Statistical Tests for Bullet Lead Comparisons

NRC Report to the FBI

Karen Kafadar

University of Colorado at Denver

kk@math.cudenver.edu

http://math.cudenver.edu/~kk

Clifford H. Spiegelman, Texas A&M

Acknowledgements:

Bullet Lead Committee Members; M. Cohen, NRC

OUTLINE

1. Introduction and Background
   Charge to *Committee on Scientific Assessment of Bullet Lead Elemental Composition Comparison*
2. Bullet Manufacturing Process
3. FBI tests on individual elements
4. Available data sets
5. False Positive Probabilities
6. Probative impact of CABL
7. Alternative tests for matching bullets
8. Committee Recommendations; Further work

1. Introduction

Background

- Crime → evidence → bullets
- Gun recovered: match striations on bullet and gun barrell (separate NRC committee)
- *No gun*: **Compositional Analysis of Bullet Lead** (CABL)
- "Working hypothesis": chemical concentration of lead used to make a "batch" of bullets provides a "unique signature" ⇒ "equal" concentrations of elements in bullet from Crime Scene (CS) and bullets from Potential Suspect (PS) may indicate "guilt"

- Local police dept sends CS, PS bullets sent to FBI lab
- FBI measures (in triplicate) concentrations of 7 elements
- Reports "analytically indistinguishable concentrations" between CS and PS bullets if "mean $\pm$ 2·SD intervals overlap for *all* 7 elements" (**2-SD-overlap**)
- FBI also uses "**range overlap**" and "**chaining**"
- FBI court testimony when requested

Charge to the Committee

1. Is analytical procedure (ICP-OES) sound, best available? Choice of elements, use of isotopes?
2. **"Are the statistical tests used to compare two samples appropriate?"**
3. **"Can known variations introduced in manufacturing process be used to model specimen groupings and provide improved comparison criteria?"**
4. Interpretation issues (probative value): "What are the appropriate statements that can be made to assist the requester in interpreting the results of compositional bullet lead comparison? Can significance statements be modified to include effects of such factors as the analytical technique, manufacturing process, ...

2. Manufacturing Process

- Most bullets made from lead in recycled batteries (5%)
- Process involves removal of impurities (Cu, Se, Zn, S, ...) by cooling, heating, crystalizing, addition of chemicals (e.g., charcoal, Zn, Sb for hardness)
- Formed into blocks ("pigs," "ingots," "slugs")
- Extruded into wire of dimension depending upon caliber
- Cut into pieces for bullets; poured into bins
- Bullets sent to cartridge manufacturer
- Mixed in bins; placed in boxes
- Boxes shipped to customer (many to 1 store in small town, or many to many stores in large city, or ...)

'Homogeneity' of bullets within a 'source'

- "melt": some oxidation of elements, but likely insignificant
- pig, ingot, billet: some segregation of solutes during solidification
- wires, slugs, bullets: "uniformity" along length of wire (Randich et al. 2002; Koons and Grant 2002)
- CIVL = "compositionally indistinguishable volume of lead"
- All discussion refers to chemical composition of a CIVL, which could yield 12,000–35 million bullets, depending upon manufacturing consistency, volume, bullet size

3. FBI Tests on Individual elements

2 bullets, 3 measurements of 7 elements on each bullet (As, Sb, Sn, Bi, Cu, Ag, Cd)

Measurement = average of triplicates of one element:
Each bullet [bullet fragment] has three measurements

Crime Scene (CS) bullet: $\mathbf{X_i} = (X_{i1}, ..., X_{i7})'$
$\bar{X}$ and s_X = vector of means, SDs

Potential Suspect (PS) bullet: $\mathbf{Y_i} = (Y_{i1}, ..., Y_{i7})'$
$\bar{Y}$ and s_Y = vector of means, SDs

FBI presented three procedures for assessing "match":

- 2-SD-overlap
- range overlap
- "chaining"

Committee reformulated question as follows:

Do they come from populations (CIVLs) with same mean concentration on each of the 7 elements?

Three issues:

1. Components of variance
   - σ_e = measurement variation
   - σ_w = variation between bullets **within** batch
   - σ_b = variation between bullets **between** batches
2. False positives:
   P{claim 'match' | bullet mean concentrations differ by δ}
3. Sensitivity (and specificity):
   - P{bullet mean concentrations $< \delta$ | test claims "match"}
   - P{bullet mean concentrations $> \delta$ | claims "no match"}

2-SD-overlap test: Claim "match" if "2-SD-intervals overlap":

$$\bar{X} + 2s_X > \bar{Y} - 2s_Y \text{ or } \bar{Y} + 2s_Y > \bar{X} - 2s_X$$
$$\text{i.e., } |\bar{X} - \bar{Y}| < 2(s_X + s_Y) \text{ on each element}$$

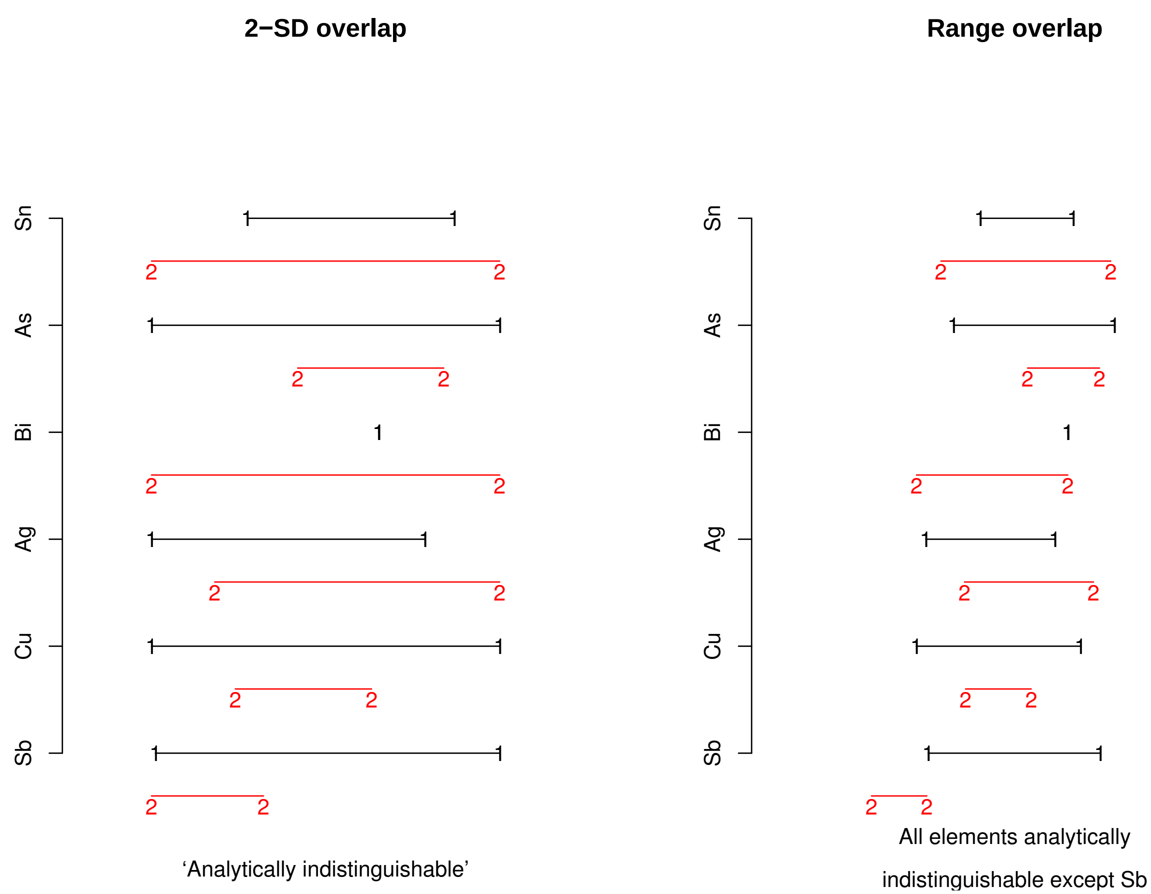
1. σ better estimated by s_p =pooled SD from many bullets: comparable error in measuring concentrations in CS, PS bullets (close in time, same sets of standards, etc)
2. Chemical concentrations *lognormal*, not normal: $\log(X) \sim \mathrm{N}(\mu_X, \sigma^2)$, $\log(Y) \sim \mathrm{N}(\mu_Y, \sigma^2)$
3. $\mathrm{SD}(X) \sim \mathrm{SD}(\log(X))$ if $\sigma/\mu < 0.05$
4. $\mathrm{SD}(\bar{X} - \bar{Y}) \approx \sigma\sqrt{2/3}$, not 2σ
5. FBI allowance of $\approx 4\sigma$ is too wide

