The Role of Statistics in Forensic Science: Past, Present, Future

Karen Kafadar

Department of Statistics

University of Virginia

kkafadar@virginia.edu

http://www.stat.virginia.edu/faculty.shtm

Acknowledgements:

Adele Peskin, NIST-Boulder; CSAFE (NIST)

OUTLINE

1. Where have statistical methods been useful in science?
2. Past: A statistical success in Forensic Science:
   Interpreting DNA evidence (NRC 1996)
3. Where statistics *might have been used* in Forensic Science
   CABL (NRC 2004), Anthrax (NRC 2009), EWI (NRC 2014)
4. Present: Where statistics *is being used*
   LRs for latent prints, Interpretation
5. Future: Where statistics *can be used* in Forensic Science
6. From Research to Implementation
7. Final comments: Broad Role of Statistician

1. Statistical Methods in Science

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
- **Chemistry**: discovery of argon (Lord Rayleigh); source attribution via MSMS (mass spec); environmental contamination levels (San Juan River contamination)
- **Physics**: data analysis of high-energy physics (HEP) experiments to discover new particles; estimating 'big G' with uncertainty; global warming (IPCC)
- **Medicine**: clinical trials of new drugs; evaluation of treatment and screening programs; estimating disease prevalence, incidence, spread

2. A statistical success in Forensic Science: Interpreting DNA evidence

- "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)
- Well-characterized error rates:
  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; process control
- Clear guidelines for interpreting/reporting results

Statistics involved in all steps

Challenges ahead:

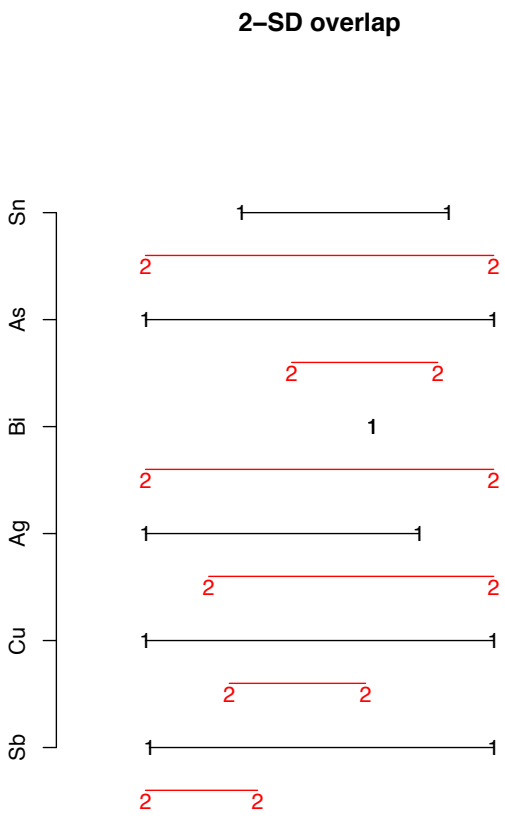
- Managing CODIS database (millions of records)
- Resolving mixtures
- Missing data (allelic drop-out/drop-in, etc.)
- Different distributions by gender / ethnic / racial groups
- Investigating independence assumptions
- Robustness in calculations of correlations
- Multiplicity of estimates and uncertainties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

- **Statisticians working with geneticists**
  (D.P. Byar: "A statistician working alone is a statistician making mistakes")
- **We do *not* expect that *all* forensic methods will have the same high accuracy as DNA**
- **We *do* expect that statistics can *contribute* to *characterizing* sources of uncertainty in the methods and begin to *quantify* their effects on accuracy**
- **Statisticians must work with Forensic Scientists**
- **Goal of Scientific Method: Continuously update knowledge**
- Another success: Use of statistical methods to assess reliability of polygraph evidence (NRC 2003; Fienberg)

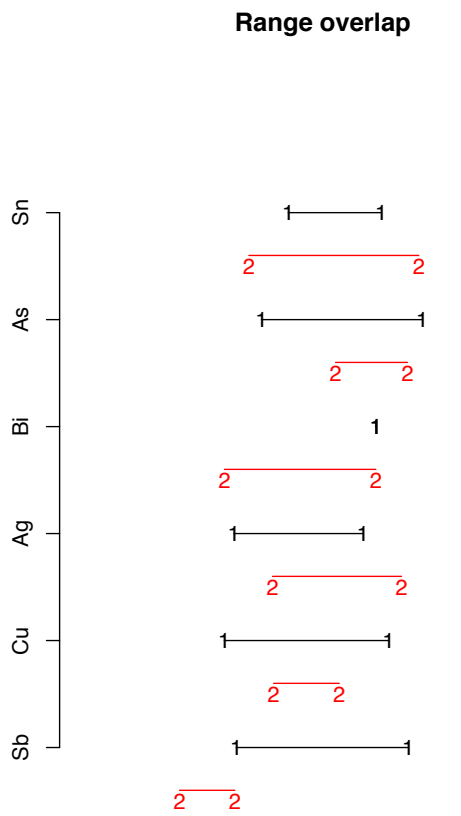
Where Statistics might have been used in Forensic Science: CBLA

- Scenario: Crime → evidence → bullets
- Gun recovered: match striations on bullet and gun barrell (separate problem: CHS, ALC)
- *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%



‘Analytically indistinguishable’



All elements analytically indistinguishable except Sb

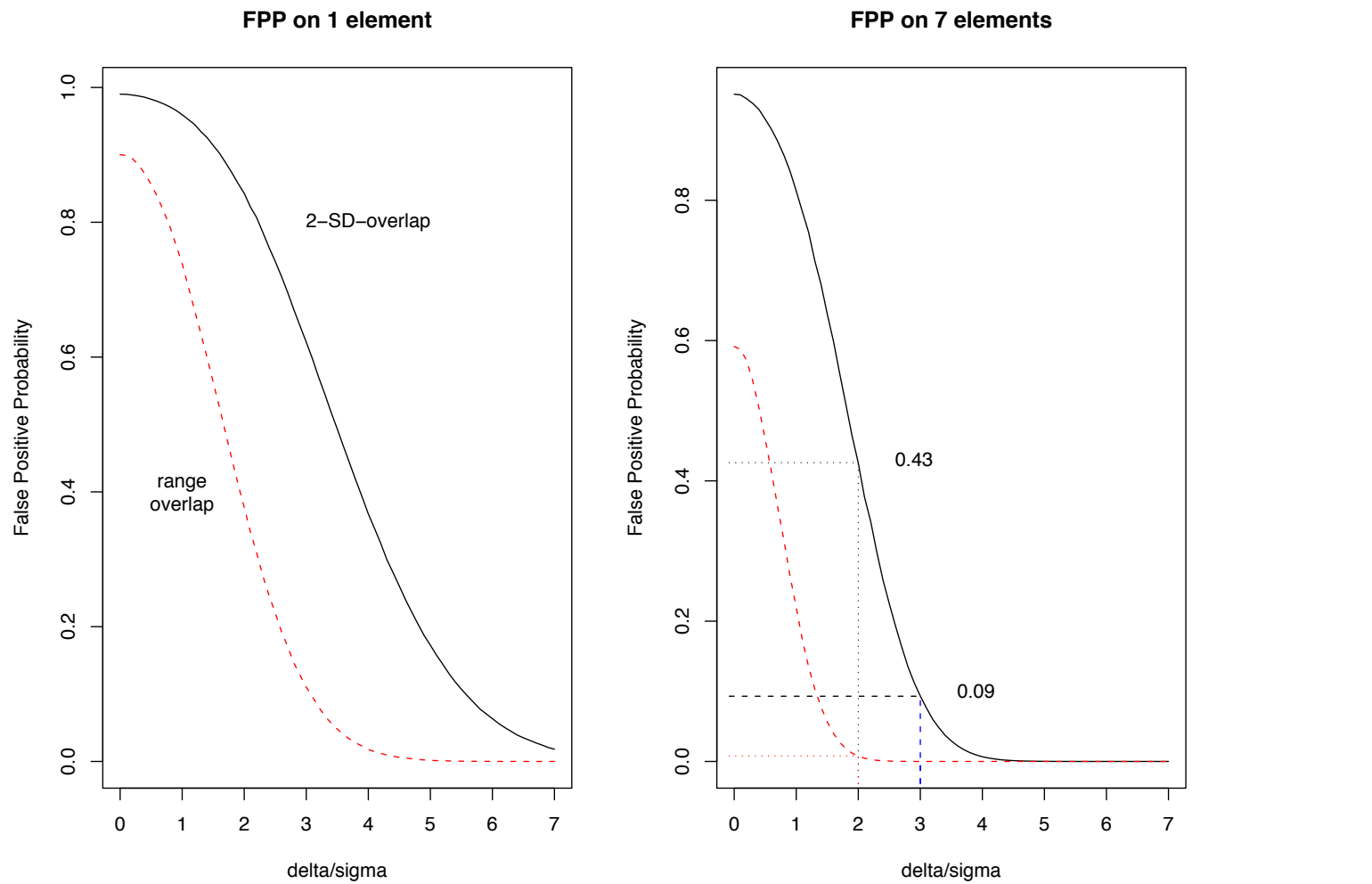
Statisticians on NRC Committee (NRC 2004)

