the bullet evidence. Formally E could contain evidence of many types but we restrict attention to trace element measures from the bullets. One reasonable goal of a forensic examination is a measure of the probability of the hypothesis G . Using the probability tool Bayes’ Theorem we find an expression for the probability of guilt based on the evidence E ,

$$\Pr(G|E) = \frac{\Pr(E|G) \Pr(G)}{\Pr(E|G) \Pr(G) + \Pr(E|\bar{G}) \Pr(\bar{G})}.$$

This expression can be written more simply in terms of the odds in favor of the hypothesis G given the evidence E ,

$$\frac{\Pr(G|E)}{\Pr(\bar{G}|E)} = \frac{\Pr(E|G) \Pr(G)}{\Pr(E|\bar{G}) \Pr(\bar{G})}.$$

This last expression gives the odds of the hypothesis G given the evidence E (sometimes known as the posterior odds) as the product of two terms, the odds in favor of the hypothesis G before looking at the evidence, $\Pr(G)/\Pr(\bar{G})$ (sometimes known as the prior odds), and the *likelihood ratio*, $\Pr(E|G)/\Pr(E|\bar{G})$. The numerator of the likelihood ratio measures the probability (or likelihood) of the evidence under the hypothesis G while the denominator measures the probability (or likelihood) of the evidence under the hypothesis \bar{G} .

The two quantities that define the likelihood ratio are related to the two steps of the traditional significance test/coincidence probability approach. The numerator is related to determining whether the two objects match; as with a significance test, the evidence is evaluated under the null hypothesis that the two objects do in fact have a common source (our hypothesis G). The denominator is related to the step in which the significance of the match is assessed by determining the probability of a coincidental match, i.e., a match under the hypothesis \bar{G} . One nice feature of the likelihood ratio is that it provides a single quantitative measure without requiring a binary decision as to whether the two objects match or not. Uncertainty about the match status is factored in to the single measure along with the probability of alternative explanations of the evidence. It turns out however that calculation of the likelihood ratio requires that a number of assumptions be made. Moreover, we show below that even under these assumptions computing the likelihood ratio in the context of trace element data for bullet evidence becomes increasingly difficult as the amount of evidence increases.

5.2 Application to bullet trace element data

We now explain the likelihood ratio in the context of trace element data for bullet lead. We assume that a crime has been committed and that the evidence includes k bullet fragments found at the crime scene and m bullets found with a suspect. Often k is small; we start by trying to find a likelihood ratio for the case with $k = 1$. The fragments and bullets are assumed to be analyzed yielding a set of trace element concentrations (perhaps several measurements of the trace element concentrations for each). The number of trace elements can vary depending on the bullet manufacturer and perhaps other factors. Formally then, the evidence E includes all of the available trace element measurements for the $k + m$ bullets and fragments.

The hypothesis G is that the fragments originate from the same box (or even set of boxes) as the bullets found on the suspect. The hypothesis \bar{G} is that the fragments do not originate from the same box as the bullets found on the suspect. One complication with bullet data is a consequence of the manufacturing process. It is assumed that bullets manufactured at the same time (from the same vat of raw material) share relatively similar trace element concentrations. It is known that bullets manufactured at the same time can

end up in different boxes. Thus it is possible that under the hypothesis G , which specifies that the fragments come from the same box as the suspect's bullets, the fragments will not have trace element concentrations that match those of the bullets in the box. This differs from other forms of evidence where under G we can be reasonably certain (barring contamination) of a close match (DNA, blood type). Of course, it is also possible that the fragments' trace element concentrations will match those of the suspect's bullets even though the true hypothesis is \bar{G} (this is the coincidental match we worry about). The relative chance of these two events must be determined as part of the likelihood ratio calculation.

Based on the preliminary analysis reported in Section 4, we take it as given that the manufacturer of the bullet fragment and the manufacturer of the suspect's bullets can be determined without error. If so, then it is reasonable to only work on the case where the fragments and suspect's bullets have the same manufacturer. If they do not, then the likelihood ratio is near zero (nonzero because of the possibility that one manufacturer buys bullet lead from one another).

5.3 The one fragment/one bullet case

To begin, we consider the case with $k = 1$ fragment and $m = 1$ bullet on the suspect. This scenario is not terribly realistic in that the suspect is likely to have more than a single bullet, most likely the unused portion of a box. The scenario is however very useful in explaining how the likelihood ratio is constructed and in pointing out some of the difficulties that occur when there is more evidence. Let x denote the vector of trace element measurements on a bullet fragment found at a crime scene and y denote the vector of trace element measurements on a bullet found with a suspect. The evidence E in this case is the combined data x, y . It is possible to simplify the form of the likelihood ratio in this case because the likelihood or probability associated with just a single measurement, say y , ignoring the other is the same under either hypothesis (in the sense that the likelihood is governed completely by the manufacturing process since there is no other information to be taken into account). Then the likelihood ratio can be expressed as $p(x|y, G)/p(x|y, \bar{G})$, where we use $p(x|y, H)$ denote the probability or likelihood of observing x given that y has been observed and hypothesis H is true. If the evidence in question can only take a limited number of values, as in the case of blood type, then p will represent actual probabilities, but when the evidence is a continuous measurement it is more natural to think of p as measure of likelihood (because it is not formally possible to assign probabilities to the essentially infinite number of possible values).

A difficulty with the form of the likelihood ratio specified above in terms of G and \bar{G} is that, as mentioned earlier, boxes consist of bullets from a number of compositional groups. It is reasonably straightforward to think about the likelihood of x given y if the two lead pieces are hypothesized to come from the same compositional group or from different compositional groups, but harder to think about the likelihood given only the hypothesis G or \bar{G} . Fortunately this additional structure can be taken advantage of by rewriting the

likelihood ratio as

$$\begin{aligned}
 LR &= \frac{p(x|y, G)}{p(x|y, \bar{G})} \\
 &= \frac{p(x|y, \text{same group})p(\text{same group}|G) + p(x|y, \text{diff group})p(\text{diff group}|G)}{p(x|y, \text{same group})p(\text{same group}|\bar{G}) + p(x|y, \text{diff group})p(\text{diff group}|\bar{G})}
 \end{aligned}$$

where “same group” (or “diff group”) refer to the hypothesis that the two fragments come from the same (or different) vat of raw material during the manufacturing process. This last expression involves two different types of terms: terms describing the variation in two bullets’ measurements given that they come from the same or different compositional groups, and terms describing the makeup of bullet boxes in terms of the compositional groups represented. The former have thus far been approximated by normal distributions with bullets from different compositional groups assumed to exhibit more variability than bullets from the same compositional groups. The latter do not depend on the measurement values at all, they need only be estimated once (and updated periodically as the industry changes). Ideally their estimation would be done with a large, carefully designed study. For now, we have used data from the 800 bullet FBI laboratory study to illustrate our approach to estimating these quantities. Our approach to computing the required quantities are described briefly in the next section in the context of an example and then in more detail in the technical appendices of Section 8.

Following the example, we discuss the difficulty in applying this approach to more extensive evidence sets. The main problem is that precise application of the likelihood ratio requires considering all possible arrangements of compositional groups within boxes. The work required to do this is prohibitive once we get beyond a handful of bullets.

5.4 An example

We now apply the likelihood ratio approach for the assessment of evidence for the case with $k = 1$ and $m = 1$. As we have assumed that manufacturer can be easily identified, we restrict attention to the 200 bullets produced by Cascade Cartridge Industries (CCI) analyzed in the FBI study. The data are the measurements of five trace element concentrations (antimony(Sb), copper(Cu), arsenic(As), bismuth(Bi), and silver(Ag)) for each of the 200 bullets obtained from four boxes, including two boxes (box 2 and box 4) with the same packaging date. We have taken the logarithms of the measurements (recall this makes the measurement variation more consistent) and then averaged these to obtain a single measure for each element.

