Critical Role of Statistics in Development and Validation of Forensic Methods

Karen Kafadar

Department of Statistics

Indiana University

kkafadar@indiana.edu

http://mypage.iu.edu/∼kkafadar

Acknowledgements:

Clifford H. Spiegelman, Texas A&M University

Judge H.T. Edwards, New York University

OUTLINE

1. Statistics in Science: biology, chemistry, physics, engineering, medicine (social sciences)
2. Statistics in Forensic Science to date: DNA, CBLA, Anthrax
3. Forensic science: Needs (Quality metrics, study designs)
4. Progress in fingerprint analysis (ACE-V)
5. Research agenda for forensic science (fingerprints)
6. Statisticians involvement
7. Summary

1. Statistics in Science: Examples

Science of analyzing data, characterizing uncertainties

- **Biology**: extinction/abundance of species; characterizing genetic expression (millions of SNPs) in response to stimuli; associating genotypes with phenotypes
- **Physics**: data analysis of high-energy physics (HEP) experiments to discover new particles; estimating 'big G' with uncertainty; existence of global warming
- **Engineering**: product design & development; nuclear safety programs; production efficiency
- **Medicine**: clinical trials of new drugs; evaluation of treatment and screening programs; estimating disease prevalence, incidence, spread

2. Statistics in Forensic Science to date

- DNA (NRC-2, 1996): Identification of 13 markers (presumed independent); distribution of genotypes; separating "signal" peaks (allele identification) from noise; resolving mixtures; minimizing errors in DNA lab measurements
- Comparative Bullet Lead Analysis (NRC, 2004): Statistical procedures for assessing "match" between crime scene bullet and suspect's bullets; characterizing sources of variability within/between batches of lead; error rates
- Review of Anthrax Investigation (NRC, 2009) Significance of "Smoking gun" evidence: 7 samples from USAMRIID in FBI repository of 1,070 samples; sources of variation

DNA analysis (NRC 1996: "DNA-2")

- "DNA-1" (NRC 1992) lacked statistical credibility
- "DNA-2" (NRC 1996): Statisticians' participation
- Marker selection: sensitivity (how well alleles make correct id), specificity (how well alleles distinguish individuals)
- 13 core loci (L_j, $j = 1, ...13$), each with 6–21 alleles (k_j alleles, frequency $> 0.01 \Rightarrow n_j \approx k_j(k_j + 1)/2$ genotypes at each loci)
- Calculate probabilities of "match" at 13 (independent) loci if samples come from different sources
- "Independence": Assume outcome (genotype ID) at marker location i is *independent* of outcome at marker location j
- Use CODIS database to verify "independence" assumption?

	CSF1PO	FGA	TH01	TPOX	vWA
#alleles	8	21	6	7	9
#genotypes	36	231	21	28	45

	D3S1358	D5S818	D7S820	D8S1179
#alleles	8	8	8	10
#genotypes	36	36	36	55

	D13S317	D16S539	D18S51	D21S11
#alleles	7	7	15	17
#genotypes	28	28	120	153

Why DNA analysis is a successful forensic method:

- Well-defined markers (not just any 13 loci)
- HIGH sensitivity: P{'match' | samples from same source}
- HIGH specificity: P{'no match' | different sources}
- $\Rightarrow$ HIGH Positive/Negative Predictive Value:
  PPV = P{samples came from same source | 'match' call}
  NPV = P{samples came from different sources | 'no match'}
- Well-designed experiments to validate performance
- Careful analysis of experimental data on performance
- Well-defined procedures for execution
- Clear guidelines for interpreting/reporting results

Statistics involved in all steps

Challenges ahead:

- Managing CODIS database (millions of records)
- Resolving mixtures
- Missing data (allelic drop-out/drop-in, etc.)
- Different distributions by gender / ethnic / racial groups
- Validation of independence assumptions
- Robustness in calculations of correlations
- Multiplicity of estimates and uncertainties
- Lab testing process improvement (reduce errors)

Scientific Method: Continuously update knowledge

CBLA (NRC 2004)