- "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?
- FBI's "error rate": **selected** 1837 bullets from "70,000-bullet database" (17,000?) **to be as different as possible**
- FBI found only 693 "matches" out of $(1837 \cdot 1836/2) = 1{,}686{,}366$ pairs of bullets (0.04%)
- Simulation demonstrated otherwise: Suppose difference in concentrations in all 7 elements is x; what is the probability of the 2-SD test claiming a match?

- "Innocent until proven guilty" $\Rightarrow$
  H_0: $|\mu_{CS} - \mu_{PS}| > \delta$, H_1: $|\mu_{CS} - \mu_{PS}| \leq \delta$
  (equivalence testing)
- Proper test: Hotelling's T^2, not "2-SD overlap"
- Historical data $\Rightarrow$ *correlated* measurement errors
- Simulations $\Rightarrow$ "2-SD-overlap" false positive rate $> 0.04\%$

Estimated correlation matrix (200 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


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

Using FBI "2-SD-match" criterion:
How often do bullets from different boxes "match"?

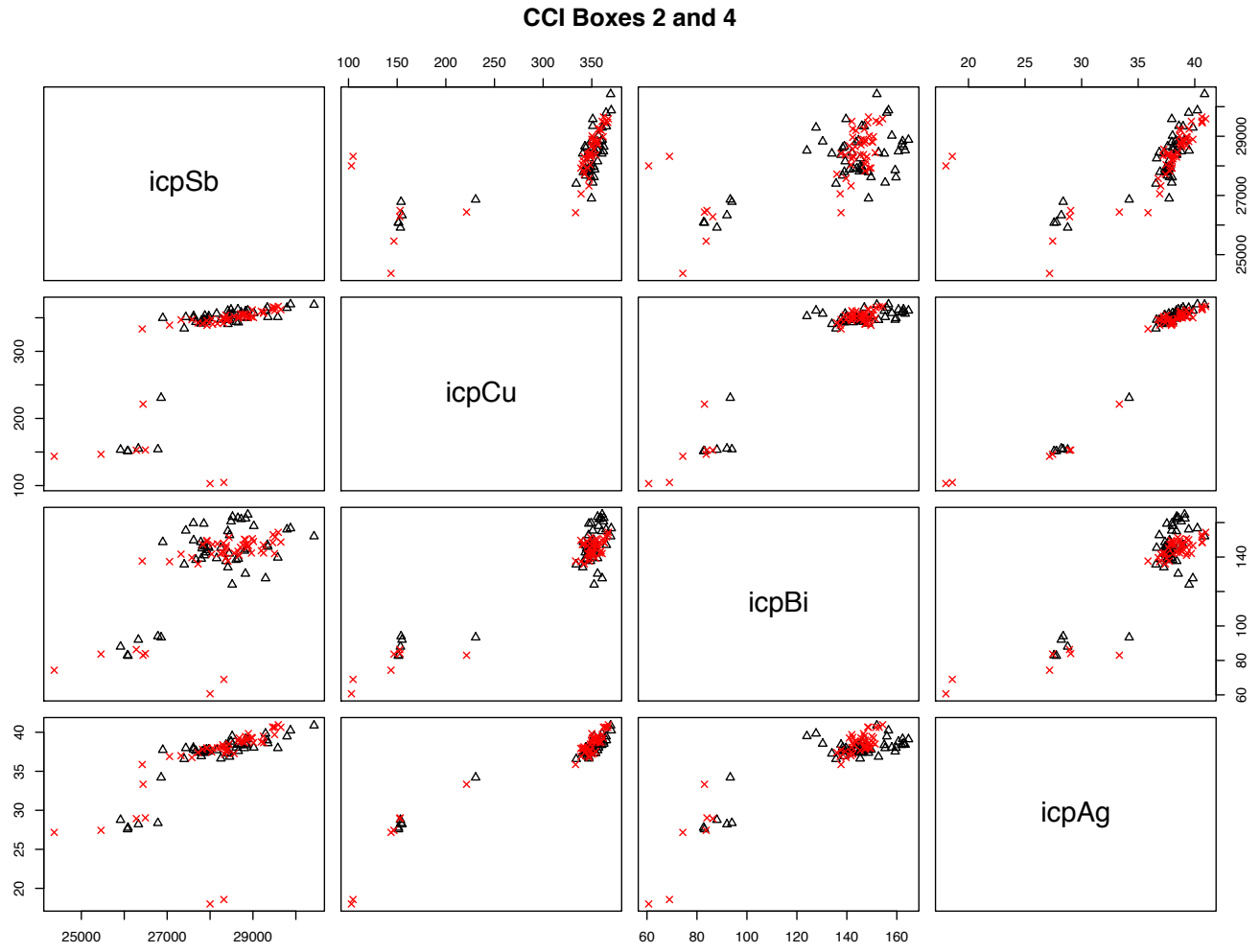
Ex: CCI bullets – 4 boxes, 50 bullets per box

Sometimes FBI-"matches" are rare:

- Box 1 with Box 2: Bullet 45(1) "matches" Bullet 93(2)
- Box 1 with Box 3: None
- Box 1 with Box 4: Bullet 45(1) "matches" Bullet 194(4)

Sometimes frequent:

- Box 2 with Box 4: 1092 "matches"!
  $(50 \times 50 = 2500$ comparisons)


CCI Boxes 2 and 4
icpSb
icpCu
icpBi
icpAg
100 150 200 250 300 350
20 25 30 35 40
25000 27000 29000
100 200 300
60 80 100 140
20 25 30 35 40
25000 27000 29000
60 80 100 120 140 160

For CCI boxes 2 and 4:

(∗) Prob{bullets come from different boxes | FBI 'match' }

Box 2: 674 'matches' from $(50)(49)/2 = 1225$ comparisons

Box 4: 573 'matches' from $(50)(49)/2 = 1225$ comparisons

Boxes 2 & 4: 1092 'matches' from 2500 comparisons

$$(*) = \frac{P\{match|dif\,box\} \cdot P\{dif\,box\}}{P\{match|dif\} \cdot P\{dif\} + P\{match|same\} \cdot P\{same\}}$$

$$\frac{(1092/2500) \cdot (2500/4950)}{(1092/2500) \cdot (2500/4950) + (674+573)/(1225+1225) \cdot (2450/4950)}$$

$= 0.4668 \Rightarrow$ "Match" does not mean "same box"!

'2-SD Matches' among bullets from other manufacturers' boxes

- Federal Box 1 with Box 2: 363 matches
- Federal Box 1 with Box 4: 347 matches
- Federal Box 2 with Box 4: 313 matches
- Remington Box 1 with Box 2: 476 matches
- Winchester Box 3 with Box 4: 134 matches
- No matches of bullets across manufacturers

Distinguishing bullets from different manufacturers is easy

Distinguishing bullets from different boxes is virtually impossible

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 variability: 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.