Our basic approach assumes that the trace element concentrations found in a bullet (or bullet fragment) can be thought of as normally distributed with mean equal to the mean trace element concentration for its compositional group and variance described by a “within-group” variance matrix. Further we assume that the compositional group means for a single manufacturer are normally distributed around a common manufacturer average with variance described by a “between-group” variance matrix. In this view of the manufacturing

process there is a typical composition for the manufacturer (perhaps the manufacturer’s specification) but the actual composition of each vat of raw material differs. Then within each vat the concentration of individual bullets varies because the mixing is not perfect. We have experimented with alternatives to the normal distribution, but use it here to keep things simple. Estimates of the variance matrices are required; this is described in the technical appendix. Once estimates are obtained for the variance matrices it is straightforward to evaluate the likelihood terms, $p(x|y, \text{same group})$ and $p(x|y, \text{diff. group})$. In these likelihood terms we take the simplest approach and assume that the bullet measurement y is the best estimate for the manufacturer (or compositional group) mean. Our approach can be modified to use other information about a manufacturer’s manufacturing process.

The remaining terms in the likelihood ratio concern the makeup of boxes of bullets in terms of compositional groups. We have used the model-based clustering approach described in Section 4 to identify putative compositional groups for Cascade (and other manufacturers). There we found 8 compositional groups for Cascade (see Table 4 in Section 4) represented among the four boxes. As expected the two boxes packaged on the same day tend to be populated with bullets from the same compositional groups. Given the makeup of these boxes, e.g., box 1 appeared to contain 3 compositional groups of size 22, 9 and 19, we calculated the probability that two randomly chosen bullets from the same box would come from the same (or different) compositional groups and the probability that two randomly chosen bullets from different boxes would come from the same (or different) compositional groups. The results (more details are provided in the appendix) indicate that for Cascade, the probability that two bullets from the same box are from the same compositional group is .67 and the probability that they are from different groups is .33. For bullets from different boxes, the probability of the same compositional group is .27 and the probability of different groups is .73. Of course these results are sensitive to the 200 bullets used – in particular since 100 of the 200 bullets were packed on the same day our estimate that the probability of a coincidental match of compositional groups is .27 is likely an overestimate. A larger study to refine these estimates would be recommended before computing likelihood ratios for courtroom use.

Now we put the pieces together, first with a single situation involving one pair of bullets and then we summarize the results of a larger study of the one-bullet, one-fragment case. For Cascade, we obtain estimated variance matrices which we denote as $\hat{\Sigma}_{within}$ and $\hat{\Sigma}_{between}$ (exact numerical values are provided in the appendix). Now consider a fragment with trace element concentrations $x = (10.219, 5.195, 4.536, 5.157, 4.065)$ and a bullet with trace element concentrations $y = (10.221, 5.202, 4.571, 5.157, 4.032)$. Then $p(x|y, \text{same group})$ is a multivariate normal distribution evaluated at the given x and y , formally $p(x|y, \text{same group}) = (2\pi)^{-5/2} |\hat{\Sigma}_{within}|^{-1/2} e^{-0.5(x-y)^T \hat{\Sigma}_{within}^{-1} (x-y)} = 172223$. The other likelihood $p(x|y, \text{diffgroup})$ is a multivariate normal with the same formula except that $\hat{\Sigma}_{between} + \hat{\Sigma}_{within}$ is used in place of $\hat{\Sigma}_{within}$, with resulting value 196.4. Then the likeli-

hood ratio is

$$LR = \frac{.67(172223) + .33(196.4)}{.27(172223) + .73(196.4)} = 2.48. \quad (1)$$

Note that the likelihood of obtaining measurements like these is much higher under hypothesis G than under hypothesis \bar{G} though both are possible explanations. In fact x and y in this case were chosen randomly from the same box so that the LR should be large, favoring that hypothesis. The likelihood ratio is greater than one but not very large because we have estimated that there is a probability .27 that two groups will have come from the same vat even if they are in different boxes. It seems that this probability should probably be much lower which would tend to increase the size of the likelihood ratio.

To illustrate the variation that one can expect in computing likelihood ratios. We have computed the likelihood ratio for 50 randomly chosen pairs, where the pairs come from a common box. Those values ranged from 0.45 to 2.48. Four of the observed LR's are less than one suggesting that the evidence favors \bar{G} when in fact the bullets come from the same box. Such values may indicate that the clusters we have formed are not accurate or they may indicate that there is sufficient variation within a vat of raw material that two bullets may look quite different. We also computed the likelihood ratio for 50 randomly chosen pairs, where the pairs do not come from a common group. Here we'd expect small likelihood ratios. All fifty values were approximately 0.45 (which is the smallest possible value given our estimated probabilities). This suggests that bullets from different clusters are very different and can be easily detected as not coming from the same cluster.

5.5 Difficulty in applying the likelihood ratio

Having explained the basic principal of the likelihood ratio in the context of a simple example, it remains only to elaborate on how the calculation changes for more complex scenarios. It is here, however, that difficulties arise. The basic difficulty is that our likelihood ratio approach for the one-bullet, one-fragment case essentially considers each of the possible scenarios for the two bullets: same compositional group and same box, same group and different box, different group and same box, different group and different box. Each possibility contributes to the likelihood ratio calculation. What happens when there are more bullets or fragments?

Extending the likelihood ratio approach to address the case when there is more than one bullet found in the suspect's partial box is difficult because it requires considering all possible configurations of the $m + 1$ bullets (m from the suspect and the fragment). The fragment may come from the same compositional group as all m of the suspect's bullets or just a subset of the suspect's bullets. To illustrate, consider the scenario with $k = 1$ and $m = 2$. Let y_1 and y_2 denote the trace element measurements from the two bullets found with the suspect and, as usual let x denote the measurements from the fragment. There are five possible scenarios: (i) x , y_1 , and y_2 all belong to different compositional groups (different manufacturing batches); (ii) y_1 and y_2 belong to the same homogeneous group which is different than x ; (iii) x and y_1 belong to the same homogeneous group but y_2 is

from a different group; (iv) x and y_2 belong to the same homogeneous group but y_1 is from a different homogeneous group; and (v) x , y_1 , and y_2 belong to the same homogeneous group. Then the numerator and denominator of the likelihood ratio each need to consider the five scenarios along with their probability of occurring under the two hypotheses. Clearly the task is more difficult than it was in the simple case. In addition, the data collection required to reliably estimate the various probabilities increases as well.

Working with additional fragments is also difficult but in that case it is possible to avoid the problem by carrying out a separate analysis for each fragment. This ignores potentially useful information (e.g., whether the two fragments appear to have come from a single compositional group) but should provide reasonable information for courtroom use.

Note that there are other difficulties with the likelihood ratio approach as well. Our likelihood ratio derivation does not account for bullet distribution patterns or bullet usage patterns in assessing the likelihood of two bullets coming from the same vat of raw material. Thus if all of the bullets in a given local area are from the same manufacturer, and were manufactured at approximately the same time, then it will not be surprising to find matching bullets with most any partial box found by investigators in this locality. One can contrast that with a situation in which the fragment and suspect's box found by investigators are unusual for a given town. Such considerations are not insignificant and could vary considerably from case to case.

The technical and logistical difficulties have led us to conclude that the likelihood ratio approach is not feasible at the current time.

6 A Non-parametric Testing Approach

We argued in the previous section that implementation of the likelihood ratio (LR) approach to assess the evidence of from a crime scene and a suspect can be difficult except in the simplest cases. In this section, we consider an alternative to the LR approach that could, in principle, be used even when the number m of bullets found on the suspect is large. As before, we let k denote the number of bullet fragments found at the crime scene.

