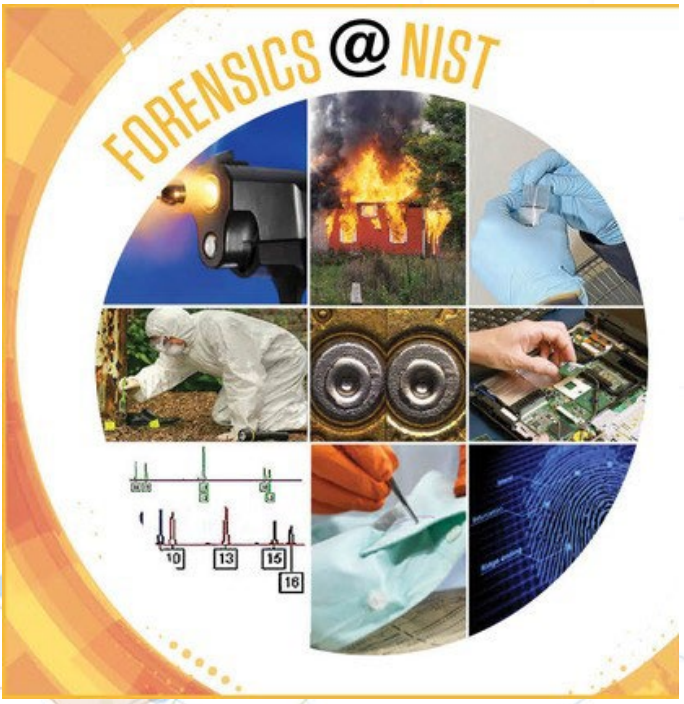


Overview of NIST Statistical Research in Forensic Science

Will Guthrie
NIST Statistical Engineering Division
will.guthrie@nist.gov

National Institute of Standards and Technology
U.S. Department of Commerce

NIST Statistical Research in Forensics – View from 10^5 m

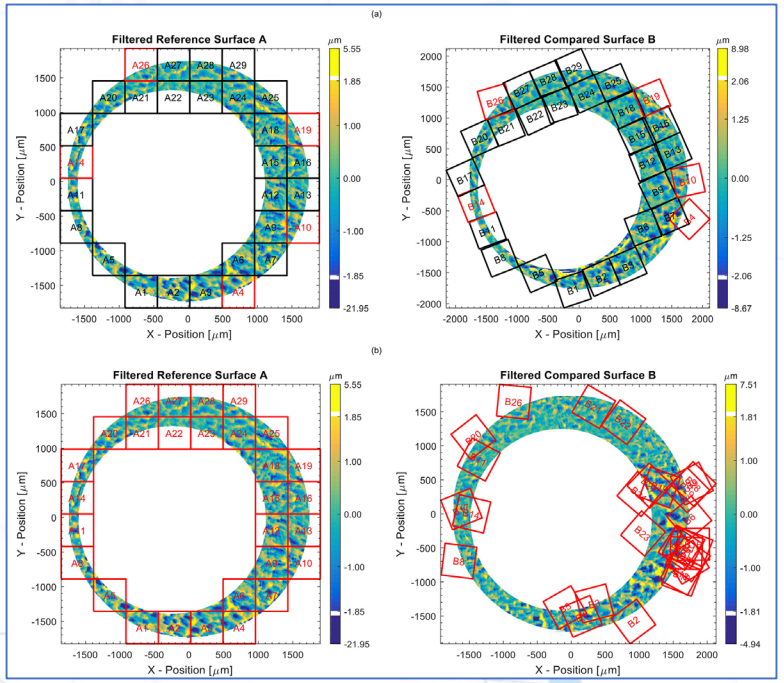
- 15 current research projects
  - Staffing level between 3 and 4 FTE/year
- Projects a mix of:
  - collaborative, individual
  - foundational, applied
- Key partnerships include:
  - NIST staff working in other Forensic Science Focus Areas
  - FBI, MSP, NFI, NFEA, IAI, FIU, CSAFE, OSAC

Research Areas

- New Reference Materials for Trace Elements in Glass
- Uncertainty of Drug Mass Measurements
- Optimization of GC/MS for Fire Debris Analysis
- Challenge Problem: Statistical Comparison of Paint Spectra

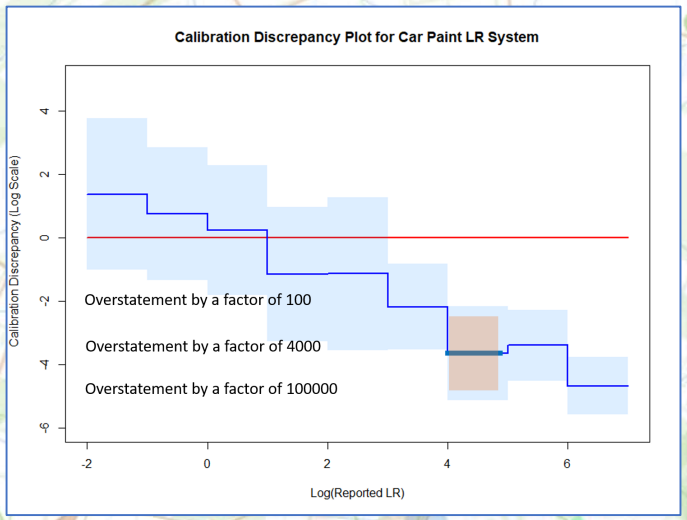
- Complex DNA Mixture Interpretation
- Characterization of Noise in Next Generation Sequencing Data
- Use of Next Generation Sequencing for DNA Mixture Analysis
- Assessment of Thresholds for CE STR Profiles

Research Areas



- Error Rate Assessment for Firearms ID
- Uncertainty Budget Framework for Automated Firearms Examination
- Quantitative Understanding of Uniqueness and Reproducibility of Firearm Toolmark Surfaces
- Reference Population Data for Firearm Toolmarks

- Likelihood Ratios as Weight of Evidence
- Calibration of Likelihood Ratios
- Quantitative Evaluation of Footwear Evidence

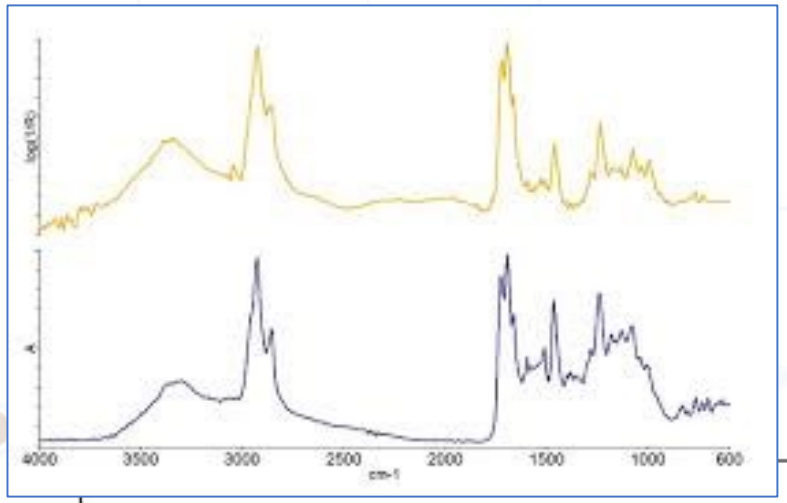


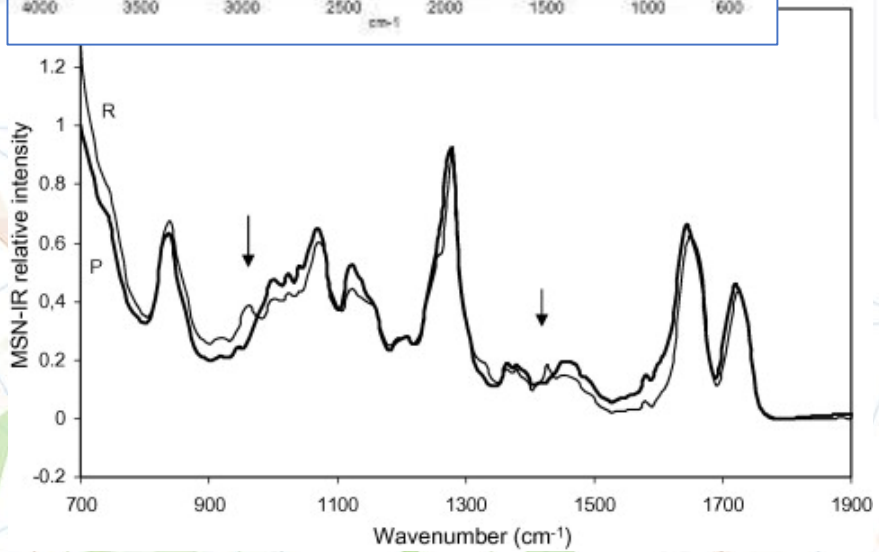
Challenge Problem: Statistical Comparison of Paint Spectra

- Working with an OSAC TG and NIST chemists to design a challenge problem on comparison of paint spectra in forensic settings
  - hit and run car accidents, other crime-related scenes with paint traces left behind

- Goals
  - Promote community development of new statistical or machine learning algorithms
  - Comparison of algorithm performance on common test data

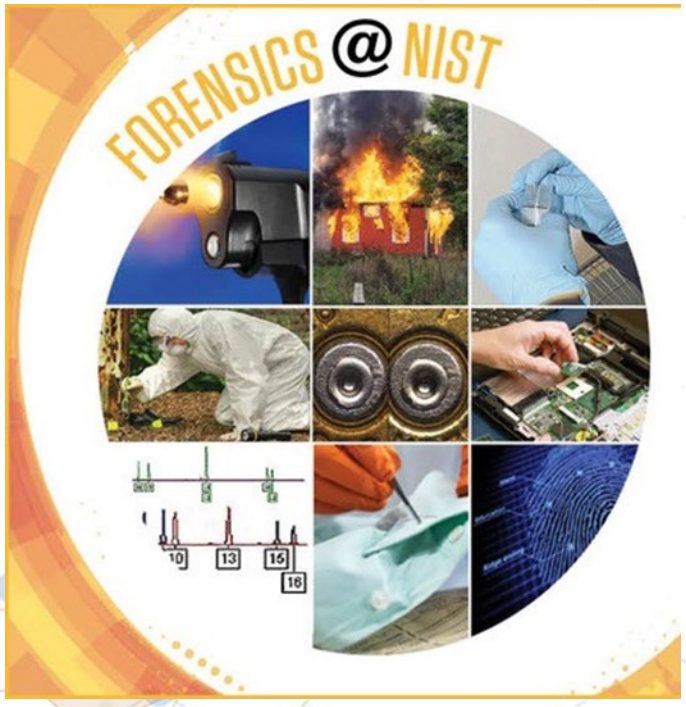
Challenge Problem:
Statistical Comparison of Paint Spectra





- Importance
  - Current methods focus on visual comparison and rely on analyst expertise and judgment
  - Statistical or machine learning methods have potential to be more consistent across labs

- Status
  - Paint spectra collected from multiple cars
  - Assessment of algorithm comparison methods currently underway
  - Stay tuned for more news!



Statistical Models for Similarity Score Comparisons in Firearm Evidence Identifications

Nien Fan Zhang

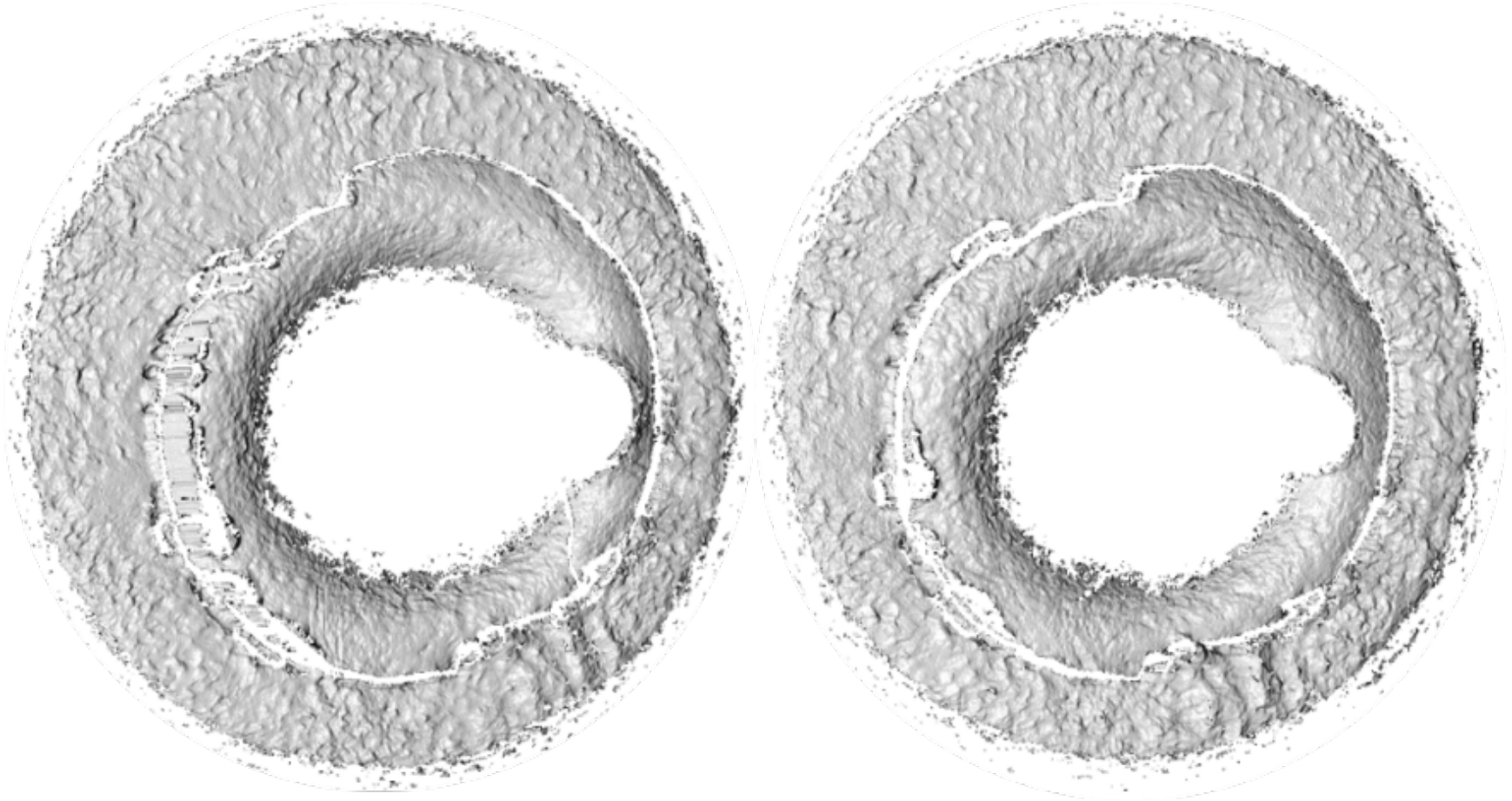
NIST Statistical Engineering Division

National Institute of Standards and Technology
U.S. Department of Commerce

Outline

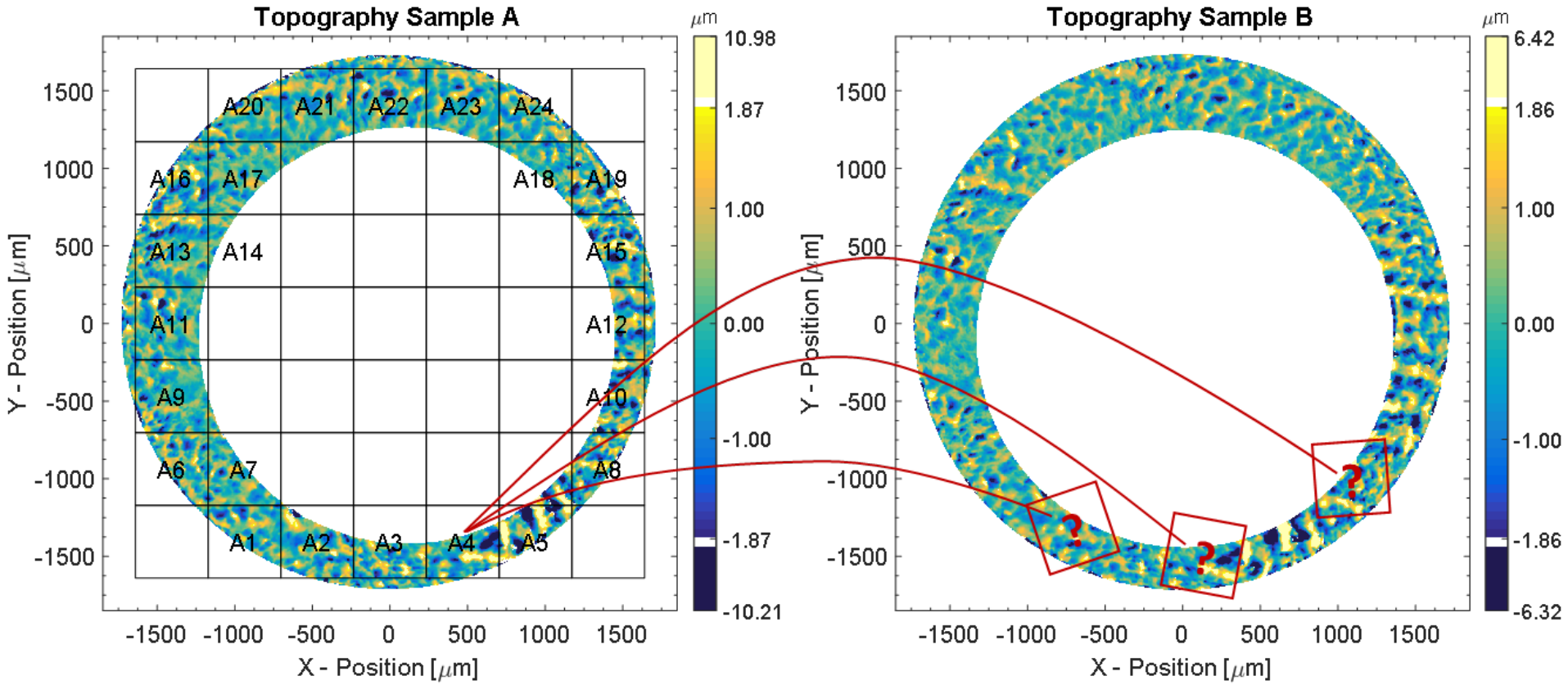
1. Introduction
2. Statistical models for CMC measurements
3. Estimating model parameters
4. A practical example
5. Discussions and conclusions
6. References

1. Introduction



3-D topographic images of breech face impressions obtained from a pair of cartridge cases ejected from the same firearm slide.

