8-13-2019

Statistical Thinking for Forensic Practitioners

Hal Stern  
*University of California, Irvine*

Naomi Kaplan  
*University of California, Irvine*

Follow this and additional works at: https://lib.dr.iastate.edu/csafe_conf

Part of the Forensic Science and Technology Commons

**Recommended Citation**  
https://lib.dr.iastate.edu/csafe_conf/51

This Presentation is brought to you for free and open access by the Center for Statistics and Applications in Forensic Evidence at Iowa State University Digital Repository. It has been accepted for inclusion in CSAFE Presentations and Proceedings by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Statistical Thinking for Forensic Practitioners

Disciplines
Forensic Science and Technology

Comments
Posted with permission of CSAFE.

This presentation is available at Iowa State University Digital Repository: https://lib.dr.iastate.edu/csafe_conf/51
Statistical Thinking for Forensic Practitioners

Hal Stern
University of California, Irvine

Naomi Kaplan Damary
University of California, Irvine

August 13, 2019
Motivation - Interesting times in forensic science

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Context for the Workshop

- Daubert standard governs admission of scientific expert testimony in federal courts
  - Judge as gatekeeper
  - Relevant factors for judge to consider include peer review, known error rate, standards, etc.
  - Some states still use Frye standard of "general acceptance" in the relevant scientific community
- FRE 702 requires testimony be "based on sufficient facts or data" and use "reliable principles and methods" "reliably applied ... to the facts of the case"
- National Academies of Science (2009) and PCAST (2016) reports raise questions about the scientific foundation of pattern matching (and other types) of evidence
- Increased attention on the role of probability and statistics
Context for the Workshop

- Many different forensic questions
- Focus of this discussion is questions of source determination
- Do evidence samples (e.g., from a crime scene and a suspect) come from the same source? Examples to be discussed include
  - DNA
  - Trace evidence (e.g., glass)
  - Pattern evidence (e.g., fingerprints, shoe prints)
Outline

- Part I - Probability and Statistics Preliminaries
  - review of probability concepts
  - review of statistical inference concepts
- Part II - Statistics for Forensic Science Applications
  - forensic examination as expert opinion
  - two-stage approach (significance test/coinccidence probability)
  - likelihood ratio / Bayes factor
The Role of Probability  
People (CA) v. Collins (1968)

- An elderly woman walking in an alley was attacked from behind and robbed
- She saw a young woman with blonde hair running away
- Other witnesses saw a woman with blonde hair in a ponytail get into a yellow car driven by a black man with a mustache and beard
- Police were eventually led to an interracial couple living in the area with a yellow Lincoln
The Role of Probability
People (CA) v. Collins (1968)

- Prosecution gave estimates of the frequency of the characteristics identified by the witnesses
  - Black man with a beard: 1 out of 10
  - Man with a mustache: 1 out of 4
  - White woman with blonde hair: 1 out of 3
  - Woman with a ponytail: 1 out of 10
  - Interracial couple in a car: 1 out of 1,000
  - Yellow car: 1 out of 10

- Prosecution did not indicate basis for the numbers
  (more on this later)
The Role of Probability
People (CA) v. Collins (1968)

- Expert witness for the prosecution: mathematics professor
  - Given all of the probabilities by the prosecutor
  - Asked to combine them all to result in the probability of finding all the characteristics in one couple
  - Multiplied them all together (citing the product rule for independent events):
    "1 out of 10" × "1 out of 4" × …
  - Resulting probability is 1 out of 12 million

- Prosecutor's conclusion:
  - A couple which matches all of the witness observations is so rare that the couple on trial must be the couple that committed the robbery
Population and Statistics Preliminaries
“The Big Picture”

- Population = universe of objects of interest
- Sample = objects available for study
- Probability: population → sample (deductive)
- Statistics: sample → population (inductive)
Probability and Statistics Preliminaries
The Big Picture in Practice

- Applications
  - Drug seizure (population = 100 bags; sample chosen for analysis)
  - Glass fragments (two populations = glass from crime scene and glass from suspect; take samples from each)
  - Forensic accounting
    (population = all transactions; sample chosen for analysis)

- Relevance to pattern evidence
  - Interested in variation among samples from a population of "same source" impressions (e.g., distortion in latent prints)
  - Interested in variation among samples from a relevant population of alternative sources

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
What is probability?

• Probability is the mathematical language of uncertainty
• The probability of an event is a number (between 0 and 1) describing the likelihood that the event occurs
• Applications are very broad. Example of events include:
  • measurement of glass refractive index is between 1.52 and 1.53
  • randomly chosen finger has a loop pattern
  • the proposition that the crime scene evidence and the suspect evidence have a common source is true
• Notation:
  \[ \Pr(E) = \text{probability of the event } E \]
  \[ \Pr(\bar{E}) = \Pr(E^c) = \text{probability } E \text{ does not occur} = 1 - \Pr(E) \]
  • \[ \Pr(E) = 1 \] - event is sure to happen
  • \[ \Pr(E) = 0 \] - event never happens
What is probability?

- Interpretations of probability
  - long run frequency of occurrence of event
    (must be a repeatable experiment such as toss of a coin or roll of a die)
  - subjective belief of likelihood of an event
    (probability Angels win the baseball World Series)

- Where do probabilities come from?
  - empirical evidence / data
  - mathematical models
  - subjective opinion
Probability
Probability and Odds

- Probabilities are related to odds
  - odds are ratios of probabilities
  - odds in favor of event \( Y \) are defined as
    \[ O_f = P(Y)/P(\bar{Y}) = P(Y)/(1 - P(Y)) \]
  - odds against event \( Y \) are defined as
    \[ O_a = P(\bar{Y})/P(Y) = (1 - P(Y))/P(Y) \]
  - if we are given the odds against event \( Y \), then \( P(Y) = 1/(O_a + 1) \)
  - e.g., if \( O_a = 4 \) ("4 to 1 against") then \( P(Y) = .2 \)
    - if you bet $1 that \( Y \) will happen then
      - 20% of the time you win $4
      - 80% of the time you lose $1
      (note: you will break even in the long run)
Probability

- Probability can be confusing because subtle differences in wording can lead to major differences in the answer.
- The birthday problem
  - What is the probability that somebody in this room shares my birthday?
  - What is the probability that two people in the room share a birthday?
- Note that the first question fixes a specific date ("my birthday")
- The second question allows for any date
Conditional Probability

- Consider the example of an individual flying from LAX to JFK and worried about a delay
- Based on historical data $\Pr(\text{delay}) = 0.27$
- Now suppose that the weather forecast calls for thunderstorms in New York
  - May now believe the probability of a delay is larger
  - This leads to the notion of conditional probability, what is the probability of a delay given that thunderstorms are forecast?
  - Notation: We write $\Pr(\text{delay} \mid \text{thunderstorms})$ . . . with the vertical bar serving as shorthand for "given" or "conditional on" or "given the condition"
  - Perhaps we conclude $\Pr(\text{delay} \mid \text{thunderstorms}) = 0.50$
Understanding Conditional Probability

- Recall the "big picture"

- For the first statement, $Pr(\text{delay}) = 0.27$, the population consists of all LAX-JFK flights

- For the second statement, $Pr(\text{delay} | \text{thunderstorms}) = 0.50$, the population consists of LAX-JFK flights on days with forecasts of thunderstorms

- Conditional probability changes the information we have and changes the population we are talking about
Understanding Conditional Probability

- Study of sentencing of 362 black convicted murderers in Georgia in the 1980s found that 59 were sentenced to death.
- Murderers categorized by race of victim and sentence received.

<table>
<thead>
<tr>
<th></th>
<th>Death Penalty</th>
<th>No DP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White victim</td>
<td>45</td>
<td>85</td>
<td>130</td>
</tr>
<tr>
<td>Black victim</td>
<td>14</td>
<td>218</td>
<td>232</td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>303</td>
<td>362</td>
</tr>
</tbody>
</table>

- \( P(\text{Death Penalty}) = \frac{59}{362} = .16 \)
- \( P(\text{Death Penalty} \mid \text{White Victim}) = \frac{45}{130} = .35 \)
- \( P(\text{Death Penalty} \mid \text{Black Victim}) = \frac{14}{232} = .06 \)
- Note: A number of important factors are not included (e.g., context of murder)
Understanding Conditional Probability

- Consider the following data regarding the use of consecutive matching striae (CMS) as a criterion for deciding whether a pair of bullets have the same source.
- Li, 2012 thesis, U Central Oklahoma - max CMS in comparing groove impressions (9mm bullets) from known matches and known non-matches.

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Matches</td>
<td>55</td>
<td>54</td>
<td>23</td>
<td>11</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>146</td>
</tr>
<tr>
<td>Known Non-Matches</td>
<td>48</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
</tbody>
</table>

- \( \Pr(\text{CMS} \geq 5 \mid \text{known match}) = \frac{14}{146} = .10 \)
- \( \Pr(\text{CMS} \geq 5 \mid \text{known nonmatch}) = \frac{0}{60} = .00 \)
- \( \Pr(\text{known match} \mid \text{CMS} = 4) = \frac{23}{24} = .96 \)
Conditional Probability and Independence

- Sometimes the additional information doesn’t change the probability of an event
- Famous classroom examples include coin flips, dice rolls
- Suppose I have pasta for dinner the day before my flight. Presumably ....
  \[ \Pr(\text{flight delay} \mid \text{pasta for dinner}) = \Pr(\text{flight delay}) \]
- We would then say that having a flight delay is independent of what I had for dinner
- A well-known example of independent events in forensic science is the independence of DNA markers found on different chromosomes
Independence and the Product Rule

- We may want to know the probability that two different events both happen
  - What is the probability that my flight is delayed and my luggage is lost?
  - What is the probability that I get a head on my first coin toss and my second coin toss?
  - What is the probability that the Yankees make the playoffs and win the World Series?