The idea is simple. Suppose we define a measure of the “distance” between trace element concentrations of two bullets, and suppose further that we have available the distribution of our distance measure among bullet pairs known to come from the same compositional group and for pairs known to come from different compositional groups. If these two distributions are available, then we could use them as a reference to decide whether the measurements for a given pair of bullets indicate that the bullets likely have come from the same compositional group or not. Below we give a more detailed intuitive justification for the approach, propose a metric, and develop a formal test for determining whether the bullets are a match. The performance of the test that we construct is also discussed.

6.1 Empirical distance distributions

Recall that due to the manufacturing process, bullets from a manufacturer can be viewed as coming from relatively homogeneous compositional groups. We hypothesize that the groupings are related to the date on which bullets were manufactured and more specifically to the vat of molten lead that was used at that time. Because we do not have data to test this hypothesis (we do not know, from our sample of 200 bullets from each of four manufacturers, which bullets come from the same vat and which do not) we use the model-based clustering algorithm described in Section 4 to empirically construct compositional groups. We take the clusters to correspond to compositional groups (or vats) in deriving the non-parametric approach to assessing bullet evidence.

If two bullets are manufactured from the same vat of molten lead, then these two bullets are likely to be “close” to each other or have a small distance measure. The squared distance between two bullets i and j is denoted by d_{ij}^2 . It is basically the sum of squared differences between the trace element concentrations standardized by a measure of the variability of the measurements. Details about the distance measure are provided in the appendix (Section 8.6). The next step is to determine the distribution of distances d_{ij}^2 for pairs of bullets known to come from the same group and then a second distribution for pairs of bullets known to come from different groups. Then these distributions can be used to take an observed distance for a pair of evidential bullets and identify which of the two distributions seem more likely to yield such a value.

Consider first all possible pairs of same-group bullets, and for each pair compute the squared statistical distance d_{ij}^2 . The d_{ij}^2 of all pairs in all groups are used to construct a *within-group distance* distribution. Intuitively, it would seem that if the squared distance between a fragment found in a crime scene and a bullet found on the suspect falls somewhere in the middle of the within-group distribution, then we would be inclined to conclude that the fragment and the bullet came from the same vat of molten lead. This would tend to *incriminate* the suspect. If the distance is too large to have come from the within-group distribution, then it would tend not to incriminate the suspect. Whether the within-distance distribution is useful as a probative tool depends also on the *between-group distance* distribution, i.e., the distribution of pairwise squared distances between bullets from different compositional groups. The “between-groups” distances are computed from all possible pairs where the two bullets are in different clusters. If the distance between the bullet fragment and the bullets on the suspect falls somewhere in the middle of the between-group distribution and is larger than would be typical in the within-group distribution, then we would have reason to believe that the fragment and the bullets come from different sources. This would tend to *decrease* the probability of guilt. The approach is likely to produce reliable results when the between-group and within-group distributions of the d_{ij}^2 do not overlap in any significant way, or in other words, when compositional groups are distinct.

Below, we show the within group and the between group distance distributions constructed from Cascade bullet data (Fig. 3). Table 4 of Section 4 lists the eight clusters

that were obtained from the Cascade bullets, and the number of members in each of the clusters. From these numbers, we find that 5042 pairwise distances were used to construct the within-group distance distribution. The number of between group pairs used to construct the distribution of between group distances is 14859. The within-group distance distribution is shown in the top pane. The between-group distance distribution is given in the middle pane, at a different scale. So that the two distributions can be more easily compared, we have overlaid them and present them together in the bottom pane of Fig. 3.

As is clear from the figure, the two distributions are quite distinct. That is, if the fragment and the bullet come from the same group, it would be very unlikely that we would conclude otherwise. Likely values (the middle 95% of values) of the within-group distance for Cascade bullets are between 2.6 and 27.4; for between-group distances, the likely values are between 74.2 and 258.3. Notice that both squared distance distributions are skewed, with a long tail to the right. This is the expected shape of these distributions. Fig. 4 - 6 display the within-group, between-group and overlaid distributions of squared distances for Federal, Remington, and Winchester bullets, respectively. In general, the within-group and between-group distance distributions look similar across manufacturers. The Remington distributions are the ones that overlap the most, and consequently these will prove to be least useful in assessing evidence.

These empirical distributions play a role similar to existing DNA databases. At present, the usefulness of these empirical distributions is limited, as the database from which they were constructed contains only 200 bullets for each manufacturer. As more compositional group data become available, the distributions of within- and between-group squared distances could be updated.

6.2 Developing a test

Suppose that we recover a fragment from a crime scene and m bullets from a suspect. A natural approach is to compute the distance between the fragment and each of the m bullets. We might then look at the smallest such distance and see how it compares to our within-group distance distribution. There is however a subtle problem here because we have considered the closest of m bullets. It is natural to expect a closer match will be found among the m bullets if m is large than if m is small. This is sometimes known as a multiple comparisons problem in statistics; there is a need to adjust for computing m distances and then restricting attention to the smallest. This is not a standard multiple comparison problem however because some of the m distances are calculated with respect to bullets from different compositional groups and hence not really relevant. We propose to adjust for the multiple comparison problem as follows. Determine how many compositional groups are found among the m bullets on the suspect (perhaps using a clustering algorithm). Then if the fragment is closest to a bullet that belongs to a compositional group with $n \leq m$ members; we claim that n is the relevant number of comparisons. We then pose the following question: is the distance from the fragment to the closest bullet in its compositional group