The congruent matching cells (CMC) method is used to compare pairs of topography images of breech face impressions and provide a basis for estimating error rates.



To estimate error rates, the key is to find appropriate probability distributions for the relative frequency distribution of the observed CMC values.

2.1 Model 1 for CMC measurements

Binomial distribution

- Matches between each of the N_1 cell pairs from first image pair, X_{1j}, are the outcomes of a sequence of independent Bernoulli trials with:

$$P(X_{1j} = 1) = p \qquad P(X_{1j} = 0) = 1 - p \qquad X_{11}, \cdots, X_{1N_1}$$

Y_1 is the number of CMCs for the first image pair.

$$Y_1 = \sum_{j=0}^{N_1} X_{1j} \quad \Rightarrow \quad P_{[1]}(Y = k) = \binom{N_1}{k} p^k (1-p)^{N_1 - k}$$

- To estimate p, we use an independent sequence of M image pairs each with potentially different values of N_i but with the same value of p.

2.2 Model 2: Correlated binomial distribution

- Matches between each of the N_1 cell pairs from first image pair, X_{1j}, are the outcomes of correlated Bernoulli trials with:

$$P(X_{1j}=1)=p \quad P(X_{1j}=0)=1-p \quad r_{(2)}=\mathrm{Cov}(X_{1i},X_{1j})\Big/\sigma^2$$

$$Y_1=\sum_{j=0}^{N_1}X_{1j} \;\Rightarrow\; P_{[2]}(Y_1)=P_{[1]}(Y_1)\{1+r_{(2)}g_2(Y_1,p)\}$$

- To estimate the parameters, we use an independent sequence of image pair comparison results with a total of M images $Y_1,,,Y_M$. The cell numbers N_i for different image pairs can be different.

2.3 Model 3: Beta-binomial distribution

- Matches between each of the N_I cell pairs from first image pair, X_{Ij}, are the outcomes of independent Bernoulli trials with:

$$P(X_{1j}=1)=p \qquad P(X_{1j}=0)=1-p \qquad X_{11},\cdots,X_{1N_1}$$

- For different image pairs, they are independent and p is random with a beta probability distribution, i.e., $p \sim Beta(\alpha,\beta)$.

$$Y_1=\sum_{j=0}^{N_1} X_{1j} \quad P(Y_1=k \mid N_1,\alpha,\beta)=\binom{N_1}{k}\frac{B(k+\alpha,N_1-k+\beta)}{B(\alpha,\beta)}$$

- M independent image pairs are used to estimate the parameters.

2.4 Model 4: Beta-correlated binomial distribution

- Matches between each of the N_1 cell pairs from first image pair, X_{1j}, are the outcomes of correlated Bernoulli trials with:

$$P(X_{1j}=1)=p \quad P(X_{1j}=0)=1-p \quad r_{(2)}=\mathrm{Cov}(X_{1i},X_{1j})\Big/\sigma^2$$

- For different image pairs, they are independent and p is random with a beta probability distribution, i.e., $p \sim Beta(\alpha,\beta)$

$$Y_1=\sum_{j=0}^{N_1} X_{1j} \quad P(Y_1=k \mid N_1,\alpha,\beta,r_{(2)})=\int_0^1 \frac{P_{[2]}(k,p)\,p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}dp$$

- M independent image pairs are used to estimate the parameters.

3. Estimating the parameters of the models

- CMC method applied to a set of cartridge cases to have certain known matching (KM) image pairs and certain known non-matching (KNM) image pairs.

- Sum of CMC values $Y_{1,N_1},\ldots,Y_{M,N_M}$ for M independent image pairs,

- Likelihood function for corr. Binomial,

$$L=\prod_{i=1}^{M}P_{[2]}(y_i\mid p,r_{(2)})=\prod_{i=1}^{M}\binom{N_i}{y_i}p^{y_i}(1-p)^{N_i-y_i}\{1+r_{(2)}g_2(y_i,p)\}.$$

The ML estimator of p and $r_{(2)}$ are obtained when log (L) reaches the maximum.

Nonlinear regression can be used to estimate parameter(s).

- Assume all the CMC values for each image pair is binomial distributed,

$$Y_i \sim Bin(N, p) \qquad i = 1, \ldots, M$$

By the law of large number, the frequency curve approaches the binomial mass function when $M \to \infty$.

- For a sample from an independent sequence $\{Y_{1,N_1}, \ldots, Y_{M,N_M}\}$, call

$$f_M(k) = \frac{number\ elements \text{ in } sample = k}{M} = \frac{\sum_{i=1}^{M} \mathbf{1}_{Y_{i,N_i}=k}}{M} \quad \text{for} \quad k = 0, \max\{N_i\}$$

a generalized frequency function. Assume there are L distinct N's.

Assume the CMC values $Y_i \sim \text{corr.}Bin(N_i, p)$

Assume for each distinct N, Nd_l $(l = 1,...,L)$, there are C_l indicators with same Nd_l. $\sum_{l=1}^{L} C_l = M$

Under some regular conditions, by the law of large number, when $M \to \infty$

$$f_M(k) = \frac{\sum_{i=1}^{M} \mathbf{1}_{Y_{i,N_j}=k}}{M} \to \sum_{l=1}^{L} W_l \cdot \text{corr. } Bin(k, Nd_l, p) \text{ almost surely}$$

where $W_l = C_l / M$. Thus, we have a nonlinear regression model based on the correlated binomial distribution

$$f_M(y) = \sum_{l=1}^{L} W_l \cdot P_{[1],Nd_l}(y)\{1 + r_{(2)} g_2(y, p, Nd_l)\} + \varepsilon \quad y = 0, \max\{N_i\}$$

4. A practical example

The Weller data set of cartridge cases has 370 known matching (KM) image pairs, but not with same N . The 370 columns representing 370 image pairs. There are L = 15 distinct N's.

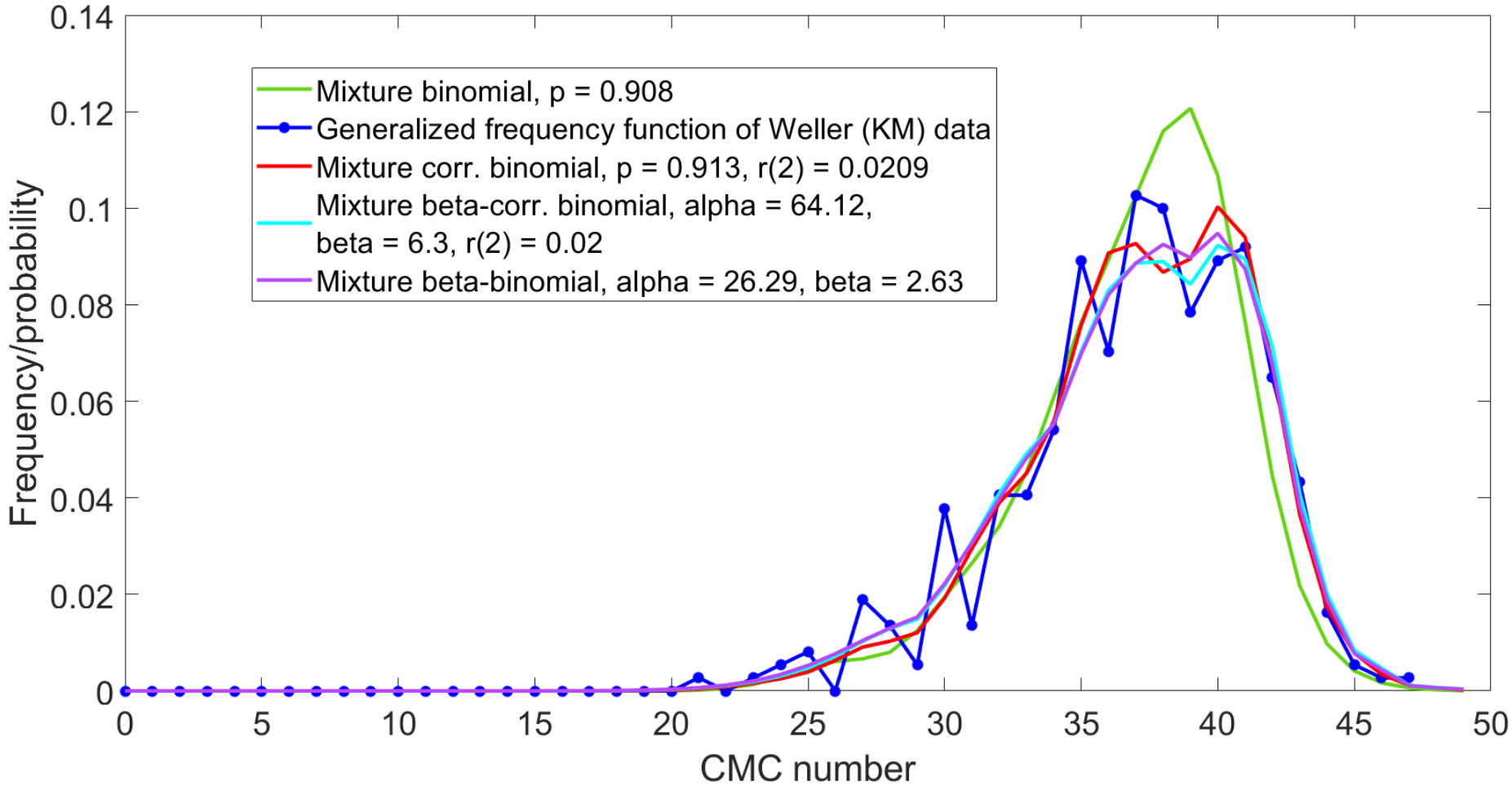
CMC (Y)	36	27	37	36	36	37	38	37	27	40	37	36	….	44	41	37	37	43
N	38	38	38	38	38	38	38	38	41	41	41	41	….	46	46	43	43	45

Based on the original data, the CMC values and the corresponding counts are given by

CMC (Y)	21	22	23	24	25	26	27	28	29	30	31	32	….	43	44	45	46	47
Count (C_j)	1	0	1	2	3	0	7	5	2	14	5	15	….	16	6	2	1	1

The generalized frequency function is obtained by count/370 since there are 370 image pairs. Note the sum of counts (C_j) = 370.

The M =370 image pairs have L = 15 distinct N's. The generalized frequency function will approach a pmf which is a mixture or a weighted mean of 15 pmf's of the underlying distribution when $M \to \infty$.



5. Discussions and conclusions

- We discussed four statistical models for the similarity score comparisons in firearm evidence identifications based on pass-or-fail tests, specifically the CMC method. In this case, binomial model seems a default or generic one.

- Because the assumption of independence among the cell pair comparisons from the CMC method is most likely invalid in practice, the correlated binomial model was proposed to relax the assumption.

- Although using the beta-binomial distribution can relax the assumption of the same p for all image pair comparisons, it still assumes that within each image pair, all cell pair comparisons are independent. Thus, the beta-correlated binomial was proposed.

- From the point view of the number of assumptions, the beta-correlated binomial, beta-binomial, and correlated binomial models are better than the binomial model.

- Using the smallest sum of squares of difference between the generalized frequency function and each of the four mixture pmf's as a performance criterion, we conclude that the correlated binomial, beta-binomial, and beta-correlated binomial models fit the example KM data much better than the binomial model.

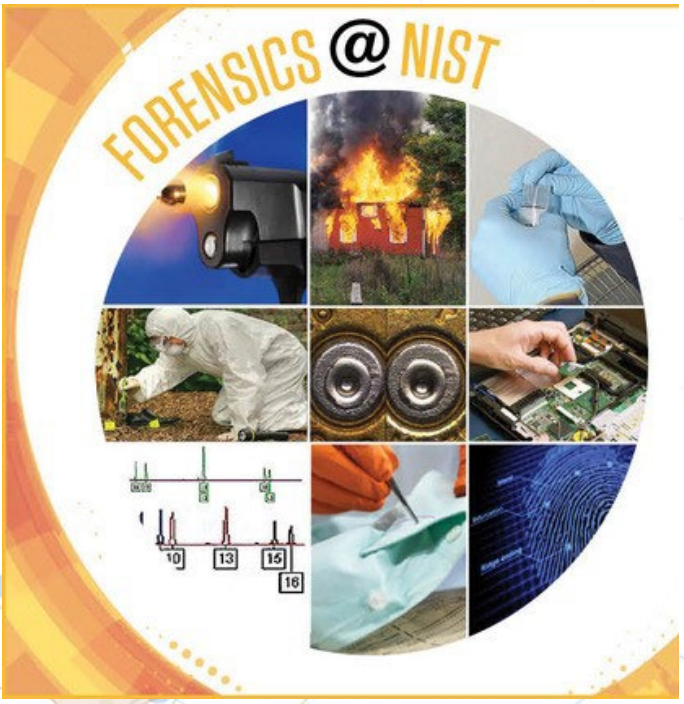
References

1. J. Song, T. V. Vorburger, W. Chu, J. Yen, J. A. Soons, D. B. Ott, and N. F. Zhang (2018) Estimating error rates for firearm evidence identification in forensic science, *Forensic Science International*, 284, 15-32.

2. N. F. Zhang (2019) The use of correlated binomial distribution in estimating error rates for firearm evidence identification, *Journal of Research of the National In statute of Standards and Technology*, 124, Article No. 124026.

3. A. W. van der Vaart (1998) Asymptotic Statistics, Cambridge University Press, Cambridge, UK, 265-266.

NIST Footwear Impression Comparison System

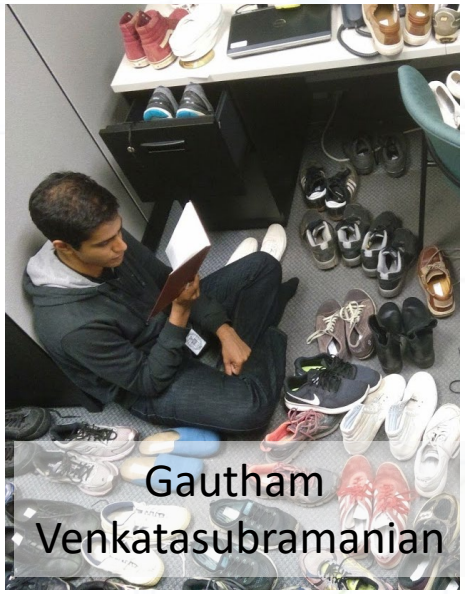


Steve Lund
Statistician
Statistical Engineering Division
steve.lund@nist.gov

National Institute of Standards and Technology
U.S. Department of Commerce

Acknowledgements

NIST Footwear Research Team:



Günay Doğan

FBI Examiners

Brian McVicker

Mike Gorn

Funding provided by

NIJ Award DJO-NIJ-17-RO-0202

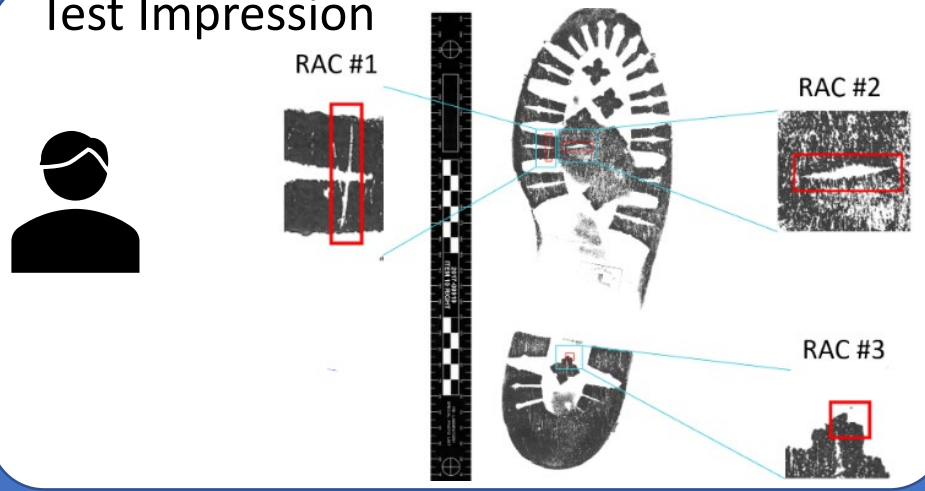
NIST SPO

NIST ITL

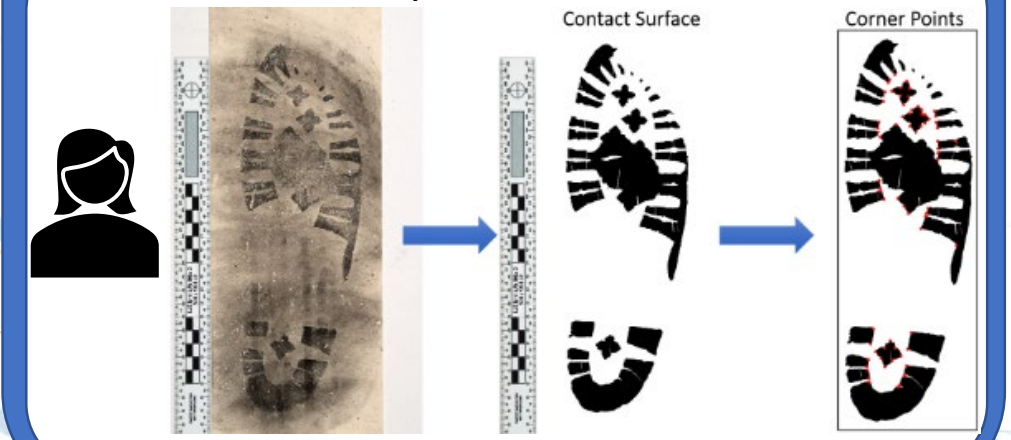
Version 0 - Workflow

Manual Annotation

Test Impression

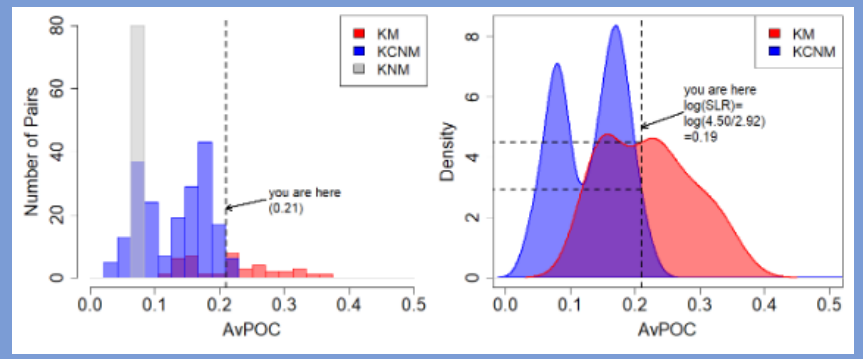


Questioned Impression



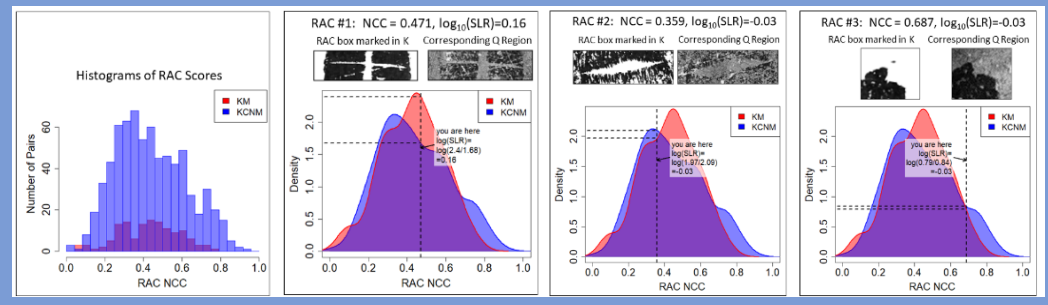
Automated Alignment and Comparison

Pattern: Design, Size, and Wear

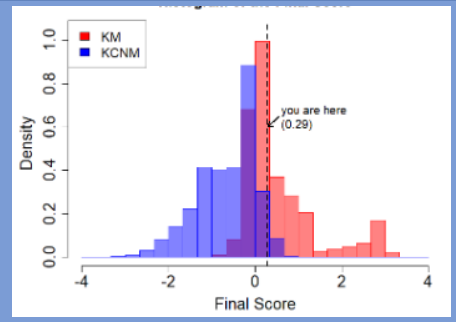


+

RACs



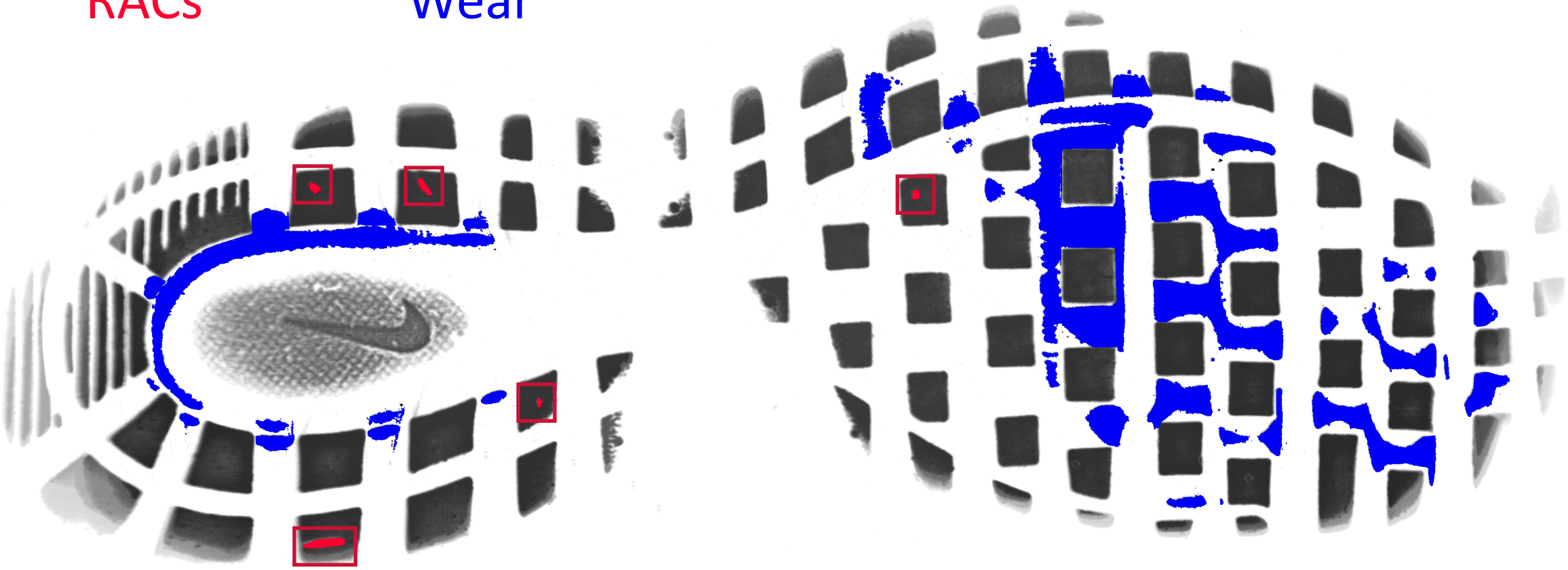
Final



Revamp with Casework Focus

Test Impression Markup

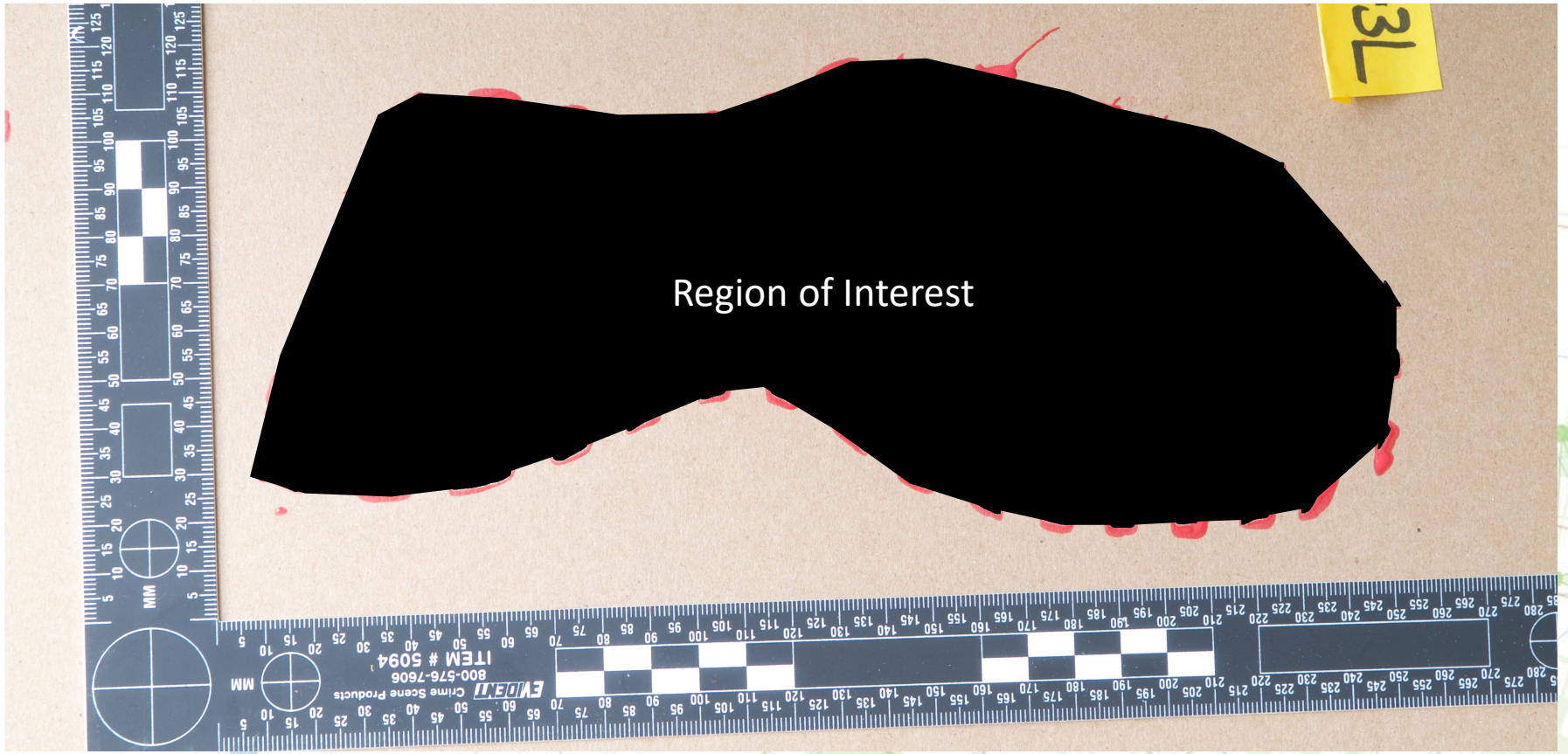
RACs Wear



Crime Scene Markup

RACs

Contact



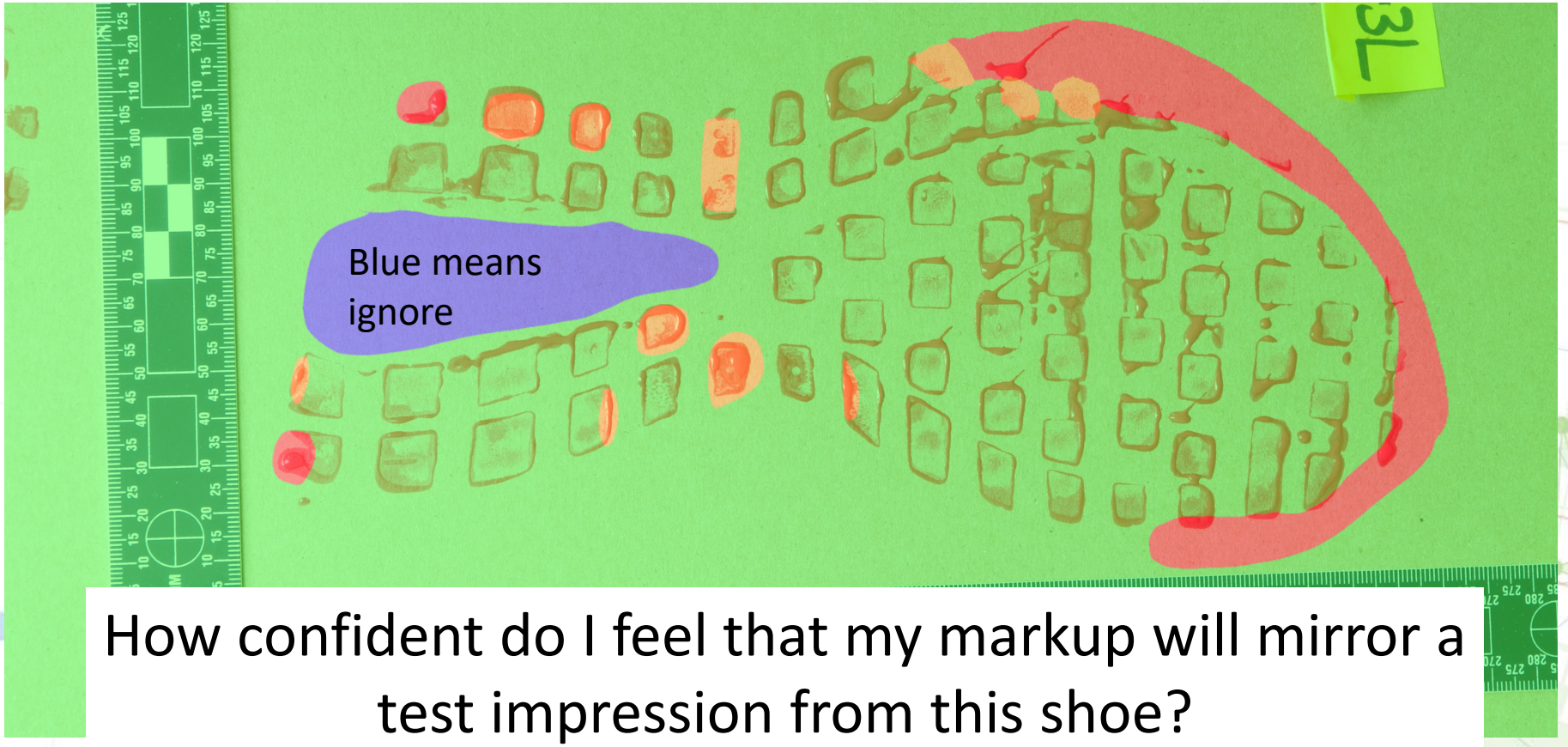
Crime Scene Markup

Red: I can hardly see anything

Orange: I can generally tell contact apart from noncontact

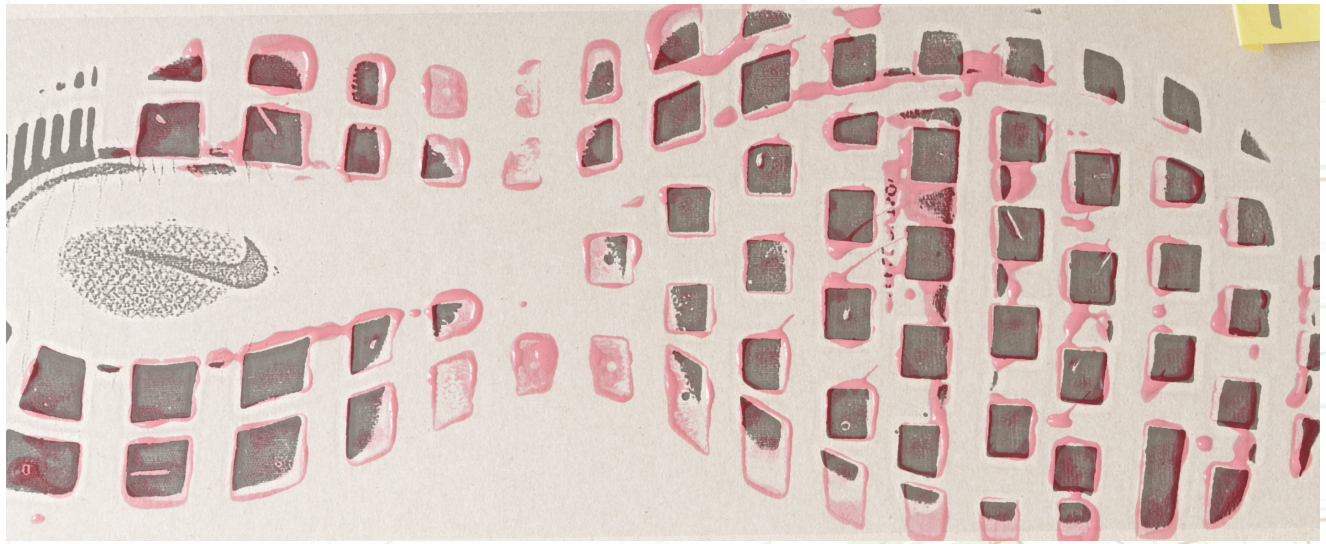
Yellow: I can even see the edges clearly

Green: I think I could see RACs!