- This can be complicated to compute because the second event may depend on the first. An extreme example:
  - Pr(Yankees win World Series) = 0.18 (according to fivethirtyeight.com)
  - Pr(Yankees win World Series | Yankees don’t make the playoffs) = 0

- In general the probability that two events both occur is found by
  - $\Pr(A \text{ and } B) = \Pr(A) \Pr(B | A) = \Pr(B) \Pr(A | B)$

- If two events are independent, then there is a simple product rule. We can just multiply probabilities
  - $\Pr(A \text{ and } B) = \Pr(A) \Pr(B)$
People (CA) vs Collins - a cautionary tale

- Collins were found guilty
- Malcolm Collins appealed claiming the probability evidence was prejudicial
- California Supreme Court reversed the conviction
  - Court indicated testimony lacked an adequate foundation
    - Inadequate evidentiary foundation for probabilities
    - Inadequate proof of statistical independence
  - Court found testimony and prosecutor’s use distracted the jury from its proper role
People (CA) vs Collins - a cautionary tale

- The Court’s concerns
  - Prosecution did not provide any sources for the probabilities supplied
  - Need to have some empirical basis for the probabilities
    (e.g., \( \Pr(\text{man with mustache}) = 0.47 \) in France 2016)
  - Suspect that some of the characteristics are not independent
    (e.g., \( \Pr(\text{beard} \mid \text{mustache}) = 0.91 \) in France 2016)
  - Dependencies of this type will lower the probability (and make the evidence less convincing)
  - Mathematics as a distraction: the prosecution’s argument provides no guidance to the jury on the critical issue of whether the Collins committed the crime
    - possibility of eyewitness error / disguise
    - possibility of more than one couple matching the description

- Interesting to note that the Court’s last concern about mathematics as a distraction has been overcome; calculations like those used in the Collins case are regularly used in analyses of DNA evidence
State (CT) vs Skipper

- Collins case introduces us to
  - probabilities of simple events (e.g., probability of blond woman)
  - conditional probability
  - product rule for independent events

- Much current discussion is focused on more sophisticated uses of probability

- Introduce these ideas through a second case, State (CT) vs Skipper (1994)
State (CT) vs Skipper

- Defendant charged with sexual assault
- State’s expert witness reported on results of a genetic paternity test
- Expert reported a paternity index (likelihood ratio) of 3496 (probability that defendant would produce a child with the given genotype is 3496 times as large as the probability that a random male would produce such a child)
- Expert indicated the paternity index could be converted into a statistic that gave the defendant’s probability of paternity
- He did so and reported the probability of paternity
  \[ \frac{3496}{3497} = 0.9997 \]
- To disentangle the statistical issues in this case we need
  - Bayes’ theorem (more advanced probability)
  - A framework for assessing forensic evidence
Bayes’ Theorem (or Rule)

- Thomas Bayes was an English mathematician, philosopher and minister.
- Famous among statisticians for his mathematical work on "inverse probability".
- Recall that in our "big picture" (below) probability tells us how to go from knowledge about the population to what we can expect to see in a sample.
- Inverse probability (now known as Bayesian statistics) refers to using our observed sample to infer (or make probability statements) about the population.

![Bayes' Theorem Diagram](image_url)
**Bayes’ Theorem - Gunshot residue example**

- Consider a diagnostic test for gunshot residue on an individual
- Let $G$ denote the event that gunshot residue is present (we will say "not $G$" to denote the opposite event)
- Let $T$ denote the event that our diagnostic test is positive (indicates gunshot residue is present) and "not $T$" to indicate a negative test

<table>
<thead>
<tr>
<th>True Status</th>
<th>Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>True Positive</td>
</tr>
<tr>
<td>Not $G$</td>
<td>False Positive</td>
</tr>
</tbody>
</table>

- We frequently have information about the performance of the test
- $P(T | G) =$ true positive rate, sensitivity
- $P$(not $T$ | not $G) =$ true negative rate, specificity
- $p(T |$ not $G) =$ false positive rate (type I error rate)
- $p$(not $T$|$G) =$ false negative rate (type II error rate)
Bayes’ Theorem - Gunshot residue example

• Bayes’ Theorem provides a means of taking information we have about the test (how the test performs on people whose status is known) to infer the status of an individual who has been tested
  • We know $\Pr(T \mid G)$ and $\Pr(T \mid \text{not } G)$
  • Bayes’ Theorem allows us to calculate $\Pr(G \mid T)$

• To do this Bayes’ Theorem also requires some ”prior” information about the prevalence of gunshot residue in the population of interest (i.e., $\Pr(G)$)

• A couple of key points
  • This ”prior” information is important and its not clear where it should come from
  • In general, $P(T \mid G) \neq P(G \mid T)$ (sensitivity of the test is not the same as our certainty given the test result)
Bayes’ Theorem - Gunshot residue example

- How does it work?
  - Assume \( \Pr(T \mid G) = 0.98 \) (the test is very sensitive - many true positives)
  - Assume \( \Pr(\text{not } T \mid \text{not } G) = 0.96 \) (the test is pretty specific - low rate of false positives)
  - For now assume \( \Pr(G) = 0.90 \) in the population of interest
  - We test an individual and get a positive test result
  - What can we say about the probability that the individual actually has gunshot residue on them

- There is a mathematical formula for this ....

\[
P(G \mid T) = \frac{P(G \text{ and } T)}{P(T)} = \frac{P(T \mid G)P(G)}{P(T \mid G)P(G) + P(T \mid \text{not } G)P(\text{not } G)} = \frac{.98 \times .9}{(.98 \times .9 + .04 \times .1)} = .995
\]

but it is easier to think about this with a picture
Bayes’ Theorem - Gunshot residue example

- Suppose we have a population of 1000 individuals

Suppose we have a population of 1000 individuals. Let's analyze the gunshot residue test results:

- 900 people have residue.
- 100 people do not have residue.
- 882 people have a positive test result.
- 18 people have a negative test result.
- 4 people have a positive test result and no residue.
- 96 people have a negative test result and no residue.

Conclusion: Note that there are 886 positive tests and 882 are "true".

\[
\Pr(G \mid T) = \Pr(\text{residue} \mid \text{positive}) = \frac{882}{886} = .995
\]

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Bayes’ Theorem - Gunshot residue example

- Can sometimes get surprising results from Bayes’ Rule
- Return to the diagnostic test for gunshot residue example
  - assume $P(T \mid G) = .98$ (sensitivity)
  - assume $Pr(\text{not } T \mid \text{not } G) = 0.96$ (specificity)
  - now assume $P(G) = .05$ (low prevalence)
    (i.e., testing in a population where gun usage is rare)
Bayes’ Theorem - Gunshot residue example

- Conclusion: Note that there are 87 positive tests and 49 are "true"
  \[ \Pr(G | T) = \Pr(\text{residue} | \text{positive}) = \frac{49}{87} = 0.56 \]
- The prior information matters a great deal when interpreting the test result
- Same phenomenon can happen with drug testing, medical diagnostics
Conditional Probability / Bayes’ Theorem in the Courtroom

- $E =$ evidence
  - DNA markers from the crime scene sample and suspect sample
  - Measurements on glass fragments from crime scene / suspect’s clothing
  - Image of bullet cartridges found at crime scene / test fire from suspect’s weapon
- $H_s =$ ”same source” proposition (two samples have same source)
- $H_d =$ ”different source” proposition (two samples w/ different sources)

Then
- $\Pr(E \mid H_s) =$ probability of seeing evidence if suspect is the source
- $\Pr(E \mid H_d) =$ probability of seeing evidence if suspect is not the source

And
- $\Pr(H_s \mid E) =$ probability suspect is the source given the evidence
- $\Pr(H_d \mid E) =$ probability suspect is not the source given the evidence
Bayes’ Theorem and the Likelihood Ratio

- \( E = \) evidence
- \( H_s = \) "same source" proposition (two samples come from the same source)
  \( H_d = \) "different source" proposition (two samples come from different sources)
- Bayes’ Theorem can be used to move from statements about the evidence to statements about the propositions/hypotheses
- But as with the gunshot residue it requires prior information about the likelihood of the propositions
- Bayes’ Theorem can be written in several different forms. Here it is helpful to write in terms of odds (ratios of probabilities)

\[
\frac{P(H_s \mid E)}{P(H_d \mid E)} = \frac{P(E \mid H_s) P(H_s)}{P(E \mid H_d) P(H_d)}
\]
Bayes’ Theorem and the Likelihood Ratio

Bayes’ Theorem

\[
\frac{P(H_s \mid E)}{P(H_d \mid E)} = \frac{P(E \mid H_s)}{P(E \mid H_d)} \frac{P(H_s)}{P(H_d)}
\]

- Term on far right is "a priori" odds in favor of the same source proposition (the prior information)
- Term in the middle is known as the likelihood ratio (LR) or Bayes factor (BF)
- Left hand side is "a posteriori" odds in favor of the same source proposition
- We will discuss in much more detail later
- For now the key point is the distinction between the likelihood ratio and the posterior odds
Bayes’ Theorem and the Likelihood Ratio

- Recall our gunshot residue example
  - $E =$ evidence = positive test
  - $H_s =$ suspect has gunshot residue
  - $H_d =$ suspect doesn’t have gunshot residue
  
  - $LR = P(E \mid H_s)/P(E \mid H_d) = .98/.04 = 24.5$
  
  - In high prevalence case
    - prior odds are 9 : 1 and posterior odds are 220.5 : 1
    - (posterior probability $= .995$)
  
  - In low prevalence case
    - prior odds are 1 : 19 and posterior odds are 24.5 : 19
    - (posterior probability $= .56$)
  
- Note: Can also compute likelihood ratio if evidence were a negative test
  - (turns out to be $.02/.96 = 1/48$ which is not the reciprocal of the LR for the positive test)
State (CT) vs Skipper - the role of prior information

- Skipper was convicted
- He filed appeal claiming the statistical evidence was improperly admitted
- State Supreme Court found the expert’s application of Bayes’ Theorem was inconsistent with the presumption of innocence and remanded for new trial
  - Court determined that the conversion done by the expert to go from LR to posterior odds assumed prior probability of paternity was 0.50
  - Found this to violate presumption of innocence
State (CT) vs Skipper