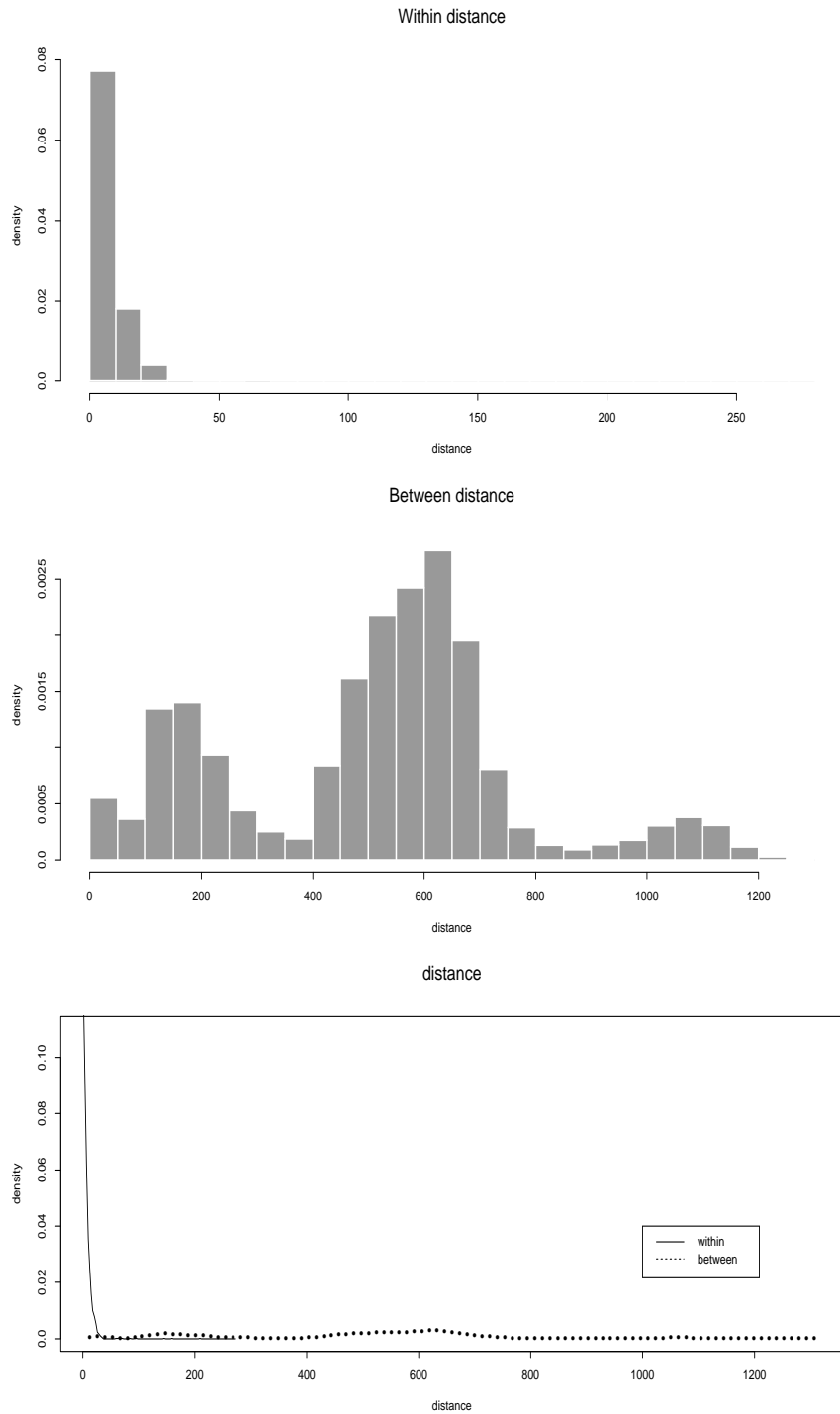


Figure 3: Within-group and between-group distance distributions for Cascade bullets. Final plot shows smooth density estimates of the two distributions in a single graph.

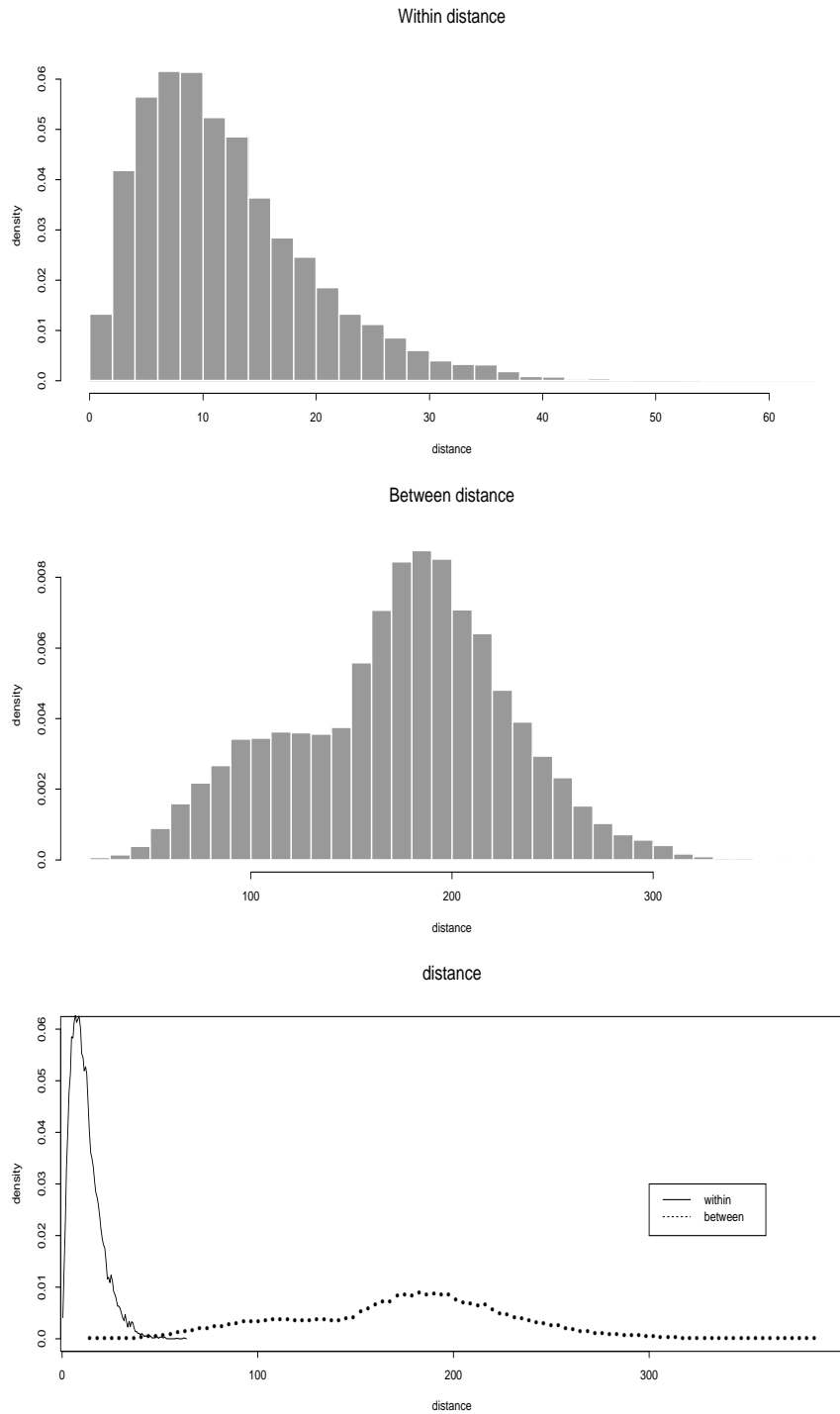


Figure 4: Within-group and between-group distance distributions for Federal bullets. Final plot shows smooth density estimates of the two distributions in a single graph.