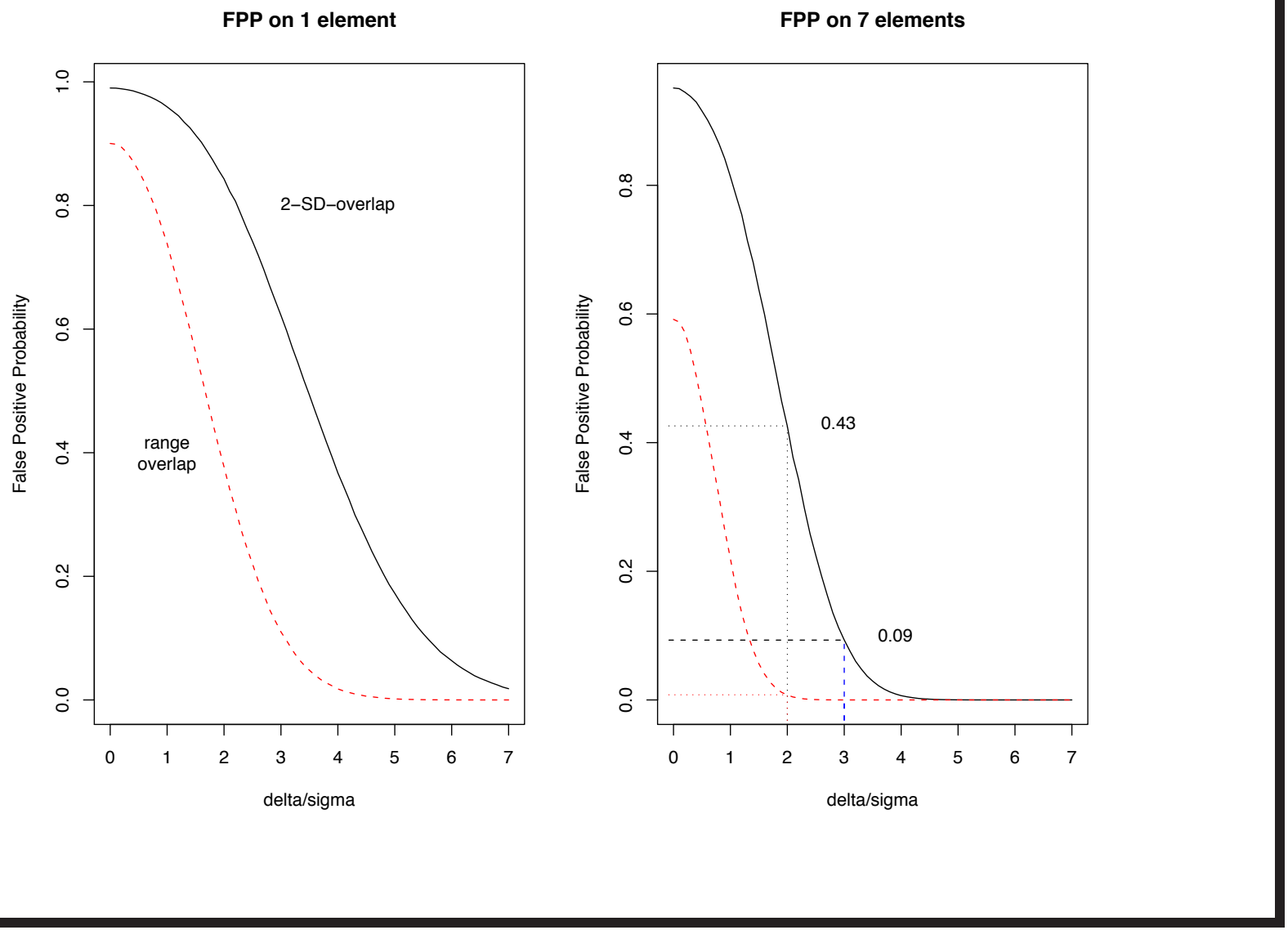
Scenario:

- Crime → evidence → bullets
- Gun recovered: match striations on bullet and gun barrell (separate NRC committee)
- *No gun*: **Comparative Bullet Lead Analysis** (CBLA)
- "Working hypothesis": chemical concentration of lead used to make "batch" of bullets provides "unique signature" ⇒ "equal" concentrations of elements in Crime Scene (CS) bullets and Potential Suspect (PS) bullets may indicate "guilt"
- Local police dept sends CS, PS bullets to FBI lab
- FBI measures (in triplicate) concentrations of 7 elements

- Reports "analytically indistinguishable concentrations" between CS and PS bullets if "mean ± 2·SD intervals overlap for *all* 7 elements" (**2-SD-overlap**), provides court testimony when requested (As, Sb, Sn, Bi, Cu, Ag, Cd)
- FBI "validates" process on "1837-bullet database": "*one specimen from each combination of bullet caliber, style, and nominal alloy class was* **selected**" for database; found 693 "matches" out of (1837·1836/2) = 1,686,366 pairs of bullets
- i.e., **bullets selected to be different (not representive)**, so actual false probability rate is higher than 0.04%