Alignment

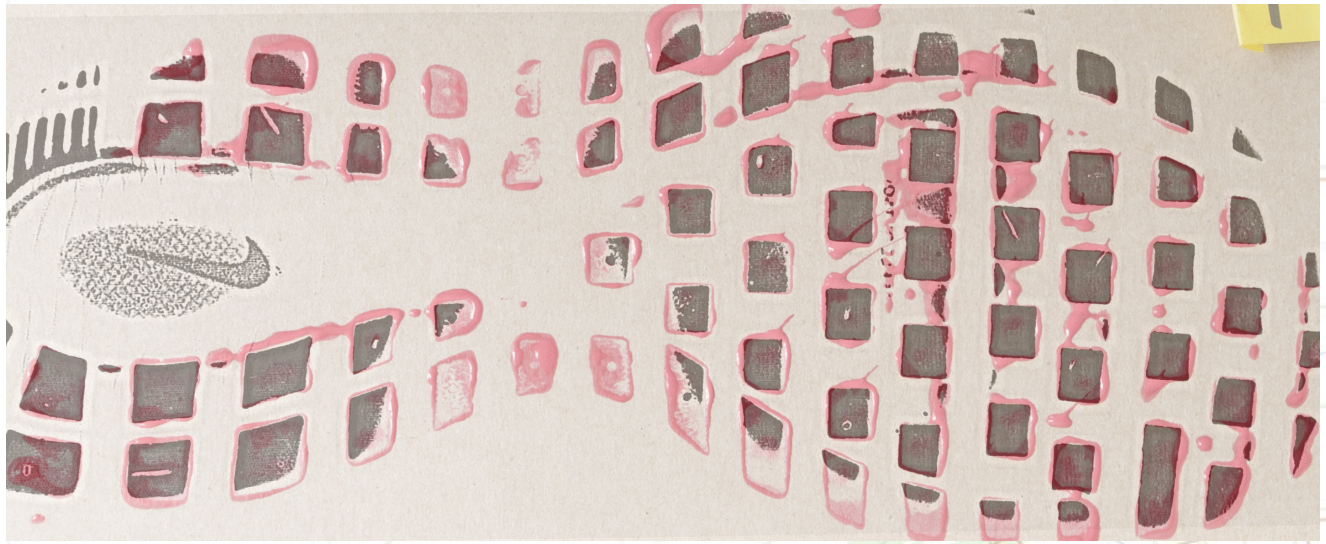
- First conduct a rigid alignment (slide and rotate, like doing a puzzle)



Rigid Alignment

Alignment

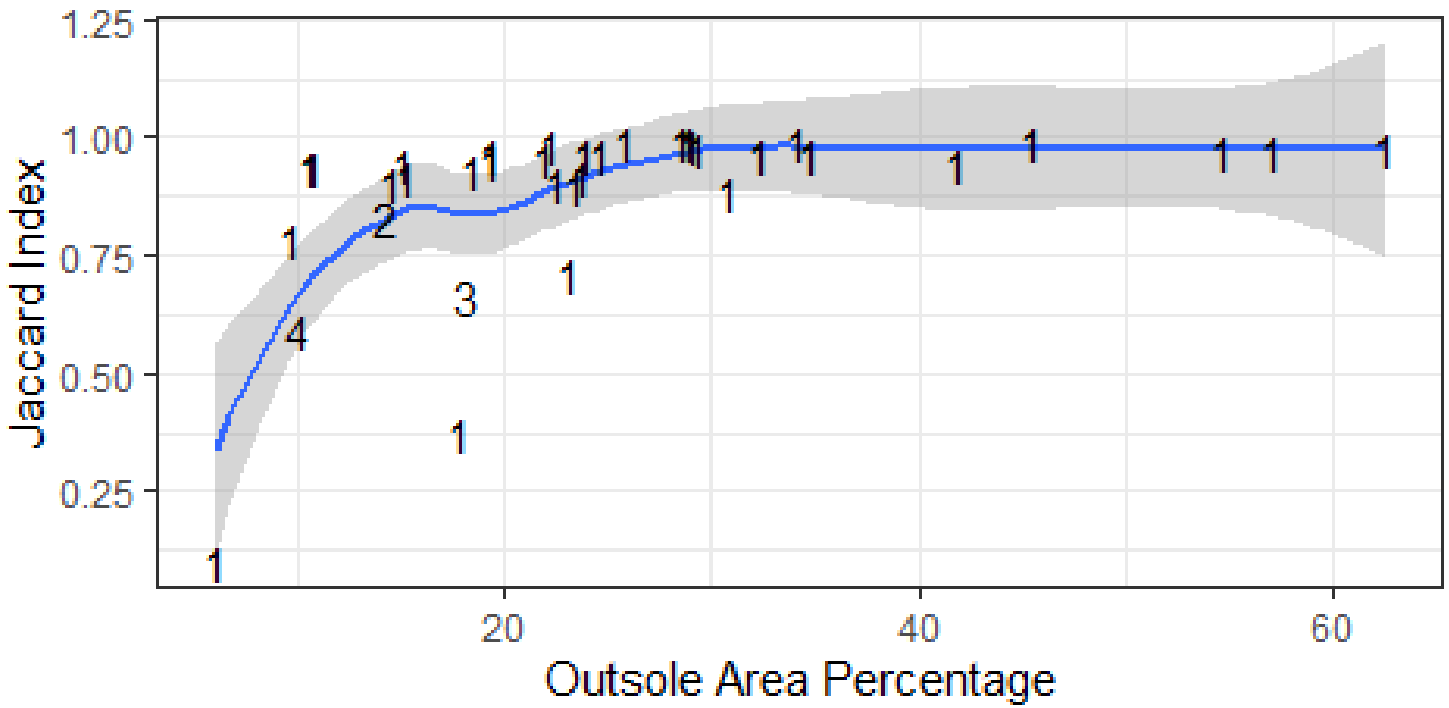
- First conduct a rigid alignment (slide and rotate, like doing a puzzle)
- Then allow moderate stretching to allow for distortions between impressions



Flexible Alignment

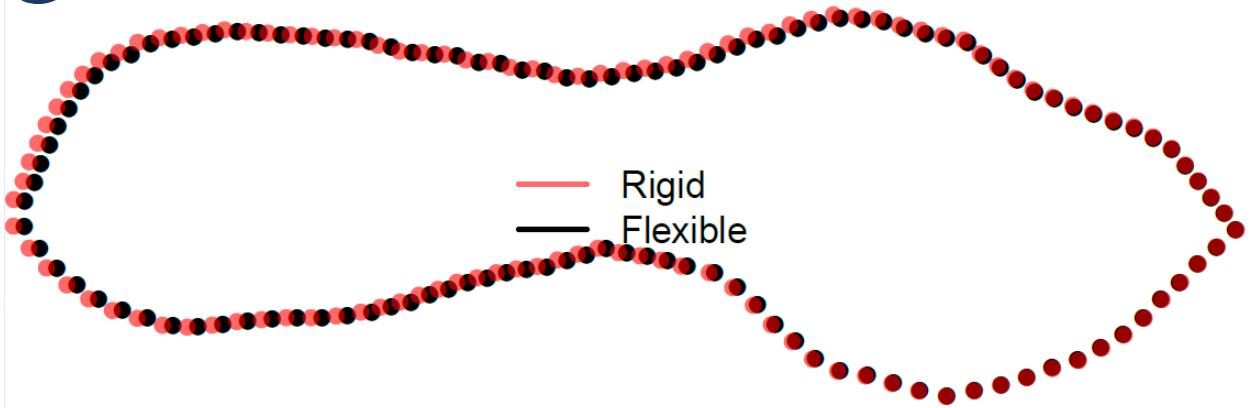
Alignment

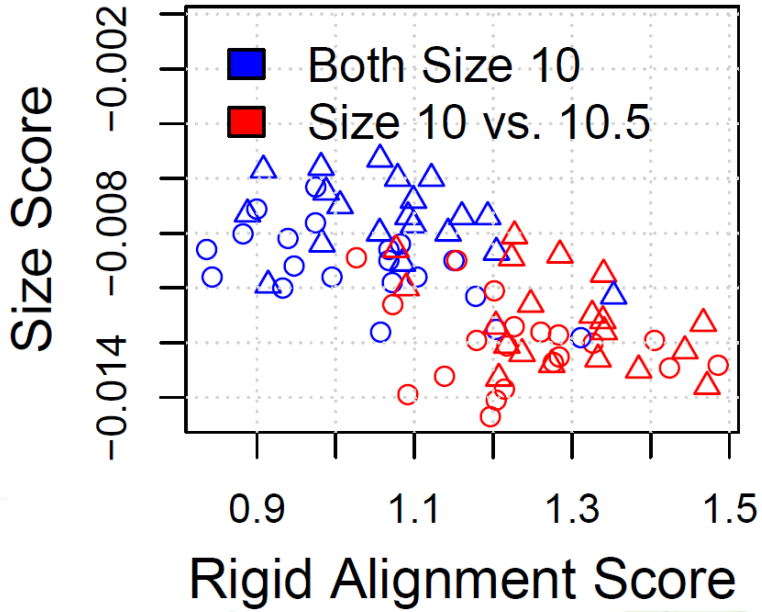
- First conduct a rigid alignment (slide and rotate, like doing a puzzle)
- Then allow moderate stretching to allow for distortions between impressions
- Promising results with partial impressions representing as little as 20% of the outsole surface

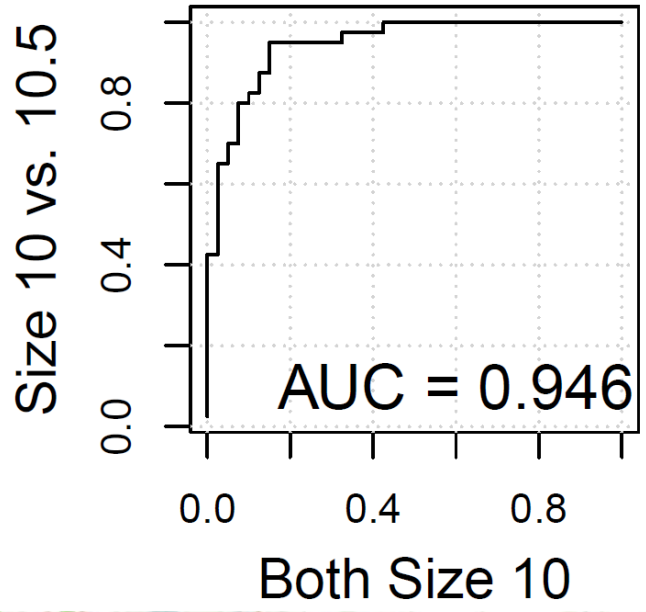


Size Metric

- Contracting boundary → K size > Q size
- Expanding boundary → K size < Q size
- Stable boundary → depends on shoe manufacturing process





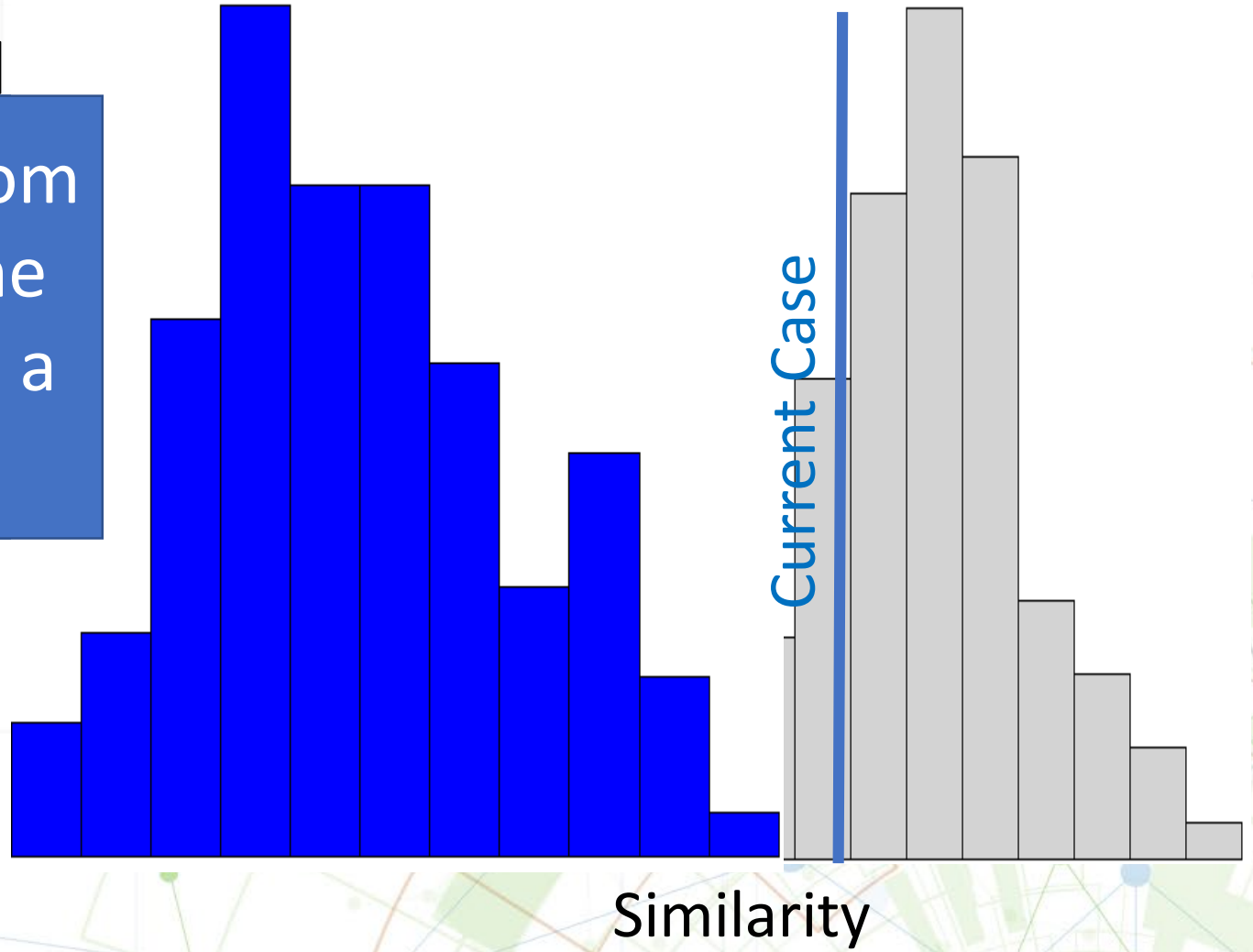


Design and Wear Metrics

- Several Similarity Metrics
  - Agreement of contact and co
  - ed
  - Re
  - Pix
- Each metric can be placed in different contexts…

Could Q and K be from shoes with the same side and design but a different size?

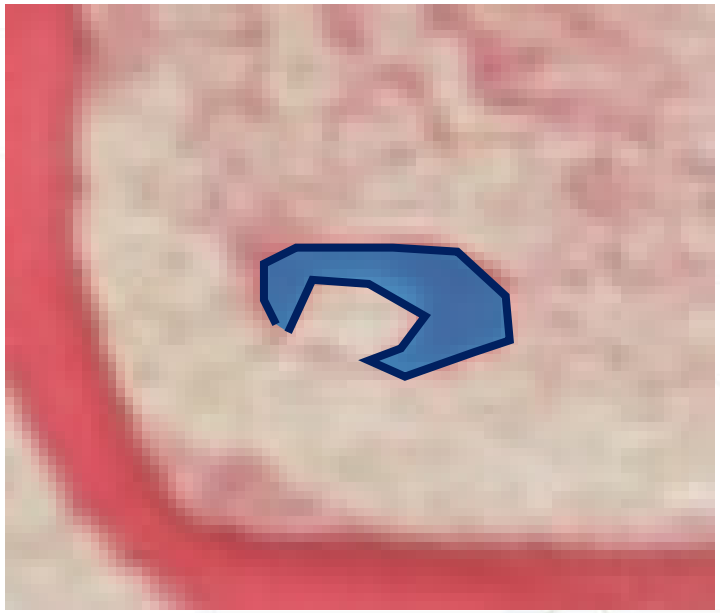
Results From Matched Comparisons with Reference Comparisons



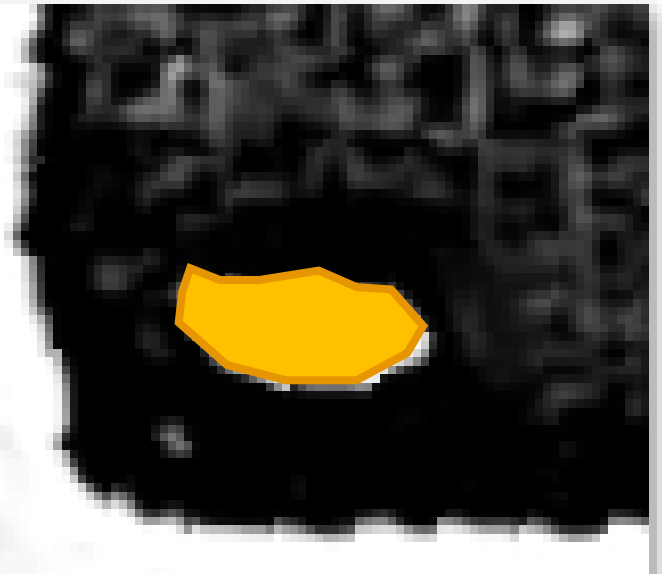
RAC Metrics

Any apparent RACs seen in the questioned impression are marked

After alignment, use overlap % (a.k.a. Intersection over union) with any RAC regions marked in test impression



Crime Scene

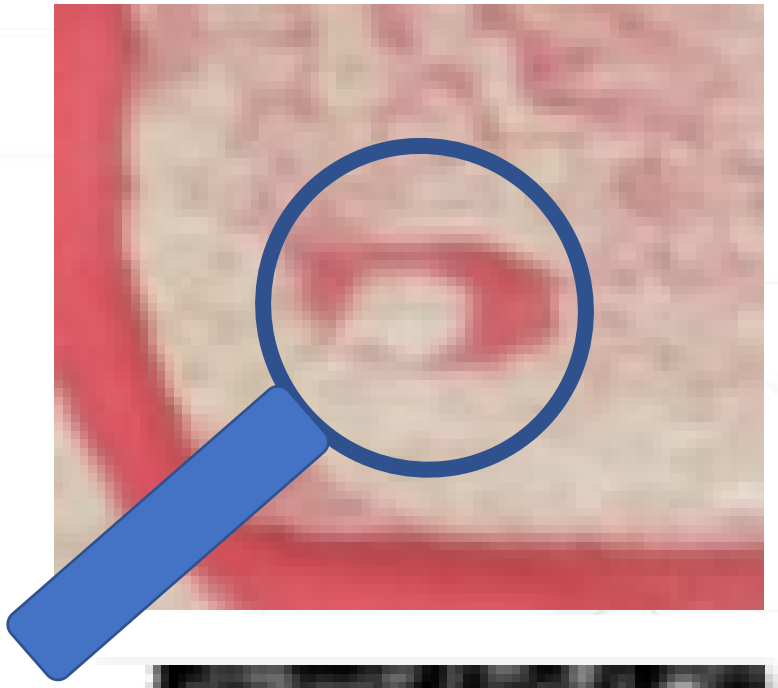


Test Impression

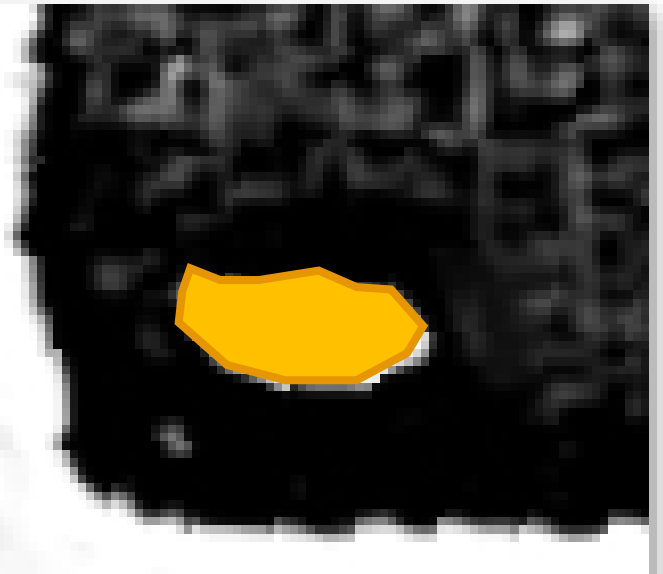
RAC Metrics

Credit: Pixabay

Developing RAC metric that uses regions marked in test impression to initiate local pattern comparison