Federal bullet F001

	icpSb	icpCu	icpAg	icpBi	icpAs	icpSn
a	29276	285	64	16	1415	1842
b	29506	275	74	16	1480	1838
c	29000	283	66	16	1404	1790
mean	29260.7	281.0	68.0	16	1433.0	1823.3
SD	253.4	5.3	5.3	0	41.1	28.9
mean-2SD	28754.0	270.4	57.4	16	1350.8	1765.5
mean+2SD	29767.4	291.6	78.6	16	1515.2	1881.2
minimum	29000	275	64	16	1404	1790
maximum	29506	285	74	16	1480	1842

Federal bullet F002

	icpSb	icpCu	icpAg	icpBi	icpAs	icpSn
a	28996	278	76	16	1473	1863
b	28833	279	67	16	1439	1797
c	28893	282	77	15	1451	1768
mean	28907.3	279.7	73.3	15.7	1454.3	1809.3
SD	82.4	2.1	5.5	0.6	17.2	48.7
mean-2SD	28742.5	275.5	62.3	14.5	1419.8	1712.0
mean+2SD	29072.2	283.8	84.4	16.8	1488.8	1906.7
minimum	28833	278	67	15	1439	1768
maximum	28996	282	77	16	1473	1863


2−SD overlap
Range overlap
Sn
As
Bi
Ag
Cu
Sb
‘Analytically indistinguishable’
All elements analytically
indistinguishable except Sb

5. FPP = False Positive (Match) Probability (1 element)

Mean concentrations differ by δ, measurement error SD σ_e:

$$P\{(\bar{x}+\bar{y}) < 2(s_x+s_y) \mid |\mu_x-\mu_y| = \delta\}$$
$$P\{(\bar{x}+\bar{y}) < 2(s_x+s_y) \mid |\mu_x-\mu_y| = \delta\}$$
$$= P\{(\bar{x}+\bar{y})/(s_p\sqrt{2/3}) < 2(s_x+s_y)/(s_p\sqrt{2/3}) \mid |\mu_x-\mu_y| = \delta\}$$

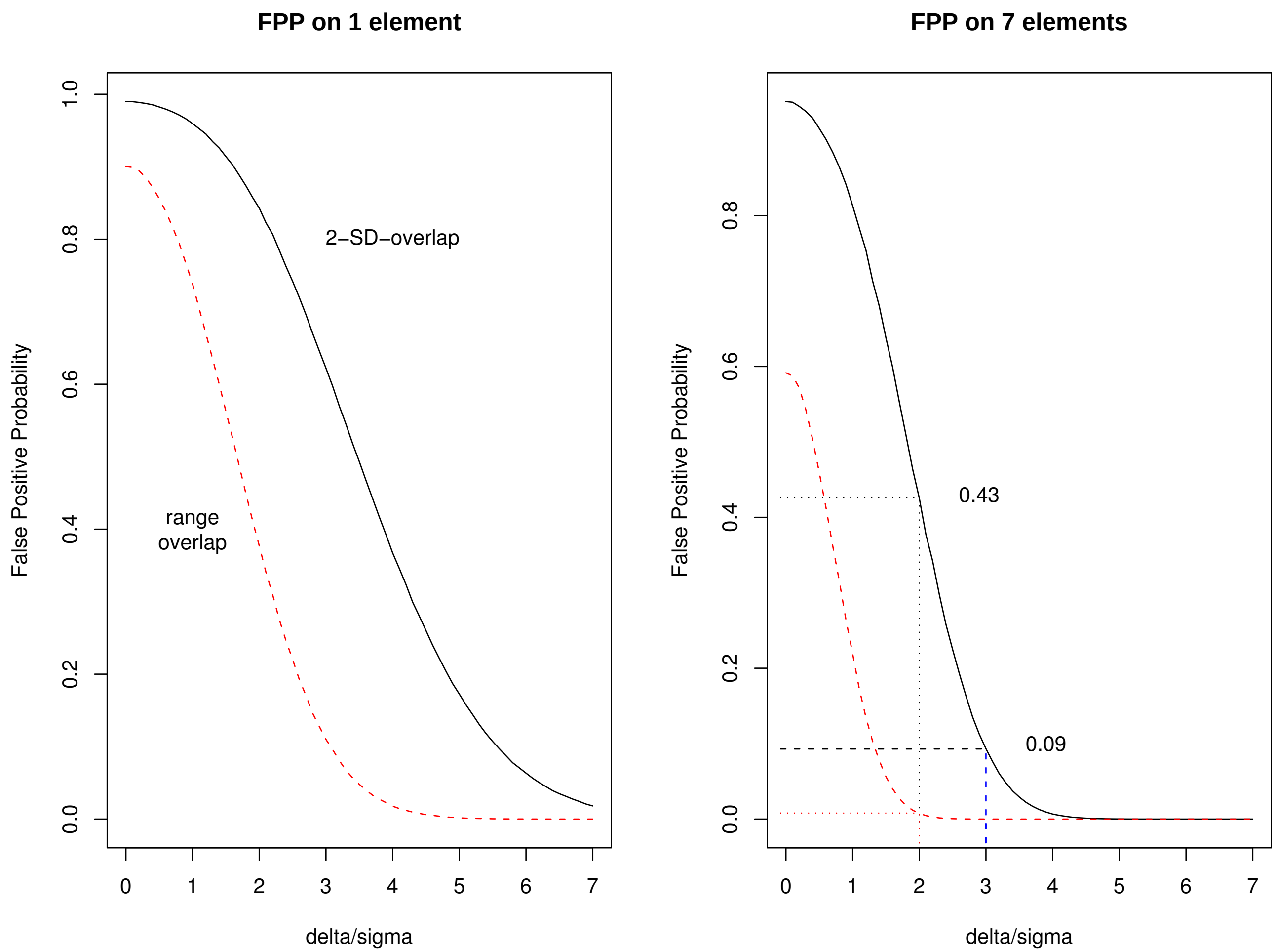
FPP is a function of only δ/σ_e:

$E(s_x) = E(s_y) = 0.8812\sigma$, $E(s_p) \approx \sigma$ (many d.f. in s_p):

$$P\{(\bar{x}+\bar{y})/(s_p\sqrt{2/3}) < 4.317 \mid |\mu_x-\mu_y| = \delta\}$$

[rough approximation: $E(P\{t < r.v.\}) \neq P\{t < E(r.v.)\}$]

- Distribution of $\frac{\bar{X}-\bar{Y}}{s_x+s_y} = \frac{N(\delta,2\sigma_e^2)}{\sqrt{\chi_2^2}+\sqrt{\chi_2^2}}$ is theoretically possible but messy
- Easier: 100,000 simulations
- Plot $FPP(\delta)$ for given levels of σ_e
- FPP = function of δ/σ, $\sigma = \sqrt{\sigma_e^2 + \sigma_b^2 + \sigma_w^2}$
- Large $\delta \Rightarrow$ small probability of false match
- Small δ ($\delta < 3\sigma_e$) $\Rightarrow$ high probability
- Smaller FPP for range overlap (E(range of 3) $\approx 1.6\sigma_e$)


FPP on 1 element
False Positive Probability
0.0
0.2
0.4
0.6
0.8
1.0
2−SD−overlap
range
overlap
0
1
2
3
4
5
6
7
delta/sigma
FPP on 7 elements
False Positive Probability
0.0
0.2
0.4
0.6
0.8
0.43
0.09
0
1
2
3
4
5
6
7
delta/sigma