Statisticians on NRC Committee (report, 2004):

- "hypothesis" *bullets came from same box*:
  not sensible (manufacturing process: bullets from different batches in same box, bullets from same batch in many boxes)
- "hypothesis" *bullets came from same batch*:
  feasible (two-sample test on means) — but not probative?
- "Innocent until proven guilty" $\Rightarrow$
  H_0: $|\mu_{CS} - \mu_{PS}| > \delta$, H_1: $|\mu_{CS} - \mu_{PS}| \leq \delta$
- Proper test: Hotelling's T^2, not "2-SD overlap"
- Historical data to estimate $\Sigma \Rightarrow$ *correlated* measurement errors
- Simulations $\Rightarrow$ "2-SD-overlap" false positive rate $> 0.04\%$


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

Anthrax investigation (NRC 2011)

Sep-Oct 2001: Anthrax letters mailed to NYC (ABC, CBS, NBC*, NYPost*), FL (AMI), DC (Daschle*, Leahy*)

- 4 morphotypes of specific anthrax *Ames* strain found in Leahy* letter (A1, A3, D, E)
- 5 assays (present/absent); 2 for D (D_M, D_I)
- Feb'02: FBI subpoenas labs for samples of *B. anthracis-Ames*
- 1,070 samples in FBI Repository, *believed* complete
- "Smoking gun": Only 8 samples showed all 4 morphotypes; 7 from one lab at USAMRIID, 8th sent to BMI from that lab
- Inference: "Anthrax came from that lab"

"Statistics means never having to say you're certain"

- 1,070 samples came from 20 labs (17 U.S.)
- 11 samples not viable $\Rightarrow$ 1,059
- Lab-to-lab variation since "D" assayed by 2 labs
  $\Rightarrow$ Concordance: 975/1059 = 0.921 (0.903, 0.937) (not 1.000)
- Ignored D_I for vague reasons
- 947 samples had "conclusive" measurements A1,A3,D_M,E
- One suspect sample assayed 30 times $\Rightarrow$ measurement varibility: 16 of 30 reps showed all 4 morphotypes
- Dilution studies: sudden "appearance" of morphotype at higher dilution rates after disappearance at lower dilution rates

Distribution of #samples by Lab:

F	S	N	P	T	G	E	H	Q	A
598	74	62	50	49	31	24	18	15	6

J	K	I	M	O	R	B	C	D	L	F*
4	3	2	2	2	2	1	1	1	1	1

One Lab F submitted 598 samples (63%)

$\Rightarrow$ P{7 or 8 from Lab F} = 0.14 (hypergeometric distn)

Not an everyday occurrence, but certainly not rare.

3. Forensic Science: Needs

- Development of quality thresholds for evidence
- Construction of objective analysis steps
- Design of validation studies for process
- Design of reliability studies:
  repeatability, sensitivity, specificity
- Appropriate data analysis, report, access
- Focus: Fingerprint analysis (ACE-V)

Development of quality thresholds for evidence and objective analysis

- **A**CE-V: **Assessment**: Quality of latent print
- Sufficiently discernible minutiae (*subjective*)
- Working premise: better print $\Rightarrow$ correct inferences
- More *objectivity* ensures greater repeatability
- A**C**E-V: **Comparison**: Examiner identifies points of similarity (minutiae); subjective but can be done by AFIS

Design of validation studies for process

- **Accuracy**: Does method provide accurate results?
- **Precision**: Does method provide consistent results?

Unrealistic validation study: "50K fingerprint study"

- One, and only one, true match
- Process: "find the closest match of the 50,000"
- Prints were from 10-print cards (high quality)
- One examiner; one AFIS system
- "Latent" print = 25% of a "perfect" print

Design of reliability studies:
repeatability, sensitivity, specificity

- Identify relevant populations (examiners, prints, etc.)
- **Randomly** select units from populations
  (practical issues: participation, data access, ...)
- Include factors as possible sources of variation
- **Double-blind** (cf. clinical trials)
- Incorporate repetition (same examiners/prints)
- Include both true matches & true non-matches