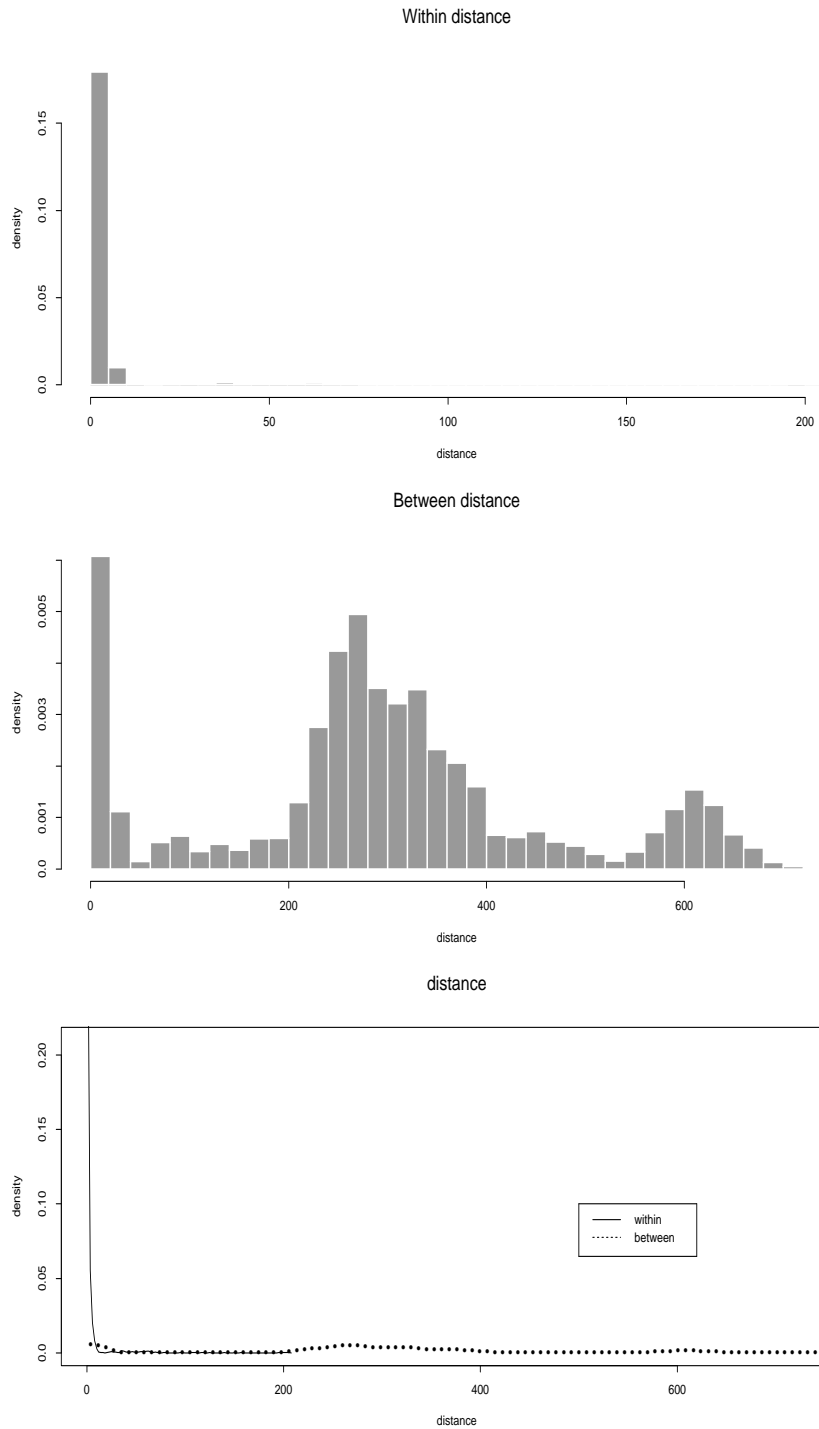


Figure 5: Within-group and between-group distance distributions for Remington bullets. Final plot shows smooth density estimates of the two distributions in a single graph.

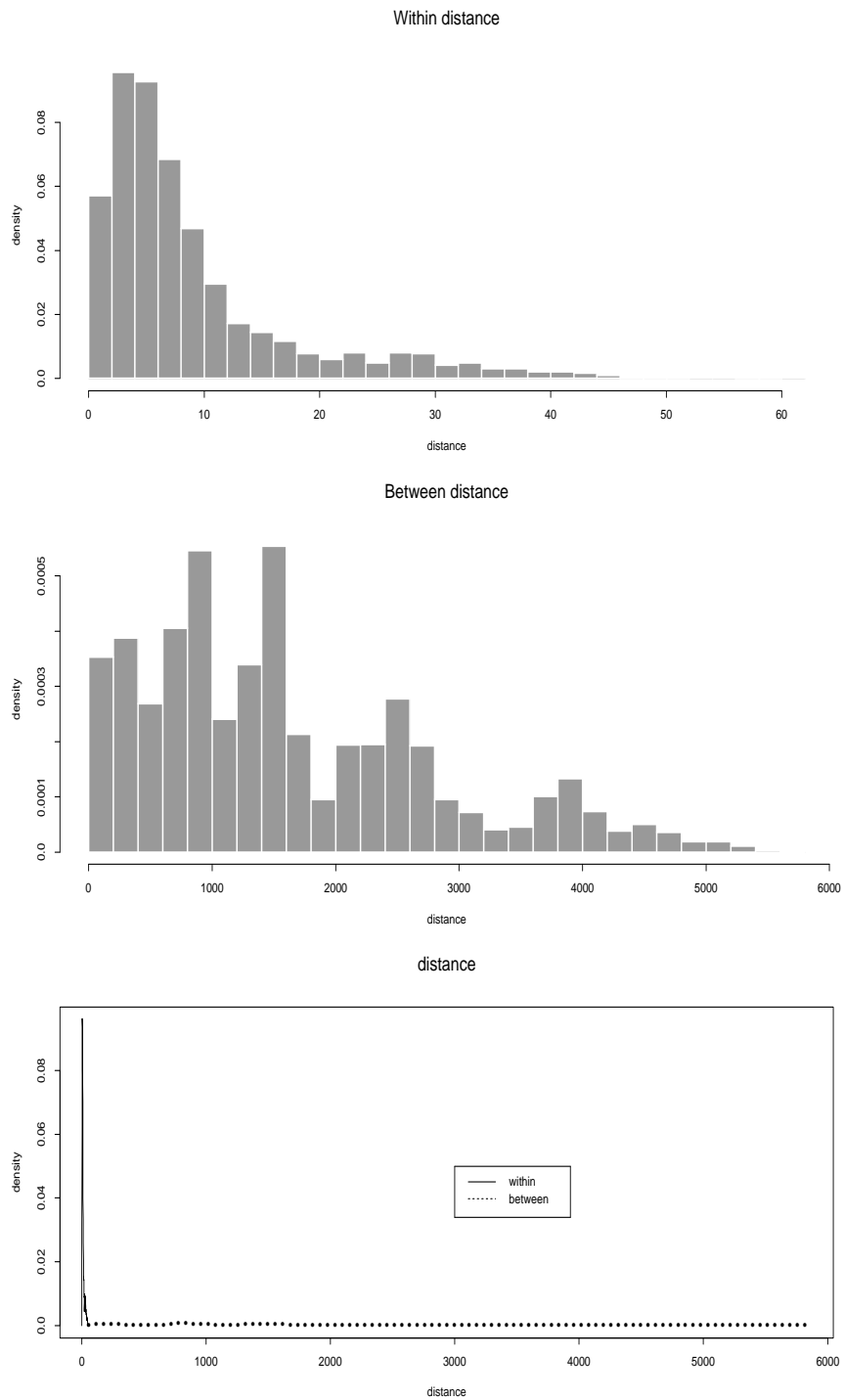


Figure 6: Within-group and between-group distance distributions for Winchester bullets. Final plot shows smooth density estimates of the two distributions in a single graph.

“typical” given the empirical distribution of within-group distances. If so, then we tend to believe that the fragment and (some of the) bullets came from the same batch. Else, we have no evidence to believe that the fragment and the suspect’s bullets were manufactured from the same raw material. The problem is now to devise a decision tool (or a test statistic) to decide what is “typical”. We develop such a tool by simulation.

We compute a reference distribution for our measure (the smallest distance), which we call d^2 , in the situation described above (1 fragment, m bullets on the suspect, fragment is closest to bullet whose compositional group has $n - 1$ other members) as follows. We repeat the steps below a very large number (10,000) of times.

1. Randomly sample n distances from the within-group distribution of squared distances, call these $d_1^2, d_2^2, \dots, d_n^2$.
2. Let $d_{min(n)} = \min\{d_1^2, \dots, d_n^2\}$.

This produces a sample of values for $d_{min(n)}$. Fig. 7 provides an example of this process. The top panel of Fig. 7 shows the within-group distance distribution for Winchester (the same as in the top panel of Fig. 6). The next panel shows the relevant distribution from our simulation for the smallest of two draws from the Winchester within-group distribution. Note that it is concentrated on smaller values. The bottom panel of Fig. 7 shows what the smallest of 5 distances would be expected to look like.

These empirical distributions play a role similar to existing DNA

To get a summary statistic that we can use in our test, we take the observed distance d^2 and compare it to the appropriate simulated reference distribution. The test statistic that we use to make a decision is the proportion of samples in our simulated reference distribution that are less than or equal to our observed d^2 . Call this \hat{p} . The statistic \hat{p} is adjusted for the fact that we are looking at the smallest distance and not just the distance between a random pair of bullets. If \hat{p} is small, then our observed distance is small compared to what would be expected if we randomly examined n bullets from a common group. This supports the hypothesis G that the bullets have a common origin or came from the same box. On the other hand if \hat{p} is large, then the observed distance is relatively large.