FPP for 2-SD-overlap, independent measurement errors

σ	$\delta = 0$	1	2	3	4	5	6	7
0.5	0.931	0.298	0.001	0.000	0.000	0.000	0.000	0.000
1.0	0.931	0.749	0.298	0.036	0.001	0.000	0.000	0.000
1.5	0.931	0.849	0.612	0.303	0.084	0.013	0.001	0.000
2.0	0.931	0.883	0.747	0.535	0.302	0.125	0.036	0.007
2.5	0.931	0.903	0.817	0.669	0.487	0.302	0.151	0.062
3.0	0.931	0.911	0.850	0.748	0.615	0.450	0.298	0.175

FPP for 2-SD-overlap, Federal correlation matrix

σ	$\delta = 0$	1	2	3	4	5	6	7
0.5	0.950	0.426	0.007	0.000	0.000	0.000	0.000	0.000
1.0	0.950	0.813	0.426	0.093	0.007	0.000	0.000	0.000
1.5	0.950	0.884	0.713	0.426	0.163	0.048	0.007	0.001
2.0	0.950	0.916	0.813	0.638	0.426	0.225	0.093	0.029
2.5	0.950	0.929	0.864	0.754	0.599	0.426	0.258	0.135
3.0	0.950	0.938	0.884	0.813	0.713	0.553	0.426	0.298

4. Data Sets Available to Committee

- '800-bullet'
- '1837-bullet' (subset 854 bullets)
- Tables 1+2 in Randich et al. (2002)
- Table 3 in Koons and Grant (2002)
- Others in published articles (but not analyzed here)

800-bullet data set

- Peele, Havekost, Peters, Riley, Halberstam, Koons (1991), "Comparison of Bullets Using the Elemental Composition of the Lead Component," *Proceedings of the International Symposium on the Forensic Aspects of Trace Evidence*
- 4 manufacturers (CCI, Federal, Remington, Winchester)
- 4 boxes per manufacturer
- 50 bullets per box
- Replicates a, b, c per bullet

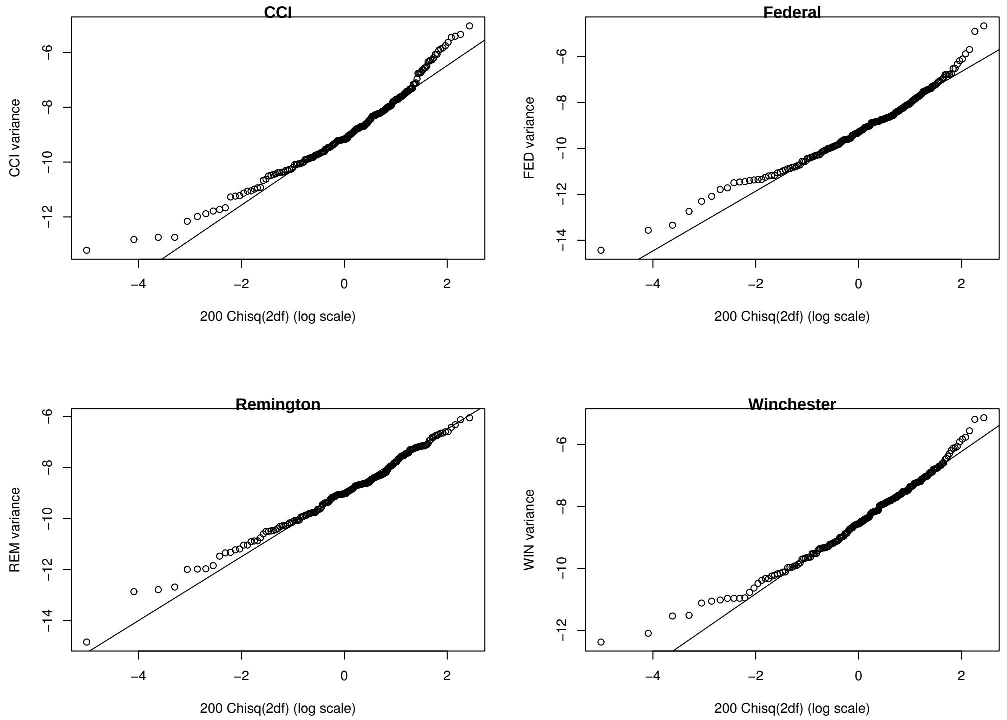
What this data set can provide:

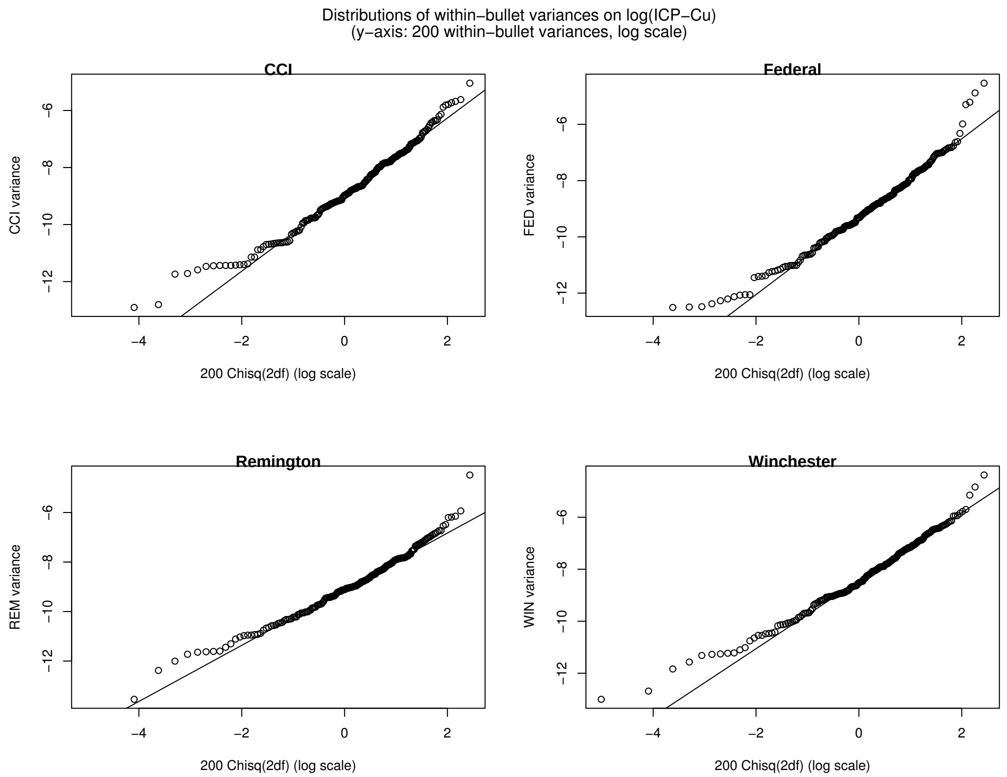
- Rough idea of distributions of concentrations across bullets
- measurement error variance σ_e^2
- correlations between errors in measuring two different elements on same bullet (only Federal data measured 6 elements by ICP-OES; none measured Cadmium)

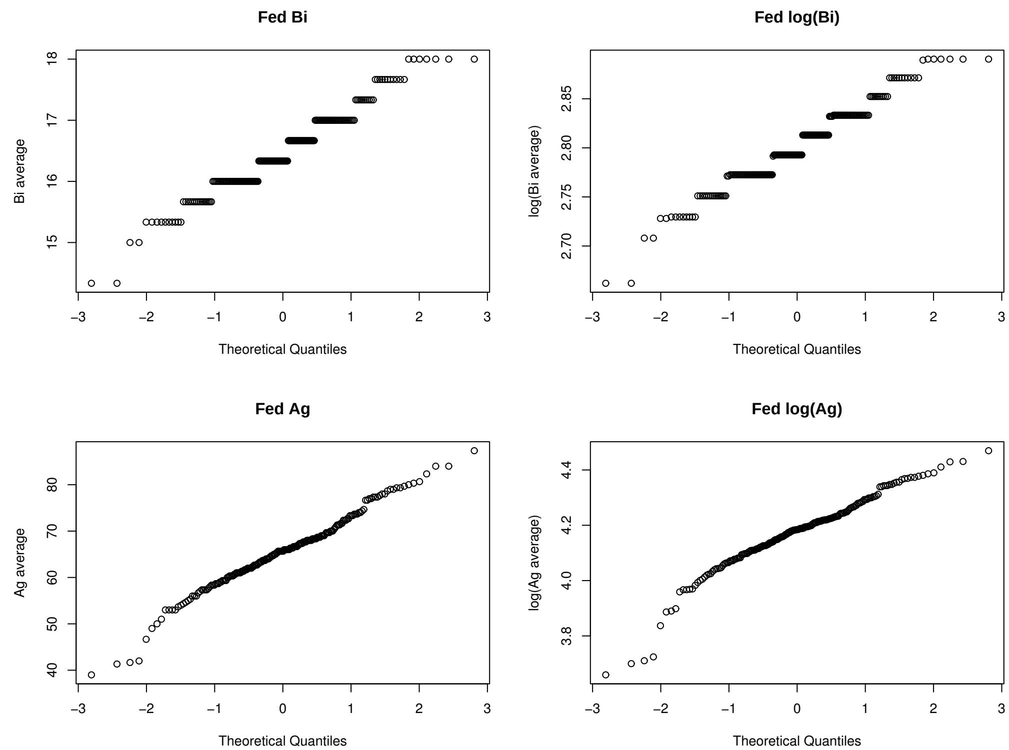
What this data set not provide:

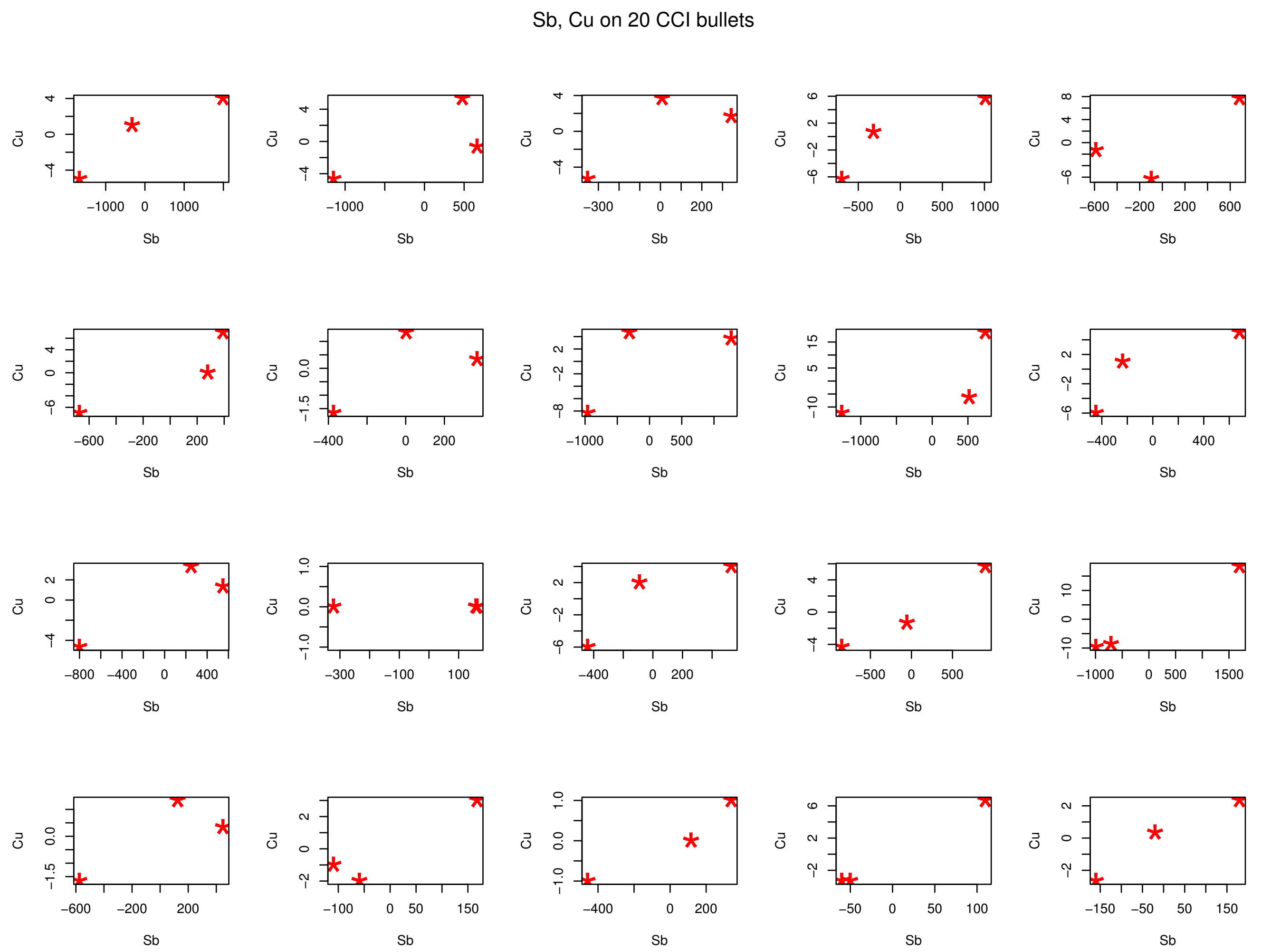
- Does not provide estimates of within-batch homogeneity σ_w
- Does not provide estimates of between-batch variability σ_b – unless one believes that the "batches" defined by one of thousands of possible clustering algorithms are "homogeneous" (e.g., ISU Tech Report, FBI "chaining").

Distributions of within–bullet variances on log(ICP–Sb)
(y–axis: 200 within–bullet variances, log scale)




Distributions of within−bullet variances on log(ICP−Cu)
(y−axis: 200 within−bullet variances, log scale)
CCI
CCI variance
Federal
FED variance
Remington
REM variance
Winchester
WIN variance
200 Chisq(2df) (log scale)
−4 −2 0 2
−6 −8 −10 −12


Fed Bi
Bi average
Theoretical Quantiles
Fed log(Bi)
log(Bi average)
Theoretical Quantiles
Fed Ag
Ag average
Theoretical Quantiles
Fed log(Ag)
log(Ag average)
Theoretical Quantiles


Sb, Cu on 20 CCI bullets
Cu
Sb

Sample correlation matrix: Federal bullets

	As	Sb	Sn	Bi	Cu	Ag	(Cd)
As	1.000	0.320	0.222	0.236	0.420	0.215	0.000
Sb	0.320	1.000	0.390	0.304	0.635	0.242	0.000
Sn	0.222	0.390	1.000	0.163	0.440	0.154	0.000
Bi	0.236	0.304	0.163	1.000	0.240	0.179	0.000
Cu	0.420	0.635	0.440	0.240	1.000	0.251	0.000
Ag	0.215	0.242	0.154	0.179	0.251	1.000	0.000
(Cd)	0.000	0.000	0.000	0.000	0.000	0.000	1.000

Consequence:

$$\mathrm{P}\{7\ |t|\ \text{statistics} < K_\alpha(\nu, n)\} \neq \alpha^7\ (\text{maybe } \alpha^5)$$