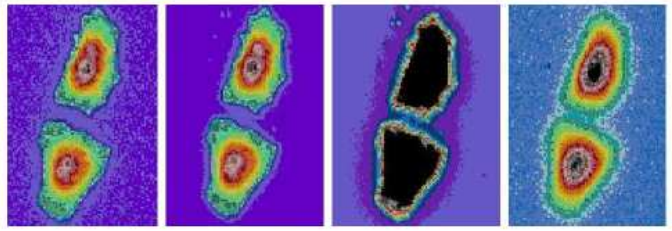
Box, Hunter, Hunter, *Statistics for Experimenters*, 2005

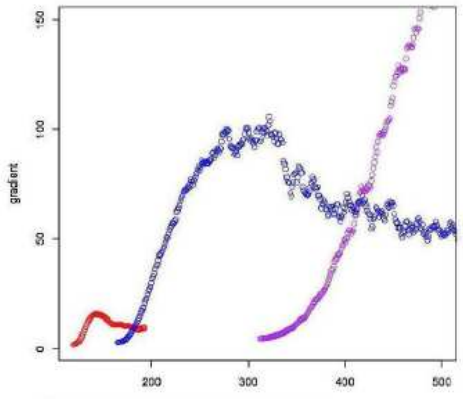
4. Progress in ACE-V fingerprint analysis

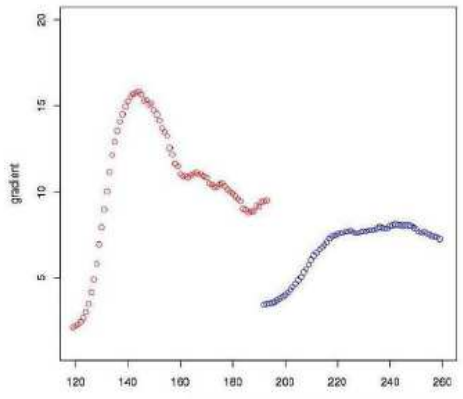
- 'A': Quality standards (Yoon et al. 2011, Peskin et al. 2012)
- 'C': AFIS-identified minutiae
- 'E': Likelihood ratios (Neumann et al. 2011)
- 'V': Sources of variation in verification stage

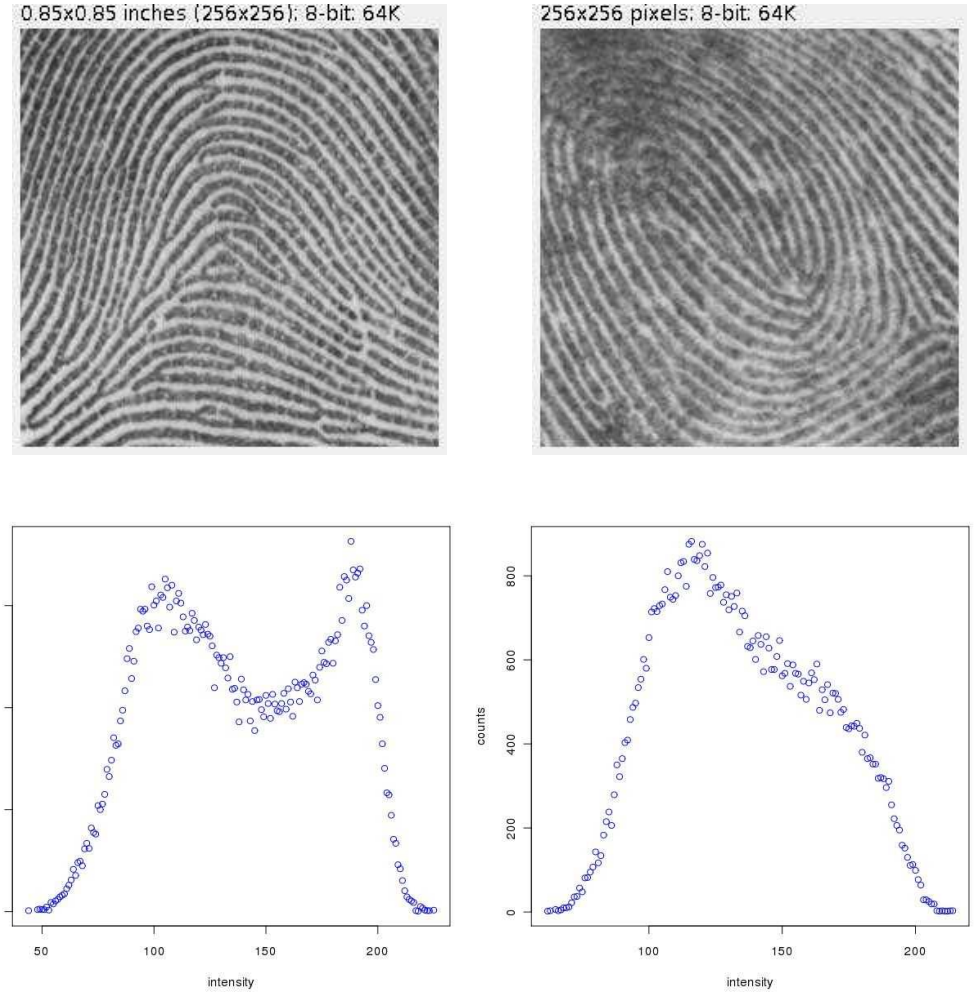
Peskin et al. (2010): Measure cell image quality (NIST)