Statisticians' contribution: Identify sources of uncertainty

Eyewitness Identification

Background:

- Eyewitness testimony can be very useful and incredibly powerful in the courtroom
- But ... the memory can play tricks, hence not always accurate nor reliable
- What procedures are used in eyewitness identification (EWI)?
- Which procedures lead to accurate identifications?
- **How to compare procedures in terms of accuracy?**

Situational aspects of EWI (*Estimator variables*):
Beyond the control of the criminal justice system

1. Eyewitness' level of stress or trauma at incident
2. Conditions affecting visibility
3. Distance between witness and perpetrator
4. Presence/absence of threat (e.g., weapon)

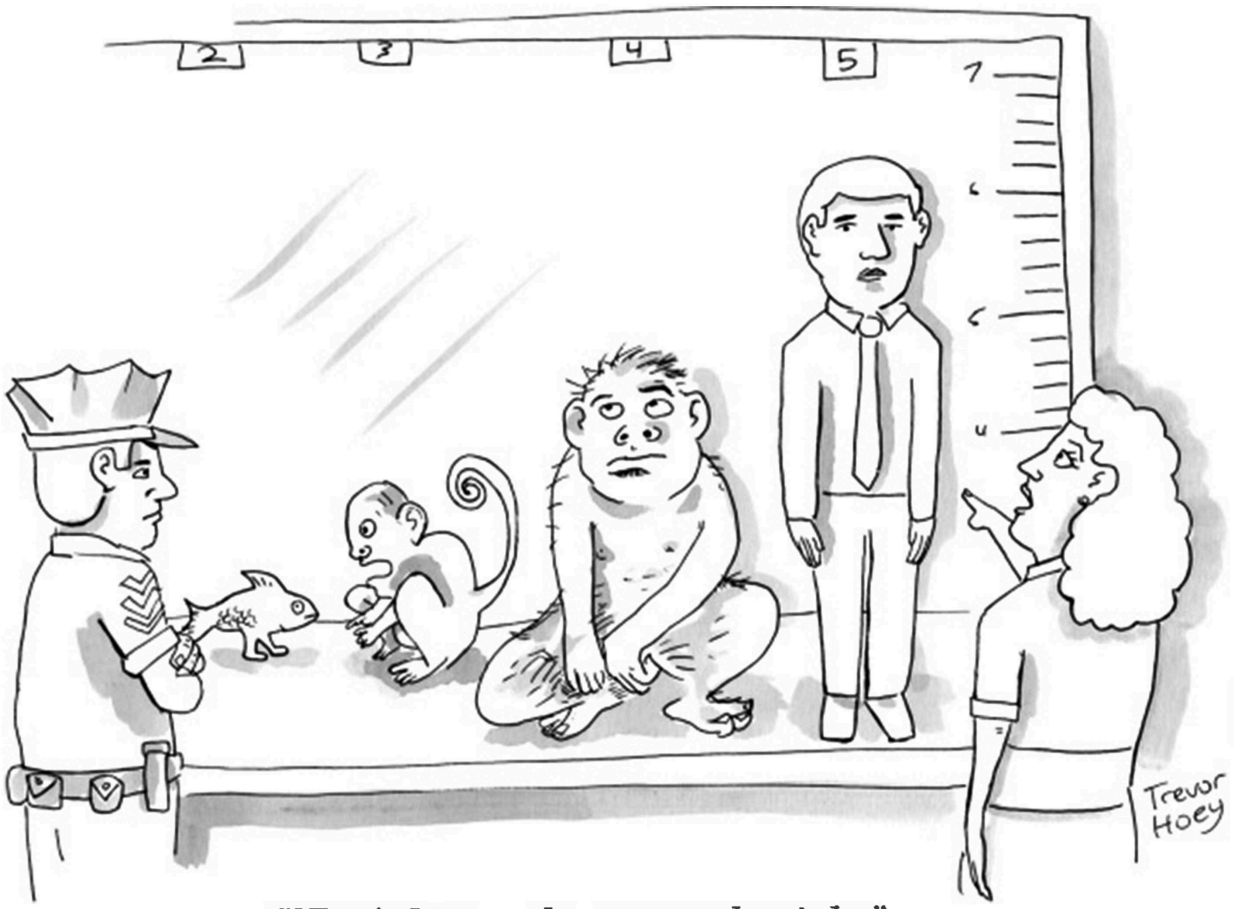
etc.

Procedural aspects of EWI (*System variables*):

1. Conditions & protocols for **lineups**
   (e.g., *sequential vs simultaneous*)
2. Nature of instructions (oral or written, short or long, ...)
3. Presence/absence of feedback
4. Number and similarities of fillers with "target"
5. Retention interval (longer ⇒ less reliable)

etc. *Which factors matter most to accuracy?*

Focus: Compare accuracy between two lineup procedures —
but methods apply to comparing *any* two procedures



"That's him—the one on the right."

From THE NEW YORKER, March 7, 2011

Sequential vs Simultaneous?

- *Sequential*: Present each photograph, one at a time
- *Simultaneous*: Present all six photographs at once
- Early research: "Sequential is more accurate"
- Later research: "Metric for comparison is incomplete; Simulaneous is more accurate"
- Which was correct?

Lab tests and proposed metrics

Lab tests: Present participants (usually Psych 1 students) a scenario, followed by lineup (sequential or simultaneous); count proportions of correct IDs (HR = *hit rate*) and mistaken IDs (FAR = *false alarm rate*)

1. *Diagnosticity Ratio*: Collapse all participants, all scenarios:

   $$\begin{aligned} &\textit{diagnosticity ratio} = \textit{hit rate / false alarm rate} \\ &= \textit{Sensitivity / (1 – Specificity)} \ (= LR^{+}) \end{aligned}$$

2. Some participants express more *confidence* in their choices; *confidence* is related to *accuracy*; therefore, we should look at HR and FAR as functions of *levels of expressed confidence*.

Which approach is correct?

- *Sensitivity*: When shown the *true* perpetrator, what is the probability that the "witness" identifies him/her?
- *Specificity*: When shown an *imposter*, what is the probability that the "witness" excludes him/her?
- *Sensitivity, Specificity* can be estimated only in studies *where truth is known* (by design)
- Real life: All you have is response:
  "Yes, that's the one" or "No, not that one"

- *Positive Predictive Value (PPV)*: If claim is "Yes, that's the one", what is the probability that the identified person is the perpetrator?
- *Negative Predictive Value (NPV)*: If claim is "No, not the one", what is the probability that the excluded person is not the perpetrator?
- *PPV, NPV* are functions of *Sensitivity, Specificity, and odds that the suspect is the true perpetrator*
- *Diagnosticity Ratio* is related to PPV:

  $$PPV = 1 \,/\, (1 + Odds/DR)$$

  so *higher DR* $\Rightarrow$ *higher PPV*
- What about correct exclusions, NPV?
  (cf. $LR^{-} = (1 - sens)/spec$)