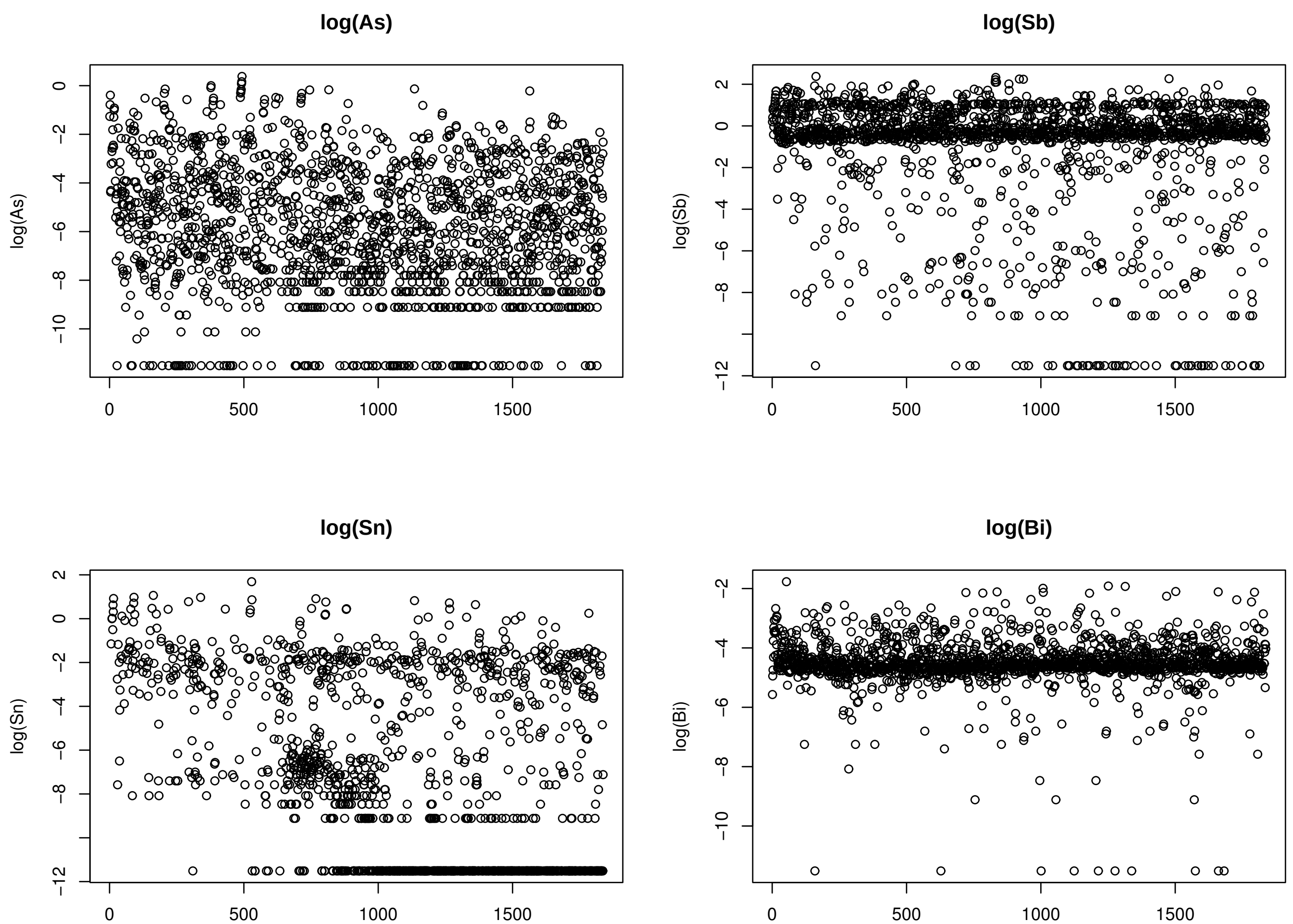
1837-bullet data set

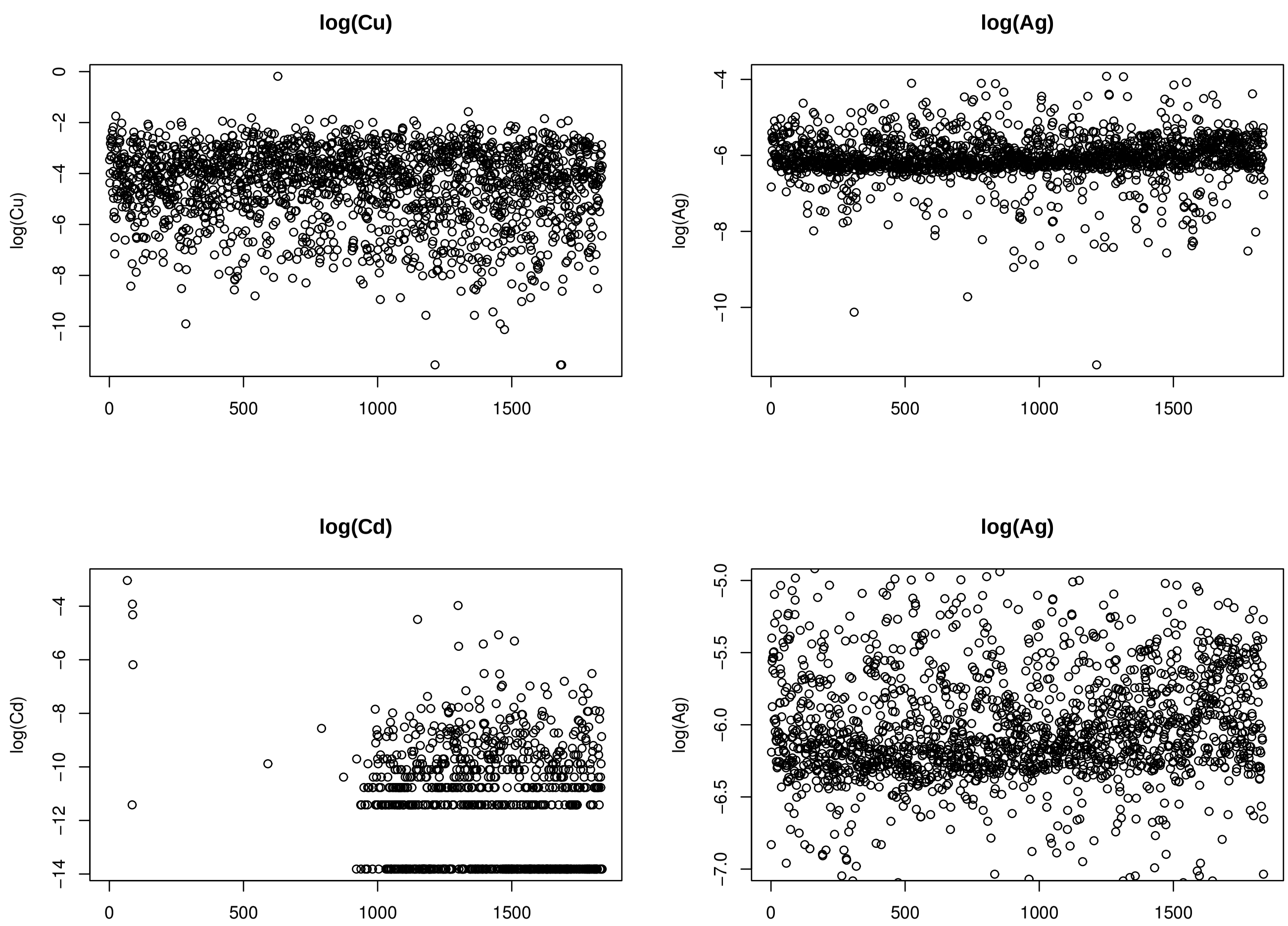
- Part of the complete data log containing chemical analyses on 71,000+ bullets
- FBI "selected" 1837 bullets that were believed to be "different"
- 1837-bullet set = FBI's attempt at different "melts"
- Only 854 of 1837 had all 7 elements (1997 or later)