- $E = \text{genetic evidence}$
- $H_d = \text{"defendant is the father" proposition}$
  $H_r = \text{"random man is the father" proposition}$
- Bayes' Theorem
  \[
  \frac{P(H_d | E)}{P(H_r | E)} = \frac{P(E | H_d) P(H_d)}{P(E | H_r) P(H_r)}
  \]
- Expert testified that $\Pr(E | H_d)/\Pr(E | H_r) = LR = 3496$
  (evidence is much more likely if defendant is father than if a random man is the father)
- Expert assumed prior odds of 1-to-1 (50% probability for $H_d$)
- Expert computed posterior odds are 3496-to-1 which gives
  $\Pr(H_d | E) = 3496/3497 = .9997$
State (CT) vs Skipper

- The difference between $Pr(E \mid H_d)$ and $Pr(H_d \mid E)$ is critical!
- $Pr(E \mid H_d)$ is a statement about the probability of seeing the evidence if suspect is the father
- $Pr(H_d \mid E)$ is a statement about the probability the defendant is the father based on the observed evidence
  - It seems like we want this quantity
  - But getting it depends on specifying the pre-evidence probability of the defendant being the father
  - The Court here found it is not appropriate for forensic expert to form pre-evidence opinions about the hypothesis of guilt/paternity
- Statements about the evidence (i.e., the components of the LR) are where the forensic expertise lies

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Probability
A short recap

- Probability is the mathematical language of uncertainty
- Provides a common scale (0 to 1) for describing the chance that an event will occur
- Conditional probability is a key concept ... the probability of an event depends on what information is considered
- Independent events can be powerful (allows us to multiply probabilities as is common in DNA analysis)
- Bayes' Rule is a mathematical result showing how we should update our probabilities
  - leads naturally to thinking about the likelihood ratio as a summary of the evidence (more later)
**Probability**

**Probability and the Courts: Sally Clark**

- Sally Clark was the only person in the house when her first child died unexpectedly at 3 months. The cause of death was identified as sudden infant death syndrome (SIDS). A year later Sally had a second child, who died at 2 months under similar circumstances.

- Sally was convicted of murder. During the trial a pediatrician testified that the probability of a single SIDS death was 1/8500, so the probability of two SIDS deaths was 1/73 million (1/8500²).

- Several problems with this approach to the evidence ... what do you think?
Probability
Probability and the Courts: Sally Clark

- Sally Clark was the only person in the house when her first child died unexpectedly at 3 months. The cause of death was treated as sudden infant death syndrome (SIDS). A year later Sally had a second child, who died at 2 months under similar circumstances.

- Sally was convicted of murder. During the trial a pediatrician testified that the probability of a single SIDS death was $1/8500$, so the probability of two SIDS deaths was $1/73$ million ($1/8500^2$).

- Several important issues associated with this probability argument:
  - Where does $1/8500$ come from? Is it correct for "families like the Clarks"?
  - The use of multiplication assumes independence for the two children.
  - Also need to consider alternative hypotheses (and their likelihood).
Probability to Statistical Inference
Collecting data

- Assessment of forensic evidence generally requires data collection
- Assessment and validation of forensic procedures requires careful study
- Where do data come from and how do we carry out studies
- Statistics has important ideas to contribute
- Two fundamental ideas
  - experimentation - carrying out a study of procedure/method (e.g., a black-box study)
  - sampling - getting a subset of the population of interest to study
**Motivation - ASTM 2548-16**

- ASTM 2548-16: Standard Guide for Sampling Seized Drugs for Qualitative and Quantitative Analysis

- Note the similarity to our "big picture"
- This makes us thing about how we sample or collect data

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Motivation - ASTM 2548-16

- ASTM 2548-16: Standard Guide for Sampling Seized Drugs for Qualitative and Quantitative Analysis
  - Section 4.2.1. Sampling may be statistical or non-statistical
    - 4.2.1.1. In many cases, a non-statistical approach may suffice. The sampling plan shall provide an adequate basis for answering questions of applicable law. For example, Is there a drug present in the population?
    - 4.2.1.2. If an inference about the whole population is to be drawn from a sample, then the plan shall be either statistically based or have an appropriate statistical analysis completed and limits of the inference shall be documented.
Sampling
- we sample because it is too costly or time-consuming to study the entire population
- a random sample allows us to use the laws of probability to describe how certain we are that calculations based on our sample will reflect the population
- many famous failures with non-probabilistic sampling (e.g., Truman vs Dewey election)

Relevance to forensic science
- which and how many bags of suspect powder need to be tested
- choosing example cases to incorporate in studies
- what is the correct shoe database for assessing shoe print evidence (manufactured shoes vs police database?)
**Probability to Statistical Inference**

**Collecting data**

- **Experimental design**
  - compare treatments
  - randomly assign treatments to units
  - make sure sample size is large enough
  - make environment as realistic as possible
  - role of blinding to avoid bias

- **Principles of good experimental design are relevant to forensic science**
  - can use these ideas in evaluating process improvements in the lab
  - for black box studies these principles suggest
    - integrating test cases with actual casework
Probability to Statistical Inference
Collecting data

- Statisticians distinguish between different types of data
  - qualitative
    - categorical (blood type: A, B, AB, 0)
    - ordinal (grades: A, B, C, D, F)
  - quantitative
    - discrete (consecutive matching striae)
    - continuous (refractive index of a glass fragment)
Probability to Statistical Inference

- Once data have been collected that are relevant to the scientific question of interest, the focus shifts to measurement and analysis.

  - Introduction. "One objective of a forensic glass examination is to compare glass samples to determine if they may be discriminated using their physical, optical or chemical properties (for example, color, refractive index (RI), density, elemental composition). If the samples are distinguishable in any of these observed and measured properties, it may be concluded that they did not originate from the same source of broken glass. If the samples are indistinguishable in all of these observed and measured properties, the possibility that they originated from the same source of glass may not be eliminated. The use of an elemental analysis method such as laser ablation inductively coupled plasma mass spectrometry yields high discrimination among sources of glass."
Probability to Statistical Inference
Measurement, variability and uncertainty

- ISO 1725: 7.6.1 Laboratories shall identify the contributions to measurement uncertainty. When evaluating measurement uncertainty, all contributions that are of significance, including those arising from sampling, shall be taken into account using appropriate methods of analysis.
- Key point: any measurement process involves some degree of uncertainty
- If you measure the same item multiple times you will not get exactly the same answer
- This reflects natural variability in the measurement process
- A measure of the resulting uncertainty should be provided to the user
Scientists focused on physical measurements often use uncertainty to refer to the intrinsic uncertainty in a measurement:
- a thermometer may only be accurate to within 0.1 degrees

Statisticians tend to use uncertainty more broadly to address all kinds of things that we don't know.

Uncertainty is usually addressed with probability, a probability distribution, or some summary of a probability distribution:
- e.g., probability of rain tomorrow is 60%
- e.g., weights on this scale are normally distributed with standard deviation 0.1 kg
- e.g., measurement is accurate to ±0.5 in
**Probability to Statistical Inference**

Measurement, variability and uncertainty

- Variability refers to the fact that variation is observed in repeated measurements
  - repeated measurements of a given object by the same individual
  - repeated measurements of a given object by different individuals
  - repeated measurements of different (related) objects by the same individual
  - repeated measurements of different (related) objects by different individuals

- Variability is usually measured with quantities like the standard deviation or range

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Probability to Statistical Inference
Thinking about data

- For qualitative data (like blood type) we usually summarize by providing a table of frequencies/proportions:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.42</td>
<td>.10</td>
<td>.04</td>
<td>.44</td>
</tr>
</tbody>
</table>
Probability to Statistical Inference
Thinking about data

- For discrete data (e.g., CMS) we may summarize with a table or numerical summaries (mean, standard deviation)
- Example: CMS measures from an automatic algorithm for striation patterns on known non-matching bullet lands (Chu et al., For. Sci. Int., 2013)
  - mean = 0.77, std.dev. = 0.58

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>3887</td>
<td>8219</td>
<td>782</td>
<td>70</td>
<td>2</td>
<td>0</td>
<td>12960</td>
</tr>
<tr>
<td>Proportion</td>
<td>.2999</td>
<td>.6342</td>
<td>.0603</td>
<td>.0054</td>
<td>.0002</td>
<td>.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Probability to Statistical Inference
Thinking about data

- For continuous data (e.g., refractive index of glass) we may summarize with graphs and numerical summaries
- Example: refractive index measurements of 49 fragments from a single source
- Numerical summaries include: mean=1.51999, std.dev.=0.00004, min=1.51979, 25%ile=1.51998, median=1.51996, 75%ile=1.52001, max=1.52007

![Histogram of refractive index measurements](image)
Questions about uncertainty and variability are central to thinking about forensic evidence.

For example, it is believed that signature complexity is relevant to the assessment of signature evidence.

How reliably can we measure complexity?
Probability to Statistical Inference
Reliability in forensics - handwriting complexity

- Five forensic document examiners (FDE) rated 123 signatures in terms of difficulty to simulate on a 5-point scale (easy - fairly easy - medium - difficult - very difficult)

<table>
<thead>
<tr>
<th>ID</th>
<th>FDE1</th>
<th>FDE2</th>
<th>FDE3</th>
<th>FDE4</th>
<th>FDE5</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>002</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>003</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>004</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>005</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

- Can be used to assess reproducibility (similarity of assessments by two different examiners)
- Correlation (between -1 and 1) is often used to measure degree of association between two sets of scores (with one indicating a perfect linear relationship)
- Correlation of ratings of pairs of FDEs vary with typical value .65
- A subset of five examiners were shown a subset of 7 signatures twice
- Can be used to assess repeatability (similarity of assessments by same examiner at two different times)
- Statistical approach estimates repeatability with the intra-rater correlation of .68

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Probability to Statistical Inference

Probability distributions

- Suppose we are to collect data on some characteristic for a sample of individuals or objects (weight, trace element concentration)
- Probability distribution is used to describe possible values and how likely each value is to occur
- Knowing about probability distributions is relevant for understanding how likely observed evidence is under a given hypothesis
Probability to Statistical Inference
Probability distributions

Examples of distributions

- Binomial: number of successes in n trials
  (e.g., test n bags of contraband and record the number found to contain drugs)
- Poisson: count number of events
  (e.g., number of consecutive matching stria
- Normal: bell-shaped curve
  (e.g., measure of weight of packages of drugs found on suspect)
- Log normal: logarithm of observations follow a normal distribution
  (e.g., measure of concentration of chemical in glass)
Probability to Statistical Inference

Probability distributions

- **Binomial(10,0.75)**
- **Poisson(5)**
- **Normal(100,15)**
- **LogNormal(0,1)**

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Statistical Inference
Recall “The Big Picture”

- Population = universe of objects of interest
- Sample = objects available for study
- Probability: population → sample (deductive)
- Statistics: sample → population (inductive)
- Often use both together
  1. build/assume model for population
  2. assess model by comparing sample to what is expected under model
  3. refine model; go back to step 2
Motivation - ASTM 2927-16


- Introduction. "One objective of a forensic glass examination is to compare glass samples to determine if they may be discriminated using their physical, optical or chemical properties (for example, color, refractive index (RI), density, elemental composition)...... The use of an elemental analysis method such as laser ablation inductively coupled plasma mass spectrometry yields high discrimination among sources of glass."

