NCFS Statistics Panel – Sep 13, 2016

Outline

1. Karen Kafadar, University of Virginia:
   Quantifying Weight of Pattern Evidence: General Concepts
2. Hari Iyer, NIST:
   Probabilistic Statements regarding Weight of Evidence
3. Alicia Carriquiry, Iowa State University:
   Proposed "Views on Statistical Statements in Forensic Testimony": Comments, Issues, and the Role of Databases
4. David Kaye, Pennsylvania State University:
   Statistical Statements that Satisfy Daubert
5. Discussion

1. Quantifying Weight of Pattern Evidence: General Concepts

1. Scenario: Weight of Evidence

2. Analogs for Pattern Evidence

3. Quantifying Weight of (Statistical) Evidence

Scenario for Weight of Evidence (Frans Alkemade, NL)

- One bag, one target marble in the bag.
- The initial target marble is either red or white.
- No other information.
- 2 hypotheses: (A) Marble is red; (B) Marble is white.

What is the probability of (A), Marble is red?

- No data so far.
- No reason to prefer one or the other.
- “50-50.”

Collect data:

1. Trial #1: **Add 1 red marble to the bag.**
   Shake bag, draw out 1 marble. It is red.
   Probability that initial target marble is red?

2. Repeat trial: **Add 1 red marble to the bag.**
   Shake bag, draw out 1 marble. It is red.
   Probability that target marble is red?

3. Repeat trial: **Add 1 red marble to the bag.**

Shake bag, draw out 1 marble. It is red.

Probability that target marble is red?

4. Repeat trial: **Add 1 red marble to the bag.**

Shake bag, draw out 1 marble. It is red.

Probability that target marble is red?

5. Repeat trial: **Add 1 red marble to the bag.**

Shake bag, draw out 1 marble. It is red.

Probability that target marble is red?

Based on these 5 “features”:

1. **If** target marble were red, outcome is completely expected.
2. **If** target marble were **white**, outcome *could* happen:

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}$$

   Not impossible – but less likely than if target marble were red.

3. On each trial, each “feature” (observed color of marble) had a 50-50 chance of happening under either hypothesis.
4. Imagine if we had a more sensitive feature, that had a 90-10 chance of happening. Then the probability of the data is

$$\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = \frac{1}{100{,}000}$$

   It *could* happen - but less likely than if target marble were red.

Several ways to characterize “weight of evidence”:

1. Compare:

   (1) How likely is the outcome if target marble is red
   (very likely: probability = 1)

   (2) How likely is the outcome if target marble is **white**
   (less likely: probability = 1/32)

   Ratio = 1 / (1/32) = 32: "Data outcome is 32 times more likely to have occurred under (A) (red) than (B) (**white**)."

   Does not mean: "(A) is 32 times more likely than (B)."

2. Report both probabilities:

   Probability of Data if (A) is true = 1.00
   Probability of Data if (B) is true = 1/32, or 0.03

3. Methods to quantify consistency of evidence with hypotheses (Iyer)

All assume statistical models from which valid probabilities can be calculated.

Weight of Pattern Evidence requires DATA:

- Data: Measurements on many features (ideally, independent) (Ex: bifurcation, ridge ending, ridge spacings, ...)
- How likely are these features to correspond in 2 prints if: (A) same source, or (B) different sources?
- Want features are highly sensitive (Sensitivity): high probability of matching features IF **same** source
- Want features are highly specific (Specificity): high probability that **features** do **not** match IF different sources
- Then one hypothesis will be more plausible than the other.

From sensitivity/specificity to Weight of Evidence: Compare

(1) Probability of matching features IF same source
(Sensitivity)

(each marble draw is red if target marble is red)

(2) Probability of matching features IF different sources
(1 – Specificity)

(each marble draw is red if target marble is white)

If (1) is **much** bigger than (2), then we are inclined to believe (A) ("same color") is more probable than (B) ("different colors").

This does **not** mean that "same" is more likely than "different."

Sensitivity and Specificity characterize how common/rare the features are in the population, **NOT that "same" is more likely than "different"**

Challenges for Pattern Evidence: **Which Features?**

- Features are not pre-identified
- Features often are **selected** by examiner (latent prints, ballistics, tool marks, ...)
- Which features are highly sensitive to same-source samples?
- Which features are highly specific to distinct sources?
- Before seeing the evidence, examiner does not even know which features (s)he will identify — much less whether they are *independent* of each other, whether they are sensitive (to same source) or specific (to different sources)

From Ulery et al, *PLoS One* 9(11), 2014:

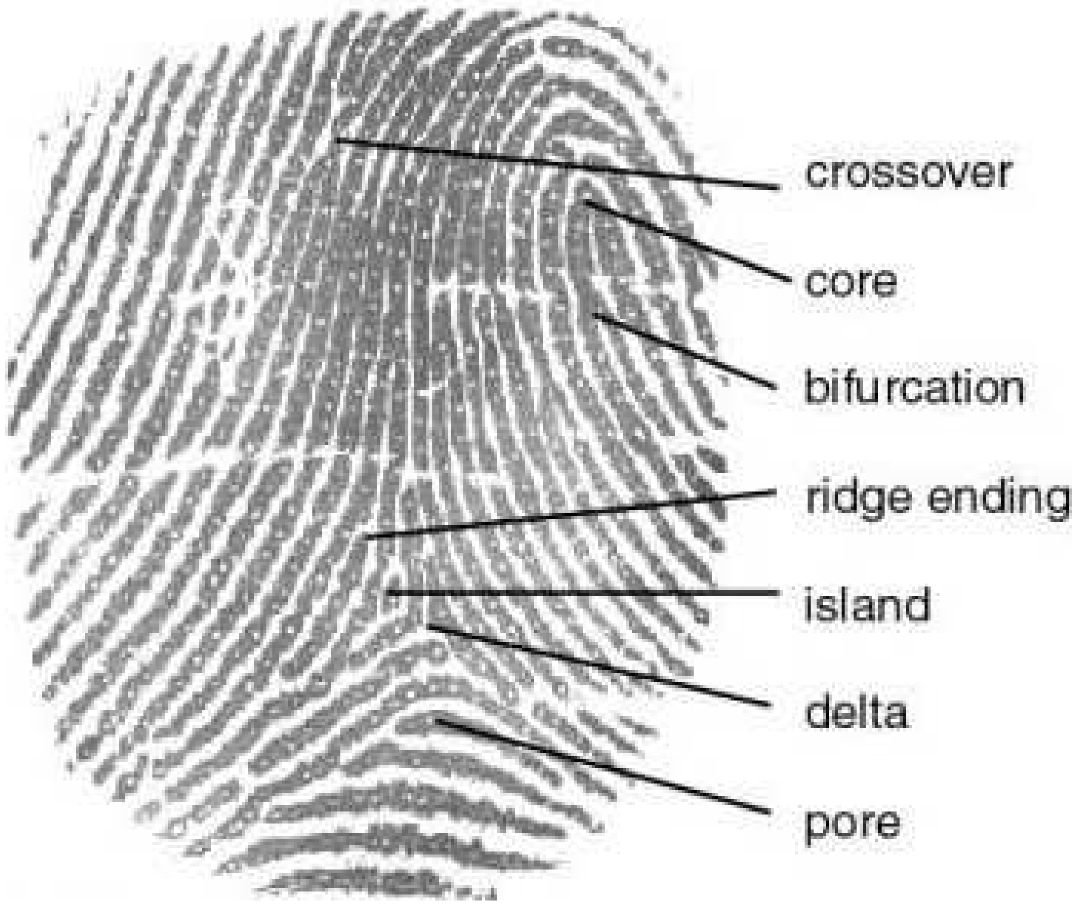
"*ACE relies upon the examiner's skills, training and experience, not upon formal criteria. In the absence of such criteria, the only available method ... is by consensus among examiners. Compounding this issue is the fact that there are no generally accepted, rigorous definitions of features or clarity, and therefore no generally accepted systematic approaches to indicate confidence in features, to define ridge detail (level-3) features, or even consistent definitions of what exactly constitutes a minutia. The lack of such rigorous definitions and systematic approaches contributes to a lack of reproducibility (interexaminer agreement) and repeatability (intraexaminer agreement).*"

How do we go forward? Questions to consider:

1. **Latent Prints, Shoeprints, ...**:

- What possible features? (ridge ending, bifurcation, ...)
- How sensitive/specific are the features?
- Spatial arrangements / constellations of features?
- Image quality & resolution of sensitive/specific features or constellations of features?

(Recent research with attempts to quantify "rarity" of features and take into account their spatial arrangments)


crossover
core
bifurcation
ridge ending
island
delta
pore

2. **Glass and other Trace Evidence**:

- Which elemental concentrations?
- Are measurements independent? (e.g., Mg & Ca)
- Consistency of manufacturing process: How many other pieces of evidence could have the same sets of measurements?
- What sources contribute to variability in measurements? (glass type, manufacturer, within pane, different panes from same manufacturer, ...)
- How representative, and how large, was the data set on which these variations were estimated?
- Hint: Estimating variability takes much larger sample sizes than you might suspect.

Risks of ignoring these considerations: Bullet Lead

For **all** evidence, consider **Measurement Uncertainty**

- If evidence were measured 1 hour / 1 day / 1 week later?
- If evidence were measured by different equipment/examiners?
- If evidence came from different parts of same source?
- In estimating feature sensitivity and specificity?
- Other sources of measurement uncertainty?

Precise, repeatable, reproducible measurements are ideal, but:

"**Better an approximate measure of something important than a precise measure of something unimportant**." (Byar)

Science & Justice 2009: Proposed **qualitative** statements for

$$\frac{\text{Prob of matching features if } \textbf{same} \text{ source}}{\text{Prob of matching features if } \textbf{different} \text{ sources}}$$

Support for Association	**Ratio**
Extremely Strong support	over 1 million
Very Strong support	10,000 – 1 million
...	...
Moderate support	10 – 100
Weak support	1 – 10

- Differing interpretations of degree of support for the Ratio.
- The Ratio depends on models, relevant populations, databases.
- Ratio is informative but is **not** what we really want to know.

Final comments:

1. **The Ratio is rarely as easy to calculate as for marbles.** When it **can** be calculated, it has **uncertainty**: it depends on assumptions, models, relevant populations, databases.

2. What we really want to know is *not* what we can expect *if the truth is known*. In real cases, the truth is never known.

3. In real life, we want to know: **Given the evidence at hand**, what is: probability that samples came from same source; probability that samples came from different sources.

4. Those probabilities depend on sensitivity, specificity, *and how probable are "same" / "different" in the relevant population.*

5. Pattern Evidence will be only one piece of information.

6. We don't ever expect Pattern Evidence features to be as sensitive and specific as DNA features.

7. But just having *some* idea of the features' sensitivity and specificity is a **big** step forward.

8. Probabilities depend on models & assumptions, and the data used to develop them, **which should be stated** (Hari, Alicia).

9. "*All models are wrong, but some are useful.*" (George Box)

10. Admitting our areas of uncertainty is better than pretending that they don't exist.

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 ... 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.*"