6.3 Deriving a threshold for \hat{p}

How do we decide whether \hat{p} is large or small relative to the simulated reference distribution? We propose that a threshold p^* be developed with \hat{p} 's smaller than the threshold being declared matches and larger values being declared nonmatches. We describe the approach for selecting a threshold here but the final choice would be made in concert with law enforcement officials. There is a natural tradeoff to be made in determining a threshold. If p^* is small then we will rarely find in favor of the hypothesis G — all innocent suspects are protected but evidence about some guilty ones will be held back unnecessarily. If p^* is large, then we will almost always find \hat{p} to be smaller and decide in favor of the hypothesis G . This makes our test extremely sensitive but will tend to take questionable evidence and treat

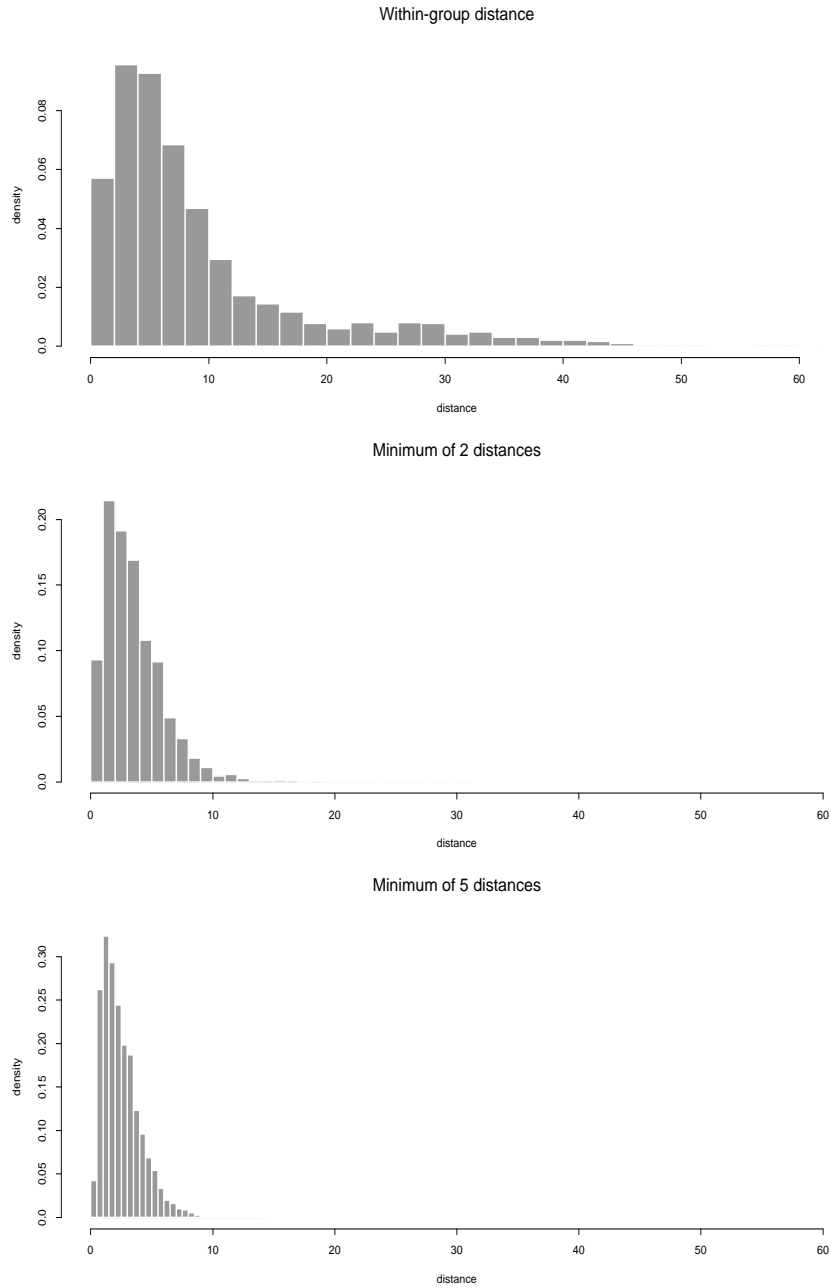


Figure 7: Reference distributions for assessing the observed distance. Top plot is the within-group distance distribution for Winchester bullets. The middle plot is the simulated reference distribution when the observed distance is the smaller of two distances. The final plots is the simulated reference distribution when the observed distance is the smallest of five distances. Note that we expect smaller distances when we make more comparisons.

it as being highly informative. In the medical world this is said to be a tradeoff between sensitivity (being able to detect a disease or identify a guilty suspect) and specificity (not identifying healthy people as diseased or not incriminating innocent suspects).

Once again we have turned to simulation to quantify the tradeoff involved. Specifically we simulate the accumulation and analysis of evidence under the two hypotheses and see how the test performs for different thresholds. We first describe the simulations under hypothesis G – that the fragment and bullets have a common source. We randomly sample $m = 15$ bullets from among the 200 for a given manufacturer. Then we pick one of the 15 bullets at random and choose a “fragment” from the same compositional group (and box). It is this last choice that insures we are operating under hypothesis G . Then for this random sample we carry out our test and record whether we have correctly identified the fragment as matching the suspect’s group of bullets or not. We can do this for every possible threshold. We repeat this 1000 times to estimate the sensitivity of the test for each threshold. We estimate specificity in a similar manner. Once again we randomly sample $m = 15$ bullets from among the 200 for a given manufacturer. This time we pick a fragment at random from a different box which insures that the true hypothesis is \bar{G} and carry out our test. We repeat this 1000 times to estimate the specificity of the test for each threshold. The end result is a table identifying the sensitivity and specificity for each threshold. The results for Cascade are provided in Table 8 below. Note that here it is easy to develop a high quality test. The specificity remains high (meaning innocent suspects are protected) for even high thresholds. It is possible to obtain sensitivity .9 (90% of matching cases are correctly identified) while maintaining specificity .95 (only 5% of non-matching cases are incorrectly identified as matching). Similar tables for the other manufacturers are presented as Tables 9-11. In general the test performs well. The weakest performance occurs for Remington where if we wish to maintain specificity .95, then we can attain only a sensitivity of .66.

Threshold	Sensitivity	Specificity
.10	.114	.998
.20	.224	.998
.30	.328	.997
.40	.393	.996
.50	.472	.994
.60	.548	.993
.70	.621	.992
.80	.707	.990
.90	.841	.987
.95	.876	.980
.96	.879	.973
.97	.897	.972
.98	.921	.950
.99	.928	.935

Table 8: Sensitivity and specificity for empirical test procedure using Cascade bullets.

Threshold	Sensitivity	Specificity
.10	.080	1.0
.20	.169	1.0
.30	.254	1.0
.40	.331	1.0
.50	.417	1.0
.60	.518	1.0
.70	.612	1.0
.80	.698	1.0
.90	.809	1.0
.95	.859	1.0
.99	.939	1.0

Table 9: Sensitivity and specificity for empirical test procedure using Federal bullets.

Threshold	Sensitivity	Specificity
.10	.188	.998
.20	.307	.998
.30	.391	.998
.40	.434	.998
.50	.470	.998
.60	.515	.998
.70	.553	.996
.80	.584	.991
.88	.656	.950
.90	.663	.932
.95	.718	.857
.99	.785	.688

Table 10: Sensitivity and specificity for empirical test procedure using Remington bullets.