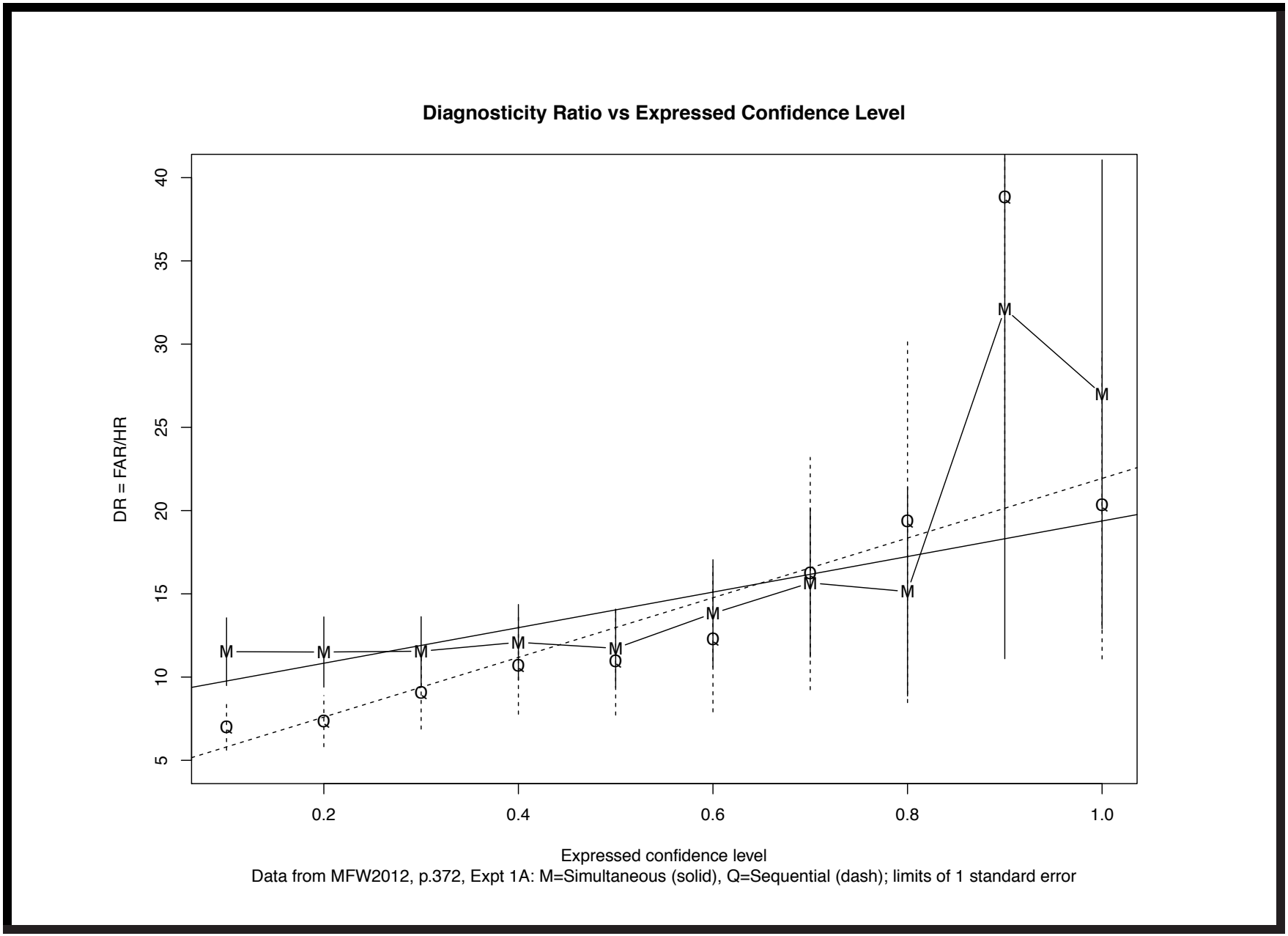
Relationship between Confidence & Accuracy?

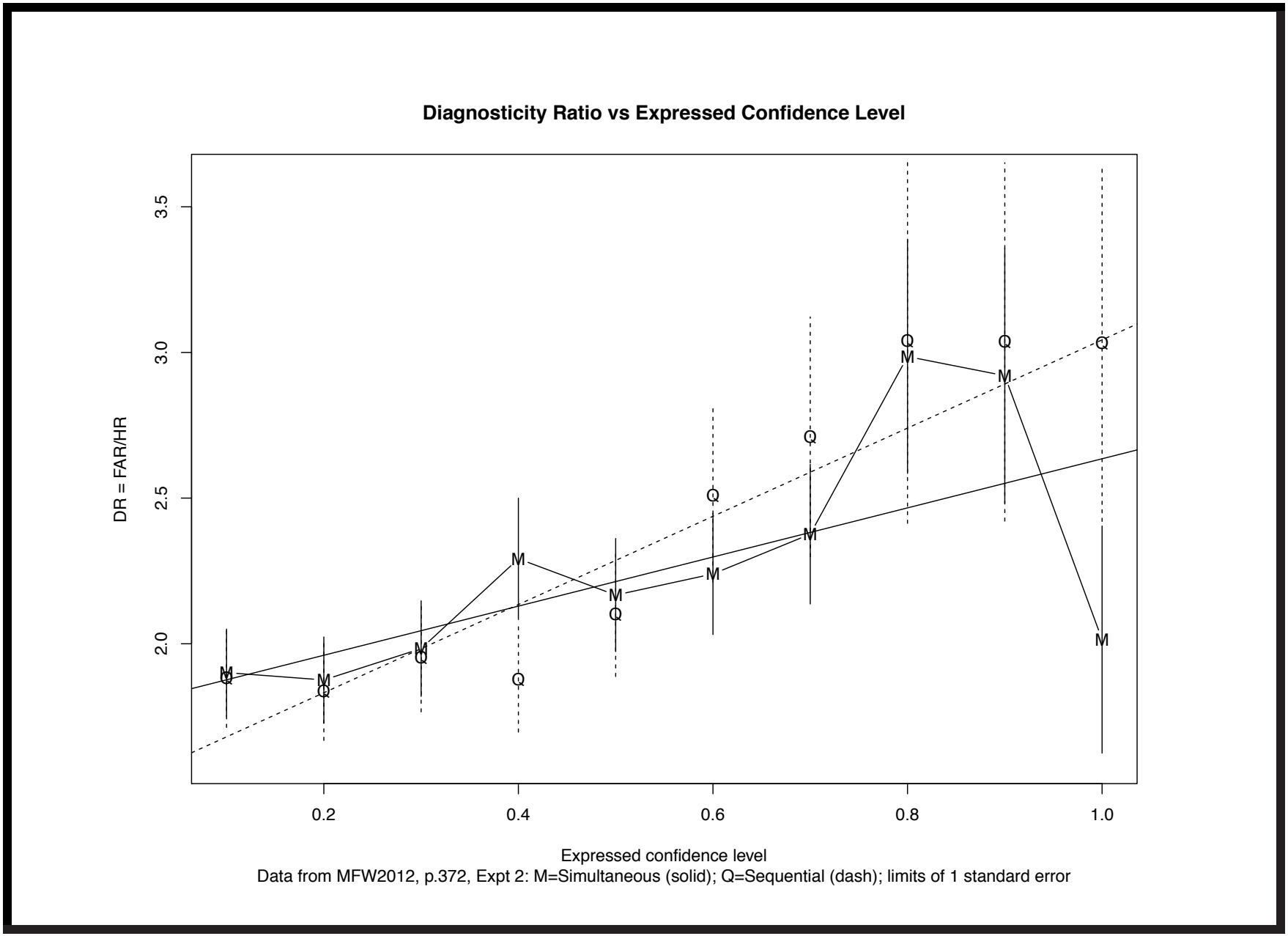
If you believe confidence is related to accuracy:

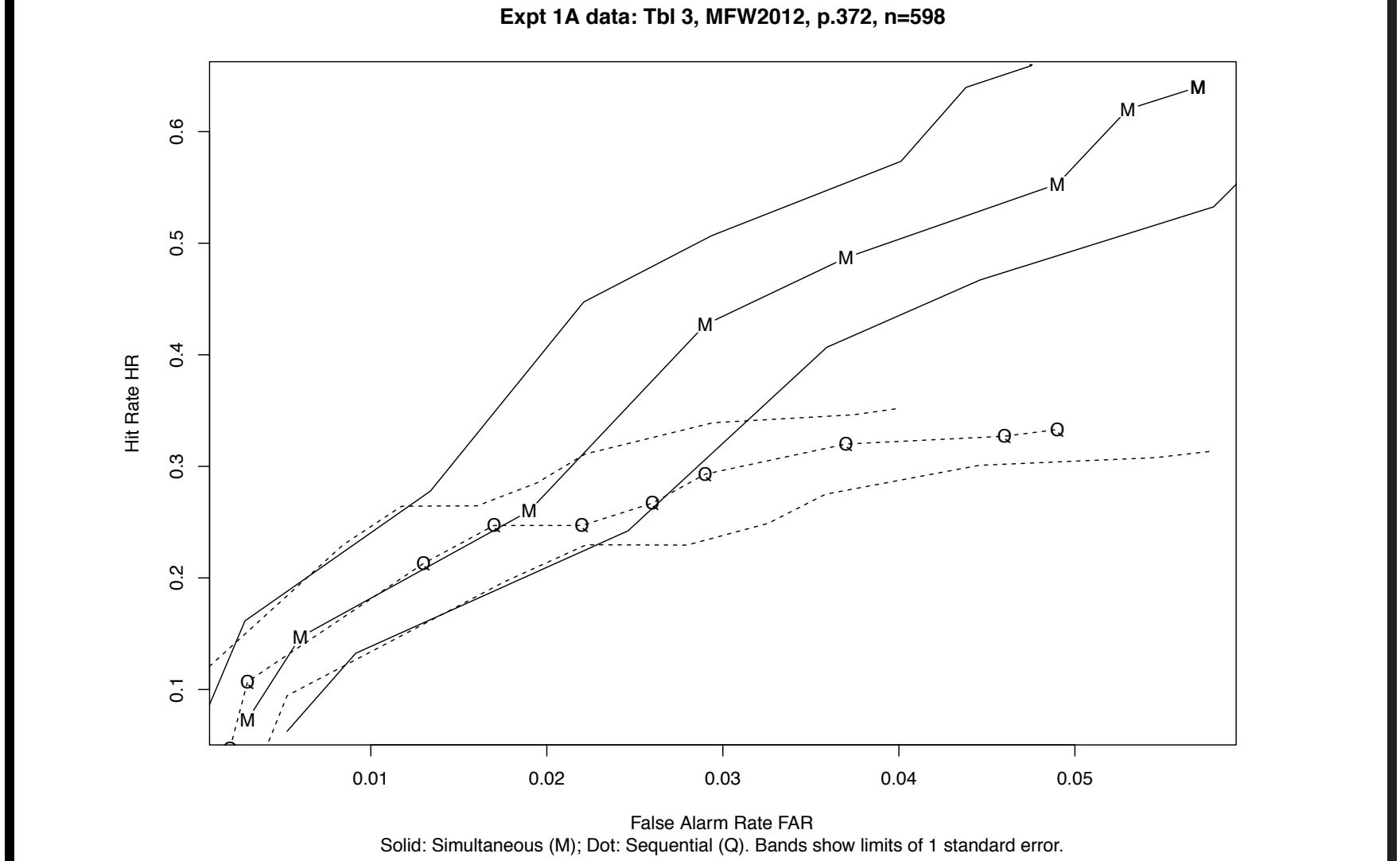
- consider calculating $DR = HR/FAR$ as a function of *Expressed Confidence Level (ECL)*
- Split the sample participants into categories of ECL (those who expressed 10%, ..., 90% confidence); calculate DR for each ECL category
- even better: Plot HR vs FAR for different ECLs
- ROC curve = Receiver Operating Characteristic
- Used in quality control and comparing medical diagnostic procedures

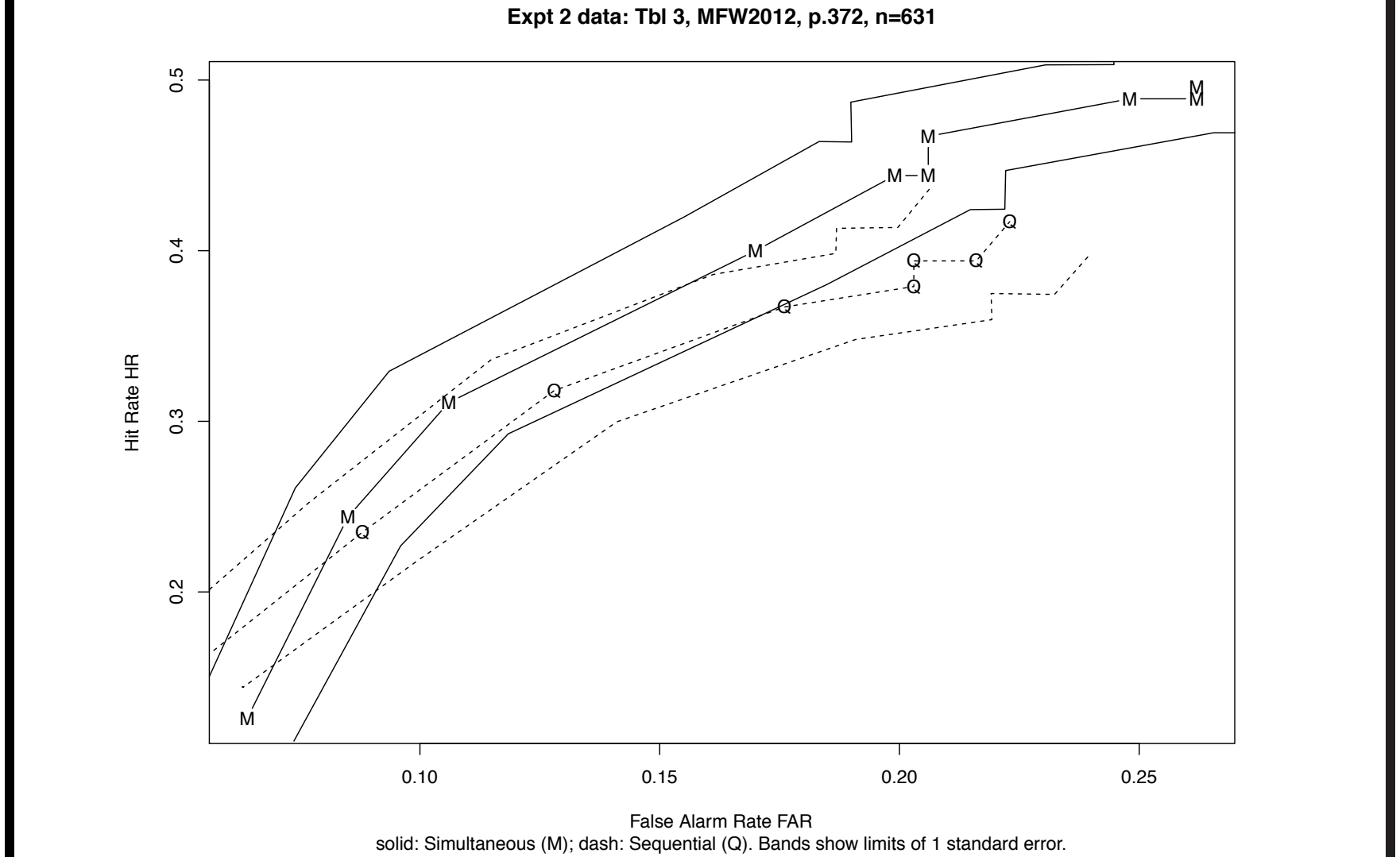
Problem: Data points (HR, FAR) have uncertainty!

- John Tukey (in discussing uncertainty in rates at NCI):
  "*What has happened is history. What might have happened is science and technology. So what you are really interested in is what might have happened if you could do it all over again.*"
- Simulate what would happen if you calculated all the HRs and FARs (for different ECLs) *as if* you repeated the same experiment all over again
- DR vs ECL for Sequential and for Simultaneous:
  How different are they?
- How different do the two ROC curves look for Sim vs Seq?
- Resulting uncertainty is underestimated, because ECLs can change (e.g., "40%" today; "20%" tomorrow)


Diagnosticity Ratio vs Expressed Confidence Level
DR = FAR/HR
5
10
15
20
25
30
35
40
0.2
0.4
0.6
0.8
1.0
Expressed confidence level
Data from MFW2012, p.372, Expt 1A: M=Simultaneous (solid), Q=Sequential (dash); limits of 1 standard error


Diagnosticity Ratio vs Expressed Confidence Level
DR = FAR/HR
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8
1.0
Expressed confidence level
Data from MFW2012, p.372, Expt 2: M=Simultaneous (solid); Q=Sequential (dash); limits of 1 standard error


Expt 1A data: Tbl 3, MFW2012, p.372, n=598
Hit Rate HR
0.1
0.2
0.3
0.4
0.5
0.6
0.01
0.02
0.03
0.04
0.05
False Alarm Rate FAR
Solid: Simultaneous (M); Dot: Sequential (Q). Bands show limits of 1 standard error.