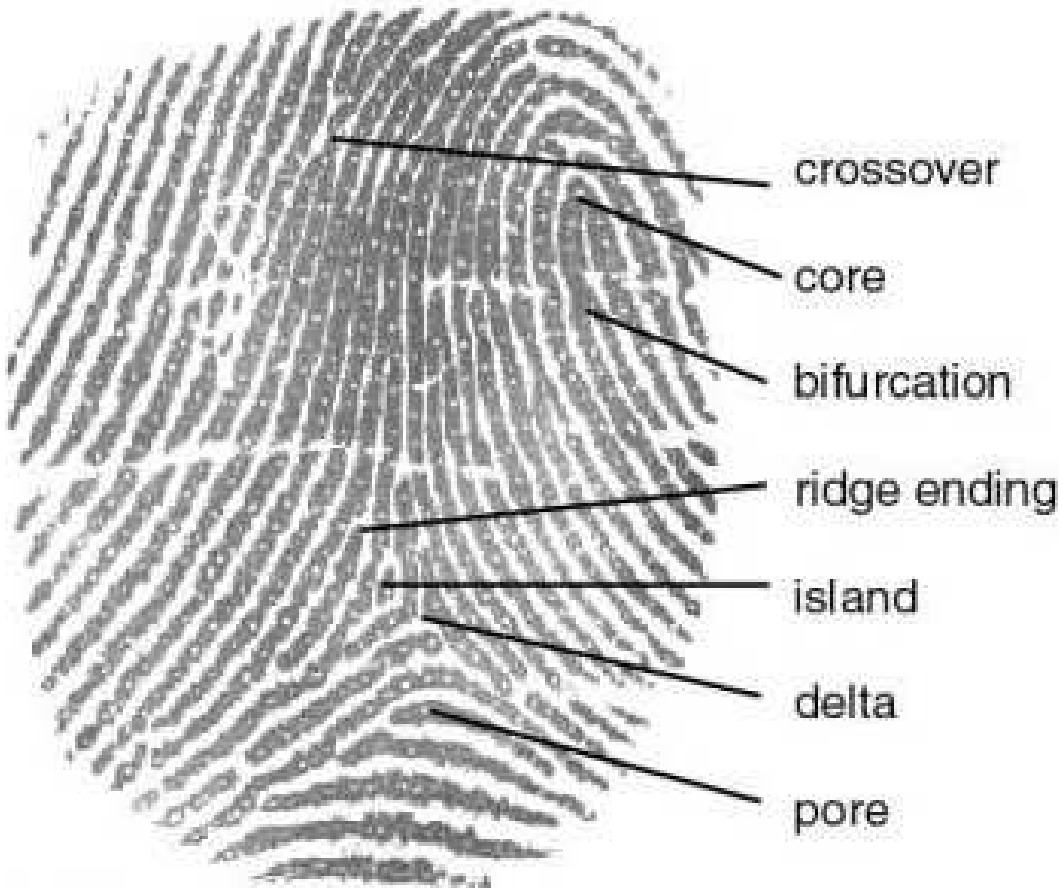
- Choice of segmentation based on edge quality
- Calculate numerical gradients from cell to background
- Sharper peak corresponds to clearer images
- Apply concepts (gradient, contrast) to fingerprint images
- Simulate increasingly degraded print via blurring; fingerprint quality score decreases accordingly
- **How does quality score relate to accuracy of identification?** (*quality threshold*)
- Yoon et al. (2011) for alternative quality metric