- The "Big Picture" applies in this situation as well

- Now two populations (one corresponding to known source and one corresponding to questioned source)

- Question of interest is whether these populations differ in important ways (are distinguishable)
Motivation - ASTM 2927-16

11. Calculation and Interpretation of Results

11.1. The procedure below shall be followed to conduct a forensic glass comparison when using the recommended match criteria:

11.1.1. For the Known source fragments, using a minimum of 9 measurements (from at least 3 fragments, if possible), calculate the mean for each element.

11.1.2. Calculate the standard deviation for each element. This is the Measured SD.

11.1.3. Calculate a value equal to at least 3% of the mean for each element. This is the Minimum SD.

11.1.4. Calculate a match interval for each element with a lower limit equal to the mean minus 4 times the SD (Measured or Minimum, whichever is greater) and an upper limit equal to the mean plus 4 times the SD (Measured or Minimum, whichever is greater).

11.1.5. For each Recovered fragment, using as many measurements as practical, calculate the mean concentration for each element.

11.1.6. For each element, compare the mean concentration in the Recovered fragment to the match interval for the corresponding element from the Known fragments.

11.1.7. If the mean concentration of one (or more) element(s) in the Recovered fragment falls outside the match interval for the corresponding element in the Known fragments, the element(s) does not "match" and the glass samples are considered distinguishable.

This is a statistical inference procedure
**Statistical Inference**

**Background**

- Definition - a **parameter** is a numerical characteristic of the population, e.g., a population mean.
- Statistical methods are usually concerned with learning about population parameters from sample data.
- A key point - the mean of a sample and the mean of a population are different concepts.
- Idea: apply laws of probability to draw inferences from a sample.
  - observe sample mean.
  - if we have a “good” sample, then this should be close to the population mean.
  - the laws of probability tells us how close we can expect them to be.
**Statistical Inference**

**Background**

- **Goal:** inference about a parameter
- **Possible parameters**
  - mean concentration of aluminum in population of glass fragments from a given source
  - proportion of bags containing illicit substances
- **Different kinds of inferential statements**
  - estimate of parameter (point estimate)
  - range of plausible values for parameter (interval estimate)
  - test a specific hypothesis about the value of a parameter
**Statistical Inference**

Point estimation

- Estimator is a rule for estimating a population parameter from a sample
- Evaluate estimator by considering certain properties
  - bias - how close on average to population value
  - variability - how variable is the estimate
- Example - suppose we are interested in estimating the population mean or average
  - the mean of a random sample from the population is one possible estimator (spoiler alert: it is a very good estimator)
  - the median of a random sample is an alternative (less sensitive to wild measurements)
  - 47 is another possible estimator (not very good – unless we are very lucky!)
Statistical Inference

Performance of different estimators for unknown $\theta$

- Figures below show what would happen in many repeated attempts for estimators with different properties
  - $\theta$ is the "true" population parameter (the center of the target)
A limitation of just providing a point estimate is that it doesn't provide any indication of uncertainty. We can do better than this. The standard error of an estimator measures the uncertainty in our estimate.

- The standard deviation is a measure of the spread (variability) in a sample or in a population (describes uncertainty about a single observation).
- When we look at a summary statistic (mean, median, percentile) it is also a random quantity (would give different value in different samples).
- The standard error is how we measure the variability of an estimator.
**Statistical Inference**

**Standard errors**

- Consider a normally distributed population with mean 100 and s.d. 15
  - review meaning of s.d
  - expect 95% of observations to be between 70 and 130

![Graph showing the distribution of a single observation](image)
**Statistical Inference**

**Standard errors**

- Consider a normally distributed population with mean 100 and s.d. 15
- Now demonstrate standard error for the mean of 25 responses
  - standard error is 3 (s.d. divided by square root of sample size)
  - mean of this distribution is still 100
  - 95% of the time it will be between 94 and 106

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Statistical Inference
Standard errors and sample size

- **n=10**, mean=100.08, sd=4.62
- **n=25**, mean=100.25, sd=2.74
- **n=50**, mean=99.9, sd=2.07
- **n=100**, mean=99.9, sd=1.57

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
**Statistical Inference**

Interval estimation

- A confidence interval is an interval based on sample data that contains a population parameter with some specified confidence level.

- Essentially a confidence interval takes a point estimate and then adds some information about uncertainty.

- Typically we get an approximately 95% confidence interval for a quantity by taking point estimate \( \pm 2 \) std errors (often use 1.96 which is correct for large samples from the normal distribution).

- Most common example is trying to estimate the population mean:
  - natural point estimate is the sample mean
  - approximate 95% confidence interval is sample mean \( \pm 2 \) standard error
**Statistical Inference**

Interval estimation - example

- Example: 10 glass fragments from crime scene
- Measure concentration of aluminum
- Mean = 0.730, standard deviation = 0.04
- Standard error = 0.040 / sqrt(10) = 0.013
- Approximate 95% confidence interval for the mean aluminum concentration in the crime scene window is 0.73 +/- 2*0.013 = (.704,.756)
- Interpretation of confidence interval is important - 95% of intervals built in this way will contain the true population parameter
- Note this type of interval (with higher confidence) is sometimes used in the analysis of glass evidence (ASTM 2926)
**Statistical Inference**

**Hypothesis testing**

- Sometimes we wish to formally test a hypothesis about a population parameter.
- The hypothesis to be evaluated is known as the null hypothesis and usually refers to an assumption of no difference or no change. We look for evidence against the null hypothesis.
- There is an alternative hypothesis that helps us to design the test.
- The test is usually summarized by a *p*-value measuring the strength of the statistical evidence against the null hypothesis (more on this later).
- Historically it has been common to use a threshold (say *p* < .05) to decide whether to accept or reject the null hypothesis.
  - If we reject the null hypothesis then we say we have a statistically significant result.
  - This approach has some problems and is currently a topic of much discussion.
  - Because it is still popular in some quarters we spend some time on it now.
Statistical Inference
Hypothesis testing

- Two types of errors
  - type I: reject the null hypothesis when it is true (false positive)
  - type II: fail to reject the null when it is false (false negative)

- Type I error often considered more serious: we only want to reject the null if strong evidence against it

- Note that these statistical testing ideas are closely related to concepts in the justice system
  - null hypothesis = innocent, alternative = guilty
  - type I error is to decide guilty when person is innocent
  - type II error is to decide innocent when person is guilty

- The two types of errors are relevant to forensics as well (but it is generally better to avoid using the type I/II terminology)
Statistical Inference

Hypothesis testing

- Basic idea of hypothesis testing is to compute a test statistic that measures 'distance' between the data we have collected and what we would expect under the null hypothesis
- Typically use a statistic of the form
  \[
  \frac{\text{point estimate} - \text{null hypothesis value}}{\text{SE(estimate)}}
  \]
  where \( SE \) is a standard error
- Can be interpreted as the number of standard errors the sample estimate is from the hypothesized value under the null hypothesis
**Statistical Inference**

Hypothesis testing

- Common to summarize test by attaching a probability to the test statistic
- Definition: a *p*-value gives the probability that we would get data like the data we have observed in the sample (or something even more extreme) given that the null hypothesis is true
- Small *p*-values mean unusual data that lead us to question the null hypothesis (since sample data are unlikely to happen by chance)
- However, the *p*-value only addresses the null hypothesis. It does not speak to the likelihood of the alternative hypothesis being true
Statistical Inference
Hypothesis testing - comparing two means

- In practice, we are often interested in comparing two samples (or more precisely two populations)
- Assume random samples from each of the two populations are available
- Test for equivalence of parameters of the two populations
- Forensic example
  - suppose we have broken glass at a crime scene and glass fragments on the suspect
  - define $\mu_{\text{scene}}$ to be mean trace element level for the “population” of glass at the scene
  - define $\mu_{\text{suspect}}$ to be the mean trace element level for “population” of glass on the suspect
  - compare means to address if glass fragments on suspect could plausibly have come from the crime scene (i.e., $\mu_{\text{suspect}} = \mu_{\text{scene}}$)
Statistical Inference
Hypothesis testing - comparing two means

- Most well established procedure is for testing a hypothesis about means of normal populations
- Null hypothesis is $H_0 : \mu_{scene} = \mu_{suspect}$
- Alternative hypothesis is $H_a : \mu_{scene} \neq \mu_{suspect}$
- Suppose we have 10 glass fragments from glass at the scene (call these data $Y$) and 9 glass fragments found on the suspect (call these data $X$)
- Test looks at the difference in the two means ($\bar{Y} - \bar{X}$)
- Expect this difference to be near zero
- Reject the null hypothesis if the difference is large compared to the standard error for the difference in two means
- Procedure is known as the $t$-test and the $p$-value is easily obtained from a $t$ distribution (software will compute)
- Key statistical result is that these procedures work well even if population is not normally distributed as long as the sample size is large
Statistical Inference
Hypothesis testing - normal data

- Left figure: observed test statistic = 1.3, p-value = 0.19
- Right figure: observed test statistic = 2.3, p-value = 0.02
Statistical Inference
Comparing two means - example