Expt 2 data: Tbl 3, MFW2012, p.372, n=631
Hit Rate HR
0.2
0.3
0.4
0.5
0.10
0.15
0.20
0.25
False Alarm Rate FAR
solid: Simultaneous (M); dash: Sequential (Q). Bands show limits of 1 standard error.

A "more complicated model": Data from Carlson & Carlson 2014, *J Appl Research in Memory and Cognition*:

- 12 conditions:
  - 3 Procedures (SIM, target in #4; SEQ, #2; SEQ, #5)
  - 2 Weapon conditions (present, absent)
  - 2 Distinctive Feature conditions (present, absent)
- Compute confidence-based ROC for each condition
- Compare "Partial Area under ROC curve" (bigger = better)

Model:

$$\log(pAUC) = \text{Proc Effect} + \text{Weapon Effect} + \text{Feature Effect} + (\text{all 3 pairwise interactions}) + \text{error}$$

Source	df	SS	MS	F-stat	p-value
Procedure	2	8.04	4.02	1.129	0.470
Weapon	1	2.94	2.94	0.826	0.460
Feature	1	14.72	14.72	4.138	0.179
Proc × Weapon	2	0.59	0.30	0.083	0.923
Proc × Feature	2	10.41	5.21	1.463	0.406
Weapon × Feature	1	34.80	34.80	9.780	0.089
Residuals	2	7.12	3.56		

Accuracy likely to be related to *many* variables

Fienberg & student investigating other methods of comparison

4. Present: Where Statistics is being used in FS

1. *Interpretation*:
   - Significance of data (convincingly identification or exclusion)
   - Calculating a likelihood of "guilt" versus "innocence"
   - RSS (Aitken et al.): Guides for judicial personnel to assess forensic evidence and interpret statistical/probabilistic reasoning
2. *Fingerprint identification method via likelihood ratios using configurations of minutiae* (Neumann 2011, Champod and others)

(Present, cont'd)

3. *Designing tests to estimate error rates in current ACE-V process* (Ulery et al. 2012): 744 pairs of latent prints with 168 examiners (6 false positive IDs among 4083 exams $\Rightarrow$ false + error probability = 0.15%, upper 95% CL 0.29%)

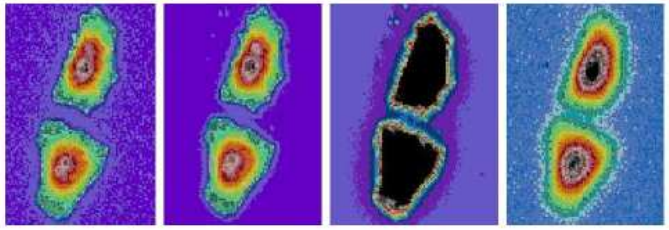
4. *Quality Metrics*: Development of objective (vs subjective) assessment of latent print quality, to be correlated with accuracy of call (ID or exclusion; Tabassi et al. 2004; Yoon et al. 2012; Swofford & Peskin et al.) — but not yet being used

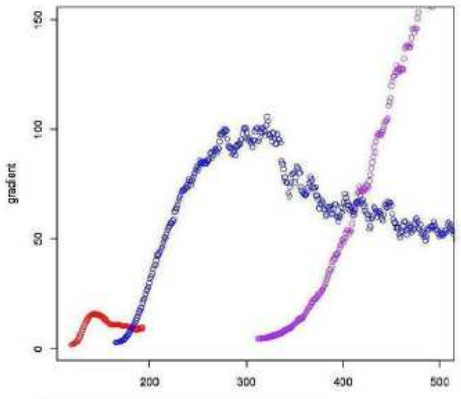
5. Future: Where Statistics can be used in FS

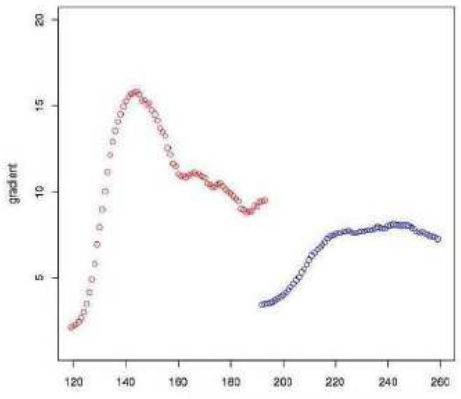
- **Problem Identification**: Compare two items? Method for analysis? Pattern generation?
- **Evidence Examination**: Is the evidence suitable for examination (quality)?
- **Process Identification**: What procedures are used for comparison? Are they objective, measureable, repeatable?
- **Research**: What alternative approaches may be appropriate?
- **Comparison**: What metrics characterize the approaches?
- **Design**: What kind of experiment will offer a valid comparison of approaches?

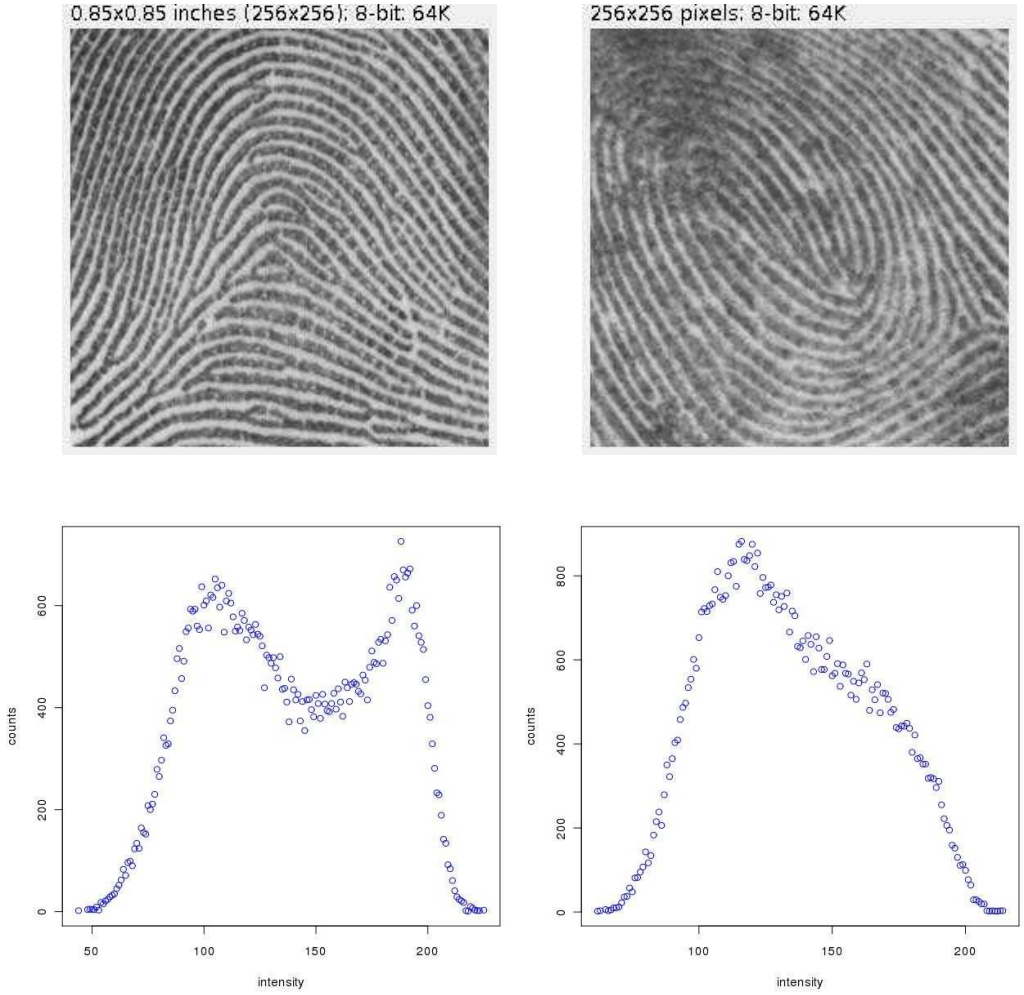
Ex: Research on quality metrics ("C" in "ACE-V")