Threshold	Sensitivity	Specificity
.10	.102	1.0
.20	.168	1.0
.30	.270	1.0
.40	.352	1.0
.50	.418	1.0
.60	.500	1.0
.70	.648	1.0
.80	.709	1.0
.90	.821	1.0
.95	.872	.999
.99	.969	.979

Table 11: Sensitivity and specificity for empirical test procedure using Winchester bullets.

6.4 The probability of a coincidence

To be useful in a courtroom setting one needs to provide a measure of how likely it would be to obtain a positive test result just by coincidence. In essence this requires determining the number of boxes circulating that contain bullets that would match the fragments found at the crime scene. One approach to answering this question would be try and estimate this quantity from patterns of bullet manufacture, distribution and use. As there is no direct information available to us of this type we propose an alternative based on the distance between bullets.

Some information about the likelihood of a coincidental match is contained in the between-group distribution described earlier but not used in our test procedure. Recall that our situation is as follows: we have found the smallest distance between the fragment at the crime scene and the suspect's m bullets (this is d^2) and we have found that our test statistic \hat{p} is below our threshold p^* thus finding in favor of the hypothesis G . Here again we propose the use of simulation, this to time to assess the probability of a coincidental match. We now want to imagine that the true hypothesis is \bar{G} and ask how often our test would favor G in that case. To do this in our situation we need to simulate m fragment-to-bullet distances that might be expected under \bar{G} . This is difficult because we must remember that it is possible that some bullets from different boxes might have originated from the same compositional group due to the manufacturing process. Our discussion of the likelihood ratio in Section 5 describes how one might estimate the probability that this event would occur (see also appendix Section 8.4) – for Cascade we estimated this probability as .27 though it is also suggested this is likely an overestimate. Then our simulation approach for estimating the probability of a coincidence is as follows:

1. For each of m bullets randomly decide for purposes of this simulation whether it is from the same group as the fragment or different (using the probability that bullets from different boxes come from the same compositional group).
2. Simulate m fragment-to-bullet distances by sampling randomly from the within-group or between-group distance distribution according to the decisions in Step 1.
3. Find the smallest of the m simulated distances and determine if it is less than or equal to our observed d^2 .
4. Repeat Steps 1-3 many times to estimate the probability of a coincidental match as significant as the one we have.

This proposal has not yet been implemented.

6.5 Limitations

Note that there are still some issues to be resolved before this method can be implemented. Ideally the within-group and between-group distribution would be established from a large

set of accurately classified data. Hopefully the distributions would remain relatively constant over time as they are not sensitive to the manufacturer's specifications, only the variability in the manufacturing process. Our simulations have used the same data to build the within-group and between-group distribution and for assessing the threshold. This means that the sensitivities and specificities in Tables 5-8 are likely higher than would be expected in actual operation. Additional work is required to assess performance in the more realistic case in which distributions developed on given database are used to analyze new, unseen bullet data. In addition, more work is clearly required for the second step of determining the probability of a coincidental match under our testing procedure.

7 Summary and Discussion

The goal of this project was to develop a means for assessing bullet evidence, especially to be able to quantify the significance of matching bullet lead. Two approaches were considered, a likelihood ratio approach and an empirically developed test procedure. The likelihood ratio approach is feasible at the current time only for small evidence sets, e.g., a single fragment and a single bullet. Moreover there are a number of issues associated with bullet distribution and usage patterns that were not factored into the current discussion. These also would tend to complicate the application of the likelihood ratio approach to bullets.

The empirically developed test procedure has great promise. It is limited at the moment by the relatively small amount of data that we have used to construct the test procedure. A more focused data collection effort, collecting data from manufacturers, is likely to provide more reliable information for improving this test procedure. A difficulty with the empirical approach is that a method for providing an estimate of the probability of a match by coincidence has not yet been tested.

A key point is that the bullet manufacturing process itself makes the analysis of forensic evidence difficult. Bullets manufactured from different batches of raw material may end up in the same box of bullets; similarly bullets manufactured from the same batch can end up in different boxes. It has been difficult from the limited data available to estimate the relative frequency of these events. This appears to be crucial information for developing quantitative methods for the assessment of bullet evidence.

8 Technical Appendices

At various points in the report we have intentionally omitted details that would make for difficult reading. Some of these issues are revisited here for purposes of completeness.

8.1 Identifying compositional groups using model-based clustering

Model-based clustering was described informally in Section 4. Here we provide a bit more detail. In model-based clustering, the basic assumption is that the trace element measure-

ments for each bullet comes from one of a number of clusters. The number of clusters as well as their defining characteristics (mean and variance) are to be determined. In the model-based approach that we used each cluster is represented by a Gaussian or normal model with the k th cluster having mean vector μ_k , and variance matrix Σ_k . The probability density for such a cluster, which specifies the relative likelihood of observing a value z is

$$\phi_k(z|\mu_k, \Sigma_k) = (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (z - \mu_k)^T \Sigma_k^{-1} (z - \mu_k) \right\}.$$

These clusters are ellipsoidal, centered at the mean μ_k . The variance matrix Σ_k determines the other geometric features of the cluster (e.g., whether it is more or less spherical). If there are G clusters and we take E_k to be the elements of cluster k , then the probability density for the observed data z_1, \dots, z_n (also known as the likelihood function for estimating the parameters μ_k, Σ_k and the cluster memberships) is

$$L = \prod_{k=1}^G \prod_{i \in E_k} \phi_k(z_i|\mu_k, \Sigma_k).$$

We obtained compositional groups using MCLUST, a software package for model-based cluster analysis, available from <http://www.stat.washington.edu/fraley/mclust/soft.shtml>. That software package includes an automatic procedure for estimating the number of clusters and the parameters describing each cluster. According to the results, there are 8 compositional groups represented in the 200 bullets for the manufacturer Cascade, 4 compositional groups for Federal, 9 compositional groups for Remington, and 36 compositional groups for Winchester (see Tables 4-7 in Section 4 of this report).

It is worth noting that these groups are similar to those reported by the FBI study in their analysis of the same data but typically somewhat smaller. The algorithm used in the FBI study which analyzed the bullets one at a time, comparing each to all existing groups, is a bit more conservative in that it forms new groups more often than the model-based approach.

8.2 Distribution of measurements within compositional group

For each manufacturer, there are measurements of p (typically 5 or 6) trace element concentrations for each bullet. We assume that the vectors of trace element measurements for bullets from the k th compositional group have a multivariate normal distribution with mean μ_k and variance matrix Σ_{within} (the density for the multivariate normal is given above in the discussion of model-based clustering). Note that this is a bit more restrictive than the most general clustering model in which each cluster has its own variance matrix. In practice we found it useful to combine the information from all of the clusters to estimate a common variance matrix for all of the compositional groups for a single manufacturer (though different manufacturers have different variance matrices). It is difficult to reliably estimate separate variance matrices for each different compositional group. The normality

assumption is made more tenable by the fact that the trace element measurements are usually an average of three or more readings per bullet.

We have considered alternatives to the multivariate normal assumption. If we think of the different compositional groups as perhaps having different variance matrices, then we are led to a multivariate t -distribution in place of the normal distribution. This yields additional variation which should yield more conservative likelihood ratios in practice.

8.3 Distribution of group means within a manufacturer

If we further assume that the means for each compositional group (the μ_k 's) have a normal distribution with mean μ_{manuf} and variance $\Sigma_{between}$, then the distribution of all bullets from that manufacturer can be viewed as a normal distribution with mean μ_{manuf} and variance $\Sigma_{between} + \Sigma_{within}$. This is the relevant distribution when we are looking at a bullet with no knowledge about its compositional group. If we know its compositional group (perhaps because we hypothesize it to be the same as another bullet's group), then we rely on the distribution for bullets within a compositional group. It is natural to assume that the manufacturer mean differs for different manufacturers. Once again alternatives to the normal distribution can be considered.