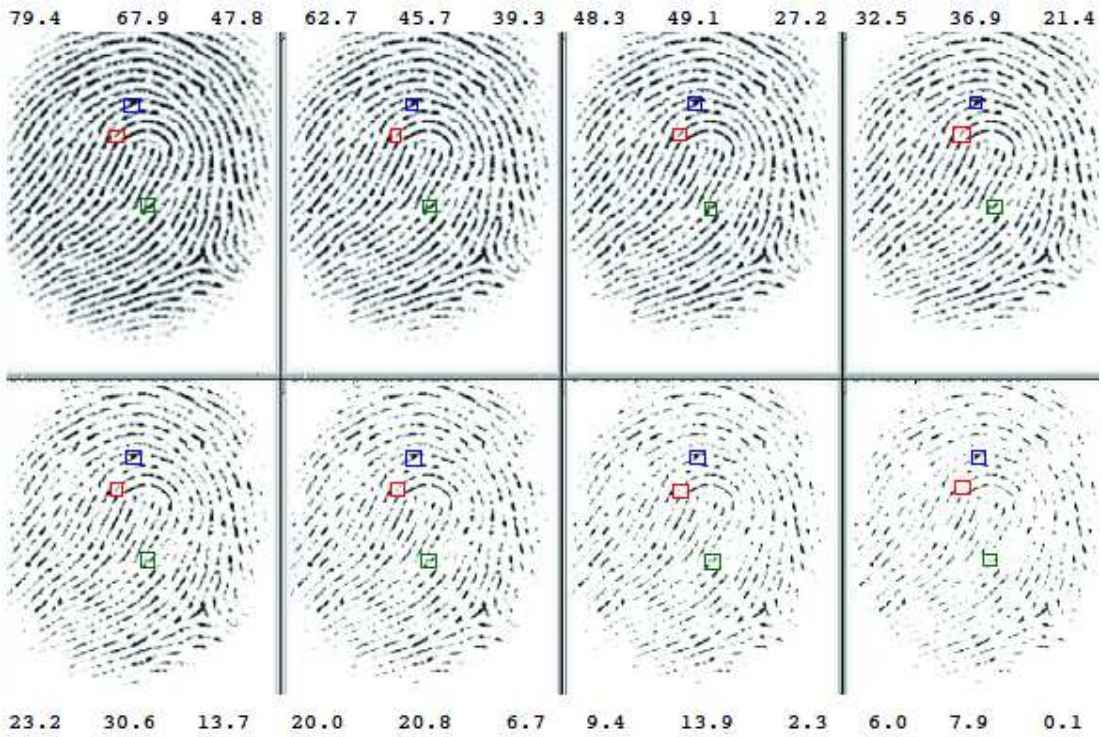


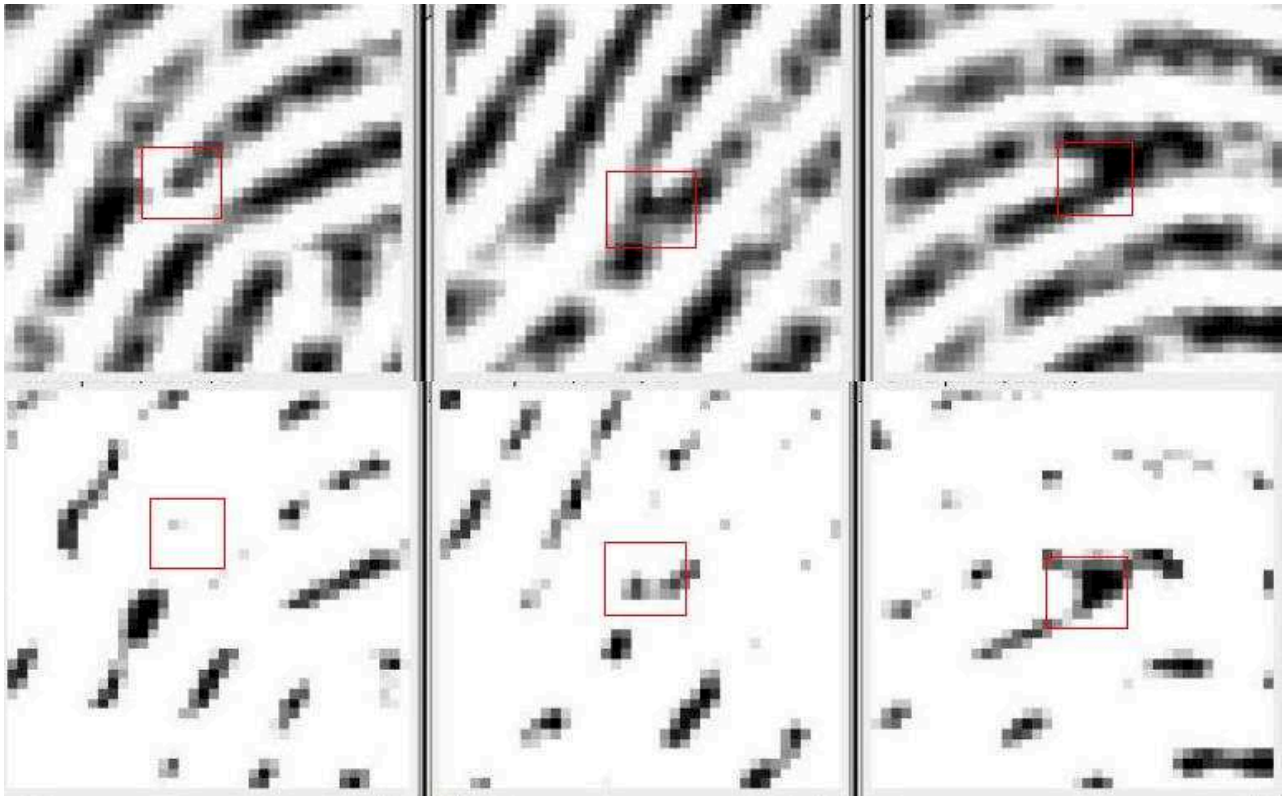

gradient
0
50
100
150
200
300
400
500


gradient
5
10
15
20
120
140
160
180
200
220
240
260


0.85x0.85 inches (256x256); 8-bit; 64K
256x256 pixels; 8-bit; 64K
counts
intensity
0
200
400
600
50
100
150
200
counts
intensity
0
200
400
600
800
100
150
200


crossover
core
bifurcation
ridge ending
island
delta
pore


79.4 67.9 47.8
62.7 45.7 39.3
48.3 49.1 27.2
32.5 36.9 21.4
23.2 30.6 13.7
20.0 20.8 6.7
9.4 13.9 2.3
6.0 7.9 0.1



Quality scores for 3 minutiae in 8 increasingly degraded images:

Image#	left (red)	center (blue)	right (green)
1	79.4	67.9	47.8
2	62.7	45.7	39.3
3	48.3	49.1	27.2
4	32.5	36.9	21.4
5	23.2	30.6	13.7
6	20.0	20.8	6.7
7	9.4	13.9	2.3
8	6.0	7.9	0.1

Objective analysis steps: AC**E**-V (Evaluation)

- Subjective identification of corresponding minutiae
- Neumann et al. (2011): Calculate for each minutiae: radius, side length, angle, area, type (5-dimensional vector)
- Distribution of features in matching vs non-matching prints
- "Likelihood ratio":

$$LR = \frac{P\{\text{`}match'|dif<\delta\}}{P\{\text{`}match'|dif>\delta\}}$$

- Objective comparison *if minutiae ID were objective*
- Note LR is not "real life"; need *posterior odds*
- Posterior odds = Prior odds × LR

$$\frac{P\{dif<\delta|\text{`}match'\}}{P\{dif>\delta|\text{`}match'\}} = \text{Prior odds} \times \text{LR}$$

5. Research Agenda for Forensic Science

- Associate quality metrics & outcomes (match/non-match)
- Use of databases for model development
- Dependence of process on minutiae identification
- Experimental design for estimating process metrics (*Sensitivity*, *Specificity*)
- Apply to other forms of pattern evidence; e.g., tool mark (*Ballistic Imaging*, NRC 2008, Ch.3: `www.nap.edu`)

Designing a better study of fingerprint accuracy:

- Identify population of examiners, latent prints
- Randomly select examiners and prints for study
- Construct test sets (includes matches, non-matches)
- Include factors: print quality, digitization, #minutiae, AFIS system, lab, repetition, ...
- **Double-blind**: Neither examiner nor administrator knows
- Collect concomitant information (experience, #points, ...)
- Obtain information on non-responders

Randomize to balance any factors not considered in design (e.g., presentation of sets to examiners)

Barriers: Culture, Confidentiality, Costs, Data Access

6. Statistician involvement

- Finding statisticians
- Interdisciplinary education programs
- Internships

We still have a long way to go ...