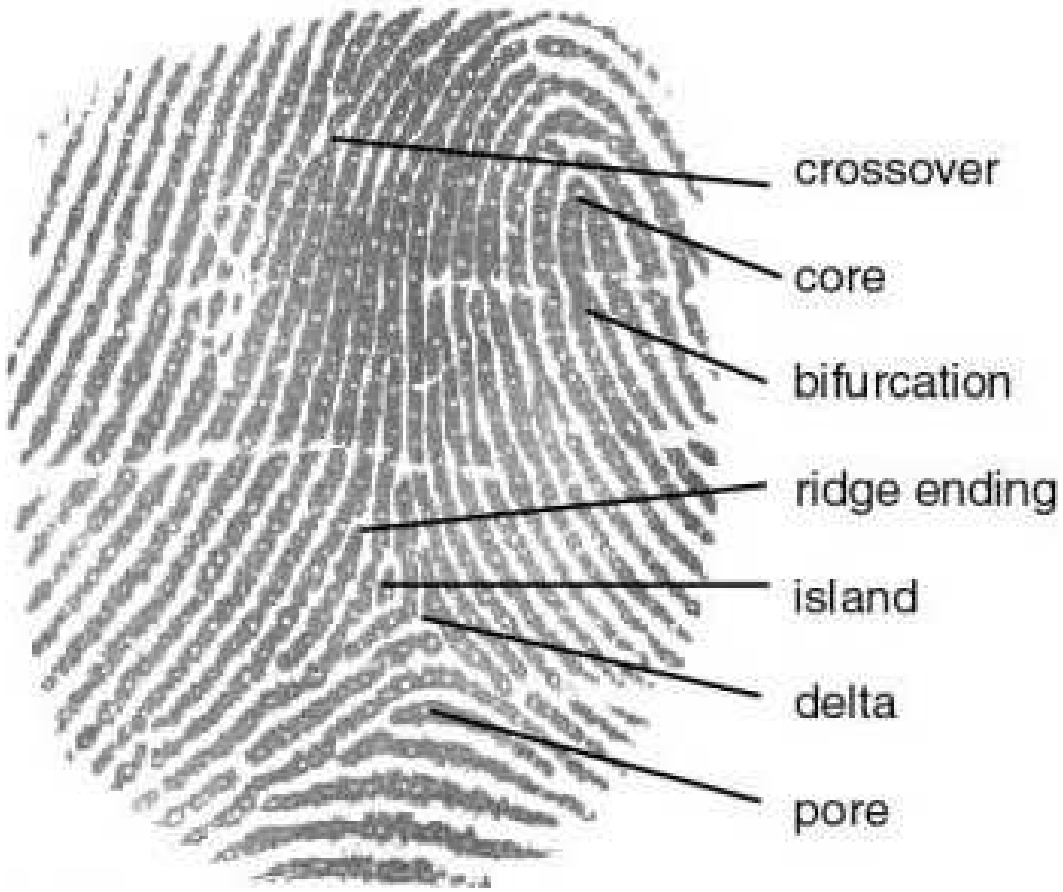
- Peskin et al. (2010): Measure cell image quality (NIST)
- Choice of segmentation based on edge quality
- Calculate numerical gradients from cell to background
- Sharper peak corresponds to clearer images
- Apply concepts (gradient, contrast) to fingerprint images
- Simulate increasingly degraded print via blurring; fingerprint quality score decreases accordingly
- **How does quality score relate to accuracy of identification?** (*quality threshold*)
- Alternative quality metrics: Yoon et al. (2012); Tabassi et al. (2004); Swofford (in progress)