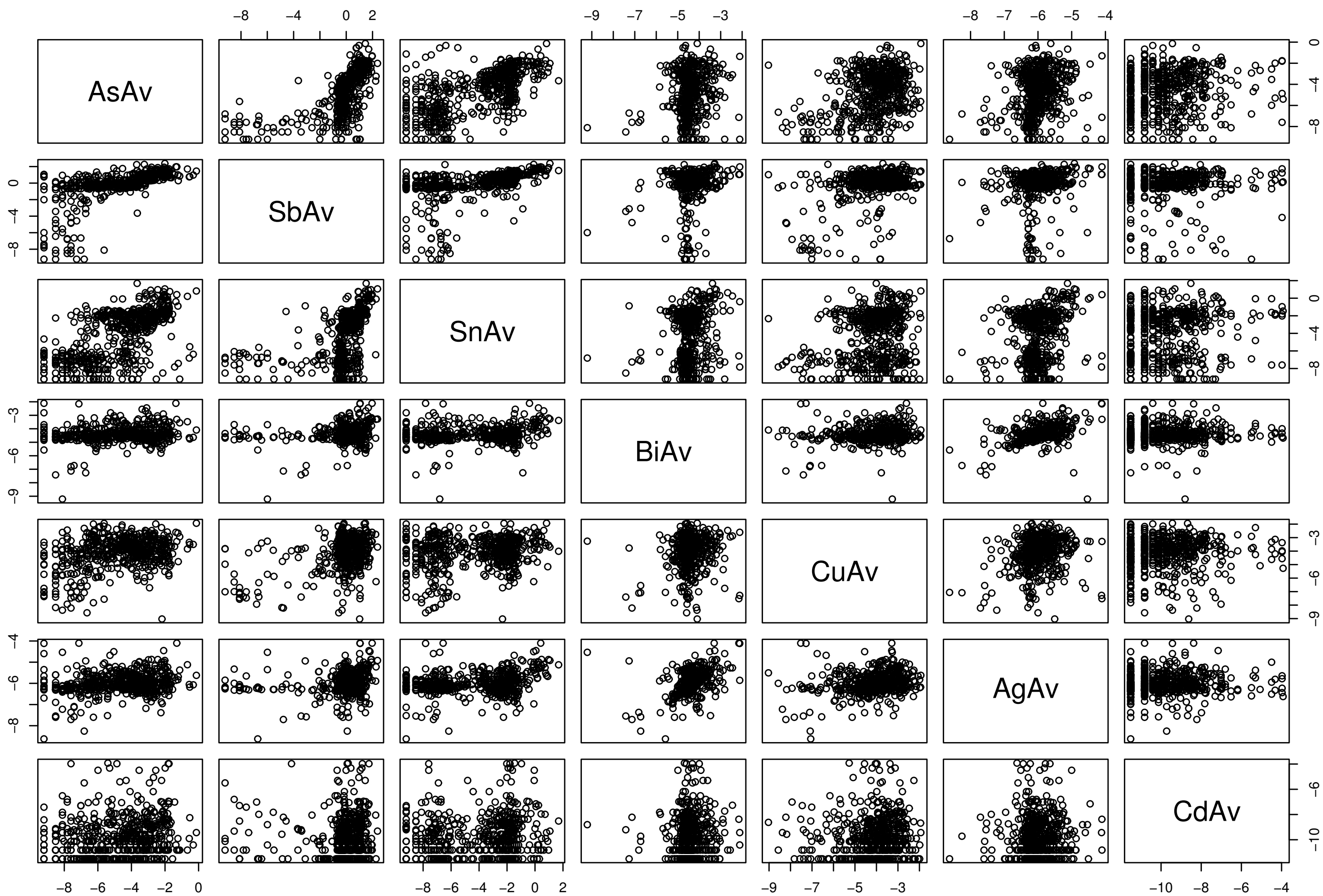
FBI Notes on 1837-bullet data set

"To assure independence of samples, the number of samples in the full database was reduced by removing multiple bullets from a given known source in each case. To do this, evidentiary submissions were considered one case at a time. For each case, one specimen from each combination of bullet caliber, style, and nominal alloy class was **selected** *and that data was placed into the test sample set. In instances where two or more bullets in a case had the same nominal alloy class, one sample was randomly selected from those containing the maximum number of elements measured. ... The test set in this study, therefore, should represent an unbiased sample in the sense that each known production source of lead is represented by only one randomly selected specimen."* (Notes on 1837-bullet dataset)

- FBI used it to estimate FPP=False Positive Probability: 693 2-SD-overlap "matches" among 1,686,366 comparisons $\Rightarrow$ "about 1 in 2500"
- Committee: This FPP (1 in 2500) is not valid (useless)
- 1837-bullet data set is not a random sample
- See Cochran, Mosteller, Tukey (1954), "Principles of Sampling" (*JASA*)
- Does provide some information on distributions of concentrations across unspecified collection of bullets


log(As)
log(Sb)
log(Sn)
log(Bi)
log(As)
log(Sb)
log(Sn)
log(Bi)
0
500
1000
1500


log(Cu)
log(Ag)
log(Cd)
log(Ag)

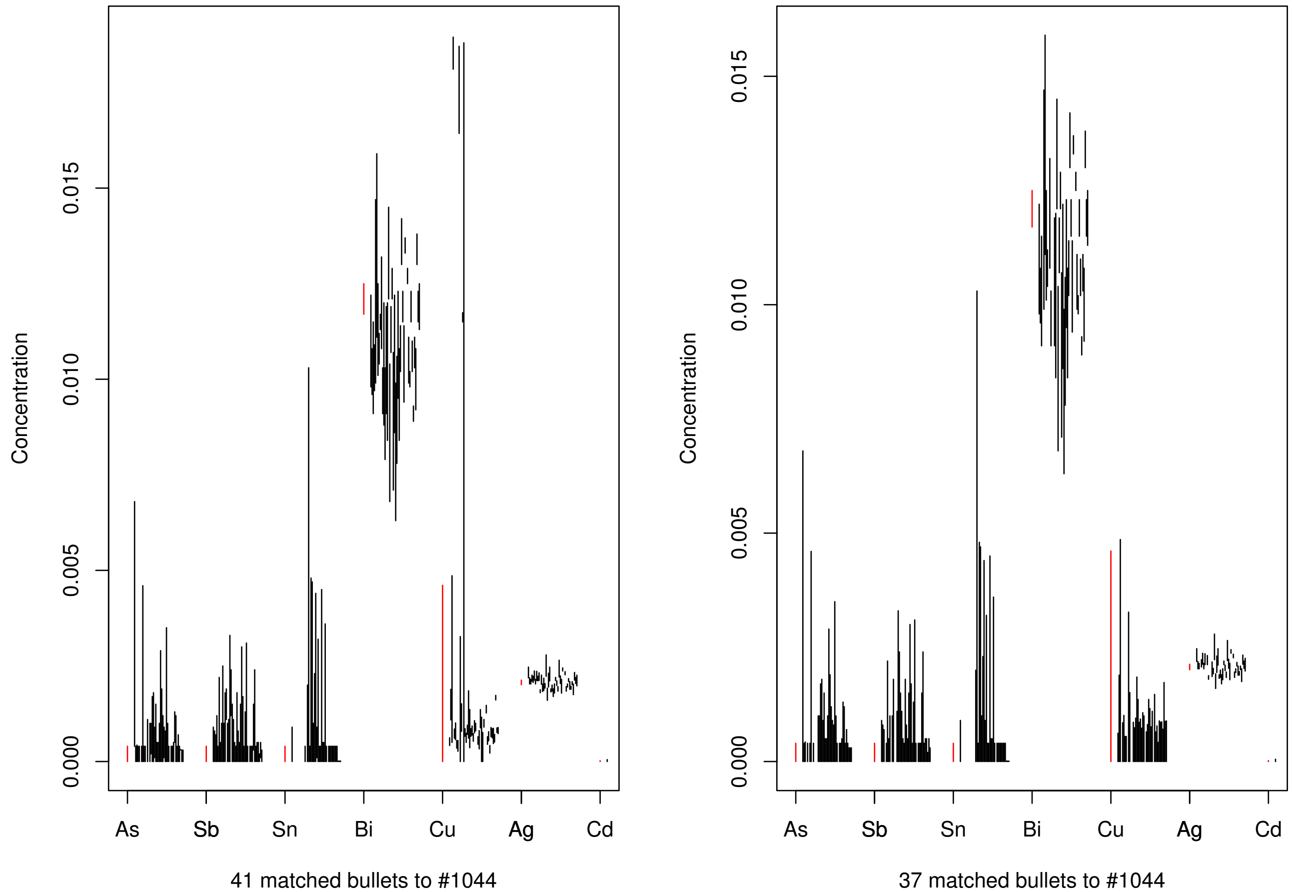


545 bullets from 1837–bullet data set with 7 measureable elements

Chaining (FBI's version of "clustering"):

"The mean element concentrations of the first and second specimens in the known material population are compared based upon twice the measurement uncertainties from their replicate analysis. If the uncertainties overlap in all elements, they are placed into a composition group; otherwise they are placed into separate groups. The next specimen is then compared to the first two specimens, and so on, in the same manner until all of the specimens in the known population are placed into compositional groups." (Peters, C.A.: *Comparative Elemental Analysis of Firearms Projectile Lead By ICP-OES*, FBI Laboratory Chemistry Unit. Issue date: October 11, 2002.)