- Suppose 10 glass fragments are taken from glass at the scene (Y) and 9 fragments are found on the suspect (X).
  - \( \bar{X} = 5.3, s.d. = 0.9, SE(\bar{X}) = 0.9/\sqrt{10} = .28 \)
  - \( \bar{Y} = 5.9, s.d. = 0.85, SE(\bar{Y}) = 0.85/\sqrt{9} = .28 \)
  - observed difference is \( \bar{Y} - \bar{X} = 0.6 \)
  - standard error for this difference is \( SE(\text{diff } \bar{X} - \bar{Y}) = \sqrt{.28^2 + .28^2} = 0.4 \)
  - test statistic is \( 0.6/0.4 = 1.5 \) which yields a p-value of 0.15
  - Do not reject the hypothesis that the two glass populations agree
  - The ASTM procedure is closely related to a test of this form
  - If the test accepts the null hypothesis, then according to ASTM standard we would say the two populations are indistinguishable

- Interpretation is a key issue (can’t reject the hypothesis of equal means and the possibility of a common source ... but this doesn’t mean it is definitely true)
Statistical Inference
Hypothesis tests and confidence intervals

- There is a very close relationship between tests and interval estimates
- Confidence interval (CI) gives range of plausible values (e.g., for the difference in two means)
- Test evaluates whether a specific value (e.g., zero in the two-sample test) is a plausible value
- Statistical hypothesis tests are very popular in practice
  - sometimes they address the scientific question of interest
  - but often they do not
- It is important to be aware of the limitations of statistical tests
Hypothesis testing - discussion

- Hypothesis testing does not treat the two hypotheses symmetrically (null is given priority)
  - This is appropriate if there is reason to prefer the null hypothesis until there is significant evidence against it
  - We don’t always want this to be the case (more on this later in forensic context)
- $P$-values depend heavily on the sample size
  - If you have the same means and standard deviations and increase the sample size the result will be more significant
- Interpretation can be tricky
  - Rejecting the null hypothesis does not mean that one has found an important difference
  - Important to consider the size of the observed difference
  - Failing to reject the null hypothesis does not mean that the null hypothesis is true
  - Important to consider the "power" of the test (how often would it reject if the alternative were true)
Statistical Preliminaries

- Reviewed basics of statistical inference
  - Statistical inference uses sample data to draw conclusions about population
  - Point estimation, interval estimation, hypothesis tests are main tools
  - Critical that procedures account for variation that could be observed due to chance
Statistics for Forensic Science

- Next .... how do ideas from probability and statistics apply to forensic science
- Outline
  - Brief review of probability/statistics
  - The forensic examination
  - Forensic evidence as expert opinion
  - Two-stage approach to forensic evidence
  - Likelihood ratio approach
Brief Review of Probability and Statistics

- Probability
  - language for describing uncertainty
  - assigns number between 0 and 1 to events
  - depends on information available (information conditioned upon)
  - useful for deducing likely values for individuals or samples from given (or hypothesized) information about the population

- Probability distributions
  - suppose we have a random quantity (e.g., trace element concentration in a glass fragment)
  - probability distribution gives possible values and relative likelihood of each value
Brief Review of Probability and Statistics

Statistics
- drawing inferences about a population (i.e., learning about some characteristic of the population) based on sample data
- need to carefully define “population"
- method used for data collection is very important
- there are a variety of inference procedures
  - point estimates
  - confidence intervals
  - hypothesis tests
The Forensic Examination

- There are a range of questions that arise in forensic examinations - source conclusions, timing of events, cause/effect
- Focus today on source conclusions
  - topics addressed (e.g., need to address uncertainty, logic of the likelihood ratio) apply more broadly
  - Evidence $E$ are items/objects found at crime scene and on suspect (or measurements of items)
    - occasionally write $E_c$(crime scene), $E_s$(suspect)
    - may be other information available, $I$ (e.g., evidence substrate)
  - Two hypotheses
    - $S$ - items from crime scene and suspect have common source (or suspect is source of crime scene item)
    - $\bar{S}$ or "not S" - no common source
  - Goal: assessment of evidence
    - do items appear to have a common source
    - how unusual is it to find observed evidence / observed agreement by chance

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
The Forensic Examination

- Evidence types
  - biological evidence (blood type, DNA)
  - glass fragments
  - fibers
  - latent prints
  - shoe prints / tire tracks
  - and others

- Different issues arise for different evidence types
  - available measurements (categorical/discrete/continuous variables)
  - information about the probability distribution of measurements
  - existence of reference database
  - role of manufacturing process
The Forensic Examination

- Daubert standard identifies the judge as “gatekeeper” to determine admissibility of expert scientific testimony
- Daubert decision provides illustrative factors that a judge may apply
  - theory/method should be testable
  - subject to peer review / publication
  - error rates
  - existence of standards and controls
  - generally accepted by a relevant scientific community
The Forensic Examination

Federal Rules of Evidence 702:
A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:

1. the expert’s scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
2. the testimony is based on sufficient facts or data;
3. the testimony is the product of reliable principles and methods; and
4. the expert has reliably applied the principles and methods to the facts of the case.
The Forensic Examination

- National Research Council (2009) findings
  - heterogeneous provider community (federal, state, local)
  - heterogeneity across disciplines
  - lack of standardization in practices
  - insufficient resources
  - questions underlying scientific basis for some conclusions
    (single source DNA’s emergence as a ”gold standard”)

- PCAST (2016) report
  - focused on validity of pattern matching disciplines
  - foundational validity - ”black box studies”
    (studies treating examiners as decision-making ”black boxes” and using
     examples with known ground truth to evaluate performance)
  - validity as applied - information about specific examiner
Logic of the Forensic Examination

- Examine the evidence \((E_c, E_s)\) to identify similarities and differences
- Assess the observed similarities and differences to see if they are expected (or likely) under the same source hypothesis
- Assess the observed evidence (including similarities and differences) to see if they are expected (or likely) under the different source hypothesis
  - Note that this includes assessing how unusual the matching features are
Common Approaches to Assessing Forensic Evidence

- There are multiple approaches to carrying out an examination of this type to assess the evidence.
- Many different categorizations of the approaches.
- For today we focus on three common approaches:
  - Expert assessment based on experience, training, accepted methods.
  - Two-stage approach:
    - Determination of similarity (often based on a statistical procedure).
    - Identification (assessing likelihood of a coincidental match).
  - Likelihood ratio / Bayes factor.
Forensic Conclusions as Expert Opinion

- Status quo in pattern disciplines
  (fingerprints, shoe prints, toolmarks, questioned documents, etc.)
- Expert analyzes evidence based on
  - Experience
  - Training
  - Use of accepted methods in the field
- Assessment of the evidence reflects examiner’s expert opinion
- Opinions typically reported as categorical conclusions
  - Identification, inconclusive, exclusion
  - Multi-point scales (some support, strong support, very strong support, etc.)
  - Some court decisions have asked what should be allowed
    (e.g., U.S. vs Glynn (2008) allowed firearms testimony to say only that
    the same source is "more likely than not")
Forensic Conclusions as Expert Opinion

Occasionally conclusions are expressed as statements about the hypotheses, e.g., based on the evidence, the author of the known samples ....

- Wrote the questioned sample
- Highly probably wrote the questioned sample
- Probably wrote the questioned sample
- Indications may have written the questioned sample
- Inconclusive
- (and similar statements on the negative side)

This is logically problematic as we saw in Skipper et al. case

Statements like this implicitly require that the examiner had an "a priori" (pre-evidence) opinion about the same source proposition
Forensic Conclusions as Expert Opinions

- What does it take to establish that testimony is
  - "based on sufficient facts or data"
  - "the product of reliable principles and methods"
- Note that the use of the word "reliable" in the legal sense (trustworthy) differs from its technical use in statistics
- Statisticians / measurement scientists distinguish between reliability (consistency) and validity (accuracy)
- Reliability / consistency is about the ability to measure well
- Validity / accuracy is about the ability to match some outcome
Forensic Conclusions as Expert Opinions

- In measurement / assessment, statisticians focus on a number of related questions in thinking about reliability (in the legal sense):
  - Would the same analyst draw the same conclusion in a new examination of the evidence (repeatability)?
  - Would different analysts draw the same conclusion given the same evidence (reproducibility)?
  - Repeatability and reproducibility are both components of reliability
  - Do analysts get the right answer in studies where ground truth is available (accuracy / validity)?
Forensic Conclusions as Expert Opinions

- PCAST-style "black box" studies of performance can be used to assess reliability and validity
  - examiner is treated as a "black box" that produces conclusions
  - examiners given cases with known "ground truth" to assess frequency of different types of errors
- PCAST report is controversial but its focus on "black box" evaluations is extremely useful
- Ulery et al. (PNAS, 2011) fingerprint study
  - false positive rate was 0.15% (non-mates declared to be mates)
  - false negative rate was 7.5% (mates declared to be non-mates)
- There are limitations in this and any study (similarity to casework? test environment?)
Forensic Conclusions as Expert Opinions

- Ulery et al. carried out a number of other studies
  - Repeatability (same examiner, 7 months apart)
    - for known mated pairs: 90% repeated decisions (including the inconclusives)
    - for known non-mated pairs: 86% repeated decisions
  - Reproducibility (different examiners)
    - for true mated pairs - 80% agreement on decisions
    - for true non-mated pairs - 80% agreement on decisions
  - Also carried out a series of "white box" studies to assess reliability of different steps in the analysis (e.g., markup of latent print)
Forensic Conclusions as Expert Opinions

- Reproducibility, reliability and validity are likely to depend on characteristics of the evidence, e.g.,
  - Quality of latent prints
  - Complexity of signature
- Ideally such characteristics can be integrated into reliability/validity studies
- This would enable reports of the kind “for evidence of this type.....”
- Handwriting example w/ signatures (Sita et al., JFS, 2002)
  - high complexity: 64% correct, 3% incorrect, 33% inconclusive
  - medium complexity: 41% correct, 4% incorrect, 55% inconclusive
Forensic Conclusions as Expert Opinions

- A few final remarks
  - Information on reliability and accuracy for forensic analyses is extremely helpful and will be increasingly expected.
  - As per FRE 702, there is also a need to address the application of the method or the technique in the current case (e.g., N.C. vs McPhaul, 2017).
  - There will always be unique situations (e.g., did this typewriter produce this note?) for which there are no relevant validation/reliability studies.
    - Not a problem ... but the conclusions expressed by the expert must acknowledge uncertainty about the likelihood of a coincidental agreement.
The Two-Stage Approach