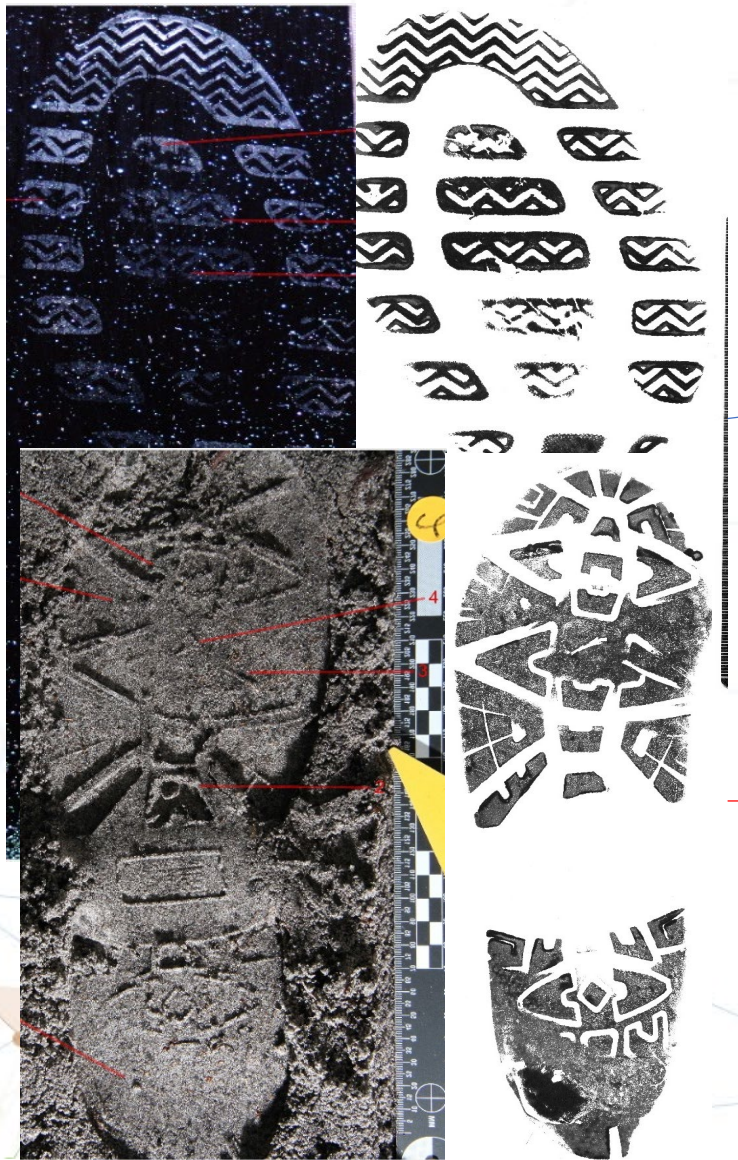
Crime Scene

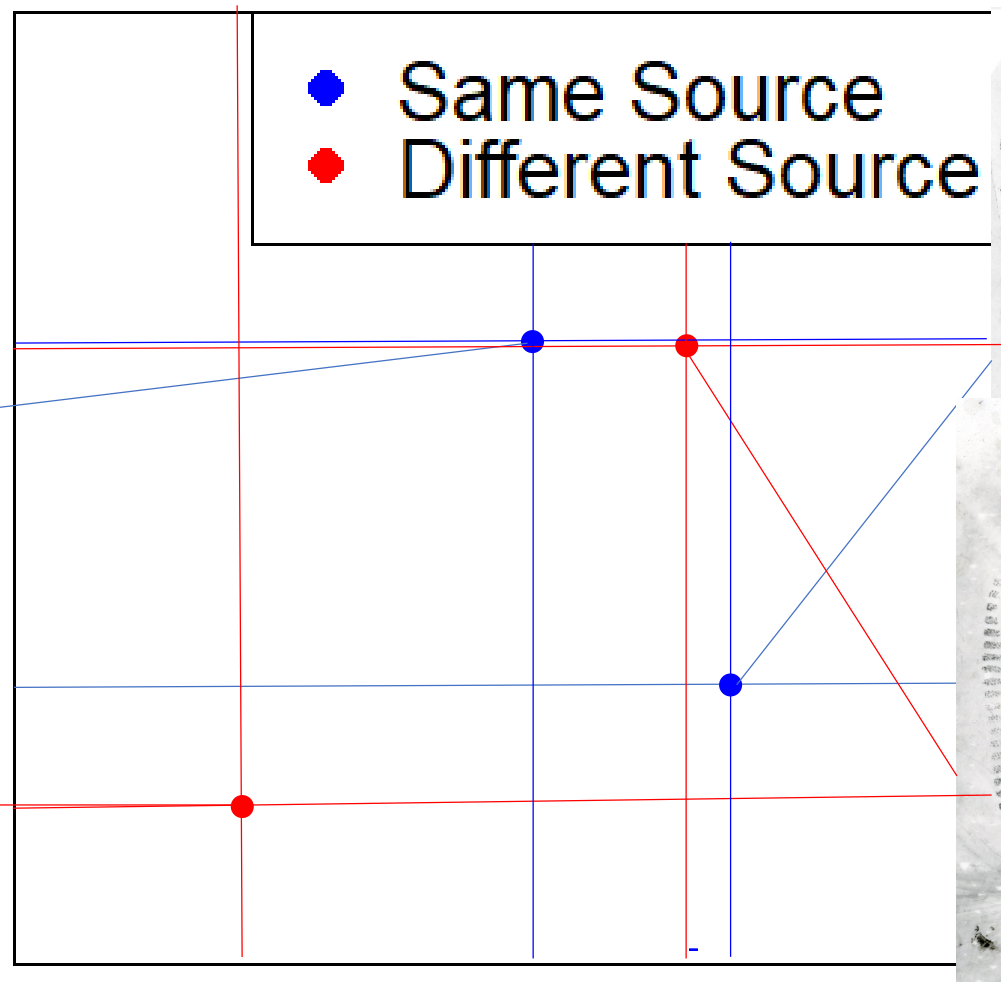


Test Impression

Relevance Metrics

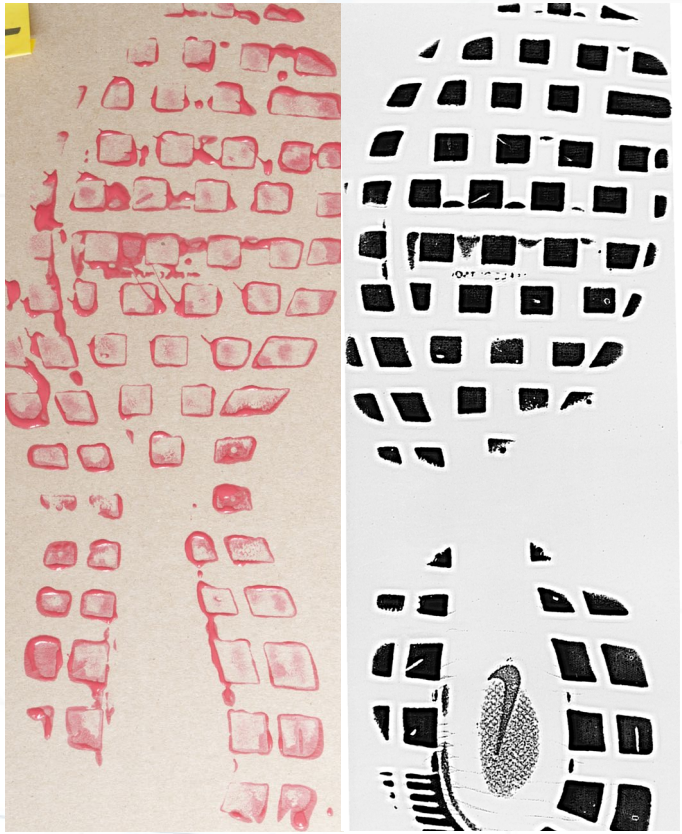
Credit: Pixabay



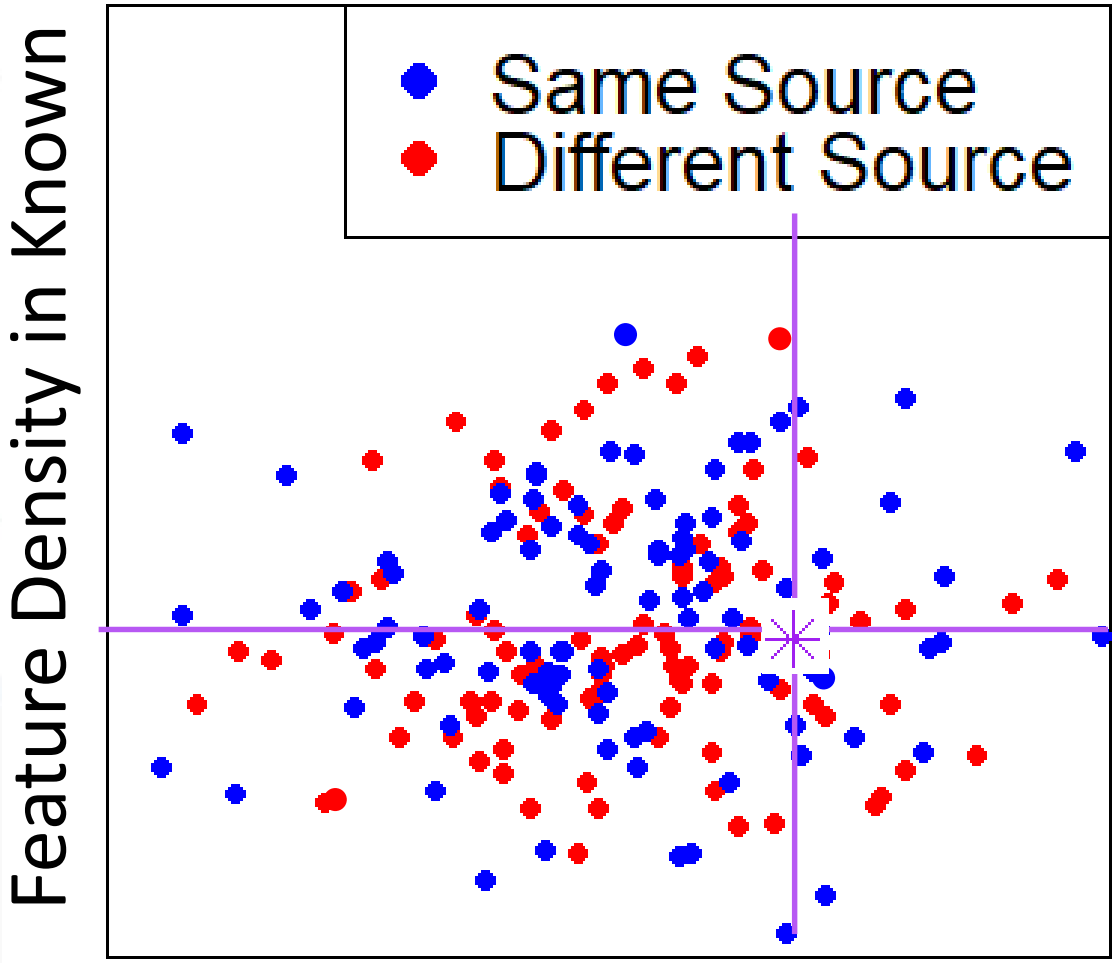


Relevance Metrics

redit: Pixabay

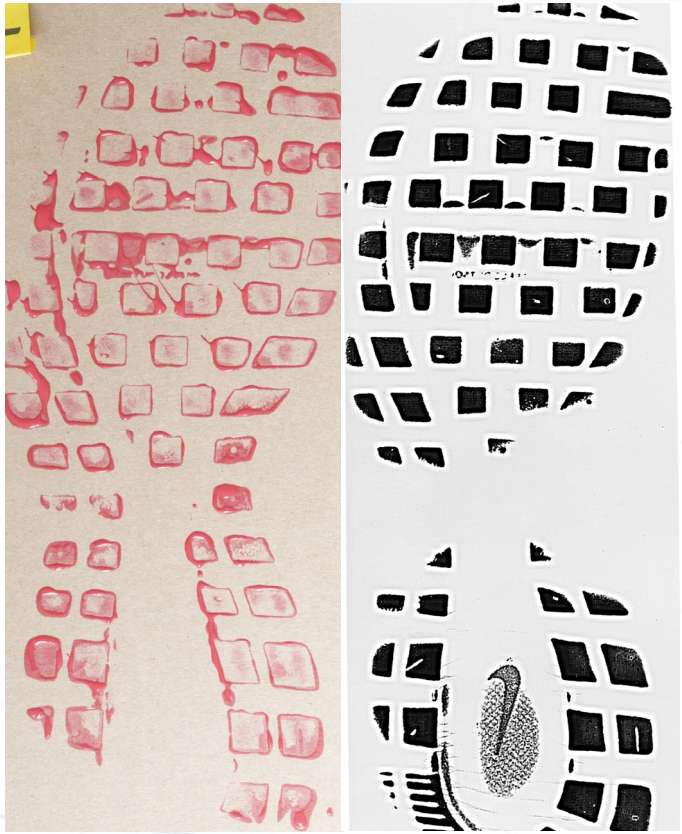


Case Comparison

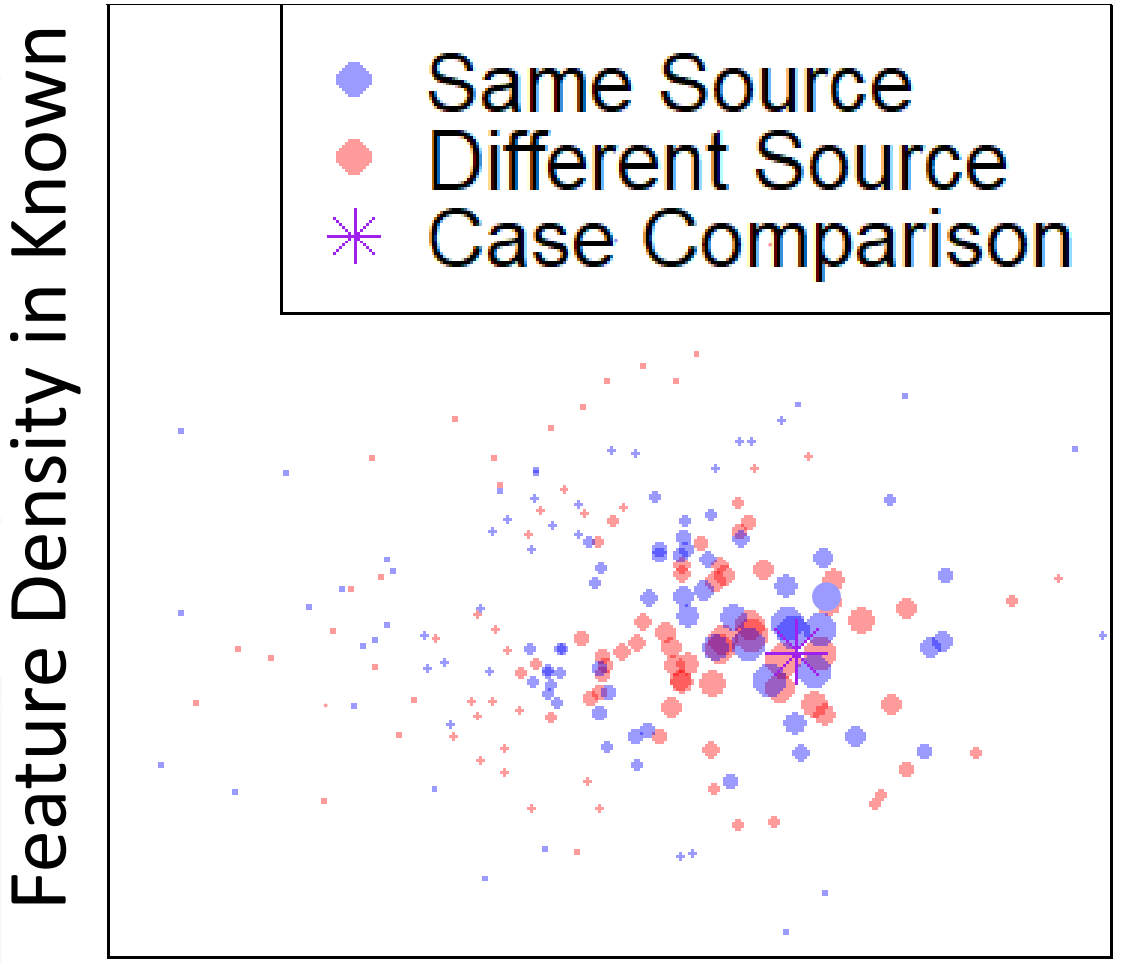


Relevance Metrics

redit: Pixabay

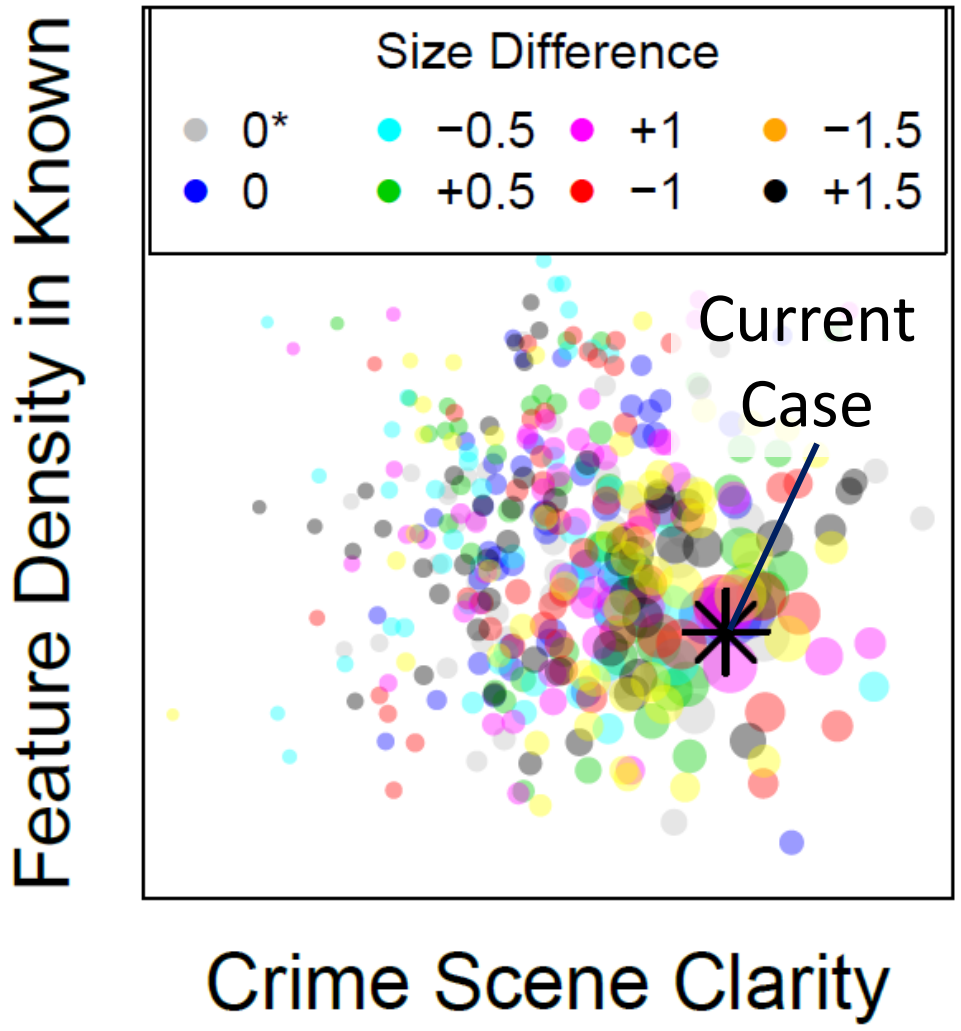


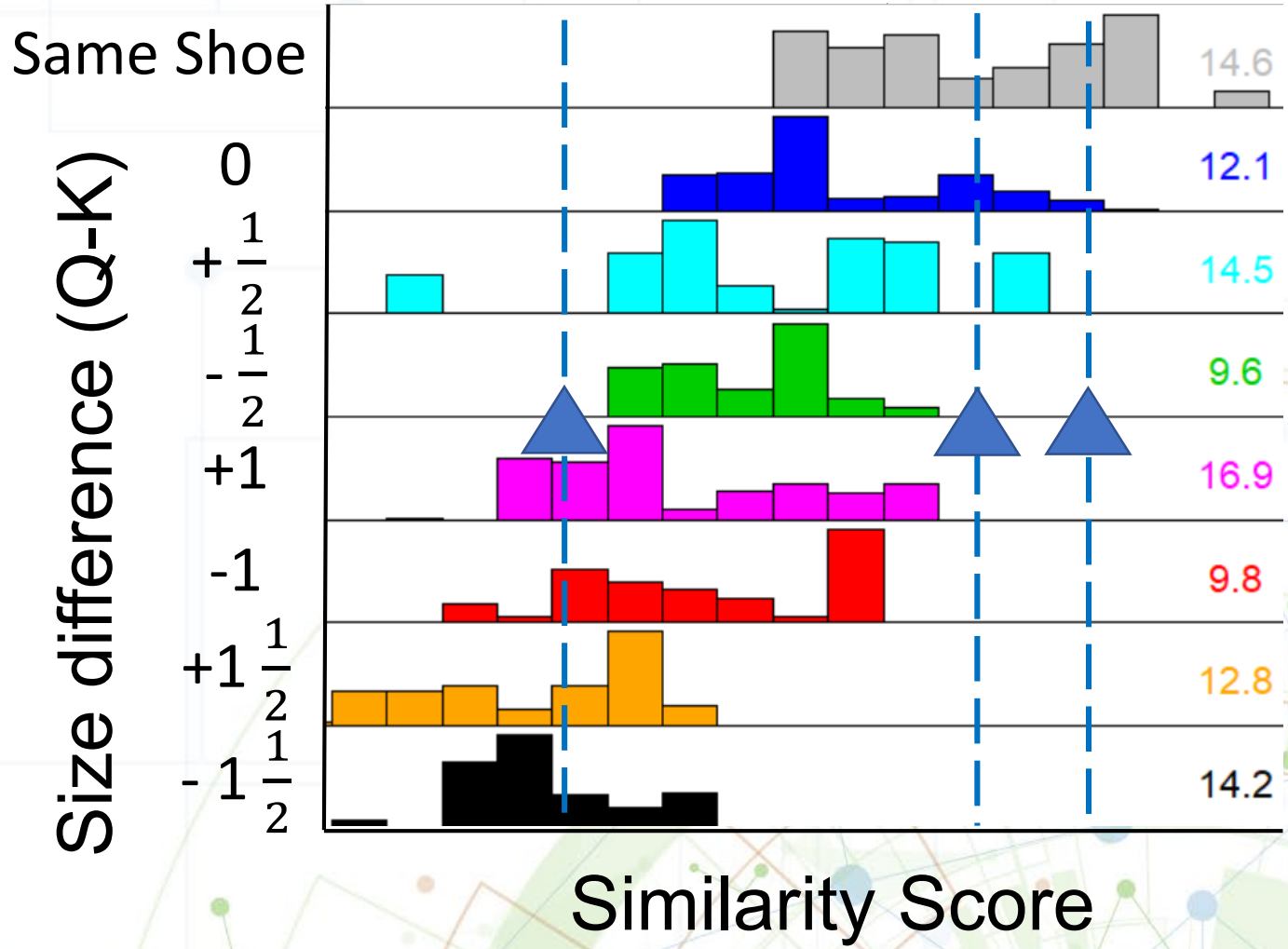
Case Comparison



Data Visualization

Credit: Pixabay

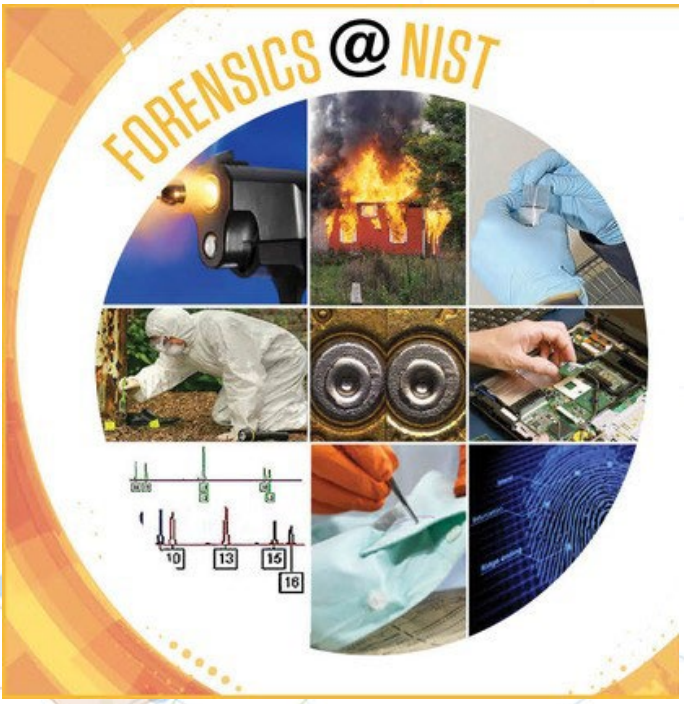




- Develop automated tools to assist with markup
- Use images from black box studies to evaluate and improve FICS
- Use NIST FICS outputs and black box responses to predict distribution of conclusions across examiners

Credit: Nick Youngson of NYPhotographic.com

A New Statistical Procedure to Assess Calibration Accuracy of Likelihood Ratio Systems



Hari Iyer
Statistician
Statistical Engineering Division
hari@nist.gov

National Institute of Standards and Technology
U.S. Department of Commerce

Joint work with Prof. Jan Hannig

University of North Carolina – Chapel Hill &

Faculty Appointee, SED/ITL/NIST

Acknowledgements and Disclaimers

- I would like to thank Steve Lund and William Guthrie for many useful discussions on topics related to performance assessment of LR systems.

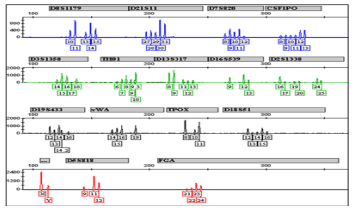
- Points of view are of the presenter and do not necessarily represent the official position or policies of the National Institute of Standards and Technology.

- Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

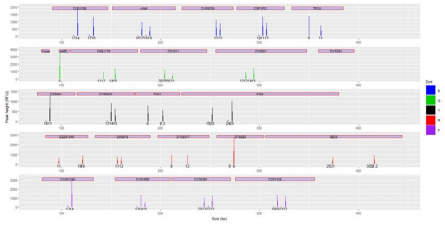
OUTLINE

- Likelihood Ratio (LR) & LR Systems
- Calibration accuracy of LRs
- Use of validation data in judging LR calibration accuracy
- Proposed method for inference regarding LR calibration discrepancy
- Example(s)

Likelihood Ratio



$$LR = \frac{Pr(E|H_p, I)}{Pr(E|H_d, I)}$$



H_p: DNA from POI is in the crime sample

H_d: DNA from POI is not in the crime sample

I = Background information available prior to examining crime sample (known or assumed to be true)

Likelihood Ratio

$$LR = \frac{Pr(E|H_p, I)}{Pr(E|H_d, I)}$$

H_p: Glass fragments recovered from POI's clothing is from the broken glass door at the scene of the burglary

H_d: The recovered fragments are not from the crime scene broken glass door

I = Background Information available prior to examining sample obtained from the POI's shirt

Likelihood Ratio Systems

- Measurement
  - Sample processing & Numerical quantification of informative features

- Interpretation
  - Statistical/Mathematical Modeling, data analysis, expert's judgements

Likelihood Ratio Systems (DNA Evidence)

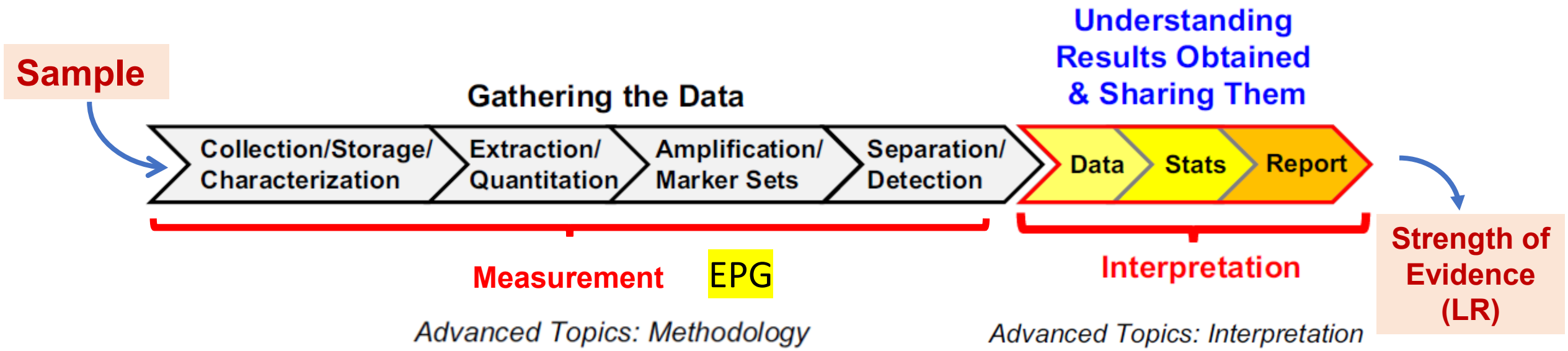


FIGURE 1.1 Steps involved in the overall process of forensic DNA typing. This book focuses on understanding the data through data interpretation and statistical interpretation.