Resulting "compositional group" could be very diverse:


Concentration
0.000
0.005
0.010
0.015
As
Sb
Sn
Bi
Cu
Ag
Cd
41 matched bullets to #1044
Concentration
0.000
0.005
0.010
0.015
As
Sb
Sn
Bi
Cu
Ag
Cd
37 matched bullets to #1044

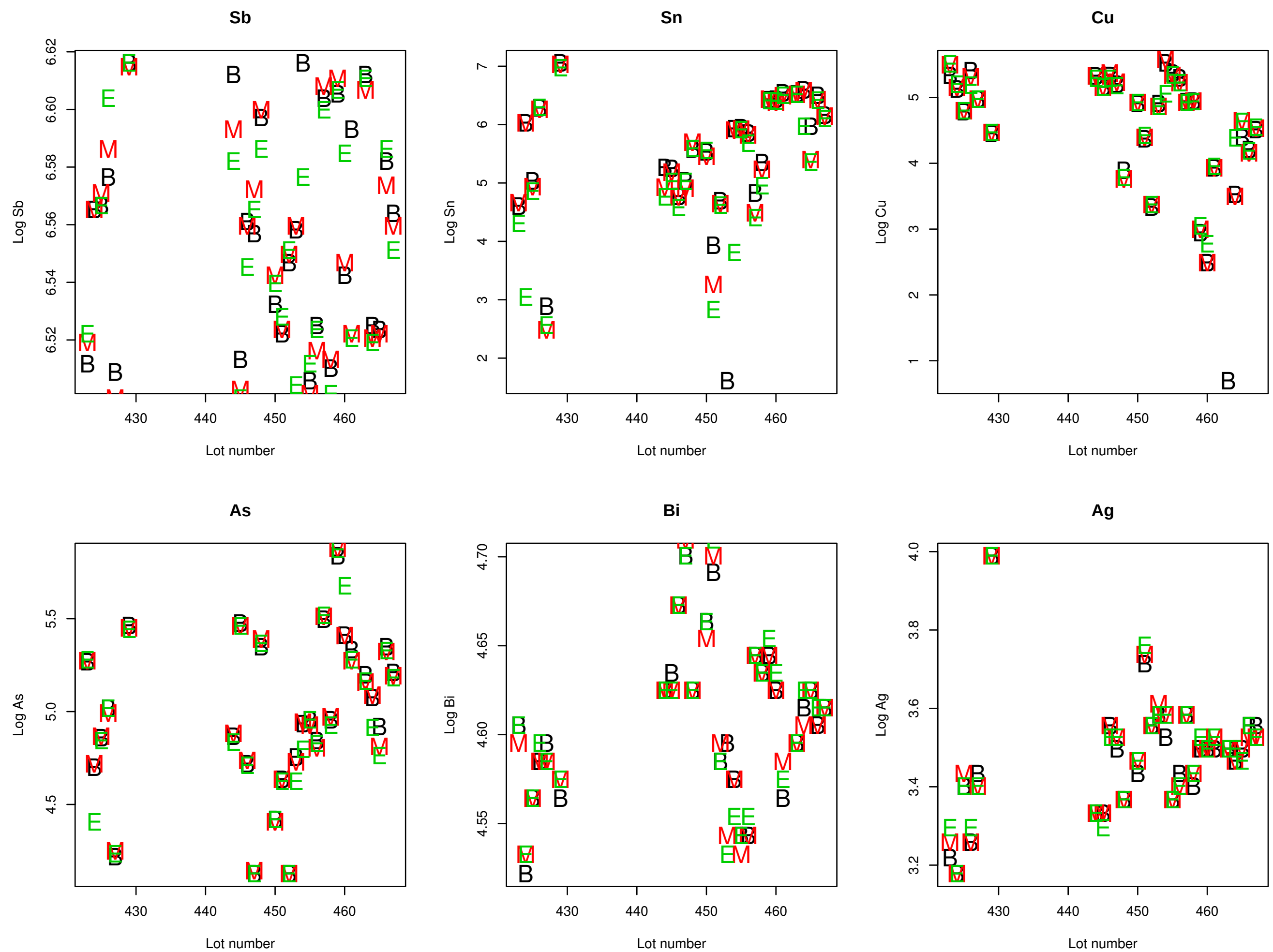
Do we need to measure all 7 elements?

Principal components analysis on 854 bullets on which all 7 elements were measured:

- Total variance (sum of eigenvalues of $X'X$ matrix): 136.944
- Percent of total variance explained with:
  - 3 elements: 83.6% (Sb, Sn, Cd)
  - 4 elements: 96.1% (Sb, Sn, Cd, As)
  - 5 elements: 97.5% (Sb, Sn, Cd, As, Cu)
  - 6 elements: 98.5% (Sb, Sn, Cd, As, Cu, Ag)
- Conclusion: Little gained by adding Bi+Ag, but little cost; need to confirm with complete 71,000+ bullet data set.

Randich et al. (2002) data set

- Erik Randich, Wayne Duerfeldt, Wade McLendon, William Tobin (2002), "A metallurgical review of the interpretation of bullet lead compositional analysis," *Forensic Science International* 127: 174–191.
- 28 lots of nominal 0.7wt.% alloy, manufactured Jan'99–Mar'00
- 3 samples per lot (Beginning, Middle, End of pour)
- Six elements (all but CD), Tables 1–2
- One reported value per sample, no standard errors
- Limited information of σ_b (between lots) compared to σ_w (within lots)
- Compare $\hat{\sigma}_w$ for these lots to FBI's σ_e


Sb
Sn
Cu
As
Bi
Ag
Log Sb
Log Sn
Log Cu
Log As
Log Bi
Log Ag
Lot number
430
440
450
460
6.52
6.54
6.56
6.58
6.60
6.62
2
3
4
5
6
7
1
4.5
5.0
5.5
4.55
4.60
4.65
4.70
3.2
3.4
3.6
3.8
4.0

Variation among "B", "M", "E" consistent with FBI measurement error variation, random ordering

Comparing within-bullet, within-lot variances

	NAA-As	ICP-Sb	ICP-Cu	ICP-Bi	ICP-Ag
between lots: Randich	4981.e-04	40.96e-04	17890e-04	60.62e-04	438.5e-04
within bullet: 800-bullet set	26.32e-04	4.28e-04	4.73e-04	18.25e-04	20.88e-04
within lot: Randich et al.	31.32e-04	3.28e-04	8.33e-04	0.72e-04	3.01e-04
within lot to within bullet	1.2	0.8	1.8	0.04	0.14

Koons and Grant (2002) data set

- R.D. Koons, D.M. Grant (2002), "Compositional variation in bullet lead manufacture," *J. Forensic Sci.* 47: 950–958
- Homogeneity within a "melt"
- 2 smelters (+ 1 ammunition manufacturer), Sep'97–Aug'99
- Lead poured into disks (7.4cm diameter; 1.6cm thick)
- Subdivided into 3 or 5 vertical layers, 3 wedges
- Reported mean & SD, triplicate measurements
- Limited information of σ_b compared to σ_w
- Most CVs $< 2\%$, except with low concentrations
- Measurement error $\approx$ 3–5%

Pour date	AsAv	SbAv	SnAv	BiAv	CuAv	AgAv
09.06.97	61.5	0.9	25.0	0.4	1.4	1.5
08.13.99	10.3	0.6	28.6	0.5	1.0	1.1
01.04.99	29.4	0.6	9.1	0.3	0.8	0.9
04.09.98	46.2	1.2	20.0	1.0	2.3	1.3
11.29.98	NA	0.9	76.9	0.8	0.9	1.4
10.07.97	81.8	2.1	1.4	1.9	3.4	2.8
10.30.98	1.3	0.8	14.3	0.4	1.3	1.3
11.22.97	0.8	0.0	100.0	0.1	1.4	0.8
07.03.99	0.8	0.7	1.7	0.7	0.9	0.6
08.12.99	0.8	0.6	20.0	0.9	0.8	0.8
11.04.97	0.2	0.3	100.0	0.4	0.3	0.4
10.21.97	11.1	0.8	0.8	1.2	1.7	0.8
01.02.98	62.5	0.6	100.0	1.0	1.0	0.8
02.22.99	2.2	1.7	33.3	1.4	2.7	0.2
02.01.99	0.9	0.5	100.0	0.7	0.7	0.8

Conclusions:

- σ_w (homogeneity within "CIVL") $\approx$ or $< \sigma_e$ (FBI measurement error)
- Concentration distributions $\approx$ lognormal
- Measurement error SDs can be pooled
- Measurement errors are not uncorrelated
- No reliable information on "within-batch" vs "between-batch" variances
- No "honest" data set from which to compute false positive probability of match using FBI 2-SD or range overlap methods, so resort to simulation

False Positive Probability:

Given the difference between mean concentrations of the CIVLs from which the bullets were manufactured (δ) and the measurement error variation (σ_e), what is P{match $\mid \delta, \sigma_e$}?