- One common statistical approach solves the forensic problem in two stages

  - **Stage 1 (Similarity)**
    - determine if the crime scene and suspect objects agree on one or more characteristics of interest (typically using a hypothesis/significance test)
    - two samples "are indistinguishable", "can't be distinguished", "match"

  - **Stage 2 (Identification)**
    - assess the significance of this agreement by finding the likelihood of such agreement occurring by chance

- Has a long history (Parker and Holford papers in the 1960s)
- Also known as the comparison/significance approach
- Used in assessment of trace evidence (e.g., glass)
- Conceptually many other disciplines appear to act in this way
The Two-Stage Approach

Stage 1 - Similarity

- Determining agreement is straightforward for discrete data like blood type or DNA alleles.
- Statistical significance tests (or other procedures) can be used for continuous data like trace element concentrations in glass.
- Note that there is a loss of information in summarizing the evidence by a binary decision.
  - "Can't be distinguished" might mean an exact match of measurements.
  - "Can't be distinguished" might mean difference between samples just misses being statistically significant.
The Two-Stage Approach

- Testing procedure for continuous data
  - characterize each object by mean value (e.g., mean trace element concentration in population of glass fragments)
  - this is the “population mean” in statistics terminology (one for glass from crime scene, one for glass on suspect)
  - obtain sample values from crime scene object
  - obtain sample values from suspect’s object
  - use sample values to test hypothesis that two objects have the same population mean
  - common tool is $t$-test demonstrated earlier
  - summary is $p$-value, probability of data like the observed data, assuming population means are the same
  - small $p$ (less than .05 or .01) indicates there is strong evidence of a difference in population means
  - otherwise can’t reject the hypothesis that the two means are equal (.... but is this evidence that they came from the same population?)
The Two-Stage Approach

- Example: Two glass samples (from Curran et al. 1997)
- Five measurements of aluminum concentration in crime scene sample
  
  .751, .659, .746, .772, .722

- Five measurements of aluminum concentration in recovered sample
  
  .752, .739, .695, .741, .715

- Control: mean = .730, std.err. = \( \frac{.0435}{\sqrt{5}} = .019 \)
- Sample: mean = .728, std.err. = \( \frac{.0230}{\sqrt{5}} = .010 \)
- Test statistic = \( \frac{.730-.728}{\sqrt{.019^2+.010^2}} = \frac{.002}{.0215} \approx 0.1 \)
- \( p \)-value = .70 … no reason to reject hypothesis of equal means
- We would say these two samples are "indistinguishable"
- In fact, these are 10 measurements from same bottle
The Two-Stage Approach

- Alternative related methods exist
  - 4-sigma methods create interval for each element in each sample (mean conc. +/- 4 standard errors) and check for overlap
  - range overlap uses "control" sample to obtain an expected range and check with "test" samples are in/out of range
  - Hotelling’s $T^2$ test compares all elements simultaneously (take account of dependence)
The Two-Stage Approach

- There are a number of technical statistical issues associated with the use of these procedures
  - the formal test procedures (t-test, Hotelling’s test) require assumptions about the probability distribution of the data
  - univariate procedures are repeated on multiple elements and the existence of multiple comparisons should be accounted for
  - univariate procedures ignore information in the correlation of elements
  - multivariate procedures (like Hotelling’s test) require large samples

- But we do not focus on these

- The more important concerns are conceptual
The Two-Stage Approach

- Stage 1 - The role of the null hypothesis
  - Significance tests do not treat the two hypotheses (equal means, unequal means) symmetrically
  - The null hypothesis (equal means) is assumed true unless the data rejects
  - Acceptance of null in this setting is taken as evidence against the suspect
  - Thus the asymmetry (null vs alternative) is an important issue
The Two-Stage Approach

- **Stage 1 - Using a binary decision**
  - A binary decision (to reject the null hypothesis or not) requires the selection of a cutoff or threshold (e.g., .05 p-value or 4-sigma interval)
  - Choice of threshold impacts the error rates associated with the test
    - a low "threshold" makes it easy to reject ... risks a type I error which rejects a true match and potentially fails to include important evidence
    - a high "threshold" makes it easy to accept the null ... risks a type II error which declares non-matching populations as indistinguishable and could thus incriminate incorrectly
  - It is also the case that inferior measurement protocols (i.e., those with larger standard errors) will make it easier to accept the null hypothesis
The Two-Stage Approach - thresholds / error rates

- thresholds / error rates
  - Black = type I error = false exclusion;
  - Red = type II error = false inclusion

- Initial threshold → low type I error, higher type II error
- Lowering the threshold → increases type I errors, lowers type II errors

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
The Two-Stage Approach

- Stages 1 and 2 - Separation of the match/non-match decision from the assessment of the probative value
  - Separation into two steps is not optimal
  - Much recent attention to this issue in the statistics community (especially around medical studies)
    - failing to find a significant difference is not the same as finding that the null hypothesis is true
    - major statistical association recently recommended not using the term "statistically significant"
  - Example where this could be relevant:
    - Consider a DNA case in which evidence is unlikely from suspect (suggesting not a match)
    - But even more unlikely from all other possible suspects
The Two-Stage Approach

- There are approaches that may address these concerns
  - Equivalence testing instead of significance testing (changes the null hypothesis and addresses the first concern)
    - Requires us to specify a “practically” important difference $\Delta$
      \[
      H_0 : |\mu_{\text{scene}} - \mu_{\text{suspect}}| > \Delta \\
      H_A : |\mu_{\text{scene}} - \mu_{\text{suspect}}| < \Delta
      \]
  - The null hypothesis (now assuming distinguishable items) is assumed true until proven otherwise
  - Bayesian approach and the likelihood ratio address the other concerns (avoids binary decision, avoids separation of match/significance)
The Two-Stage Approach

- Now return to the usual significance testing approach and assume we have found a “match” or found the glass to be “indistinguishable” (i.e., could not reject the null hypothesis)
- Stage 2 - Identification
  - Assessment of the probability that two samples from different sources would "match" or be found "indistinguishable" by chance / coincidence
    - "The Fugitive" - the one-armed man
  - Stage 2 information is rarely provided at present
    - This is a problem ....
      Evidence presented is that two glass samples "can not be distinguished" without further information
The Two-Stage Approach

- Discrete data (e.g., blood type, DNA)
  - Want to find the probability of a match by chance
  - Several important considerations
    - usually crime-scene centered: material from scene is considered fixed and want likelihood that individual would have similar object
    - depends on relevant “information” (suspect is male, suspect is of Asian descent, etc.)
    - where do data come from (population records, convenience sample)
  - This is equally relevant to likelihood ratio approach so don’t discuss the discrete data case further here
The Two-Stage Approach

- Stage 2 ... the idea for continuous data
  - Figure below shows means of refraction index measurements for different windows
  - Suppose control sample has mean 1.522 (red line)
  - We consider each possible source (in the figure below)
    - for each source we ask the probability that a sample drawn from the source would be found indistinguishable from our control sample
    - we total up (actually average) these probabilities
The Two-Stage Approach

- Stage 2 ... a bit more detail
  - suppose for the moment we know the "population" mean of a randomly chosen glass source
  - can find probability that t-test based on a sample from this random object will result in agreement with the "control" sample
  - then total coincidental agreement probability is an average over all possible choices for the random source

\[
\text{Coinc.Prob.} = \sum_{\text{means}} \Pr(\text{mean}) \Pr(\text{match} \mid \text{mean})
\]

- technically challenging .... but it can be done
- a key issue .... where does information about the set of possible random sources (i.e., the relevant population) come from?
The Two-Stage Approach

- Illustrate with example (aluminum concentration in glass)
- Control sample: $\bar{X} = .730, s.d. = .04, n = 5$
- Assume we will apply a standard statistical test (with 5 samples from the unknown) with a cutoff corresponding to a $p$-value of .05
- To start, suppose there are only three types of glass in the population
  - Some randomly chosen sources have means equal to .73
    these will be hard to distinguish from the control sample
    (indistinguishable with probability .95)
  - Some randomly chosen sources have means equal to .78
    it is possible but not certain that we can distinguish these from the control sample
    (indistinguishable with probability .49)
  - Some randomly chosen sources have means equal to .83
    it will be easy to distinguish these from the control sample
    (indistinguishable with probability .02)
The Two-Stage Approach

- More realistic to assume many different types of glass
- To investigate we use the same control sample and assume the population of sources follows a normal distribution
  - if mean is .73 and s.d. .20 then coincidence prob is .19
  - if mean is .73 and s.d. .10 then .... .37
  - if mean is .73 and s.d. .05 then .... .62
  - if mean is .83 and s.d. .20 then .... .17
  - if mean is .83 and s.d. .10 then .... .24
  - if mean is .83 and s.d. .05 then .... .18
  - if mean is .93 and s.d. .20 then .... .12
  - if mean is .93 and s.d. .10 then .... .06
  - if mean is .93 and s.d. .05 then .... .004

- Probability of a coincidental match is high when:
  - small difference between control sample and population of randomly chosen sources (i.e., control sample is "ordinary")
  - large amount of heterogeneity among the potential sources in the population
  - large amount of variability among the fragments in an individual source
The Two-Stage Approach

Summary

- First stage determines if the known and unknown samples appear to "match" or "be indistinguishable"
  - important to recognize the asymmetry in testing a null hypothesis
  - important to design a procedure with appropriate error rates
- Second stage attempts to quantify the probability of a coincidental match
  - requires careful consideration of the relevant population
  - can be challenging to compute (no standard procedure)
  - this step is important but unfortunately sometimes omitted
Likelihood Ratio Approach

Introduction

- Goal for trier of fact in courtroom setting is decision about the relative likelihood of two hypotheses (e.g., same or different source) given data
- In statistical terms this is a Bayesian formulation (asks for probabilities about the state of the world given observed data)
- Recall that Bayes’ rule is a way of reversing direction of conditional probabilities
  - We can go from statements about the likelihood of the evidence given the hypotheses to statements about the likelihood of the hypotheses given the evidence
Likelihood Ratio Approach

Introduction

- Formally using $E$ (evidence) and $S$ (same source)
  (with $\bar{S}$ denoting different source)
- Recall Bayes' Theorem written in terms of the odds in favor of the same source hypothesis

\[
\frac{\Pr(S|E)}{\Pr(\bar{S}|E)} = \frac{\Pr(E|S)}{\Pr(E|\bar{S})} \times \frac{\Pr(S)}{\Pr(\bar{S})}
\]