JOHN M. BUTLER

National Institute of Standards and Technology
Gaithersburg, Maryland, USA

Advanced Topics in Forensic DNA Typing: Interpretation
http://dx.doi.org/10.1016/B978-0-12-405213-0.00001-4

Likelihood Ratio – What is it in plain English?

Suppose a forensic expert determines that an LR value of 1000 provides the best assessment of the value of the evidence.

When asked to explain what LR = 1000 actually means, they might say something like

The evidence is 1000 times more likely if Hp is true than if Hd is true (or something to this effect)

Meaning of Calibration Property of LR Systems

When considering RELIABILITY of expert witness testimony (or report), a key question is:

Is there empirical data to support the expectation that the system used by the expert is accurately assessing the value of evidence?

- What does empirical validation data say about ‘how much more likely the evidence is IF Hp is true than IF Hd is true?
- How does this compare with the expert’s assessment?

Some Available Diagnostic Tools

Turing's Lemma:

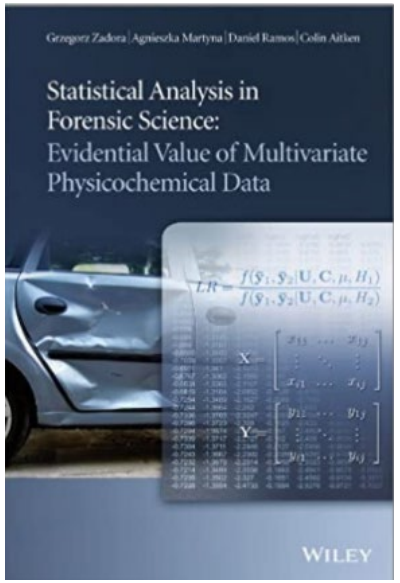
If the LR system is providing proper probabilistic assessments of value of evidence, then

the average of LR values obtained from Hd true cases must equal 1.

(this is a necessary condition but not sufficient)

Some Available Diagnostic Tools

Empirical Cross-entropy Analysis (ECE Plots)



Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, Colin Aitken

6 Performance of likelihood ratio methods

6.2 Empirical measurement of the performance of likelihood ratios

6.5 Accuracy equals discriminating power plus calibration: Empirical cross-entropy plots

- the discriminating power is poor. This means that the validation set of *LR* values is poor at separating *LR* values for which H_1 is true from *LR* values for which H_2 is true.
- the calibration is poor. This means that the *LR* values provide poor probabilistic measures of the value of the evidence. Even if the *LR* values have high discriminating power, poor calibration can degrade the accuracy considerably.

An R-package called **comparison** **can be used to apply their method**

David Lucy, James Curran, Agnieszka Martyna

Calibration discrepancy is assessed as part of an overall prediction accuracy analysis by comparing posterior odds (across all possible prior odds) obtained using expert's LR.

Our Approach

LR systems that provide accurate probabilistic assessments of value of evidence (relative to empirical observations) must satisfy the following property:

LR of LR = LR

Green and Swets, 1966, page 26, section 1.8, equation (1.32)

If $g(x)$ is the probability density function for Hp-true LRs

and

$f(x)$ is the probability density function for Hd-true LRs

then

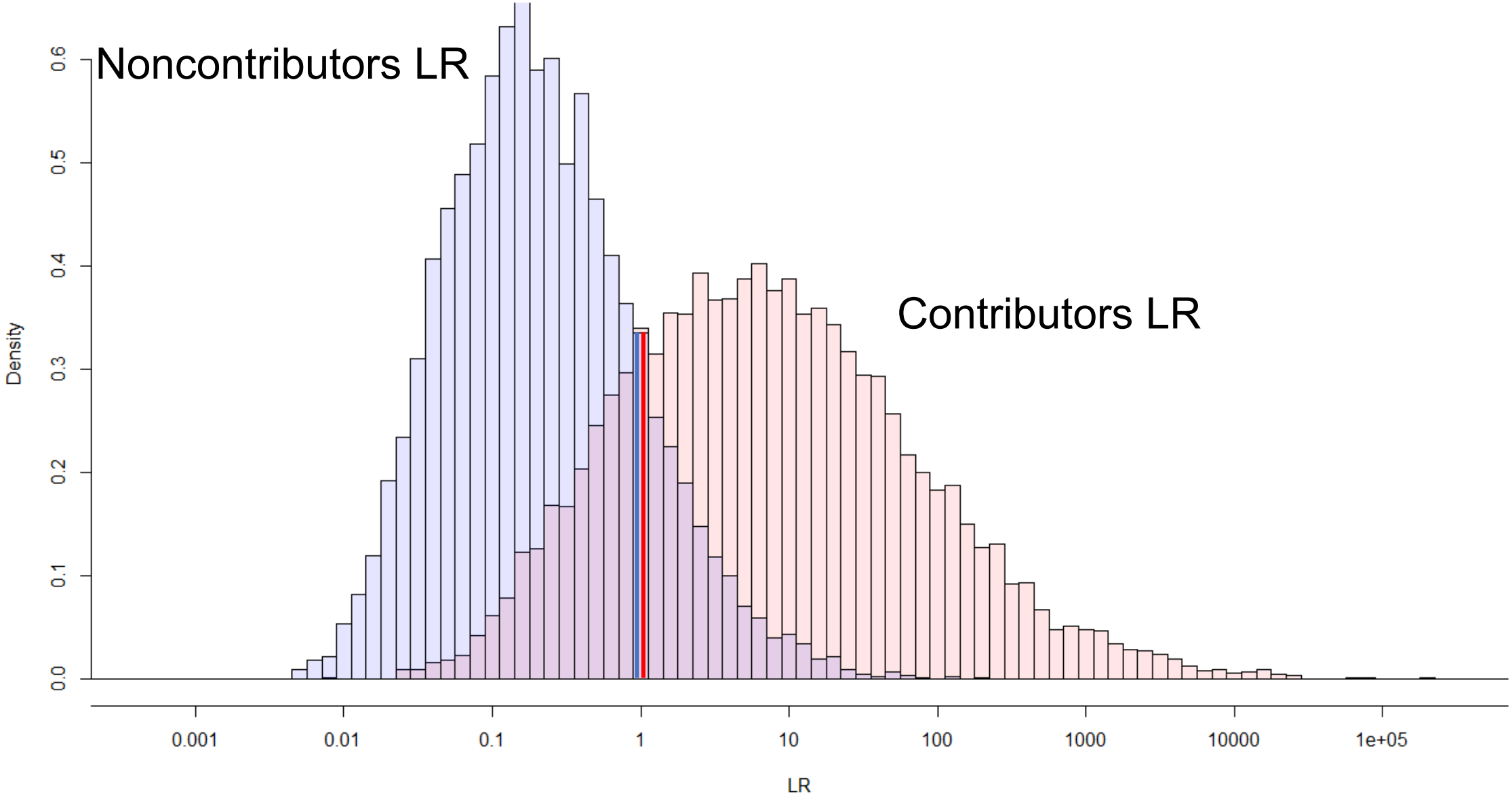
$$\frac{g(x)}{f(x)} = x \quad \text{for all } x$$

What does LR of LR = LR mean?