- Training statisticians in design, process control
- Cultural awareness of need for improvement
- Sources of funding for internships
- **Education of courtroom personnel**

Judge Richard Posner for 7th Circuit U.S. Court of Appeals (No. 11-2894, Jan 8, 2013):

"*Matching ... fingerprint evidence, is less rigorous than the kind of scientfic matching involved in DNA evidence ... But expert evidence is not limited to "scientific" evidence ... It includes any evidence created or validated by expert methods and presented by an expert witness that is shown to be reliable ... Ultimately the matching depends on "subjective judgments by the examiner," ... but responsible fingerprint matching is admissible evidence, in general and in this case.*" [pp 11-14]

In other words,

- fingerprint matching is admittedly a subjective process;
- an "Expert" can claim that evidence has been "created or validated by expert methods" and claimed, as an expert witness, that it "is shown to be reliable";
- so, "fingerprint matching is admissible evidence" (even if reliability studies do not exist).

7. Summary

- Methods & standards in forensic **science** must be developed by **scientists with statisticians**, not by law enforcement
- Collaborations between forensic scientists and statisticians can bear fruit, as seen with DNA analysis, CBLA, Anthrax investigation, and some advances in fingerprint analysis
- Many more areas in need of research (e.g. tool marks)
- Need education of law enforcement and court officials

From Judge H.T. Edwards' testimony to Senate Judiciary Committee 18 Mar 2009:

"*Although the report offers no proposals for law reform, the committee believes, that with more and better educational programs, mandatory accreditation and certification, sound operational principles and procedures, and serious research to establish the limits and measures of performance in each discipline, forensic science experts will be better able to analyze evidence and coherently report their findings in the courts.*"

References

Box, G.E.P.; Hunter, W.F.; Hunter, J.S. (2005), *Statistics for Experimenters, 2nd ed.*, Wiley.

Bradford, T; Ulery, R; Hicklin, A; Buscaglia, J; Roberts, M (2011), "Accuracy and reliability of forensic latent fingerprint decisions, 108 *PNAS*, `http://www.pnas.org/content/108/19/7733.full.pdf`.

Cole, S. (2005), "More than Zero: Accounting for Error in Latent Fingerprint Identification," *The Journal of Criminal Law and Criminology* 96(3):985–1078.

Haber L.; Haber R.N.: Scientific validation of fingerprint evidence under Daubert, *Law, Probability, and Risk* 7(2):87–109 (2008).

Kafadar, K: Statistical Issues in Assessing Forensic Evidence, *International Statistical Review* (in revision)

Mearns, G.S. (2010): The NAS report: In pursuit of justice, *Fordham Urban Law Journal*, December 2010.

National Research Council (1996), *The Evaluation of DNA Evidence*, National Academies Press, ISBN-10: 0-309-12194-9.

National Research Council (2004), *Forensic Analysis: Weighing Bullet Lead Evidence*, National Academies Press, ISBN-10: 0-309-09079-2.

National Research Council (2008), *Ballistic Imaging*, National Academies Press, ISBN-10: 0-309-11724-0.

National Research Council (2009), *Strengthening Forensic Science in the United States: A Path Forward*, National Academies Press, ISBN-10: 0-309-13135-9.

Neumann, C.; Evett, I.W.; Skerrett, J. (2011), "Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm," *J. Royal Statistical Society A*175(2), 1–26.

Peskin et al (2010), A Quality Pre-Processor for Biological Cell Images

Peskin, A.P.; Kafadar, K.: A new measurement of quality for minutiae in latent fingerprints, preprint.

Spiegelman C.S.; Kafadar K (2006), Data Integrity and the Scientific Method: The Case of Bullet Lead Data as Forensic Evidence, *Chance* 19(2):17–25.

Ulery B.T.; Hicklin, R.A.; Buscaglia, J.; Roberts, M.A. (2011), Accuracy and reliability of forensic latent fingerprint decisions, *Procedings of the National Academy of Sciences* 108(19), 7733–7738.

Ulery B.T.; Hicklin, R.A.; Buscaglia, J.; Roberts, M.A. (2012), Repeatability and reproducibility of decisions by latent fingerprint examiners, *PLoS ONE* 7(3).

Yoon, S; Liu, E; Jain, A.K. (2012), "On latent fingerprint image quality," *Proceedings of the Fifth International Workshop on Computational Forensics*, Tsukuba, Japan, 11 Nov 2012.

Zabell, S.L. (2006), "Fingerprint Evidence", *Journal of Law and Policy*, 143–179