- In words: Posterior odds = Likelihood ratio $\times$ Prior odds
- The likelihood ratio is the summary of the evidence that is relevant to applying Bayes' Theorem
- The likelihood ratio already plays a role outside of forensics (e.g., in medical diagnostic tests)
- Europe has moved decisively in this direction (ENFSI Guideline)
Likelihood Ratio Approach
Introduction

- Reminder: $LR = \frac{Pr(E|S)}{Pr(E|\bar{S})}$

- Some observations
  - The LR speaks to the relative likelihood of the evidence under the two hypotheses
  - The LR does not make any direct statement about the probability of the hypotheses
  - If we wish to talk about the probability of the hypothesis, then we are interested in the posterior probabilities of the hypotheses
    - But to talk about posterior probabilities ....
      - we must have had prior probabilities to start with
      - and we should be willing to say what they are
    - Does not seem that this is the role of the forensic examiner
Likelihood Ratio Approach

Introduction

- Reminder: \( LR = \frac{\Pr(E|S)}{\Pr(E|\bar{S})} \)

- Some observations
  - numerator assumes common source and asks about the likelihood of the evidence in that case
    - somewhat related to finding a \( p \)-value for testing the hypothesis of equal means
    - but ... no binary decision regarding match
    - instead a quantitative measure of likelihood of evidence under \( S \)
  - denominator assumes no common origin and asks about the likelihood of the evidence in that case
    - analogous to finding coincidence probability
    - here too, doesn’t require a binary decision regarding match
    - a quantitative measure of likelihood of evidence under \( \bar{S} \)
Likelihood Ratio Approach

Introduction

- Reminder: \( LR = \frac{\Pr(E|S)}{\Pr(E|\bar{S})} \)

- Some more technical observations
  - the term likelihood is used because if \( E \) includes continuous measurements then can’t talk about probability
  - could in principle be used with \( E \) equal to “all” evidence of all types
  - other available information (e.g., background) can be incorporated into the LR
Likelihood Ratio Approach
LR and Bayes Factor - terminology

- There is some confusion about terminology
- How do the LR approach and Bayesian approach relate?
- The LR (often called the Bayes Factor) plays a central role in a Bayesian approach to forensic evidence
- LR (Bayes Factor) is the quantity used to update a priori odds and obtain posterior odds
- LR vs Bayes Factor
  - Distinction is technical and has to do with how statistical parameters are treated
Likelihood Ratio Approach

Introduction

- Makes explicit the need to consider the evidence under two different hypotheses
- Separates "objective" information about the evidence from "subjective" assessments about the hypothesis
- Interpretation
  - LR is the factor we should use to update our odds
  - LR statement: "The evidence (e.g., level of agreement) is LR times more likely if the objects have the same source than if the objects have different sources"
  - There are proposals (e.g., ENFSI) that map LRs to verbal scales (2-10: weak support; 10-100: moderate support; ....)
- There is considerable subtlety here ...
  - Prosecutor's fallacy: interpreting $Pr(E|\bar{S})$ as $Pr(\bar{S}|E)$
    - Evidence is unlikely under $\bar{S}$ is interpreted as saying that $\bar{S}$ is unlikely
    - Prosecutor in the Collins case did this!

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Likelihood Ratio Approach
Some notes on implementation

- Reminder: \( LR = \frac{\Pr(E|S)}{\Pr(E|\bar{S})} \)

- Assume \( E = (x, y) \) where \( y \) is measurement of evidence from unknown or questioned sample and \( x \) is measurement of evidence from known or control sample

- Glass ...
  \( y \) are measurements from the glass on the suspect
  \( x \) are measurements from the crime scene glass

- Latent prints ...
  \( y \) are the observations on the latent print at the crime scene
  \( x \) are the observations on the fingerprints of the suspect
**Likelihood Ratio Approach**

Some notes on implementation

- Reminder: \( LR = \frac{Pr(E|S)}{Pr(E|\bar{S})} \)
- Replace \( Pr() \) by \( p() \) to cover discrete and continuous cases
- Then we have from the laws of probability

\[
LR = \frac{p(x, y|S)}{p(x, y|\bar{S})} = \frac{p(y|x, S) p(x|S)}{p(y|x, \bar{S}) p(x|\bar{S})}
\]

- Generally the likelihood of \( x \) (control measurements) is same for \( S \) and \( \bar{S} \) so that \( p(x|S) = p(x|\bar{S}) \)
- If so ..... \( LR = \frac{p(y|x,S)}{p(y|x,\bar{S})} \)
  and the LR focuses on the difference between \( y \) and \( x \)
Likelihood Ratio Approach
Some notes on implementation

- Assume we can start with

\[ LR = \frac{p(y|x, S)}{p(y|x, \bar{S})} \]

- Discrete case
  - Numerator is typically one or zero (or nearly so)
    (We should consider the possibility of lab error or contamination)
  - Denominator is the probability of a coincidental match

- Continuous case
  - Numerator is a measure of how likely it is to observe \( y \) if it is from the same source that gave us the measurement \( x \)
  - Denominator is a measure of how likely it is to observe \( y \) if it is from a different source
Likelihood Ratio Approach
A simple example

- Suppose evidence is blood types for a crime scene sample \((y)\) and suspect sample \((x)\)
- We have information about the distribution of blood types in the population
  
<table>
<thead>
<tr>
<th>Type</th>
<th>A</th>
<th>B</th>
<th>AB</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. Freq</td>
<td>.42</td>
<td>.10</td>
<td>.04</td>
<td>.44</td>
</tr>
</tbody>
</table>

- Suppose both samples are observed to be of blood type O
- \(\Pr(y = O|x = O, S) \approx 1\)
  (expect crime scene sample to match suspect’s type O if \(S\) is true)
- \(\Pr(y = O|x = O, \bar{S}) = \Pr(y = O|\bar{S}) = 0.44\)
  (type O blood is relatively common in the U.S.)
- \(LR \approx 1/0.44 \approx 2.3\)
- Evidence provides weak support for the “same source” hypothesis
Likelihood Ratio Approach
Where it works ..... DNA

- A DNA profile identifies alleles at a number of different locations along the genome (e.g., alleles at location TH01 are 7,9)
- As with blood type, we may see matching profiles (crime scene and suspect)
- Numerator is approximately one (as in blood type example)
- Can determine probability of a coincidental match for each marker or location

<table>
<thead>
<tr>
<th>TH01</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>9.3</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>.001</td>
<td>.001</td>
<td>.266</td>
<td>.160</td>
<td>.135</td>
<td>.199</td>
<td>.200</td>
<td>.038</td>
<td>.001</td>
</tr>
</tbody>
</table>

- For TH01 agreeing on alleles 7, 9, the probability of a random agreement is $2 \times .16 \times .199 = .064$ so $LR \approx 15$
Likelihood Ratio Approach
Where it works .... DNA

- DNA evidence consists of data for a number of locations (CODIS used 13 locations pre-2017 and more now)
- Locations on different chromosomes are independent
- Recall that if events are independent, then we can multiply probabilities (which basically means multiplying likelihood ratios)
- A match at all locations can lead to likelihood ratios in the billions (or even larger)
Likelihood Ratio Approach

Where it works .... DNA

- Underlying biology is well understood
- Probability model for the evidence follows from genetic theory
- Population databases are available
- Peer-reviewed and well accepted by scientific community
- Note: Even with the above information, there are still issues in the DNA world
  - Allele calling still has some subjective elements
  - Samples containing multiple sources (i.e., mixtures)
Likelihood Ratio Approach
Where it can work .... Trace evidence

- Glass and bullet lead are examples
- Can measure chemical concentrations of elements in glass (or bullet lead)
- May have broken glass at crime scene and glass fragments on suspect
- Can we construct a likelihood ratio for evidence of this type?
  - Perhaps .... motivate with some pictures of distributions of refractive indices of glass
Likelihood Ratio Approach

Where it can work .... Trace evidence

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Likelihood Ratio Approach
Where it can work .... Trace evidence
Likelihood Ratio Approach
Continuous measures for trace data - example

- Take \( y \) and \( x \) to be measurements (element, refractive index) from several glass fragments at the crime scene (the control \( x \)) and the subject (the questioned \( y \))
- Assume normal distribution for trace element concentrations (may be more reasonable for logarithms)
- Under the same source hypothesis \( S \)
  - \( x \) and \( y \) are two sets of measurements from a single source (i.e., from a single “within source” normal distribution)
- Under the different source hypothesis \( \overline{S} \)
  - \( x \) and \( y \) are sets of measurements from two different sources (i.e., from two different normal distributions with means drawn from the relevant “between source” distribution of possible sources)
Likelihood Ratio Approach
Continuous measures for trace data - example (cont’d)

- It is possible compute a likelihood ratio in this scenario if we have information about
  - variability of repeated measurements from "within" a single source
  - variability among the mean measurements of sources in the population of interest (i.e., the "between" source variability)

- Key findings:
  - LR is small if \( y \) and \( x \) are very different (i.e., no match)
  - LR is big if \( y \) and \( x \) are similar and \( y \) is unusual for the population of interest (i.e., a match on an unusual value)

- Aitken and Lucy examples find typical LRs for glass evidence in 100s or 1000s
Likelihood Ratio Approach
Where it can work .... Trace evidence

- Well-defined set of measurements (e.g., chemical concentrations)
- Plausible probability models to describe variation within a sample (e.g., normal distribution or less restrictive models)
- Possible to sample from a population (e.g., other windows) to assess variation across different sources
- Can and has been done
  - Aitken and Lucy (2004) - glass
  - Carriquiry, Daniels and Stern (2000 technical report) - bullet lead
- But ...
  - Assessing the relevant “population” is hard (and may vary from case to case)
  - Likelihood ratios can be very sensitive to assumptions that are made (Lund and Iyer, NIST 2017)
Likelihood Ratio Approach
Where it might work .... Pattern evidence