The practical issue:

Given that the 2-SD-overlap test claims "match", what is the probability that the two bullets really did come from the same or different sources?

"Likelihood" that two bullets came from CIVLs whose mean difference is no more than δ:

$$\text{Prob}\{|\mu_x - \mu_y| < \delta | \text{ 2-SD-overlap "match" }\}$$
$$\text{Prob}\{|\mu_x - \mu_y| > \delta | \text{ 2-SD-overlap "match" }\}$$

Practical issue (cont'd):

- Bayes rule $\Rightarrow$ need to know Prob$\{\delta\}$; i.e., typical δ's (distribution of distances between bullets, SD $= \sigma_b$).
- Depends upon caliber, manufacturer, geographical location, ... None of the available data sets provides reliable, unbiased information.
- We don't know how often δ may be BIG or small. It depends on how "different" the sources are.

What we **can** say:

Scenario 1: Two bullets.
No tests on them, just two bullets. What is the probability they came from same source? (tiny)

Scenario 2: Two bullets.
We measure them; "2-SD-overlap" says "match".
What is the probability they came from same source? (probably higher than in scenario 1, whose bullets were never measured)

Probability that two bullets came from the same CIVL is increased by a finding that they are analytically indistinguishable, versus no evidence of match status.

6. Probative impact of matching evidence

From the *FBI Handbook of Forensic Sciences 36* (rev 1999):

"Differences in the concentrations of manufacturer-controlled elements and uncontrolled trace elements provide a means of differentiating among the lead of manufacturers, among the leads in individual manufacturer's production lines, and among specific batches of lead in the same production line of a manufacturer."

Accordingly, FBI testimony has included statements such as:

- "Could have come from the same box"
- "Could have come from the same box or a box manufactured on the same day"
- "Were consistent with their having come from the same box of ammunition"

- "Probably came from the same box"
- "Must have come from the same box or another box that would have been made by the same company on the same day"
- "had come from the same batch of ammunition: they had been made by the same manufacturer on the same day and at the same hour"
- "likely originated from the same manufacturer's source (melt) of lead"
- "The specimens within a composition group are analytically indistinguishable. Therefore, they originated from the same manufacturer's source (melt) of lead."

(*NAS Report, pp. 91–92*)

Such testimony grossly overstates probative impact of "matches"; can say only that two bullets that "match" may have come from sources with the same chemical composition — NOT that they came from the same box, wire, ingot, source, or melt.

Finding: The available evidence do not support any statement that a crime bullet came from, or is likely to have come from, a particular box of ammunition, and references to "boxes" of ammunition in any form are seriously misleading under Federal Rule of Evidence 403. Testimony that the crime bullet came from the defendant's box or from a box manufactured at the same time is also objectionable because it may be understood as implying a substantial probability that the bullet came from defendant's box.

7. Alternative analyses

(a) Equivalence tests using a series of t statistics:

$$|\bar{X} - \bar{Y}| < K s_{pool} \cdot \sqrt{2/3}$$

- $\bar{X}$, $\bar{Y}$ are the means of the **logarithms** of the three measurements on the CS and PS bullets
- s_{pool} is the root mean square of the standard deviations on the logs of the three measurements from **many** bullets
- Measurement SDs should be monitored (control chart; e.g., Vardeman and Jobe 1999)
- Equivalence test hypotheses:
  H_0: means differ by more than δ_0 (no match)
  H_1: means differ by less than δ_0 ("match")

- K (critical point) depends upon: chosen FPP (α), $n =$ sample size (here, 3 per bullet), $\nu =$ degrees of freedom in s_p, "limit" of "equivalence" δ_0, power at a value $\delta_1 < \delta_0$
- For large ν, K can be solved from:

$$\Phi(K - \delta_0/(s_p\sqrt{2/3})) - \Phi(-K - \delta_0/(s_p\sqrt{2/3})) = \alpha$$

$\Phi(\cdot) =$ cumulative standard Gaussian distribution

For small FPPs (< 0.05) and small δ_0 ($< 2\sigma_e$), $K < 2$.

Some values of K for $\alpha = 0.25$, n=3 ($\alpha^5 \approx 0.001$):

Values of δ_0/σ_e

	0.25	0.33	0.50	1	1.5	2	3
df= 3	0.3577	0.3703	0.4109	0.6814	1.1924	1.7741	2.9155
20	0.3363	0.3482	0.3866	0.6481	1.1690	1.7730	2.9816
50	0.3348	0.3466	0.3848	0.6458	1.1676	1.7742	2.9922
100	0.3344	0.3462	0.3843	0.6450	1.1672	1.7746	2.9960

(mean difference/pooled SD) $< K \cdot \sqrt{2/3}$

Gaussian probabilities of false match, false non-match

(b) Multivariate Equivalence Hotelling's T^2 test:

$$H_0 : |\mu_x - \mu_y| \geq \boldsymbol{\delta_0}$$

$$H_1 : |\mu_x - \mu_y| < \boldsymbol{\delta_0}$$

Assume:

$$\mathbf{X_i} \sim N_7(\mu_x, \Sigma), \quad \mathbf{Y_i} \sim N_7(\mu_y, \Sigma)$$

Test: Reject if

$$(\bar{\mathbf{X}} - \bar{\mathbf{Y}})'\hat{\Sigma}^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) < \frac{7\nu}{(\nu - 6)} F_{7,n-7}(ncp) \text{ (noncentral F)}$$

Hotelling's T^2 and 7 t tests:

- $\mathrm{P}\{T^2 < F_{crit} \mid |\mu_x - \mu_y| \geq \delta_0 \cdot \mathbf{1}\} = \alpha$
- t-statistics:
  $\mathrm{P}\{\ |t_j| < K \mid |\mu_x - \mu_y| > \delta_0\ \} = \alpha_j \equiv \alpha$
  (given α, K comes from noncentral Student's t distribution, or Gaussian approximation above)
- $\mathrm{P}\{$ all t-statistics $< K \mid |\mu_x - \mu_y| > \delta_0\ \} \neq \alpha^7$ (correlation)
- Compare:
  $p_1 = \mathrm{P}\{\text{indt } t\text{-statistics} < K \mid \delta_0\} = \alpha^7$
  $p_2 = \mathrm{P}\{\text{corr } t\text{-statistics} < K \mid \delta_0\} = \alpha^?$
  $\Rightarrow\ ? = 7 - [\log(p_1) - \log(p_2)]/\log(\alpha)$

Simulation (100,000 trials):

K	Values of δ_0					
	0.0	0.5	1.0	1.5	2.0	2.5
0.000	5.27	5.18				
0.645	5.42	5.15	4.77			
1.167	5.68	5.44	5.27	5.23	5.20	
2.000	6.55	5.91	5.58	5.38	5.06	5.00

Overall FPP $\approx \alpha^5$

(c) Empirical Mahalanobis distances and sampling:
Alicia L. Carriquiry, Michael Daniels, Hal S. Stern, "Statistical treatment of class evidence: Trace element concentrations in bullet lead," Iowa State Technical Report, May 4, 2000 (Table 4 corrected April 19, 2002)

- Noted many of same difficulties with using CABL
- Generate empirical distributions of Mahalanobis distances for bullet pairs in same ("within"), different ("between") batches
- Very sensible strategy, **if** manufacturers would cooperate
- Slight trends in concentrations over time (Ag)

8. Committee Recommendations

- Discontinue discussion of "boxes" in court testimony
- Monitor measurement SDs using standard SQC charts
- Replace "2-SD-overlap" with "successive equivalence t-tests" using per-element α of about $(target\,\alpha)^{1/5}$ or equivalence Hotelling's T^2 using pooled covariance matrix
- Plan to analyze the 71,000+ file of FBI-measured bullets
- **Arrange for a well designed and executed experiment to estimate σ_b, σ_w.** These values will depend on bullet manufacturer, type, caliber, geographic locale.

9. Further work

- How robust is Hotelling's T^2 to misspecified Σ?
- Is Σ easily estimated by non-statisticians (monitoring pooled variances and covariances)?
- Will non-statisticians invert 7×7 covariance matrices? (in Excel?)
- "Chaining"?
- What would the 71,000+ bullet data set show?
- Issues of testing multiple CS/PS bullets