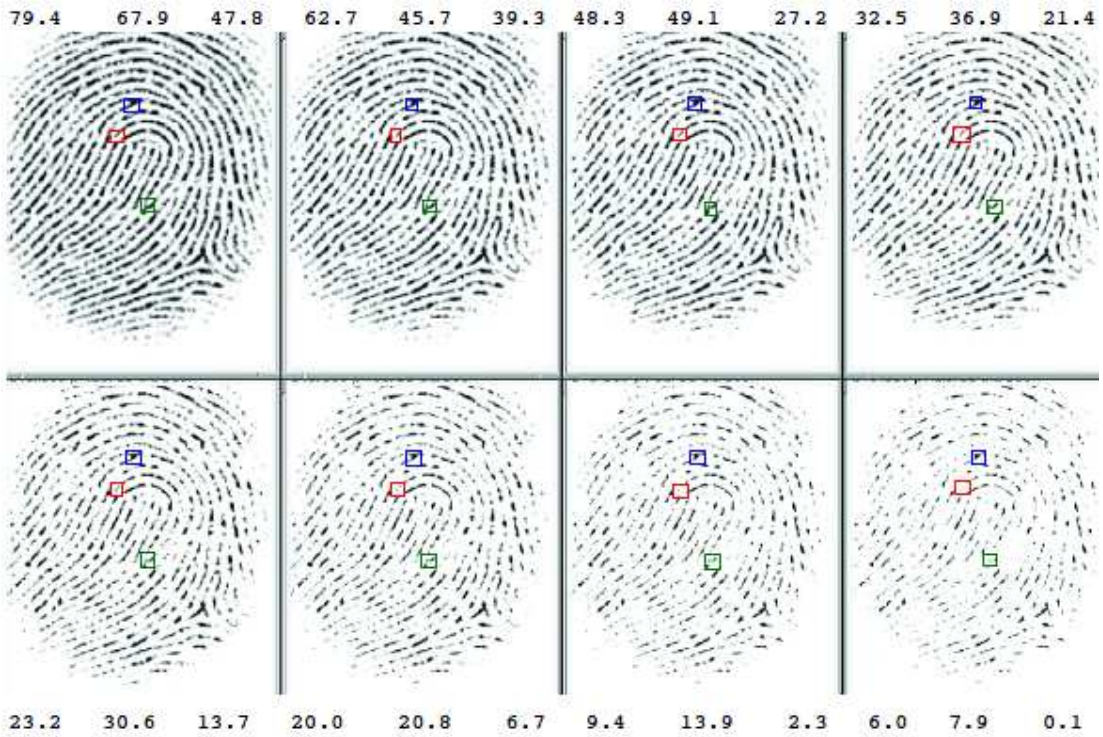


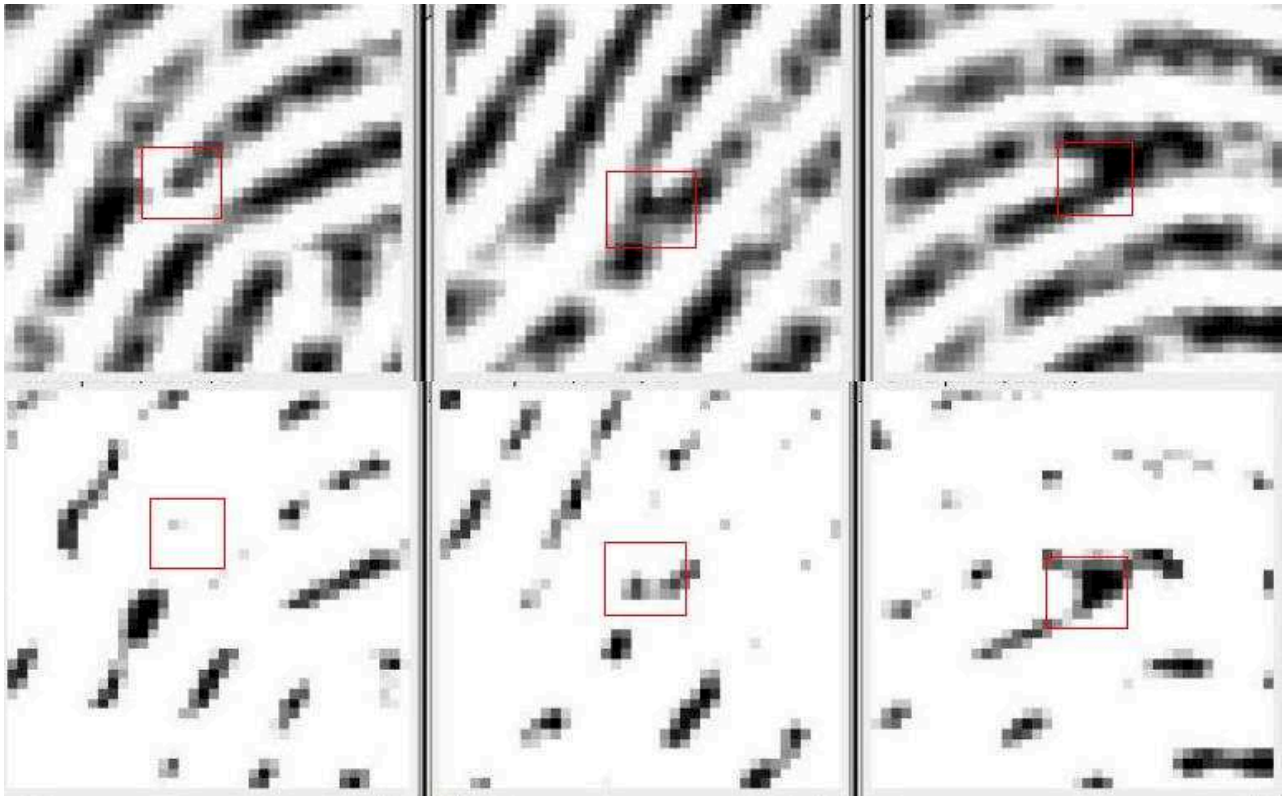

gradient
0
50
100
150
200
300
400
500


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

Design of studies to compare different metrics: 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 "same source" & "different source" pairs

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

(Future, cont'd)

- Probability models for pattern evidence:
  Need probabilistic & statistical methods associated with *collection, analysis, validation, interpretation* of general pattern & digital evidence ⇒ quantification of uncertainties, starting from evidence to "conclusion"
- Ex: "Inverse problem": What caused the outcome (blood spatter, arson, explosion, ...)? cf. AIDS incidence from HIV infection, oil wells in porous media, etc.
- Ex: for analysis, LRs for latent prints (Neumann et al. 2011)
- Shoe print comparisons (Weisner, Yekutieli); Tool marks & ballistics (Spiegelman), Hair

Issues to consider:

- How *frequent* are features used in an analysis (e.g., minutiae or pairs of minutiae)
- Can features be identified objectively?
- If not, how sensitive are "conclusions" to identified features?
- Are features sufficiently sensitive, specific?
- Role of imaging technology? (Spier, Jain)
- Combining information of different types (image, text, numeric) from different sources (patterns, descriptions, numerical metrics)

(Future, cont'd)

- Effect of bias in drawing conclusions (Thompson, Zabell)
- Effect of presentation of evidence to jury: with or without uncertainty statements, prior beliefs about reliability of evidence, etc. (Garrett & Murrie, UVa)
- How to present evidence (Wed panel)
- Educating future forensic scientists, judicial personnel
- How to design studies to validate procedures (collection, analysis, comprehensibility)
- Statistical Process Control in laboratories (Garner, Lin)
- Inspiring **change**

No one person or lab can do it alone

Some initial collaborations between Center for Statistics & Applications in Forensic Evidence (CSAFE)

- Defense Forensics & Biometrics Agency (DFBA) - Tontakski
- Houston Forensics Science Center (HFSC) - Castillo, Stout
- West Palm Beach Crime Lab - Crouse
- More to come from this year-long program

6. From Research to Implementation

R&D: "Failures because Production doesn't follow directions"
Production: "Failures because R&D designs are impossible"
Managers: "We don't have time to do studies"

7 tools of Statistical Process Control (SPC)

- Process flow chart
- Cause-and-effect diagram
- "Pareto analysis": #failures by cause
- Control chart
- Plots: histogram, scatter plot, ...
- Design of Experiments
- Continuous improvement

Designing process validation / reliability studies: repeatability, sensitivity, specificity

- **Accuracy**: Does method provide accurate results?
- **Precision**: Does method provide consistent results?
- 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 "same source" & "different source" pairs

Ex: Experimental Design for Toolmark Analysis

- Which factors can affect "accuracy"?
- "Accuracy": Correct call? (ID vs exclusion)
- Gun type, Image clarity, Experience (yrs), Time (hrs)
- # runs for each condition, #labs involved
- Focus on factors having greatest impact on accuracy

(Spiegelman)

Issues of Interpretation

- Methods for identification, comparison, etc.
- Goal: *Correct assessment* (ID of drug, person, etc.)
- Problem: *We never really know: All we have are data*
- Approach: **Given data**, make an assessment
- Performance metric: high probability of correct assessment

  Probability{Correct assessment | data} is **HIGH**
- Problem: *We don't know the true answer*
- Statistics: We can **estimate** this probability

What we *want*: HIGH Positive/Negative Predictive Value

$$\text{PPV} = \text{Prob}\{\text{same source} \mid \text{data suggest “same”}\}$$
$$\text{NPV} = \text{Prob}\{\text{different sources} \mid \text{data suggest “different”}\}$$

But we never know for sure.

What we **can do**: Design a study where “truth” is known

Connect results of designed study to PPV, NPV

Posterior Odds, Prior Odds, Likelihood Ratios:

- What we want to know: If "same source" is claimed, how likely is it that the two specimens did in fact come from the same source, versus different sources?
- In notation, we want to know:

$$\frac{P\{same\,source \mid \text{“}same''\}}{P\{different\,sources \mid \text{“}same''\}}$$

  But we will never know for sure.

- Bayes Theorem tells us that *what we want to know* involves quantities that *we can design from an experiment*:

$$\frac{P\{same\,source \mid \text{“}same''\}}{P\{different\,sources \mid \text{“}same''\}}$$
$$= \frac{P\{\text{“}same'' \mid same\,source\}}{P\{\text{“}same'' \mid different\,sources\}} \cdot \frac{P\{same\,source\}}{P\{different\,sources\}}$$
$$= (\text{Likelihood Ratio}) \cdot (\text{prior odds})$$

- P {"same" | same source } = *Sensitivity*;
  P {"same" | different sources } = 1 – *Specificity*
- Prior odds? Anyone's guess.
- **We can estimate the Likelihood Ratio – with uncertainty. But we want the Posterior Odds.**
- If we are on a desert island with only 10 people, Prior Odds might be (0.1)/(0.9) = 1/9.
- If we are in NYC, Prior Odds might be 0.000001
- Whatever we do, the quantity is only an estimate. We have to provide a range of plausible values (Lund & Iyer, 2015).

Study where "truth" is known

- **Prepare materials from same sources**
- Test administrator give materials to examiners
- Proportion of examiners correctly claim "same source"?
- **Prepare materials from different sources**
- Test administrator give materials to examiners
- Proportion of examiners correctly claim "different sources"?
- Mix up materials so only test designer knows which is which
- Double-blind

The designed study enables us to estimate

$$\text{Sensitivity} = \text{Prob}\{\text{“same source”} \mid \text{same source}\}$$
$$1 - \text{Specificity} = \text{Prob}\{\text{“same source”} \mid \text{different sources}\}$$

These estimates are used in **PPV, NPV, Posterior Odds**.

That's why we need the designed studies:

- to estimate Sensitivity & Specificity,
- which are needed to estimate PPV, NPV,
- and how likely the “same source” claims are correct.

But statistics is needed well before then (methods development, experimental design, lab process monitoring, ...).

7. Final Comments: Roles for Statisticians

- "Statistical thinking" (Hoerl & Snee: *Use what you have*)
- Quantifying vague concepts
- "Alternatives": What *might* happen
- Identifying confounding factors
- Quantifying uncertainty
- Designing methods to reduce subjectivity & uncertainty

References

Box, GEP; Hunter, WF; Hunter, JS (2005), *Statistics for Experimenters, 2nd ed.*, Wiley.

Hoerl, R; Snee, R (2002), *Statistical Thinking: Improving Business Performance*

Kafadar, K (2015), Statistical Issues in Assessing Forensic Evidence, *International Statistical Review*

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

Spiegelman CH; Kafadar K (2006), Data Integrity & Scientific Method: The Case of Bullet Lead Data as Forensic Evidence, *Chance*

Vardeman, SB; Jobe, JM (1999), *Statistical Quality Assurance Methods for Engineers*, Wiley.