- Many forensic disciplines are focused on comparing a sample (mark) at the crime scene (the “unknown” or “questioned”) and a potential source (the “known”)
- Need to assess whether two samples have same source or different source
- Many examples
  - Latent print examinations
  - Shoe prints and tire tracks
  - Questioned documents / handwriting
  - Firearms
  - Tool marks
Likelihood Ratio Approach
Where it might work .... Pattern evidence
Likelihood Ratio Approach
Where it might work .... Pattern evidence

- A number of challenges in constructing likelihood ratios
- Even defining what we mean by the evidence $E$ is challenging
  - Data are very high dimensional (often images)
  - Flexibility in defining the numbers/types of features to look at
  - Typically $E$ is taken to include observed similarities and differences
Likelihood Ratio Approach
Where it might work .... Pattern evidence

- As with trace evidence, formal evaluation here requires that we study two different types of variation
  - Require information about the variation expected in repeated impressions from the same source (e.g., distortion of fingerprints) to talk about $\Pr(E \mid S)$
  - Require information about the variation expected in impressions from different items in the population (i.e., the "coincidental match" probability) to talk about $\Pr(E \mid S')$
  - May also need information about manufacturing, distribution, wear patterns (e.g., for shoes)
Likelihood Ratio Approach
Where it might work .... Pattern evidence

- How do we measure $\Pr(\mathbf{E} | \mathbf{S})$ and $\Pr(\mathbf{E} | \bar{\mathbf{S}})$
- This is a very hard problem!!
  - Need to assign probabilities to all possible observations $\mathbf{E}$
- Some approaches:
  - Probability models for features (Neumann et al., 2015)
  - Score-based likelihood ratios
  - Subjective likelihood ratios (permitted by ENFSI Guideline)
Likelihood Ratio Approach

ENFSI Guideline for Evaluative Reporting

- ENFSI has officially endorsed likelihood ratios (ENFSI Guideline)
- Guideline cites four requirements for evaluative reporting:
  balance, logic, robustness, transparency
- Some key statements from the Guideline:
  - Evaluate findings (evidence) with respect to competing hypotheses
  - Evaluation should use probability as a measure of uncertainty
  - Evaluation should be based on the assignment of a likelihood ratio
- According to the Guideline, probabilities in the likelihood ratio are ideally based on published data but experience, subjective assessments, case-specific surveys can be used as long as justified
  - The use of experience-based or subjective probabilities has been viewed a bit more skeptically in the U.S.
Likelihood Ratio Approach
ENFSI Guideline for Evaluative Reporting

- LRs reported as numbers or as verbal equipments
- Verbal equivalents are less precise, but may be easier to understand

<table>
<thead>
<tr>
<th>Value of LR</th>
<th>Verbal equivalent: “The forensic findings ...”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>do not support one proposition over the other</td>
</tr>
<tr>
<td>2 – 10</td>
<td>provide <strong>weak support</strong> for the same source proposition relative to the different source proposition</td>
</tr>
<tr>
<td>10 - 100</td>
<td>provide <strong>moderate support</strong> for the same source proposition relative ...</td>
</tr>
<tr>
<td>100 - 1000</td>
<td>provide <strong>moderately strong support</strong> for the same source proposition relative ...</td>
</tr>
<tr>
<td>1000 - 10000</td>
<td>provide <strong>strong support</strong> for the same source proposition relative ...</td>
</tr>
<tr>
<td>10000 – 1 mill.</td>
<td>provide <strong>very strong support</strong> for the same source proposition relative ...</td>
</tr>
<tr>
<td>1 million +</td>
<td>provide <strong>extremely strong support</strong> for the same source proposition relative ...</td>
</tr>
</tbody>
</table>
Likelihood Ratio Approach
Score-based likelihood ratios

- Given the challenge in developing LRs for pattern evidence there has been recent work developing score-based approaches
- Define a score measuring the "difference" between the questioned and known samples (let’s call the score D)
- Obtain the distribution of scores for a sample of known matches (i.e., under same source hypothesis S)
- Obtain the distribution of scores for a sample of known non-matches (i.e., under different source hypothesis $\bar{S}$)
Likelihood Ratio Approach

Score-based likelihood ratios

- The score-based likelihood ratio idea
  - Fit a probability distribution to the scores of known matches \( (Pr(D \mid S)) \)
  - Fit a probability distribution to the scores of known nonmatches \( (Pr(D \mid \tilde{S})) \)
  - Score-based likelihood ratio if we observe score \( D \) is
    \[
    SLR = \frac{Pr(D \mid S)}{Pr(D \mid \tilde{S})}
    \]
Likelihood Ratio Approach
Score-based likelihood ratios

- Example - bullet land signatures
  (note: some scores measure "difference" and some "similarity")
Likelihood Ratio Approach
Score-based likelihood ratios - challenges

- Across a number of existing examples the score distribution for known matches seems relatively straightforward to characterize.
- There are challenges though in defining the relevant non-match population:
  - Is there a single non-match score distribution?
  - Should the non-match score distribution depend on characteristics of the crime scene sample?
  - If so, which characteristics?
Likelihood Ratio Approach

Score-based likelihood ratios - challenges

- FRSTATS (Swofford et al., Forensic Science International, 2018)

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Likelihood Ratio Approach
Score-based analyses - another approach

- Another approach is to use the scores (D) within the two-stage approach
  - If D (difference score) is above a threshold, then declare different source
  - If D (difference score) is below a threshold, then declare same source
  - Compute error rates for this type of procedure (perhaps for different thresholds)
- Example from a paper analyzing bullet lands (Hare et al., AOAS, 2017) (but with score a measure of the likelihood of a match, so higher scores mean more likely to be a match)
Likelihood Ratio Approach
Sensitivity of results to assumptions

- Lund and Iyer (NIST, 2017) noted that a range of different statistical models can be used in deriving the likelihood ratios.
- For a given set of observed data they considered a range of "plausible" models and explored the range of LRs observed.
- Figure below is adapted from Lund and Iyer.
  - Red lines show range of LRs obtained using distributions in the identified class (e.g., normal) that are consistent with the observed data.
  - Lund/Iyer make a critical point.
  - Lund/Iyer are (in my view) extremely generous in defining what is "plausible."

![Uncertainty Pyramid for LR](image)
Likelihood Ratio Approach
Complications

- Many issues can complicate the calculation of LRs in practice. Examples include ...
  - accounting for transfer process with glass or fibers
  - accounting for heterogeneity due to packaging of bullets into boxes
  - accounting for usage/lifetime of products (e.g., sneakers)
- Though good work is being done, it seems likely that it will be some time before LRs are available for pattern evidence
- Important to remember that there is not one LR for a given item of evidence
  - The LR calculation depends on assumptions/models for the measured data
  - The LR calculation depends on assumptions/models for the relevant population
Likelihood Ratio Approach
Summary

Advantages
- explicitly compares two relevant hypotheses/propositions
- provides a quantitative summary of the evidence
- assumptions being used are (or should be) made explicit and open to question
- no need for arbitrary match/non-match decisions when faced with continuous data
- can accommodate a wide range of factors
- flexible enough to accommodate multiple pieces and multiple types of evidence
**Likelihood Ratio Approach**

**Summary**

- **Disadvantages**
  - requires assumptions about distributions
  - calculated LR can be quite sensitive to these assumptions
  - in the US there is no requirement for defense to provide a specific alternative hypothesis
  - need for reference distributions to define denominator
    (although this needs to be done implicitly in any examination)
  - can be difficult to account for all relevant factors
  - how should this information be conveyed to the trier of fact
Expressing Source Conclusions

Before closing, we briefly address one more topic:

- How do potential jurors understand the various types of conclusion statements?
- Studied by several researchers including recent work by Thompson et al. (LPR, 2018)
Expressing Source Conclusions

<table>
<thead>
<tr>
<th>Study</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>2, 3</td>
<td>RMP4: ‘one person in 10 million’</td>
</tr>
<tr>
<td>1, 2</td>
<td>RMP3: ‘one person in 100,000’</td>
</tr>
<tr>
<td>1</td>
<td>RMP2: ‘one person in 1,000’</td>
</tr>
<tr>
<td></td>
<td>RMP1: ‘one person in 10’</td>
</tr>
<tr>
<td>3</td>
<td>LR4: ‘10,000,000 times more likely’ if suspect rather than random person is source</td>
</tr>
<tr>
<td>3</td>
<td>LR3: ‘100,000 times more likely’ if suspect rather than random person is source</td>
</tr>
<tr>
<td>3</td>
<td>CC5: ‘suspect was the source’</td>
</tr>
<tr>
<td></td>
<td>CC4: ‘individualized . . . as coming from the finger of the suspect’</td>
</tr>
<tr>
<td>2</td>
<td>CC3: ‘identified . . . to the finger of the suspect’</td>
</tr>
<tr>
<td>2</td>
<td>CC2: ‘matches the fingerprint of the suspect’</td>
</tr>
<tr>
<td>2</td>
<td>CC1: ‘suspect could have been the source’</td>
</tr>
<tr>
<td>2</td>
<td>LoS1: ‘likelihood of observing this amount of corresponding ridge detail when two fingerprints are made by different people is considered extremely low’</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>SP3: ‘a practical certainty that suspect was the source’</td>
</tr>
<tr>
<td>1</td>
<td>SP2: ‘highly probable’</td>
</tr>
<tr>
<td>1</td>
<td>SP1: ‘moderately probable’</td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>SoS4: ‘extremely strong support’</td>
</tr>
<tr>
<td>3</td>
<td>SoS3: ‘very strong support’</td>
</tr>
<tr>
<td>1</td>
<td>SoS2: ‘moderate support’</td>
</tr>
<tr>
<td>1</td>
<td>SoS1: ‘weak support’</td>
</tr>
</tbody>
</table>
Expressing Source Conclusions

Research funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) - forensicstats.org
Workshop Summary / Conclusions

- Quantitative analysis of forensic evidence requires some familiarity with concepts from probability and statistics.
- Workshop reviewed basics of probability and statistics.
- Reviewed statistical approaches to forensic evidence (two-stage approach and likelihood ratio/Bayes factor).
- Addressed statistical considerations needed for each of these approaches to forensic evidence.
- Key points
  - Any approach should account for the two (or more) competing hypotheses about how the data was generated.
  - Need to be explicit about reasoning and data on which reasoning is based.
  - Need to describe the level of certainty associated with a conclusion.