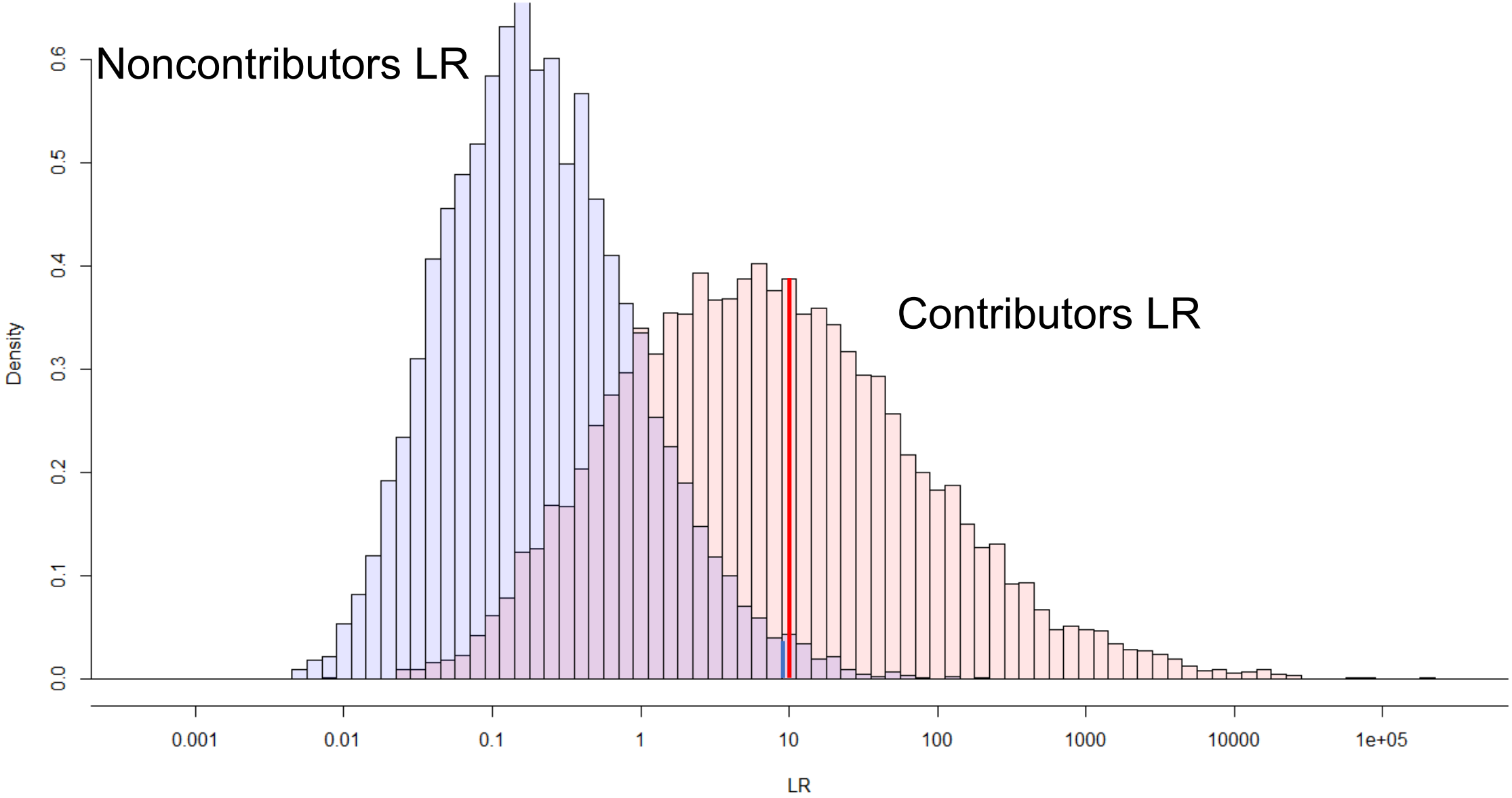
- LR value of 1 is equally likely under H_p as it is under H_d
- LR value of 10 is 10 times more likely to occur under H_p than it is under H_d.
- LR value of 100 is 100 times more likely under H_p than it is under H_d.
- LR value of 0.1 is 10 times more likely under H_d than it is under H_p.

LR value of x is x times more likely to occur under H_p than under H_d.

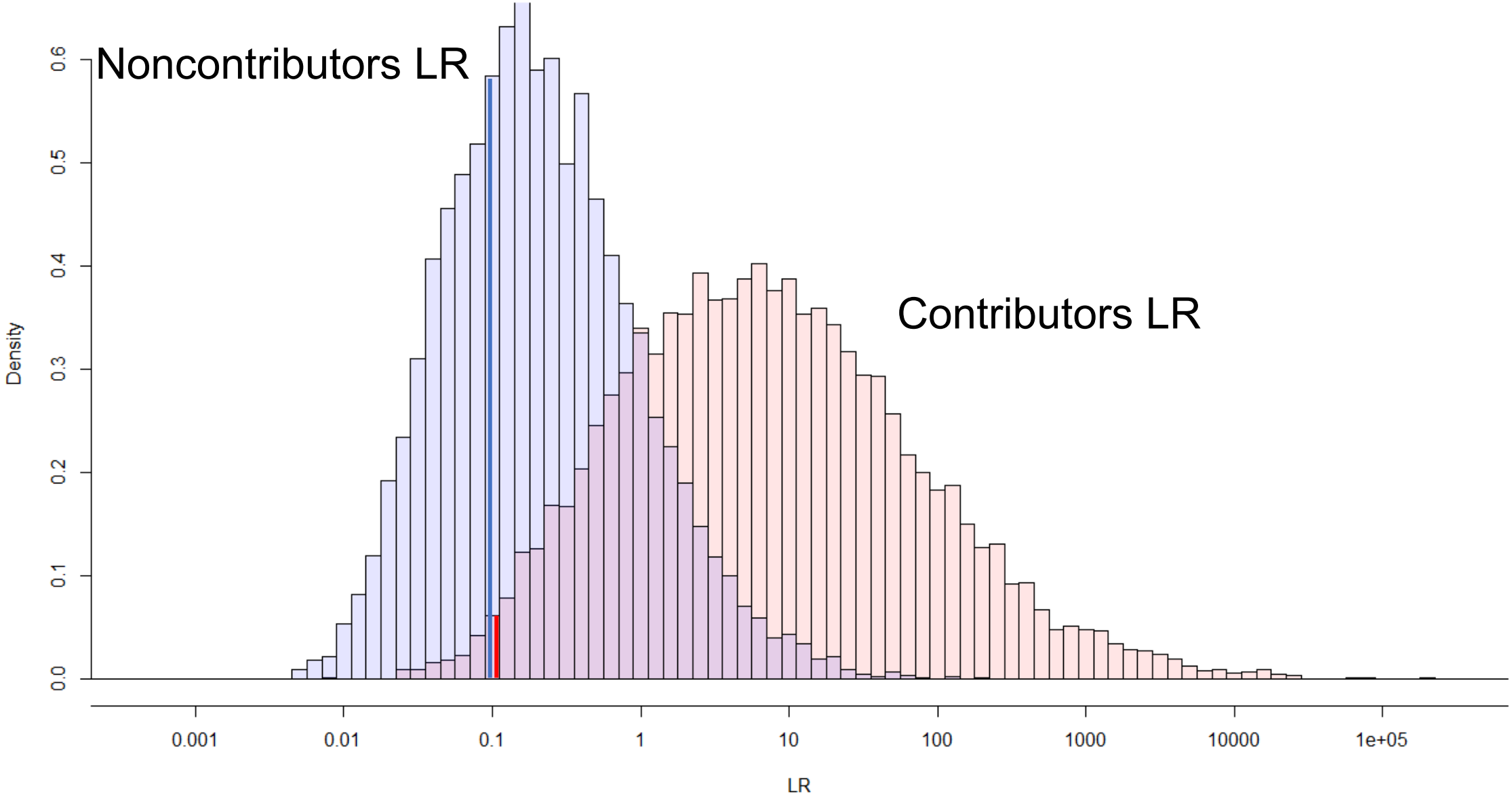
Calibration Accuracy: Empirical Assessment



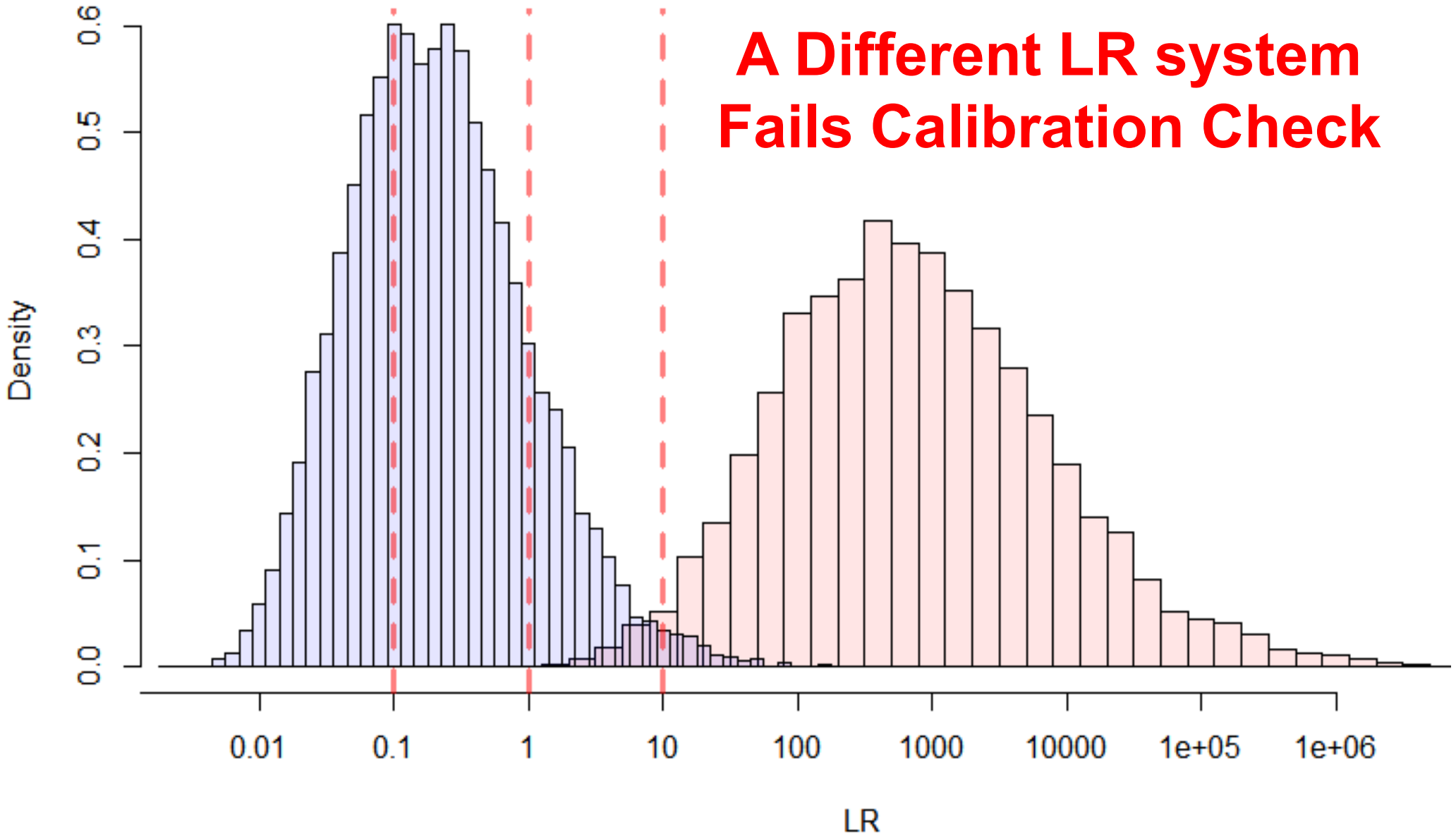
Calibration Accuracy: Empirical Assessment



Calibration Accuracy: Empirical Assessment



Calibration Accuracy: Empirical Assessment



Our Approach

Rather than compare frequencies of occurrence of a given LR value in Hp-true cases and Hd-true cases we focus on frequencies of LR values that fall within any specified interval.

For instance, we can count the number of LR values that fall in the intervals [1,100], [100,10000], [10000, 1000000], etc.

Such counts in Hp-true cases and Hd-true cases must exhibit a relationship that is dictated by the property "LR of LR = LR".

$$G(b) - G(a) = b\,F(b) - a\,F(a) - \int_a^b F(x)\,dx$$

$G(x)$ is the cumulative distribution function for Hp-true LRs

$F(x)$ is the cumulative distribution function for Hd-true LRs

Our Approach

For any given interval of interest, say (a, b), we count

(1) the actual counts for Hp-true samples in the interval

(2) expected counts if the LR system makes accurate weight of evidence assessments using

$$G(b) - G(a) = b\,F(b) - a\,F(a) - \int_a^b F(x)\,dx$$

and calculate the ratio of (1) to (2). This is the estimated calibration discrepancy.

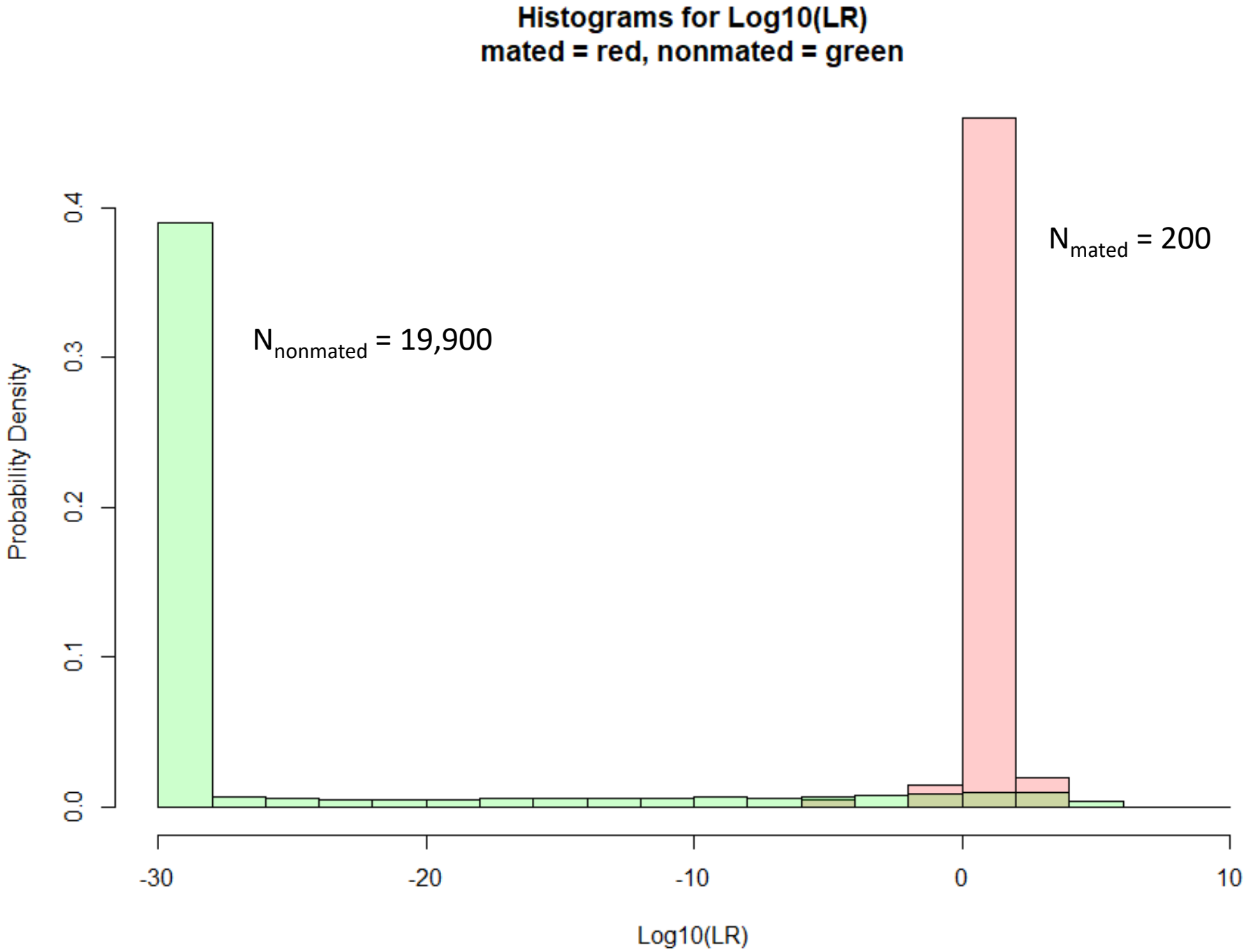
- If the ratio is less than 1 (log10 of the ratio is negative) then the LR system is **overstating the evidence** in the interval being considered.
- If the ratio is greater than 1 (log10 of the ratio is positive) then the LR system is **understating the evidence** in the interval being considered.