8.4 Combinatoric arguments for the conditional probabilities

For the conditional probabilities $p(\text{same group}|\text{same box})$ and $p(\text{same group}|\text{different box})$, we use a combinatorial argument. For example, suppose that there are b boxes of bullets from a given manufacturer, g_i homogeneous groups represented in the i th box, and n_{ij} bullets from the j th compositional group within the i th box for a manufacturer. The number of distinct compositional groups is at most $\sum_{i=1}^b g_i$ but could be less if the same groups appear in more than one box. Then one can calculate that

$$\begin{aligned} p(\text{same group}|\text{same box}) &= \frac{p(\text{same group, same box})}{p(\text{same box})} \\ &= \frac{\sum_{i=1}^b \sum_{j=1}^{g_i} \binom{n_{ij}}{2} / \binom{N}{2}}{\sum_{i=1}^b \binom{n_i}{2} / \binom{N}{2}}, \end{aligned}$$

where $n_i = \sum_{j=1}^{g_i} n_{ij}$ for $i = 1, \dots, b$ and $N = \sum_i n_i$ is the total number of bullets. Note that $n_i = 50$ for a complete box, however in our example we chose to ignore one compositional group (it was small) that contained bullets from some of the boxes so that n_i will be less than 50 in those cases. Naturally, the probability that two bullets from the same box come from different compositional groups is just one minus the previous quantity.

The probability that two bullets come from the same group though they are in different boxes can be computed in a similar way. It is a bit more difficult to write down because more detailed notation is required. We explain the approach but do not give detailed equations.

$$p(\text{same group}|\text{different box}) = \frac{p(\text{same group, different box})}{p(\text{different box})}$$

$$= \frac{p(\text{same group, different box})}{1 - p(\text{same box})}.$$

The denominator is easy to compute; we have already computed $p(\text{same box})$. The calculation of $p(\text{same group, different box})$ is a bit more complicated. It requires finding the total number of ways to select bullets from two different boxes that are from the same compositional group. We illustrate in the next section. Occasionally it is possible that additional information (like packaging date) may exist to simplify this calculation.

8.5 Worked example in more detail

Here, we consider the calculation of each component of the likelihood ratio from Section 5.4 in more detail. Recall that this example concerns the 200 bullets in the FBI study that were manufactured by Cascade. We obtained eight compositional groups using the model-based clustering method. The results are summarized in Table 4 of Section 4. Cluster (group) 6 was quite small and did not appear too homogeneous. As a result we did not use this in the remaining analysis.

If we take z_{ij} to be the vector of trace element concentrations for the j th bullet in the i th group (recall this is the average of the logarithms of three measurements), then we estimate the within group variance matrix as

$$S_{within} = \frac{1}{N - g} \sum_{i=1}^g \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)(z_{ij} - \bar{z}_i)',$$

where \bar{z}_i is the mean of the trace element measurements for the bullets in group i , $N = 195$ is the total number of bullets, $g = 7$ is the number of compositional groups, and n_i is the number of bullets in i th group. The result is the 5×5 matrix

$$\hat{\Sigma}_{within} = S_{within} = \begin{pmatrix} 0.000333 & 0.000167 & 0.000285 & 0.000071 & 0.000026 \\ 0.000167 & 0.000642 & 0.000375 & -0.000071 & 0.000080 \\ 0.000285 & 0.000375 & 0.001097 & 0.000216 & -0.000116 \\ 0.000071 & -0.000071 & 0.000216 & 0.003487 & 0.000424 \\ 0.000026 & 0.000080 & -0.000116 & 0.000424 & 0.000786 \end{pmatrix}.$$

When the number of bullets in each group/cluster is the same there are standard procedures for estimating $\Sigma_{between}$. That is not the case here so we use the following approach (described for example in Section 13.7 of the the statistical text *Statistical Methods* by G. W. Snedecor and W. G. Cochran, Iowa State University Press, 1989). The variation among cluster means can be estimated directly as

$$S_{between} = \frac{1}{g - 1} \sum_{i=1}^g (\bar{z}_i - \bar{z})(\bar{z}_i - \bar{z})',$$

where \bar{z} is the mean of all of the trace element measurements. If the sample sizes were equal this quantity would be a good estimate of $\Sigma_{between} + \Sigma_{within}/n$ where n is the number of

observations in each cluster. Instead here it is a natural estimate of $\Sigma_{between} + \Sigma_{within}/n_h$ where n_h is the harmonic mean ($n_h = g/(\sum_{i=1}^g \frac{1}{n_i})$). Then a natural estimate of $\Sigma_{within} + \Sigma_{between}$ can be constructed from S_{within} and $S_{between}$ as $\hat{\Sigma}_{within} + \hat{\Sigma}_{between} = \frac{n_h-1}{n_h}S_{within} + S_{between}$ with resulting estimate

$$\hat{\Sigma}_{within} + \hat{\Sigma}_{between} = \begin{pmatrix} 0.001903 & 0.009086 & -0.026462 & 0.009287 & 0.014005 \\ 0.009086 & 0.122653 & -0.085549 & 0.036607 & 0.016389 \\ -0.026462 & -0.085549 & 0.524556 & -0.172858 & -0.296696 \\ 0.009287 & 0.036607 & -0.172858 & 0.068801 & 0.089066 \\ 0.014005 & 0.016389 & -0.296696 & 0.089066 & 0.192721 \end{pmatrix}.$$

Thus we have estimates of the variance matrices that define the two likelihood functions or densities that contribute to the likelihood ratio.

The remaining pieces of the likelihood ratio are probabilities that a bullet and fragment randomly chosen from the same (or different) box would come from the same group. Using the clustering results for Cascade and the combinatoric arguments from the previous section we find that,

$$\begin{aligned} p(\text{same group}|\text{same box}) &= \frac{p(\text{same group, same box})}{p(\text{same box})} \\ &= \frac{\binom{22}{2} + \binom{9}{2} + \binom{19}{2} + \binom{44}{2} + \binom{5}{2} + \binom{40}{2} + \binom{9}{2} + \binom{43}{2} + \binom{4}{2}}{\binom{50}{2} + \binom{49}{2} + \binom{49}{2} + \binom{47}{2}} \\ &= 0.67 \end{aligned}$$

and

$$\begin{aligned} p(\text{same group}|\text{different box}) &= \frac{p(\text{same group, different box})}{p(\text{different box})} \\ &= \frac{2 [\binom{44}{1}\binom{43}{1} + \binom{5}{1}\binom{4}{1}]}{\binom{195}{2} - [\binom{50}{2} + \binom{49}{2} + \binom{49}{2} + \binom{47}{2}]} \\ &= 0.27. \end{aligned}$$

The calculations reported here are then used to construct the likelihood ratio as described in Section 5.

8.6 A distance measure

The testing approach in Section 6 relies on the definition of a measure of the distance between the vector of trace element concentrations for two bullets. The measure we rely on is closely related to the distribution of measurements within a compositional group. Let x_i and x_j denote the vectors of measurements for the two bullets, then

$$d_{ij}^2 = (x_i - x_j)^T \hat{\Sigma}_{within}^{-1} (x_i - x_j).$$

This choice is motivated by the multivariate normal distribution that we have assumed for measurements within a compositional group. It can be recognized as being similar to the exponent of that normal distribution (multiplied by -2).

Acknowledgements

Ying Huang, Ji-Yeon Kim, and Jennifer Schumi, research assistants on the project, provided support for the research and preparation of this report. We are also appreciative of the help provided by Robert Koons and Jo Ann Buscaglia of the Federal Bureau of Investigation's Forensics Unit.