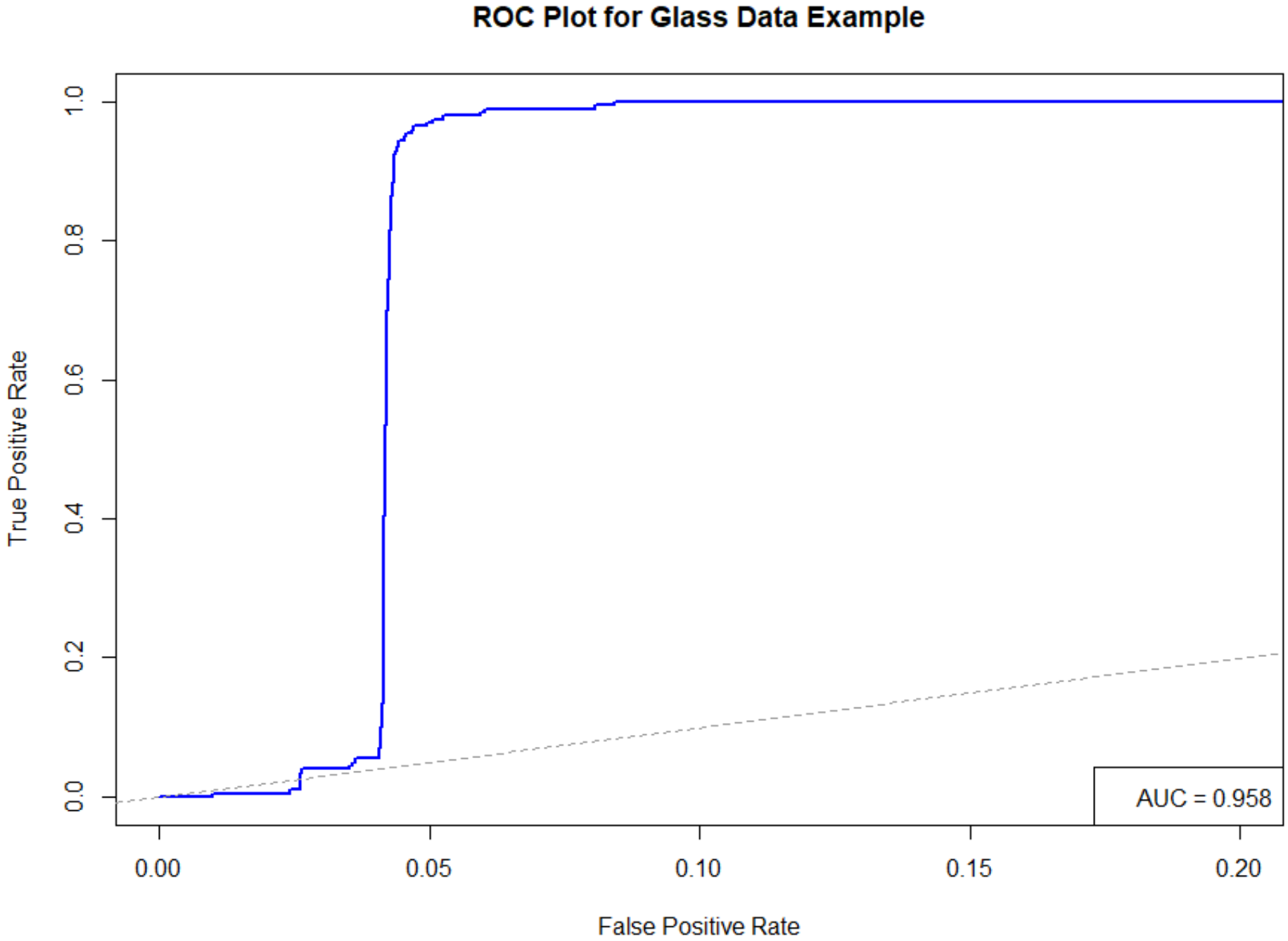
We can visualize the results in an "Interval specific calibration discrepancy plot" or simply "calibration discrepancy plot".

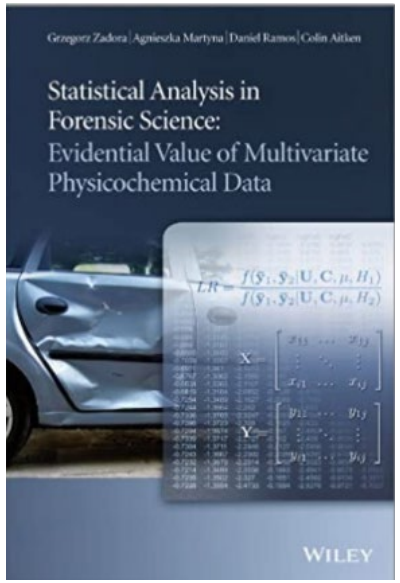
Glass Fragments Example

This example is discussed in the book by Zadora et al. (2013) (see Chapter 4, Section 4.4.6).

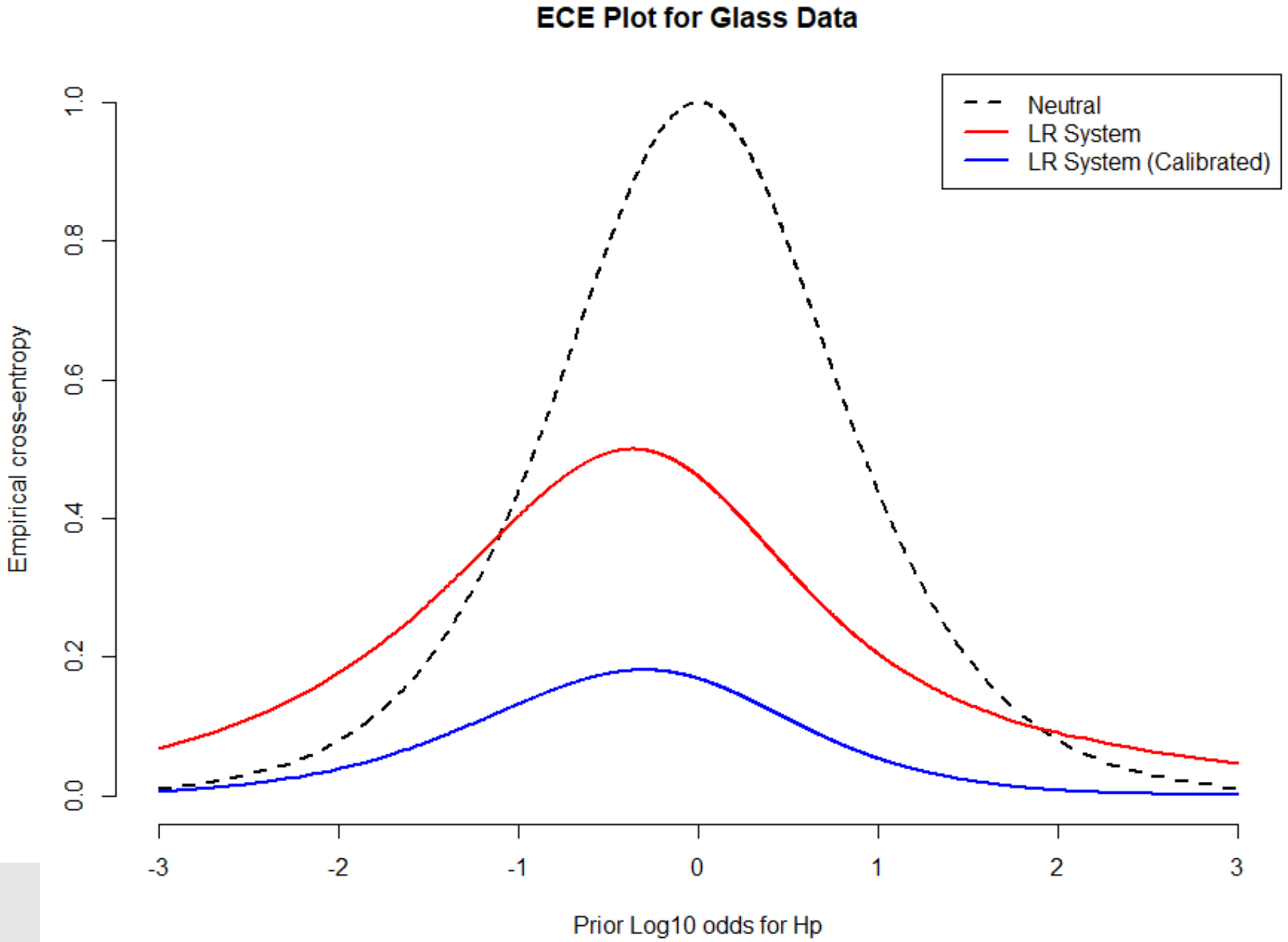
- Twelve fragments of glass were obtained from each of 200 glass objects and each fragment was subjected to an elemental analysis using scanning electron microscopy electron diffraction (SED-EDX).
- Eight elemental concentrations were measured:
  Na, Mg, Al, Si, K, Ca, Fe, and O.
- Base 10 logarithms of the ratios of the first 7 elemental concentrations to the concentration of O (Oxygen) formed the response variables in the analysis.
- Zadora et al. (2013) used a graphical model approach and kernel density estimation for computing the same-source LRs and different-source LRs (consult their book for further details).
- Here we focus on assessing how well-calibrated this particular LR system is.

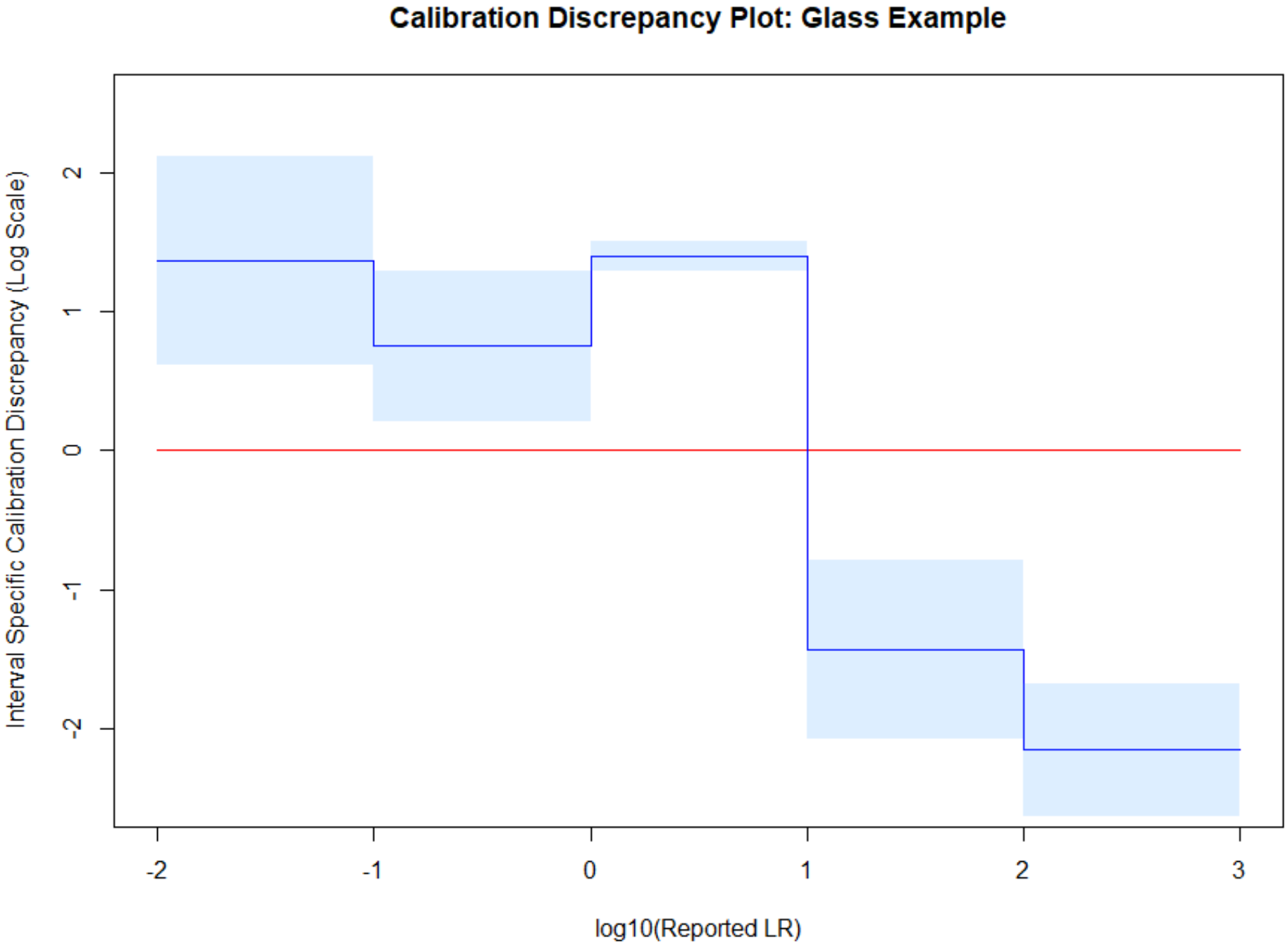

Histograms for Log10(LR)
mated = red, nonmated = green
Probability Density
0.0
0.1
0.2
0.3
0.4
N nonmated = 19,900
N mated = 200
-30
-20
-10
0
10
Log10(LR)

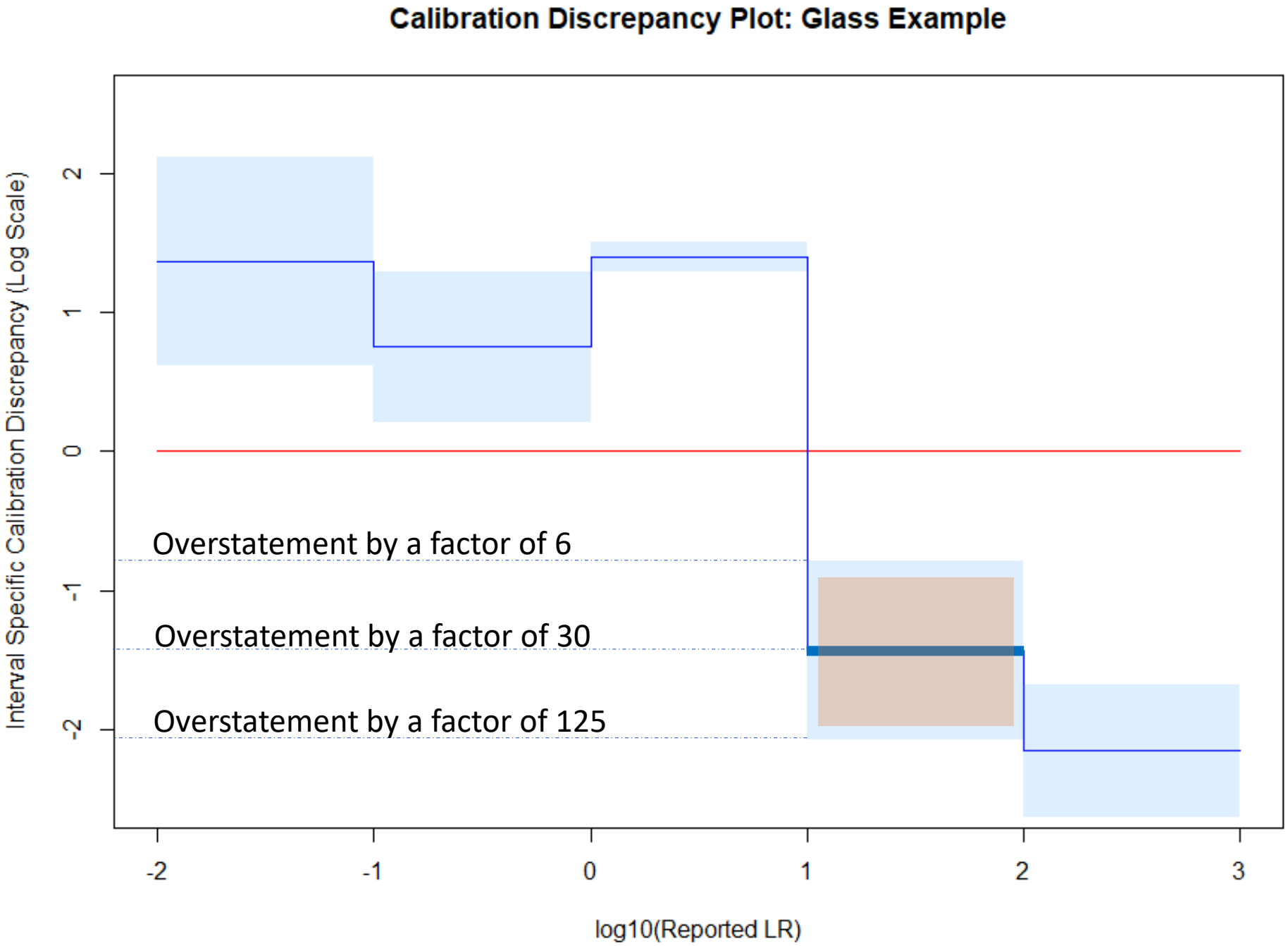

ROC Plot for Glass Data Example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
AUC = 0.958



Zadora, Martyna, Ramos, Aitken




Calibration Discrepancy Plot: Glass Example
Interval Specific Calibration Discrepancy (Log Scale)
2
1
0
-1
-2
-2
-1
0
1
2
3
log10(Reported LR)


Calibration Discrepancy Plot: Glass Example
Interval Specific Calibration Discrepancy (Log Scale)
2
1
0
-1
-2
Overstatement by a factor of 6
Overstatement by a factor of 30
Overstatement by a factor of 125
-2
-1
0
1
2
3
log10(Reported LR)

Summary

- CALIBRATION property is an important component of LR system assessment as it focuses directly on the accuracy of value of evidence assessments.

- We outlined a new method for examining potential degree of discrepancy between
  (a) evidential value inferred from validation data and
  (b) evidential value as assessed by an LR system (or an expert).

- We illustrated the ideas using an example.

BACKUP SLIDES